



# Business Statistics

COMMUNICATING  
WITH NUMBERS

Second Edition



Jaggia / Kelly

# BUSINESS STATISTICS





Second Edition

# BUSINESS STATISTICS

## Communicating with Numbers

**Sanjiv Jaggia**

*California Polytechnic  
State University*

**Alison Kelly**

*Suffolk University*



## BUSINESS STATISTICS: COMMUNICATING WITH NUMBERS, SECOND EDITION

Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121. Copyright © 2016 by McGraw-Hill Education. All rights reserved. Printed in the United States of America. Previous editions © 2013. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written consent of McGraw-Hill Education, including, but not limited to, in any network or other electronic storage or transmission, or broadcast for distance learning.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 0 DOW/DOW 1 0 9 8 7 6 5

ISBN 978-0-07-802055-1

MHID 0-07-802055-7

Senior Vice President, Products & Markets: *Kurt L. Strand*

Vice President, General Manager, Products & Markets: *Marty Lange*

Vice President, Content Design & Delivery: *Kimberly Meriwether David*

Managing Director: *Jame Heine*

Marketing Director: *Lynn Breithaupt*

Brand Manager: *Dolly Womack*

Director, Product Development: *Rose Koos*

Product Developer: *Christina Holt*

Director of Digital Content: *Doug Ruby*

Digital Product Analyst: *Kevin Shanahan*

Director, Content Design & Delivery: *Linda Avenarius*

Program Manager: *Mark Christianson*

Content Project Managers: *Harvey Yep / Bruce Gin*

Buyer: *Jennifer Pickel*

Design: *Srdjan Savanovic*

Content Licensing Specialists: *Keri Johnson / John Leland / Rita Hingtgen*

Cover Image: © Comstock/Stockbyte/Getty Images/RF; © Mitch Diamond/Photodisc/Getty Images/RF;

© Mark Bowden/iStock/Getty Images Plus/Getty Images; © Rob Tringali/Getty Images; © Image

Source, all rights reserved/RF; © Honqi Zhang/iStock/Getty Images Plus/Getty Images/RF;

© imageBROKER/Alamy/RF; © TongRo Images/Getty Images; © Yellow Dog Productions/Digital

Vision/Getty Images/RF

Compositor: *MPS Limited, A Macmillan Company*

Printer: *R. R. Donnelley*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

### Library of Congress Cataloging-in-Publication Data

Jaggia, Sanjiv, 1960-

Business statistics: communicating with numbers / Sanjiv Jaggia,

California Polytechnic State University, Alison Kelly, Suffolk University.

Second Edition.

pages cm.—(Business statistics)

ISBN 978-0-07-802055-1 (hardback)

1. Commercial statistics. I. Hawke, Alison Kelly. II. Title.

HF1017.J34 2015

519.5—dc23

2015023383

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw-Hill Education, and McGraw-Hill Education does not guarantee the accuracy of the information presented at these sites.





*Dedicated to Chandrika, Minori, John, Megan, and Matthew*

# ABOUT THE AUTHORS

## Sanjiv Jaggia



Sanjiv Jaggia is the associate dean of graduate programs and a professor of economics and finance at California Polytechnic State University in San Luis Obispo, California. After earning a Ph.D. from Indiana University, Bloomington, in 1990, Dr. Jaggia spent 17 years at Suffolk University, Boston. In 2003, he became a Chartered Financial Analyst (CFA®). Dr. Jaggia's research interests include empirical finance, statistics, and econometrics. He has published extensively in research journals, including the *Journal of Empirical Finance*, *Review of Economics and Statistics*, *Journal of Business and Economic Statistics*, and *Journal of Econometrics*. Dr. Jaggia's ability to communicate in the classroom has been acknowledged by several teaching awards. In 2007, he traded one coast for the other and now lives in San Luis Obispo, California, with his wife and daughter. In his spare time, he enjoys cooking, hiking, and listening to a wide range of music.

## Alison Kelly



Alison Kelly is a professor of economics at Suffolk University in Boston, Massachusetts. She received her B.A. degree from the College of the Holy Cross in Worcester, Massachusetts; her M.A. degree from the University of Southern California in Los Angeles; and her Ph.D. from Boston College in Chestnut Hill, Massachusetts. Dr. Kelly has published in journals such as the *American Journal of Agricultural Economics*, *Journal of Macroeconomics*, *Review of Income and Wealth*, *Applied Financial Economics*, and *Contemporary Economic Policy*. She is a Chartered Financial Analyst (CFA) and regularly teaches review courses in quantitative methods to candidates preparing to take the CFA exam. Dr. Kelly has also served as a consultant for a number of companies; her most recent work focuses on how large financial institutions satisfy requirements mandated by the Dodd-Frank Act. She resides in Hamilton, Massachusetts, with her husband and two children.

# A Unique Emphasis on Communicating with Numbers Makes Business Statistics Relevant to Students

Statistics can be a fun and enlightening course for both students and teachers. From our years of experience in the classroom, we have found that an effective way to make statistics interesting is to use timely business applications to which students can relate. If interest can be sparked at the outset, students may end up learning statistics without realizing they are doing so. By carefully matching timely applications with statistical methods, students learn to appreciate the relevance of business statistics in our world today. We wrote *Business Statistics: Communicating with Numbers* because we saw a need for a contemporary, core statistics textbook that sparked student interest and bridged the gap between how statistics is taught and how practitioners think about and apply statistical methods. Throughout the text, the emphasis is on communicating with numbers rather than on number crunching. In every chapter, students are exposed to statistical information conveyed in written form. By incorporating the perspective of professional users, it has been our goal to make the subject matter more relevant and the presentation of material more straightforward for students.

In *Business Statistics*, we have incorporated fundamental topics that are applicable for students with various backgrounds and interests. The text is intellectually stimulating, practical, and visually attractive, from which students can learn and instructors can teach. Although it is application oriented, it is also mathematically sound and uses notation that is generally accepted for the topic being covered.

*This is probably the best book I have seen in terms of explaining concepts.*

Brad McDonald, Northern Illinois University

*The book is well written, more readable and interesting than most stats texts, and effective in explaining concepts. The examples and cases are particularly good and effective teaching tools.*

Andrew Koch, James Madison University

*Clarity and brevity are the most important things I look for—this text has both in abundance.*

Michael Gordinier, Washington University, St. Louis



## Continuing Key Features

The second edition of *Business Statistics* reinforces and expands six core features that were well-received in the first edition.

**Integrated Introductory Cases.** Each chapter begins with an interesting and relevant introductory case. The case is threaded throughout the chapter, and it often serves as the basis of several examples in other chapters.

**Writing with Statistics.** Interpreting results and conveying information effectively is critical to effective decision making in a business environment. Students are taught how to take the data, apply it, and convey the information in a meaningful way.

**Unique Coverage of Regression Analysis.** Relevant coverage of regression without repetition is an important hallmark of this text.

**Written as Taught.** Topics are presented the way they are taught in class, beginning with the intuition and explanation and concluding with the application.

**Integration of Microsoft Excel®.** Students are taught to develop an understanding of the concepts and how to derive the calculation; then Excel is used as a tool to perform the cumbersome calculations. In addition, guidelines for using Minitab, SPSS, and JMP are provided in chapter appendices; detailed instructions for these packages and for R are available in *Connect*.

**Connect® Business Statistics.** *Connect* is an online system that gives students the tools they need to be successful in the course. Through guided examples and LearnSmart adaptive study tools, students receive guidance and practice to help them master the topics.

*I really like the case studies and the emphasis on writing. We are making a big effort to incorporate more business writing in our core courses, so that meshes well.*

*Elizabeth Haran, Salem State University*

*For a statistical analyst, your analytical skill is only as good as your communication skill. Writing with statistics reinforces the importance of communication and provides students with concrete examples to follow.*

*Jun Liu, Georgia Southern University*

## Features New to the Second Edition

The second edition of *Business Statistics* features a number of improvements suggested by numerous reviewers and users of the first edition.

First, every section of every chapter has been scrutinized, and if a change would enhance readability, then that change was made. In addition, Excel instructions have been streamlined in every chapter. We feel that this modification provides a more seamless reinforcement for the relevant topic. For those instructors who prefer to omit the Excel parts, these sections can be easily skipped. Moreover, most chapters now include an appendix that provides brief instructions for Minitab, SPSS, and JMP. More detailed instructions for Minitab, SPSS, and JMP can be found in *Connect*.

Dozens of applied exercises of varying levels of difficulty have been added to just about every section of every chapter. Many of these exercises include new data sets that encourage the use of the computer; however, just as many exercises retain the flexibility of traditional solving by hand.

Both of us use *Connect* in our classes. In an attempt to make the technology component seamless with the text itself, we have reviewed every *Connect* exercise. In addition, we have painstakingly revised tolerance levels and added rounding rules. The positive feedback from users due to these adjustments has been well worth the effort. In addition, we have included numerous new exercises in *Connect*. We have also reviewed every probe from LearnSmart. Instructors who teach in an online or hybrid environment will especially appreciate these modifications.

Here are some of the more noteworthy, specific changes:

- Some of the Learning Outcomes have been rewritten for the sake of consistency.
- In Chapter 3 (Numerical Descriptive Measures), the discussion of the weighted mean occurs in Section 3.1 (Measures of Central Location) instead of Section 3.7 (Summarizing Grouped Data). Section 3.6 has been renamed from “Chebyshev’s Theorem and the Empirical Rule” to “Analysis of Relative Location”; in addition, we have added a discussion of  $z$ -scores in this section.
- In Chapter 4 (Introduction to Probability), the term *a priori* has been replaced by *classical*.
- In Chapter 5 (Discrete Probability Distributions), the use of graphs now complements the discussion of the binomial and Poisson distributions.
- In Chapter 7 (Sampling and Sampling Distributions), the standard error of a statistic is now denoted as “*se*” instead of “*SD*.” For instance, the standard error of the sample mean is now denoted as  $se(\bar{X})$  instead of  $SD(\bar{X})$ .
- The discussion of the properties of estimators has been moved from Section 8.1 to an appendix in Chapter 7.
- In Section 16.1 (Polynomial Models), the discussion of the marginal effects of  $x$  on  $y$  has been expanded.
- In Section 17.1 (Dummy Variables), there is now an example of how to conduct a hypothesis test when the original reference group must be changed.
- In Chapter 18 (Time Series Forecasting), the data used for the “Writing with Statistics” example has been revised.

# Students Learn Through Real-World Cases and Business Examples . . .

## Integrated Introductory Cases

Each chapter opens with a real-life case study that forms the basis for several examples within the chapter. The questions included in the examples create a roadmap for mastering the most important learning outcomes within the chapter. A synopsis of each chapter's introductory case is presented when the last of these examples has been discussed. Instructors of distance learners may find these introductory cases particularly useful.



### SYNOPSIS OF INTRODUCTORY CASE

Vanguard's Precious Metals and Mining fund (Metals) and Fidelity's Strategic Income fund (Income) were two top-performing mutual funds for the years 2000 through 2009. An analysis of annual return data for these two funds provides important information for any type of investor. Over the past 10 years, the Metals fund posts the higher values for both the mean return and the median return, with values of 24.65% and 33.83%, respectively. When the mean differs dramatically from the median, it is often indicative of extreme values or outliers. Although the mean and the median for the Metals fund do differ by almost 10 percentage points, a boxplot analysis reveals no outliers. The mean return and



### INTRODUCTORY CASE

#### Investment Decision

Rebecca Johnson works as an investment counselor at a large bank. Recently, an inexperienced investor asked Johnson about clarifying some differences between two top-performing mutual funds from the last decade: Vanguard's Precious Metals and Mining fund (henceforth, Metals) and Fidelity's Strategic Income fund (henceforth, Income). The investor shows Johnson the return data that he has accessed over the Internet, but the investor has trouble interpreting the data. Table 3.1 shows the return data for these two mutual funds for the years 2000–2009.

*In all of these chapters, the opening case leads directly into the application questions that students will have regarding the material. Having a strong and related case will certainly provide more benefit to the student, as context leads to improved learning.*

*Alan Chow, University of South Alabama*

*This is an excellent approach. The student gradually gets the idea that he can look at a problem—one which might be fairly complex—and break it down into root components. He learns that a little bit of math could go a long way, and even more math is even more beneficial to evaluating the problem.*

*Dane Peterson, Missouri State University*



# and Build Skills to Communicate Results

## Writing with Statistics

One of our most important innovations is the inclusion of a sample report within every chapter (except Chapter 1). Our intent is to show students how to convey statistical information in written form to those who may not know detailed statistical methods. For example, such a report may be needed as input for managerial decision making in sales, marketing, or company planning. Several similar writing exercises are provided at the end of each chapter. Each chapter also includes a synopsis that addresses questions raised from the introductory case. This serves as a shorter writing sample for students. Instructors of large sections may find these reports useful for incorporating writing into their statistics courses.

*Writing with statistics shows that statistics is more than number crunching.*

Greg Cameron,  
Brigham Young University

*These technical writing examples provide a very useful example of how to take statistics work and turn it into a report that will be useful to an organization. I will strive to have my students learn from these examples.*

Bruce P. Christensen,  
Weber State University

*This is an excellent approach. . . . The ability to translate numerical information into words that others can understand is critical.*

Scott Bailey, Troy University

*Excellent. Students need to become better writers.*

Bob Nauss, University of  
Missouri, St. Louis

### WRITING WITH STATISTICS



The Associated Press reports that income inequality is at record levels in the United States (September 28, 2010). Over the years, the rich have become richer while working-class wages have stagnated. A local Latino politician has been vocal regarding his concern about the welfare of Latinos, especially given the recent downturn of the U.S. economy. In various speeches, he has stated that the mean salary of Latino households in his county has fallen below the 2008 mean of \$49,000. He has also stated that the proportion of Latino households making less than \$30,000 has risen above the 2008 level of 20%. Both of his statements are based on income data for 36 Latino households in the county, as shown in Table 9.5.

TABLE 9.5 Representative Sample of Latino Household Incomes in 2010

FILE	22	36	78	103	38	43
Latino_Income	62	53	26	28	25	31
	62	44	51	38	77	37
	29	38	46	52	61	57
	20	72	41	73	16	32
	52	28	69	27	53	46

Incomes are measured in \$1,000s and have been adjusted for inflation.

Trevor Jones is a newspaper reporter who is interested in verifying the concerns of the local politician.

Trevor wants to use the sample information to:

1. Determine if the mean income of Latino households has fallen below the 2008 level of \$49,000.
2. Determine if the percentage of Latino households making less than \$30,000 has risen above 20%.

### Sample Report— Assessing Whether Data Follow the Normal Distribution

As part of a broader report concerning the mutual fund industry in general, three-year return data for the 50 largest mutual funds were collected with the objective of determining whether or not the data follow a normal distribution. Information of this sort is particularly useful because much statistical inference is based on the assumption of normality. If the assumption of normality is not supported by the data, it may be more appropriate to use nonparametric techniques to make valid inferences. Table 12.A shows relevant summary statistics for three-year returns for the 50 largest mutual funds.

TABLE 12.A Three-Year Return Summary Measures for the 50 Largest Mutual Funds, August 2008

Mean	Median	Standard Deviation	Skewness	Kurtosis
5.96%	4.65%	3.39%	1.37	2.59

The average three-year return for the 50 largest mutual funds is 5.96%, with a median of 4.65%. When the mean is significantly greater than the median, it is often an indication of a positively skewed distribution. The skewness coefficient of 1.37 seems to support this claim. Moreover, the kurtosis coefficient of 2.59 suggests a distribution that is more peaked than the normal distribution. A formal test will determine whether the conclusion from the sample can be deemed real or due to chance.

The goodness-of-fit test is first applied to check for normality. The raw data is converted into a frequency distribution with five intervals ( $k = 5$ ). Expected frequencies are

# Unique Coverage and Presentation...

*By comparing this chapter with other books, I think that this is one of the best explanations about regression I have seen.*

Cecilia Maldonado,  
Georgia Southwestern  
State University

*The inclusion of material used on a regular basis by investment professionals adds real-world credibility to the text and course and better prepares students for the real world.*

Bob Gillette,  
University of Kentucky

*This is easy for students to follow and I do get the feeling . . . the sections are spoken language.*

Zhen Zhu, University of  
Central Oklahoma

## Unique Coverage of Regression Analysis

Our coverage of regression analysis is more extensive than that of the vast majority of texts. This focus reflects the topic's growing use in practice. We combine simple and multiple regression in one chapter, which we believe is a seamless grouping and eliminates needless repetition. This focus reflects the topic's growing use in practice. However, for those instructors who prefer to cover only simple regression, doing so is still an option. Three more in-depth chapters cover statistical inference, nonlinear relationships, dummy variables, and binary choice models.

Chapter 14: Regression Analysis

Chapter 15: Inference with Regression Models

Chapter 16: Regression Models for Nonlinear Relationships

Chapter 17: Regression Models with Dummy Variables

*The authors have put forth a novel and innovative way to present regression which in and of itself should make instructors take a long and hard look at this book. Students should find this book very readable and a good companion for their course.*

Harvey A. Singer, George Mason University

## Inclusion of Important Topics

In our teaching outside the classroom, we have found that several fundamental topics important to business are not covered by the majority of traditional texts. For example, most books do not integrate the geometric mean, mean-variance analysis, and the Sharpe ratio with descriptive statistics. Similarly, the discussion of probability concepts generally does not include odds ratios, risk aversion, and the analysis of portfolio returns. We cover these important topics throughout the text. Overall, our text contains material that practitioners use on a regular basis.

### THE SHARPE RATIO

The **Sharpe ratio** measures the extra reward per unit of risk. The Sharpe ratio for an investment  $I$  is computed as:

$$\frac{\bar{x}_I - \bar{R}_f}{s_I}$$

where  $\bar{x}_I$  is the mean return for the investment,  $\bar{R}_f$  is the mean return for a risk-free asset such as a Treasury bill (T-bill), and  $s_I$  is the standard deviation for the investment.

## Written as Taught

We introduce topics just the way we teach them; that is, the relevant tools follow the opening application. Our roadmap for solving problems is

1. Start with intuition
2. Introduce mathematical rigor, and
3. Produce computer output that confirms results.

We use worked examples throughout the text to illustrate how to apply concepts to solve real-world problems.

# that Make the Content More Effective

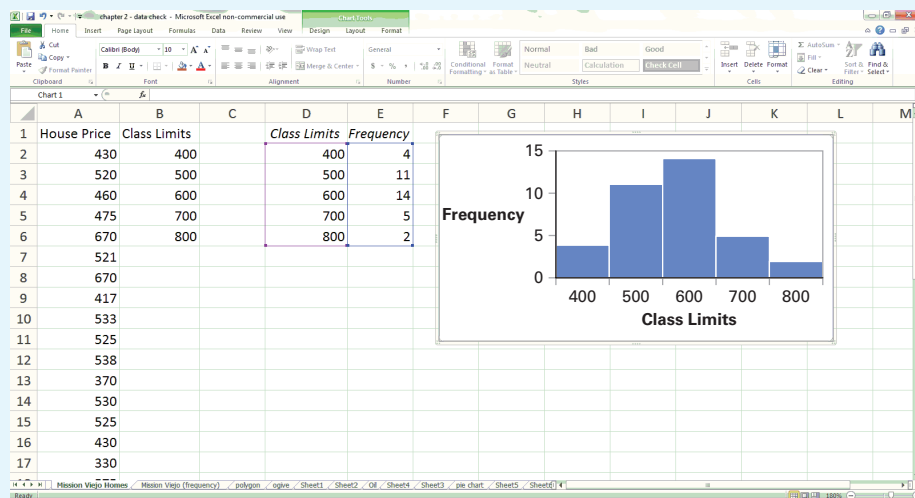
## Integration of Microsoft Excel®

We prefer that students first focus on and absorb the statistical material before replicating their results with a computer. We feel that solving each application manually provides students with a deeper understanding of the relevant concept. However, we recognize that, primarily due to cumbersome calculations or the need for statistical tables, embedding computer output is necessary. Microsoft Excel is the primary software package used in this text, and it is integrated within each chapter. We chose Excel over other statistical packages based on reviewer feedback and the fact that students benefit from the added spreadsheet experience. We provide brief guidelines for using Minitab, SPSS, and JMP in chapter appendices; we give more detailed instructions for these packages and for *R* in *Connect*.

### Using Excel to Construct a Histogram

- A. **FILE** Open *MV\_Houses* (Table 2.1).
- B. In a column next to the data, enter the values of the upper limits of each class, or in this example, 400, 500, 600, 700, and 800; label this column “Class Limits.” The reason for these entries is explained in step D. The house-price data and the class limits (as well as the resulting frequency distribution and histogram) are shown in Figure 2.8.

**FIGURE 2.8** Constructing a histogram from raw data with Excel



... does a solid job of building the intuition behind the concepts and then adding mathematical rigor to these ideas before finally verifying the results with Excel.

Matthew Dean,  
University of  
Southern Maine



# Real-World Exercises and Case Studies that Reinforce the Material

## Mechanical and Applied Exercises

Chapter exercises are a well-balanced blend of mechanical, computational-type problems followed by more ambitious, interpretive-type problems. We have found that simpler drill problems tend to build students' confidence prior to tackling more difficult applied problems. Moreover, we repeatedly use many data sets—including house prices, rents, stock returns, salaries, and debt—in the text. For instance, students first use these real data to calculate summary measures and then continue on to make statistical inferences with confidence intervals and hypothesis tests and perform regression analysis.

Applied exercises from *The Wall Street Journal*, *Kiplinger's*, *Fortune*, *The New York Times*, *USA Today*, various websites—Census.gov, Zillow.com, Finance.yahoo.com, ESPN.com; and more.

### Applications

The Department of Transportation (DOT) fields thousands of complaints about airlines each year. The DOT categorizes complaints by airline, tallies complaints, and then periodically publishes reports of airline performance. The following table presents the 2006 results for the 10 largest U.S. airlines.

	Complaints*	Airline	Complaints*
Northwest Airlines	1.82	Northwest Airlines	8.84
Delta Airlines	3.98	Delta Airlines	10.35
Alaska Airlines	5.24	American Airlines	10.87
AirTran Airways	6.24	US Airways	13.59
Continental Airlines	8.83	United Airlines	13.60

Source: Department of Transportation; \*per million passengers.

- Which airline fielded the least amount of complaints? Which airline fielded the most? Calculate the range.
  - Calculate the mean and the median number of complaints for this sample.
  - Calculate the variance and the standard deviation.
44. The monthly closing stock prices (rounded to the nearest dollar) for Starbucks Corp. and Panera Bread Co. for the first six months of 2010 are reported in the following table.

to promise good returns (*The Wall Street Journal*, September 24, 2010). Marcela Treisman works for an investment firm in Michigan. Her assignment is to analyze the rental market in Ann Arbor, which is home to the University of Michigan. She gathers data on monthly rent for 2011 along with the square footage of 40 homes. A portion of the data is shown in the accompanying table.

Monthly Rent	Square Footage
645	500
675	648
⋮	⋮
2400	2700

Source: <http://www.zillow.com>.

- Calculate the mean and the standard deviation for monthly rent.
  - Calculate the mean and the standard deviation for square footage.
  - Which sample data exhibit greater relative dispersion?
46. **FILE Largest Corporations.** Access the data accompanying this exercise. It shows the Fortune 500 rankings of America's largest corporations for 2010. Next to each corporation are its market capitalization (in billions of dollars as of March 26, 2010) and its total return to investors for the year 2009.
- Calculate the coefficient of variation for market

*I especially like the introductory cases, the quality of the end-of-section problems, and the writing examples.*

*Dave Leupp, University of Colorado at Colorado Springs*

*Their exercises and problems are excellent!*

*Erl Sorensen, Bentley University*

# Features that Go Beyond the Typical

## Conceptual Review

At the end of each chapter, we present a conceptual review that provides a more holistic approach to reviewing the material. This section revisits the learning outcomes and provides the most important definitions, interpretations, and formulas.

### CONCEPTUAL REVIEW

**LO 5.1 Distinguish between discrete and continuous random variables.**

A **random variable** summarizes outcomes of an experiment with numerical values. A random variable is either discrete or continuous. A **discrete random variable** assumes a countable number of distinct values, whereas a **continuous random variable** is characterized by uncountable values in an interval.

**LO 5.2 Describe the probability distribution for a discrete random variable.**

The **probability distribution function** for a discrete random variable  $X$  is a list of the values of  $X$  with the associated probabilities, that is, the list of all possible pairs  $(x, P(X = x))$ . The **cumulative distribution function** of  $X$  is defined as  $P(X \leq x)$ .

**LO 5.3 Calculate and interpret summary measures for a discrete random variable.**

For a discrete random variable  $X$  with values  $x_1, x_2, x_3, \dots$ , which occur with probabilities  $P(X = x_i)$ , the **expected value** of  $X$  is calculated as  $E(X) = \mu = \sum x_i P(X = x_i)$ . We interpret the expected value as the long-run average value of the random variable over infinitely many independent repetitions of an experiment. Measures of dispersion indicate whether the values of  $X$  are clustered about  $\mu$  or widely scattered from  $\mu$ . The **variance** of  $X$  is calculated as  $Var(X) = \sigma^2 = \sum (x_i - \mu)^2 P(X = x_i)$ . The **standard deviation** of  $X$  is  $SD(X) = \sigma = \sqrt{\sigma^2}$ .

**LO 5.4 Distinguish between risk-neutral, risk-averse, and risk-loving consumers.**

In general, a **risk-averse consumer** expects a reward for taking risk. A risk-averse consumer may decline a risky prospect even if it offers a positive expected gain. A **risk-neutral consumer** completely ignores risk and always accepts a prospect that offers a positive expected gain. Finally, a **risk-loving consumer** may accept a risky prospect even if the expected gain is negative.

*Most texts basically list what one should have learned but don't add much to that. You do a good job of reminding the reader of what was covered and what was most important about it.*

*Andrew Koch, James Madison University*

*They have gone beyond the typical [summarizing formulas] and I like the structure. This is a very strong feature of this text.*

*Virginia M. Miori, St. Joseph's University*

# What Technology Connects Students . . .

## McGraw-Hill *Connect*<sup>®</sup> *Business Statistics*



McGraw-Hill *Connect Business Statistics* is an online assignment and assessment solution that connects students with the tools and resources they'll need to achieve success through faster learning, higher retention, and more efficient studying. It provides instructors with tools to quickly select content for assignments according to the topics and learning objectives they want to emphasize.

**Online Assignments.** *Connect Business Statistics* helps students learn more efficiently by providing practice material and feedback when they are needed. *Connect* grades homework automatically and provides instant feedback on any problems that students are challenged to solve.

**Integration of Excel Data Sets.** A convenient feature is the inclusion of an Excel data file link in many problems using data files in their calculation. The link allows students to easily launch into Excel, work the problem, and return to *Connect* to key in the answer and receive feedback on their results.



value:  
1 points

Listed below is the rate of return for one year (reported in percent) for a sample of 12 mutual funds that are classified as taxable money market funds.

4.63	4.15	4.76	4.70	4.65	4.52	4.70	5.06	4.42	4.51	4.24	4.52
------	------	------	------	------	------	------	------	------	------	------	------

Using the .05 significance level, is it reasonable to conclude that the mean rate of return is more than 4.50 percent?

[Click here for the Excel Data File](#)

(a) What is the decision rule? (Round your answer to 3 decimal places.)

Reject  $H_0: \mu \leq 4.5\%$  and fail to reject  $H_1: \mu > 4.5\%$  when the test statistic is

(b) The value of the test statistic is . (Round your answer to 3 decimal places.)

(c) What is your decision regarding  $H_0$ ?

the mean rate of return is  4.5%.

[contact UM Publishing](#) [check my work](#) [eBook Link](#) [references](#)

# to Success in Business Statistics?

**Guided Examples.** These narrated video walkthroughs provide students with step-by-step guidelines for solving selected exercises similar to those contained in the text. The student is given personalized instruction on how to solve a problem by applying the concepts presented in the chapter. The video shows the steps to take to work through an exercise. Students can go through each example multiple times if needed.

Guided Example

**Expected value of the investment**

For a discrete random variable  $X$  with values  $x_i$  occurring with probabilities  $P(X = x_i)$

$$E(X) = \mu = \sum x_i P(X = x_i)$$

Market status	Investment, $x_i$	Probability, $P(X = x_i)$	$x_i P(X = x_i)$
Improves	\$23,000	0.25	5,750
Stays same	\$15,000	0.42	6,300
Deteriorates	\$10,000	0.33	3,300
		$\sum x_i P(X = x_i)$	15,350 ✓

The expected value of the investment is \$15,350

The expected value, \$15,350 > \$15,000, the initial investment

The investor should invest the \$15,000, if he is risk neutral

The McGraw-Hill Companies

**LearnSmart.** LearnSmart adaptive self-study technology in *Connect Business Statistics* helps students make the best use of their study time. LearnSmart provides a seamless combination of practice, assessment, and remediation for every concept in the textbook. LearnSmart's intelligent software adapts to students by supplying questions on a new concept when students are ready to learn it. With LearnSmart, students will spend less time on topics they understand and instead focus on the topics they need to master.



The pictured scatterplot depicts \_\_\_\_\_ between the x and y variable.

a positive linear relationship

a curvilinear relationship

no relationship

a negative linear relationship

Click one of the buttons below.

Do you know the answer? (Be honest.)

Yes Probably Maybe No—just guessing

Use mouse to zoom

SmartBook®, which is powered by LearnSmart, is the first and only adaptive reading experience designed to change the way students read and learn. It creates a personalized reading experience by highlighting the most relevant concepts a student needs to learn at that moment in time. As a student engages with SmartBook, the reading experience continuously adapts by highlighting content based on what the student knows and doesn't know. This ensures that the focus is on the content he or she needs to learn, while simultaneously promoting long-term retention of material. Use SmartBook's real-time reports to quickly identify the concepts that require more attention from individual students or the entire class. The end result? Students are more engaged with course content, can better prioritize their time, and come to class ready to participate.

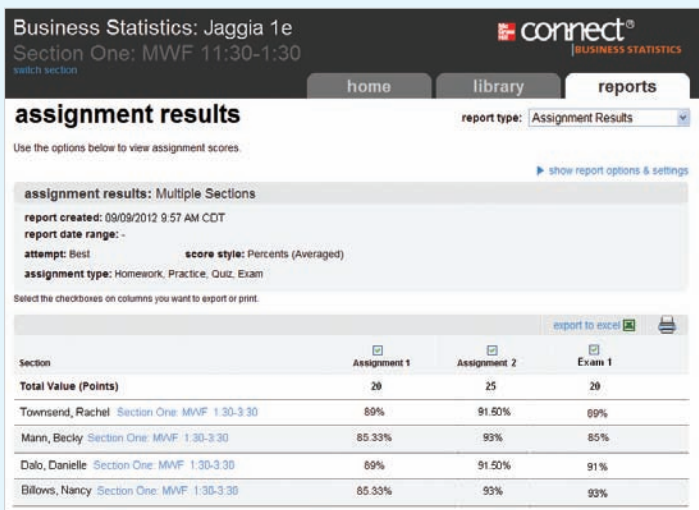


# What Technology Connects Students . . .

**Simple Assignment Management and Smart Grading.** When it comes to studying, time is precious. *Connect Business Statistics* helps students learn more efficiently by providing feedback and practice material when they need it, where they need it. When it comes to teaching, your time also is precious. The grading function enables you to

- Have assignments scored automatically, giving students immediate feedback on their work and the ability to compare their work with correct answers.
- Access and review each response; manually change grades or leave comments for students to review.

**Student Reporting.** *Connect Business Statistics* keeps instructors informed about how each student, section, and class is performing, allowing for more productive use of lecture and office hours. The progress-tracking function enables you to



Business Statistics: Jaggia 1e  
Section One: MWF 11:30-1:30  
switch section

connect<sup>®</sup>  
BUSINESS STATISTICS

home library reports

assignment results  
report type: Assignment Results

Use the options below to view assignment scores.

show report options & settings

assignment results: Multiple Sections  
report created: 09/09/2012 9:57 AM CDT  
report date range: -  
attempt: Best score style: Percents (Averaged)  
assignment type: Homework, Practice, Quiz, Exam

Select the checkboxes on columns you want to export or print.

export to excel

Section	Assignment 1	Assignment 2	Exam 1
Total Value (Points)	20	25	20
Townsend, Rachel Section One: MWF 1:30-3:30	80%	91.50%	80%
Mann, Becky Section One: MWF 1:30-3:30	85.33%	93%	85%
Dalo, Danielle Section One: MWF 1:30-3:30	89%	91.50%	91%
Billows, Nancy Section One: MWF 1:30-3:30	85.33%	93%	93%

- View scored work immediately and track individual or group performance with assignment and grade reports.
- Access an instant view of student or class performance relative to topic and learning objectives.
- Collect data and generate reports required by many accreditation organizations, such as AACSB.

**Instructor Library.** The *Connect Business Statistics* Instructor Library is your repository for additional resources to improve student engagement in and out of class. You can select and use any asset that enhances your lecture. The *Connect Business Statistics* Instructor Library includes:

- PowerPoint presentations
- Test Bank
- Instructor's Solutions Manual
- Digital Image Library

# to Success in Business Statistics?

**Connect Insight.** *Connect* Insight is *Connect*'s new one-of-a-kind visual analytics dashboard—now available for both instructors and students—that provides at-a-glance information regarding student performance, which is immediately actionable. By presenting assignment, assessment, and topical performance results together with a time metric that is easily visible for aggregate or individual results, *Connect* Insight gives the user the ability to take a just-in-time approach to teaching and learning, which was never before available. *Connect* Insight presents data that empowers students and helps instructors efficiently and effectively improve class performance.

**Mobile.** Students and instructors can now enjoy convenient anywhere, anytime access to *Connect* with a new mobile interface that's been designed for optimal use of tablet functionality. More than just a new way to access *Connect*, users can complete assignments, check progress, study, and read material, with full use of LearnSmart, SmartBook, and *Connect* Insight—*Connect*'s new at-a-glance visual analytics dashboard.

## Tegrity Campus: Lectures 24/7



*Tegrity Campus* is integrated in *Connect* to help make your class time available 24/7. With Tegrity, you can capture each one of your lectures in a searchable format for students to review when they study and complete assignments using *Connect*. With a simple one-click start-and-stop process, you can capture everything that is presented to students during your lecture from your computer, including audio. Students can replay any part of any class with easy-to-use browser-based viewing on a PC or Mac.

Educators know that the more students can see, hear, and experience class resources, the better they learn. In fact, studies prove it. With *Tegrity Campus*, students quickly recall key moments by using *Tegrity Campus*'s unique search feature. This search helps students efficiently find what they need, when they need it, across an entire semester of class recordings. Help turn all your students' study time into learning moments immediately supported by your lecture. To learn more about *Tegrity*, watch a two-minute Flash demo at <http://tegritycampus.mhhe.com>.

# What Software Is Available with This Text?

## **MegaStat® for Microsoft Excel® 2003, 2007, and 2010 (and Excel: Mac 2011)**

**Access Card ISBN: 0077426274** *Note: Best option for both Windows and Mac users.*

**MegaStat®** by J. B. Orris of Butler University is a full-featured Excel add-in that is available through the access card packaged with the text or on the *MegaStat* website at [www.mhhe.com/megastat](http://www.mhhe.com/megastat). It works with Excel 2003, 2007, and 2010 (and Excel: Mac 2011). On the website, students have 10 days to successfully download and install *MegaStat* on their local computer. Once installed, *MegaStat* will remain active in Excel with no expiration date or time limitations. The software performs statistical analyses within an Excel workbook. It does basic functions, such as descriptive statistics, frequency distributions, and probability calculations, as well as hypothesis testing, ANOVA, and regression. *MegaStat* output is carefully formatted, and its ease-of-use features include Auto Expand for quick data selection and Auto Label detect. Since *MegaStat* is easy to use, students can focus on learning statistics without being distracted by the software. *MegaStat* is always available from Excel's main menu. Selecting a menu item pops up a dialog box. Screencam tutorials are included that provide a walkthrough of major business statistics topics. Help files are built in, and an introductory user's manual is also included.



# What Resources Are Available for Instructors?

## Online Course Management

### The Best of Both Worlds



McGraw-Hill Higher Education and Blackboard have teamed up. What does this mean for you?

1. **Single sign-on.** Now you and your students can access McGraw-Hill's *Connect*® and Create™ right from within your Blackboard course—all with one single sign-on.
2. **Deep integration of content and tools.** You get a single sign-on with *Connect* and Create, and you also get integration of McGraw-Hill content and content engines right into Blackboard. Whether you're choosing a book for your course or building *Connect* assignments, all the tools you need are right where you want them—inside of Blackboard.
3. **One grade book.** Keeping several grade books and manually synchronizing grades into Blackboard is no longer necessary. When a student completes an integrated *Connect* assignment, the grade for that assignment automatically (and instantly) feeds your Blackboard grade center.
4. **A solution for everyone.** Whether your institution is already using Blackboard or you just want to try Blackboard on your own, we have a solution for you. McGraw-Hill and Blackboard can now offer you easy access to industry-leading technology and content, whether your campus hosts it or we do. Be sure to ask your local McGraw-Hill representative for details.

# What Resources Are Available for Students?

## CourseSmart

ISBN: 1259335062



CourseSmart is a convenient way to find and buy eTextbooks. CourseSmart has the largest selection of eTextbooks available anywhere, offering thousands of the most commonly adopted textbooks from a wide variety of higher-education publishers. CourseSmart eTextbooks are available in one standard online reader with full text search, notes and highlighting, and e-mail tools for sharing notes between classmates. Visit [www.CourseSmart.com](http://www.CourseSmart.com) for more information on ordering.

## ALEKS



ALEKS is an assessment and learning program that provides individualized instruction in Business Statistics, Business Math, and Accounting. Available online in partnership with McGraw-Hill/Irwin, ALEKS interacts with students much like a skilled human tutor, with the ability to assess precisely a student's knowledge and provide instruction on the exact topics the student is most ready to learn. By providing topics to meet individual students' needs, allowing students to move between explanation and practice, correcting and analyzing errors, and defining terms, ALEKS helps students to master course content quickly and easily.

ALEKS also includes an instructor module with powerful, assignment-driven features and extensive content flexibility. ALEKS simplifies course management and allows instructors to spend less time with administrative tasks and more time directing student learning. To learn more about ALEKS, visit [www.aleks.com](http://www.aleks.com).

# ACKNOWLEDGMENTS

We would like to acknowledge the following people for their help in the development of the first and second editions of *Business Statistics*, as well as the ancillaries and digital content.

John Affisco <i>Hofstra University</i>	Samathy Chandrashekar <i>Salisbury University</i>	Roman Erenshteyn <i>Goldey-Beacom College</i>
Mehdi Afiat <i>College of Southern Nevada</i>	Gary Huaite Chao <i>University of Pennsylvania—Kutztown</i>	Grace Esimai <i>University of Texas—Arlington</i>
Mohammad Ahmadi <i>University of Tennessee—Chattanooga</i>	Sangit Chatterjee <i>Northeastern University</i>	Soheila Fardanesh <i>Towson University</i>
Sung Ahn <i>Washington State University</i>	Anna Chernobai <i>Syracuse University</i>	Carol Flannery <i>University of Texas—Dallas</i>
Mohammad Ahsanullah <i>Rider University</i>	Alan Chesen <i>Wright State University</i>	Sydney Fletcher <i>Mississippi Gulf Coast Community College</i>
Imam Alam <i>University of Northern Iowa</i>	Juyan Cho <i>Colorado State University—Pueblo</i>	Andrew Flight <i>Portland State University</i>
Mostafa Aminzadeh <i>Towson University</i>	Alan Chow <i>University of South Alabama</i>	Samuel Frame <i>Cal Poly San Luis Obispo</i>
Ardavan Asef-Vaziri <i>California State University</i>	Bruce Christensen <i>Weber State University</i>	Priya Francisco <i>Purdue University</i>
Scott Bailey <i>Troy University</i>	Howard Clayton <i>Auburn University</i>	Vickie Fry <i>Westmoreland County Community College</i>
Jayanta Bandyopadhyay <i>Central Michigan University</i>	Robert Collins <i>Marquette University</i>	Ed Gallo <i>Sinclair Community College</i>
Samir Barman <i>University of Oklahoma</i>	M. Halim Dalgin <i>Kutztown University</i>	Glenn Gilbreath <i>Virginia Commonwealth University</i>
Douglas Barrett <i>University of North Alabama</i>	Tom Davis <i>University of Dayton</i>	Robert Gillette <i>University of Kentucky</i>
John Beyers <i>University of Maryland</i>	Matthew Dean <i>University of Maine</i>	Xiaoning Gilliam <i>Texas Tech University</i>
Arnab Bisi <i>Purdue University—West Lafayette</i>	Jason Delaney <i>University of Arkansas—Little Rock</i>	Mark Gius <i>Quinnipiac University</i>
Gary Black <i>University of Southern Indiana</i>	Ferdinand DiFurio <i>Tennessee Tech University</i>	Malcolm Gold <i>Saint Mary's University of Minnesota</i>
Randy Boan <i>Aims Community College</i>	Matt Dobra <i>UMUC</i>	Michael Gordinier <i>Washington University</i>
Matthew Bognar <i>University of Iowa</i>	Luca Donno <i>University of Miami</i>	Deborah Gougeon <i>University of Scranton</i>
Juan Cabrera <i>Ramapo College of New Jersey</i>	Joan Donohue <i>University of South Carolina</i>	Don Gren <i>Salt Lake Community College</i>
Scott Callan <i>Bentley University</i>	David Doorn <i>University of Minnesota</i>	Robert Hammond <i>North Carolina State University</i>
Gregory Cameron <i>Brigham Young University</i>	James Dunne <i>University of Dayton</i>	Jim Han <i>Florida Atlantic University</i>
Kathleen Campbell <i>St. Joseph's University</i>	Mike Easley <i>University of New Orleans</i>	Elizabeth Haran <i>Salem State University</i>
Alan Cannon <i>University of Texas—Arlington</i>	Erick Elder <i>University of Arkansas—Little Rock</i>	Jack Harshbarger <i>Montreat College</i>
Michael Cervetti <i>University of Memphis</i>	Ashraf ElHoubi <i>Lamar University</i>	

Edward Hartono <i>University of Alabama— Huntsville</i>	Shari Lawrence <i>Nicholls State University</i>	Khosrow Moshirvaziri <i>California State University— Long Beach</i>
Clifford Hawley <i>West Virginia University</i>	Radu Lazar <i>University of Maryland</i>	Tariq Mughal <i>University of Utah</i>
Paul Hong <i>University of Toledo</i>	David Leupp <i>University of Colorado— Colorado Springs</i>	Patricia Mullins <i>University of Wisconsin— Madison</i>
Ping-Hung Hsieh <i>Oregon State University</i>	Carel Ligeon <i>Auburn University— Montgomery</i>	Kusum Mundra <i>Rutgers University—Newark</i>
Marc Isaacson <i>Augsburg College</i>	Carin Lightner <i>North Carolina A&amp;T State University</i>	Anthony Narsing <i>Macon State College</i>
Mohammad Jamal <i>Northern Virginia Community College</i>	Constance Lightner <i>Fayetteville State University</i>	Robert Nauss <i>University of Missouri— St. Louis</i>
Robin James <i>Harper College</i>	Scott Lindsey <i>Dixie State College of Utah</i>	Satish Nayak <i>University of Missouri— St. Louis</i>
Molly Jensen <i>University of Arkansas</i>	Ken Linna <i>Auburn University— Montgomery</i>	Thang Nguyen <i>California State University— Long Beach</i>
Craig Johnson <i>Brigham Young University— Idaho</i>	Andy Litteral <i>University of Richmond</i>	Mohammad Oskoorouchi <i>California State University— San Marcos</i>
Janine Sanders Jones <i>University of St. Thomas</i>	Jun Liu <i>Georgia Southern University</i>	Barb Osyk <i>University of Akron</i>
Vivian Jones <i>Bethune—Cookman University</i>	Chung-Ping Loh <i>University of North Florida</i>	Scott Paulsen <i>Illinois Central College</i>
Jerzy Kamburowski <i>University of Toledo</i>	Salvador Lopez <i>University of West Georgia</i>	James Payne <i>Calhoun Community College</i>
Howard Kaplon <i>Towson University</i>	John Loucks <i>St. Edward's University</i>	Norman Pence <i>Metropolitan State College of Denver</i>
Krishna Kasibhatla <i>North Carolina A&amp;T State University</i>	Cecilia Maldonado <i>Georgia Southwestern State University</i>	Dane Peterson <i>Missouri State University</i>
Mohammad Kazemi <i>University of North Carolina—Charlotte</i>	Farooq Malik <i>University of Southern Mississippi</i>	Joseph Petry <i>University of Illinois— Urbana/Champaign</i>
Ken Kelley <i>University of Notre Dame</i>	Ken Mayer <i>University of Nebraska— Omaha</i>	Courtney Pham <i>Missouri State University</i>
Lara Khansa <i>Virginia Tech</i>	Bradley McDonald <i>Northern Illinois University</i>	Martha Pilcher <i>University of Washington</i>
Ronald Klimberg <i>St. Joseph's University</i>	Elaine McGivern <i>Duquesne University</i>	Cathy Poliak <i>University of Wisconsin— Milwaukee</i>
Andrew Koch <i>James Madison University</i>	John McKenzie <i>Babson University</i>	Simcha Pollack <i>St. John's University</i>
Subhash Kochar <i>Portland State University</i>	Norbert Michel <i>Nicholls State University</i>	Hamid Pourmohammadi <i>California State University— Dominguez Hills</i>
Brandon Koford <i>Weber University</i>	John Miller <i>Sam Houston State University</i>	Tammy Prater <i>Alabama State University</i>
Randy Kolb <i>St. Cloud State University</i>	Virginia Miori <i>St. Joseph's University</i>	Manying Qiu <i>Virginia State University</i>
Vadim Kutsyy <i>San Jose State University</i>	Prakash Mirchandani <i>University of Pittsburgh</i>	Troy Quast <i>Sam Houston State University</i>
Francis Laatsch <i>University of Southern Mississippi</i>	Jason Moliterno <i>Sacred Heart University</i>	Michael Racer <i>University of Memphis</i>
David Larson <i>University of South Alabama</i>	Elizabeth Moliski <i>University of Texas—Austin</i>	Srikant Raghavan <i>Lawrence Technological University</i>
John Lawrence <i>California State University— Fullerton</i>	Joseph Mollick <i>Texas A&amp;M University— Corpus Christi</i>	
	James Moran <i>Oregon State University</i>	

Bharatendra Rai	Harvey Singer	Shawn Ulrick
<i>University of Massachusetts—</i>	<i>George Mason University</i>	<i>Georgetown University</i>
<i>Dartmouth</i>	Harry Sink	Bulent Uyar
Tony Ratcliffe	<i>North Carolina A&amp;T State</i>	<i>University of</i>
<i>James Madison University</i>	<i>University</i>	<i>Northern Iowa</i>
David Ravetch	Don Skousen	Ahmad Vakil
<i>University of California</i>	<i>Salt Lake Community College</i>	<i>Tobin College of Business</i>
Bruce Reinig	Robert Smidt	Raja Velu
<i>San Diego State University</i>	<i>California Polytechnic State</i>	<i>Syracuse University</i>
Darlene Riedemann	<i>University</i>	Holly Verhasselt
<i>Eastern Illinois University</i>	Gary Smith	<i>University of</i>
David Roach	<i>Florida State University</i>	<i>Houston—Victoria</i>
<i>Arkansas Tech University</i>	Antoinette Somers	Zhaowei Wang
Carolyn Rochelle	<i>Wayne State University</i>	<i>Citizen's Bank</i>
<i>East Tennessee State University</i>	Ryan Songstad	Rachel Webb
Alfredo Romero	<i>Augustana College</i>	<i>Portland State University</i>
<i>North Carolina A&amp;T State</i>	Erland Sorensen	Kyle Wells
<i>University</i>	<i>Bentley University</i>	<i>Dixie State College</i>
Ann Rothermel	Arun Kumar Srinivasan	Alan Wheeler
<i>University of Akron</i>	<i>Indiana University—</i>	<i>University of</i>
Jeff Rummel	<i>Southeast</i>	<i>Missouri—St. Louis</i>
<i>Emory University</i>	Scott Stevens	Mary Whiteside
Deborah Rumsey	<i>James Madison University</i>	<i>University of</i>
<i>The Ohio State University</i>	Alicia Strandberg	<i>Texas—Arlington</i>
Stephen Russell	<i>Temple University</i>	Blake Whitten
<i>Weber State University</i>	Linda Sturges	<i>University of Iowa</i>
William Rybolt	<i>Suny Maritime College</i>	Rick Wing
<i>Babson College</i>	Wendi Sun	<i>San Francisco State</i>
Fati Salimian	<i>Rockland Trust</i>	<i>University</i>
<i>Salisbury University</i>	Bedassa Tadesse	Jan Wolcott
Fatollah Salimian	<i>University of Minnesota</i>	<i>Wichita State University</i>
<i>Perdue School of Business</i>	Pandu Tadikamalla	Rongning Wu
Samuel Sarri	<i>University of Pittsburgh</i>	<i>Baruch College</i>
<i>College of Southern Nevada</i>	Roberto Duncan Tarabay	John Yarber
Jim Schmidt	<i>University of</i>	<i>Northeast Mississippi</i>
<i>University of Nebraska—</i>	<i>Wisconsin—Madison</i>	<i>Community College</i>
<i>Lincoln</i>	Faye Teer	Mark Zaporowski
Patrick Scholten	<i>James Madison University</i>	<i>Canisius College</i>
<i>Bentley University</i>	Deborah Tesch	Ali Zargar
Bonnie Schroeder	<i>Xavier University</i>	<i>San Jose State University</i>
<i>Ohio State University</i>	Patrick Thompson	Dewit Zerom
Pali Sen	<i>University of Florida</i>	<i>California State University</i>
<i>University of North</i>	Satish Thosar	Eugene Zhang
<i>Florida</i>	<i>University of Redlands</i>	<i>Midwestern State University</i>
Donald Sexton	Ricardo Tovar-Silos	Ye Zhang
<i>Columbia University</i>	<i>Lamar University</i>	<i>Indiana University—Purdue</i>
Vijay Shah	Quoc Hung Tran	<i>University—Indianapolis</i>
<i>West Virginia</i>	<i>Bridgewater State University</i>	Yi Zhang
<i>University—Parkersburg</i>	Elzbieta Trybus	<i>California State</i>
Dmitriy Shaltayev	<i>California State</i>	<i>University—Fullerton</i>
<i>Christopher Newport</i>	<i>University—Northridge</i>	Yulin Zhang
<i>University</i>	Fan Tseng	<i>San Jose State University</i>
Soheil Sibdari	<i>University of</i>	Wencang Zhou
<i>University of Massachusetts—</i>	<i>Alabama—Huntsville</i>	<i>Baruch College</i>
<i>Dartmouth</i>	Silvanus Udoka	Zhen Zhu
Prodosh Simlai	<i>North Carolina A&amp;T State</i>	<i>University of Central</i>
<i>University of North Dakota</i>	<i>University</i>	<i>Oklahoma</i>

The editorial staff of McGraw-Hill/Irwin are deserving of our gratitude for their guidance throughout this project, especially Christina Holt, Dolly Womack, Doug Ruby, Harvey Yep, Bruce Gin, and Srdjan Savanovic.

# BRIEF CONTENTS

## PART ONE

Introduction

**CHAPTER 1** Statistics and Data 2

## PART TWO

Descriptive Statistics

**CHAPTER 2** Tabular and Graphical Methods 16

**CHAPTER 3** Numerical Descriptive Measures 58

## PART THREE

Probability and Probability Distributions

**CHAPTER 4** Introduction to Probability 106

**CHAPTER 5** Discrete Probability Distributions 150

**CHAPTER 6** Continuous Probability Distributions 190

## PART FOUR

Basic Inference

**CHAPTER 7** Sampling and Sampling Distributions 230

**CHAPTER 8** Interval Estimation 268

**CHAPTER 9** Hypothesis Testing 300

**CHAPTER 10** Statistical Inference Concerning Two Populations 338

**CHAPTER 11** Statistical Inference Concerning Variance 374

**CHAPTER 12** Chi-Square Tests 402

## PART FIVE

Advanced Inference

**CHAPTER 13** Analysis of Variance 432

**CHAPTER 14** Regression Analysis 476

**CHAPTER 15** Inference with Regression Models 514

**CHAPTER 16** Regression Models for Nonlinear Relationships 556

**CHAPTER 17** Regression Models with Dummy Variables 588

## PART SIX

Supplementary Topics

**CHAPTER 18** Time Series and Forecasting 622

**CHAPTER 19** Returns, Index Numbers, and Inflation 662

**CHAPTER 20** Nonparametric Tests 686

## APPENDIXES

**APPENDIX A** Tables 730

**APPENDIX B** Answers to Selected Even-Numbered Exercises 743

Glossary G-1

Photo Credits PC-1

Index I-1

## PART ONE

### Introduction

#### CHAPTER 1

### STATISTICS AND DATA 2

- 1.1 The Relevance of Statistics** 4
- 1.2 What Is Statistics?** 5
  - The Need for Sampling 6
  - Types of Data 6
  - Getting Started on the Web 7
- 1.3 Variables and Scales of Measurement** 8
  - The Nominal Scale 9
  - The Ordinal Scale 10
  - The Interval Scale 12
  - The Ratio Scale 12
  - Synopsis of Introductory Case 13
  - Conceptual Review** 14

## PART TWO

### Descriptive Statistics

#### CHAPTER 2

### TABULAR AND GRAPHICAL METHODS 16

- 2.1 Summarizing Qualitative Data** 18
  - Visualizing Frequency Distributions for Qualitative Data 19
  - Using Excel to Construct a Pie Chart 21
  - Using Excel to Construct a Bar Chart 21
  - Cautionary Comments When Constructing or Interpreting Charts or Graphs 22
- 2.2 Summarizing Quantitative Data** 25
  - Guidelines for Constructing a Frequency Distribution 26
  - Visualizing Frequency Distributions for Quantitative Data 30
  - Using Excel to Construct a Histogram 31
    - Constructing a Histogram from a Set of Raw Data 32
    - Constructing a Histogram from a Frequency Distribution 33
    - Using Excel to Construct a Polygon 34
    - Using Excel to Construct an Ogive 36
  - Synopsis of Introductory Case 37
- 2.3 Stem-and-Leaf Diagrams** 41
- 2.4 Scatterplots** 43
  - Using Excel to Construct a Scatterplot 45

Writing with Statistics 46

Conceptual Review 48

**Additional Exercises and Case Studies** 49

Exercises 49

Case Studies 52

#### Appendix 2.1: Guidelines for Other Software Packages 54

#### CHAPTER 3

### NUMERICAL DESCRIPTIVE MEASURES 58

- 3.1 Measures of Central Location** 60
    - The Mean 60
    - The Median 61
    - The Mode 63
    - Using Excel to Calculate Measures of Central Location 64
      - Excel's Formula Option 64
      - Excel's Data Analysis Toolpak Option 65
      - The Weighted Mean 66
  - 3.2 Percentiles and Box Plots** 69
    - Calculating the  $p$ th Percentile 69
    - Constructing and Interpreting a Box Plot 70
  - 3.3 The Geometric Mean** 73
    - The Geometric Mean Return 73
    - Arithmetic Mean versus Geometric Mean 74
    - The Average Growth Rate 74
  - 3.4 Measures of Dispersion** 77
    - Range 77
    - The Mean Absolute Deviation 77
    - The Variance and the Standard Deviation 78
    - The Coefficient of Variation 80
    - Using Excel to Calculate Measures of Dispersion 80
      - Excel's Formula Option 80
      - Excel's Data Analysis Toolpak Option 81
    - Synopsis of Introductory Case 81
  - 3.5 Mean-Variance Analysis and the Sharpe Ratio** 83
  - 3.6 Analysis of Relative Location** 85
    - Chebyshev's Theorem 85
    - The Empirical Rule 86
    - z-Scores 87
  - 3.7 Summarizing Grouped Data** 89
  - 3.8 Covariance and Correlation** 92
    - Using Excel to Calculate Covariance and the Correlation Coefficient 94
    - Writing with Statistics 96
    - Conceptual Review** 97
    - Additional Exercises and Case Studies** 99
    - Exercises 99
    - Case Studies 102
- #### Appendix 3.1: Guidelines for Other Software Packages 104



## PART THREE

### Probability and Probability Distributions

#### CHAPTER 4

### INTRODUCTION TO PROBABILITY 106

- 4.1 Fundamental Probability Concepts 108**
  - Events 108
  - Assigning Probabilities 111
  - Probabilities Expressed as Odds 113
- 4.2 Rules of Probability 117**
  - The Complement Rule 117
  - The Addition Rule 117
    - The Addition Rule for Mutually Exclusive Events 119
  - Conditional Probability 119
  - Independent and Dependent Events 121
  - The Multiplication Rule 122
    - The Multiplication Rule for Independent Events 122
- 4.3 Contingency Tables and Probabilities 126**
  - Synopsis of Introductory Case 129
- 4.4 The Total Probability Rule and Bayes' Theorem 131**
  - The Total Probability Rule 131
  - Bayes' Theorem 134
- 4.5 Counting Rules 138**
  - Writing with Statistics 141
  - Conceptual Review 142**
  - Additional Exercises and Case Studies 144**
    - Exercises 144
    - Case Studies 148

#### CHAPTER 5

### DISCRETE PROBABILITY DISTRIBUTIONS 150

- 5.1 Random Variables and Discrete Probability Distributions 152**
  - The Discrete Probability Distribution 153
- 5.2 Expected Value, Variance, and Standard Deviation 157**
  - Expected Value 158
  - Variance and Standard Deviation 158
  - Risk Neutrality and Risk Aversion 159
- 5.3 Portfolio Returns 162**
  - Properties of Random Variables 162
  - Expected Return, Variance, and Standard Deviation of Portfolio Returns 163
- 5.4 The Binomial Distribution 166**
  - Using Excel to Obtain Binomial Probabilities 171
- 5.5 The Poisson Distribution 173**
  - Using Excel to Obtain Poisson Probabilities 176
  - Synopsis of Introductory Case 177
- 5.6 The Hypergeometric Distribution 178**
  - Using Excel to Obtain Hypergeometric Probabilities 180
  - Writing with Statistics 182
  - Conceptual Review 184**

### Additional Exercises and Case Studies 185

Exercises 185

Case Studies 187

### Appendix 5.1: Guidelines for Other Software Packages 188

#### CHAPTER 6

### CONTINUOUS PROBABILITY DISTRIBUTIONS 190

- 6.1 Continuous Random Variables and the Uniform Distribution 192**
  - The Continuous Uniform Distribution 193
- 6.2 The Normal Distribution 196**
  - Characteristics of the Normal Distribution 196
  - The Standard Normal Variable 198
  - Finding a Probability for a Given z Value 198
  - Finding a z Value for a Given Probability 201
  - Revisiting the Empirical Rule 202
- 6.3 Solving Problems with Normal Distributions 205**
  - The Transformation of Normal Random Variables 205
  - The Inverse Transformation 207
  - Using Excel for the Normal Distribution 209
    - The Standard Transformation 209
    - The Inverse Transformation 209
  - A Note on the Normal Approximation of the Binomial Distribution 209
  - Synopsis of Introductory Case 210
- 6.4 Other Continuous Probability Distributions 213**
  - The Exponential Distribution 213
  - Using Excel for the Exponential Distribution 215
  - The Lognormal Distribution 216
  - Using Excel for the Lognormal Distribution 218
    - The Standard Transformation 218
    - The Inverse Transformation 218
  - Writing with Statistics 220
  - Conceptual Review 222**
  - Additional Exercises and Case Studies 223**
    - Exercises 223
    - Case Studies 225

### Appendix 6.1: Guidelines for Other Software Packages 227

## PART FOUR

### Basic Inference

#### CHAPTER 7

### SAMPLING AND SAMPLING DISTRIBUTIONS 230

- 7.1 Sampling 232**
  - Classic Case of a "Bad" Sample: The *Literary Digest* Debacle of 1936 232
  - Sampling Methods 233
  - The Special Election to Fill Ted Kennedy's Senate Seat 235
- 7.2 The Sampling Distribution of the Sample Mean 237**
  - The Expected Value and the Standard Error of the Sample Mean 238
  - Sampling from a Normal Population 239
  - The Central Limit Theorem 240

<b>7.3</b>	<b>The Sampling Distribution of the Sample Proportion</b>	244
	The Expected Value and the Standard Error of the Sample Proportion	244
	Synopsis of Introductory Case	247
<b>7.4</b>	<b>The Finite Population Correction Factor</b>	248
<b>7.5</b>	<b>Statistical Quality Control</b>	251
	Control Charts	252
	Using Excel to Create a Control Chart	255
	Writing with Statistics	257
	<b>Conceptual Review</b>	259
	<b>Additional Exercises and Case Studies</b>	260
	Exercises	260
	Case Studies	263
<b>Appendix 7.1:</b>	<b>Derivation of the Mean and the Variance for <math>\bar{X}</math> and <math>\bar{P}</math></b>	264
	Sample Mean, $\bar{X}$	264
	Sample Proportion, $\bar{P}$	264
<b>Appendix 7.2:</b>	<b>Properties of Point Estimators</b>	264
<b>Appendix 7.3:</b>	<b>Guidelines for Other Software Packages</b>	266

## CHAPTER 8

### INTERVAL ESTIMATION 268

<b>8.1</b>	<b>Confidence Interval for the Population Mean When <math>\sigma</math> Is Known</b>	270
	Constructing a Confidence Interval for $\mu$ When $\sigma$ Is Known	271
	The Width of a Confidence Interval	273
	Using Excel to Construct a Confidence Interval for $\mu$ When $\sigma$ Is Known	275
<b>8.2</b>	<b>Confidence Interval for the Population Mean When <math>\sigma</math> Is Unknown</b>	277
	The $t$ Distribution	277
	Summary of the $t_{df}$ Distribution	278
	Locating $t_{df}$ Values and Probabilities	278
	Constructing a Confidence Interval for $\mu$ When $\sigma$ Is Unknown	280
	Using Excel to Construct a Confidence Interval for $\mu$ When $\sigma$ Is Unknown	281
<b>8.3</b>	<b>Confidence Interval for the Population Proportion</b>	284
<b>8.4</b>	<b>Selecting the Required Sample Size</b>	287
	Selecting $n$ to Estimate $\mu$	287
	Selecting $n$ to Estimate $p$	288
	Synopsis of Introductory Case	289
	Writing with Statistics	291
	<b>Conceptual Review</b>	292
	<b>Additional Exercises and Case Studies</b>	294
	Exercises	294
	Case Studies	297

<b>Appendix 8.1:</b>	<b>Guidelines for Other Software Packages</b>	298
----------------------	---	-----

## CHAPTER 9

### HYPOTHESIS TESTING 300

<b>9.1</b>	<b>Introduction to Hypothesis Testing</b>	302
	The Decision to “Reject” or “Not Reject” the Null Hypothesis	302

Defining the Null and the Alternative Hypotheses	303
Type I and Type II Errors	305

<b>9.2</b>	<b>Hypothesis Test for the Population Mean When <math>\sigma</math> Is Known</b>	307
	The $p$ -Value Approach	308
	The Critical Value Approach	312
	Confidence Intervals and Two-Tailed Hypothesis Tests	315
	Using Excel to Test $\mu$ When $\sigma$ Is Known	316
	One Last Remark	317
<b>9.3</b>	<b>Hypothesis Test for the Population Mean When <math>\sigma</math> Is Unknown</b>	319
	Using Excel to Test $\mu$ When $\sigma$ Is Unknown	321
	Synopsis of Introductory Case	322
<b>9.4</b>	<b>Hypothesis Test for the Population Proportion</b>	325
	Writing with Statistics	330
	<b>Conceptual Review</b>	331
	<b>Additional Exercises and Case Studies</b>	333
	Exercises	333
	Case Studies	335
<b>Appendix 9.1:</b>	<b>Guidelines for Other Software Packages</b>	336

## CHAPTER 10

### STATISTICAL INFERENCE CONCERNING TWO POPULATIONS 338

<b>10.1</b>	<b>Inference Concerning the Difference between Two Means</b>	340
	Confidence Interval for $\mu_1 - \mu_2$	340
	Hypothesis Test for $\mu_1 - \mu_2$	342
	Using Excel for Testing Hypotheses about $\mu_1 - \mu_2$	344
	A Note on the Assumption of Normality	346
<b>10.2</b>	<b>Inference Concerning Mean Differences</b>	351
	Recognizing a Matched-Pairs Experiment	351
	Confidence Interval for $\mu_D$	351
	Hypothesis Test for $\mu_D$	352
	Using Excel for Testing Hypotheses about $\mu_D$	354
	One Last Note on the Matched-Pairs Experiment	355
	Synopsis of Introductory Case	356
<b>10.3</b>	<b>Inference Concerning the Difference between Two Proportions</b>	359
	Confidence Interval for $p_1 - p_2$	360
	Hypothesis Test for $p_1 - p_2$	361
	Writing with Statistics	366
	<b>Conceptual Review</b>	367
	<b>Additional Exercises and Case Studies</b>	368
	Exercises	368
	Case Studies	371

<b>Appendix 10.1:</b>	<b>Guidelines for Other Software Packages</b>	372
-----------------------	---	-----

## CHAPTER 11

### STATISTICAL INFERENCE CONCERNING VARIANCE 374

<b>11.1</b>	<b>Inference Concerning the Population Variance</b>	376
	Sampling Distribution of $S^2$	376
	Locating $\chi^2_{df}$ Values and Probabilities	377
	Confidence Interval for the Population Variance	379

Hypothesis Test for the Population Variance 380

Using Excel to Calculate  $p$ -Values 382

## 11.2 Inference Concerning the Ratio of Two Population Variances 384

Sampling Distribution of  $S_1^2/S_2^2$  385

Locating  $F_{(df_1, df_2)}$  Values and Probabilities 386

Confidence Interval for the Ratio of Two Population Variances 388

Hypothesis Test for the Ratio of Two Population Variances 388

Using Excel to Calculate the  $p$ -Value for the  $F_{(df_1, df_2)}$  Test Statistic 390

Excel's FDIST. RT Function 390

Excel's FTEST Function 391

Synopsis of Introductory Case 391

Writing with Statistics 394

**Conceptual Review** 395

**Additional Exercises and Case Studies** 396

Exercises 396

Case Studies 399

### Appendix 11.1: Guidelines for Other Software Packages 400

## CHAPTER 12

## CHI-SQUARE TESTS 402

### 12.1 Goodness-of-Fit Test for a Multinomial Experiment 404

Using Excel to Calculate  $p$ -Values 406

### 12.2 Chi-Square Test for Independence 410

Calculating Expected Frequencies 411

Synopsis of Introductory Case 414

### 12.3 Chi-Square Test for Normality 416

The Goodness-of-Fit Test for Normality 416

The Jarque-Bera Test 419

Writing with Statistics 422

**Conceptual Review** 424

**Additional Exercises and Case Studies** 425

Exercises 425

Case Studies 428

### Appendix 12.1: Guidelines for Other Software Packages 430

## PART FIVE

## Advanced Inference

## CHAPTER 13

## ANALYSIS OF VARIANCE 432

### 13.1 One-Way ANOVA 434

Between-Treatments Estimate of  $\sigma^2$  435

Within-Treatments Estimate of  $\sigma^2$  436

The One-Way ANOVA Table 437

Using Excel for a One-Way ANOVA Test 437

### 13.2 Multiple Comparison Methods 442

Fisher's Least Significant Difference (LSD) Method 442

Tukey's Honestly Significant Differences (HSD) Method 444

Synopsis of Introductory Case 447

### 13.3 Two-Way ANOVA: No Interaction 450

The Sum of Squares for Factor A, SSA 452

The Sum of Squares for Factor B, SSB 452

The Sum of Squares due to Error, SSE 452

Using Excel to Solve a Two-Way ANOVA Test without Interaction 453

### 13.4 Two-Way ANOVA: With Interaction 458

The Total Sum of Squares, SST 459

The Sum of Squares for Factor A, SSA, and the

Sum of Squares for Factor B, SSB 459

The Sum of Squares for the Interaction of Factor A and Factor B, SSAB 459

The Sum of Squares due to Error, SSE 460

Using Excel to Solve a Two-Way ANOVA Test with Interaction 460

Writing with Statistics 464

**Conceptual Review** 465

**Additional Exercises and Case Studies** 467

Case Studies 472

### Appendix 13.1: Guidelines for Other Software Packages 473

## CHAPTER 14

## REGRESSION ANALYSIS 476

### 14.1 The Covariance and the Correlation Coefficient 478

Using Excel to Calculate the Covariance and the Correlation Coefficient 480

Testing the Correlation Coefficient 480

Limitations of Correlation Analysis 481

### 14.2 The Simple Linear Regression Model 483

Determining the Sample Regression Equation 485

Using Excel to Construct a Scatterplot and a Trendline 486

Using Excel to Find the Sample Regression Equation 488

### 14.3 The Multiple Linear Regression Model 492

Determining the Sample Regression

Equation 492

### 14.4 Goodness-of-Fit Measures 497

The Standard Error of the Estimate 497

The Coefficient of Determination,  $R^2$  500

The Adjusted  $R^2$  502

Synopsis of Introductory Case 503

Writing with Statistics 506

**Conceptual Review** 507

**Additional Exercises and Case Studies** 509

Case Studies 511

### Appendix 14.1: Guidelines for Other Software Packages 513

## CHAPTER 15

## INFERENCE WITH REGRESSION MODELS 514

### 15.1 Tests of Significance 516

Tests of Individual Significance 516

Using a Confidence Interval to Determine Individual Significance 518

A Test for a Nonzero Slope Coefficient 519

Test of Joint Significance 521

Reporting Regression Results 522

Synopsis of Introductory Case 523

<b>15.2</b>	<b>A General Test of Linear Restrictions</b>	527
<b>15.3</b>	<b>Interval Estimates for the Response Variable</b>	532
<b>15.4</b>	<b>Model Assumptions and Common Violations</b>	537
	Common Violation 1: Nonlinear Patterns	538
	Detection	538
	Remedy	539
	Common Violation 2: Multicollinearity	540
	Detection	540
	Remedy	541
	Common Violation 3: Changing Variability	541
	Detection	541
	Remedy	542
	Common Violation 4: Correlated Observations	542
	Detection	543
	Remedy	544
	Common Violation 5: Excluded Variables	544
	Remedy	544
	Summary	544
	Writing with Statistics	546
	<b>Conceptual Review</b>	548
	<b>Additional Exercises and Case Studies</b>	550
	Exercises	550
	Case Studies	552

**Appendix 15.1: Guidelines for Other Software Packages** 554

**CHAPTER 16**

**REGRESSION MODELS FOR NONLINEAR RELATIONSHIPS** 556

<b>16.1</b>	<b>Polynomial Regression Models</b>	558
<b>16.2</b>	<b>Regression Models with Logarithms</b>	567
	A Log-Log Model	568
	The Logarithmic Model	569
	The Exponential Model	570
	Comparing Linear and Log-Transformed Models	574
	Synopsis of Introductory Case	575
	Writing with Statistics	578
	<b>Conceptual Review</b>	580
	<b>Additional Exercises and Case Studies</b>	581
	Exercises	581
	Case Studies	583

**Appendix 16.1: Guidelines for Other Software Packages** 585

**CHAPTER 17**

**REGRESSION MODELS WITH DUMMY VARIABLES** 588

<b>17.1</b>	<b>Dummy Variables</b>	590
	Qualitative Variables with Two Categories	590
	Qualitative Variables with Multiple Categories	593
<b>17.2</b>	<b>Interactions with Dummy Variables</b>	599
	Synopsis of Introductory Case	603
<b>17.3</b>	<b>Binary Choice Models</b>	605
	The Linear Probability Model	606
	The Logit Model	607
	Writing with Statistics	613

<b>Conceptual Review</b>	614
<b>Additional Exercises and Case Studies</b>	615
Exercises	615
Case Studies	618

**Appendix 17.1: Guidelines for Other Software Packages** 620

**PART SIX**  
**Supplementary Topics**

**CHAPTER 18**

**TIME SERIES AND FORECASTING** 622

<b>18.1</b>	<b>Choosing a Forecasting Model</b>	624
	Forecasting Methods	624
	Model Selection Criteria	625
<b>18.2</b>	<b>Smoothing Techniques</b>	626
	Moving Average Methods	626
	Exponential Smoothing Methods	628
	Using Excel for Moving Averages and Exponential Smoothing	631
<b>18.3</b>	<b>Trend Forecasting Models</b>	633
	The Linear Trend	633
	The Exponential Trend	634
	Polynomial Trends	637
<b>18.4</b>	<b>Trend and Seasonality</b>	640
	Decomposition Analysis	640
	Extracting Seasonality	641
	Extracting Trend	643
	Forecasting with Decomposition Analysis	644
	Seasonal Dummy Variables	645
	Synopsis of Introductory Case	647
<b>18.5</b>	<b>Causal Forecasting Methods</b>	650
	Lagged Regression Models	650
	Writing with Statistics	653
	<b>Conceptual Review</b>	655
	<b>Additional Exercises and Case Studies</b>	657
	Exercises	657
	Case Studies	659

**Appendix 18.1: Guidelines for Other Software Packages** 660

**CHAPTER 19**

**RETURNS, INDEX NUMBERS, AND INFLATION** 662

<b>19.1</b>	<b>Investment Return</b>	664
	The Adjusted Closing Price	665
	Nominal versus Real Rates of Return	666
<b>19.2</b>	<b>Index Numbers</b>	668
	Simple Price Indices	668
	Unweighted Aggregate Price Index	670
	Weighted Aggregate Price Index	671
	Synopsis of Introductory Case	674
<b>19.3</b>	<b>Using Price Indices to Deflate a Time Series</b>	676
	Inflation Rate	678
	Writing with Statistics	681
	<b>Conceptual Review</b>	682

## CHAPTER 20

### NONPARAMETRIC TESTS 686

#### 20.1 Testing a Population Median 688

The Wilcoxon Signed-Rank Test for a Population Median 688

Using a Normal Distribution Approximation for  $T$  691

#### 20.2 Testing Two Population Medians 693

The Wilcoxon Signed-Rank Test for a Matched-Pairs Sample 694

Using the Computer for the Wilcoxon Signed-Rank Test 695

The Wilcoxon Rank-Sum Test for Independent Samples 695

Using a Normal Distribution Approximation for  $W$  697

Using the Computer for the Wilcoxon Rank-Sum Test 698

#### 20.3 Testing Three or More Population Medians 701

The Kruskal-Wallis Test 701

Using the Computer for the Kruskal-Wallis Test 703

#### 20.4 Testing the Correlation between Two Variables 705

Using a Normal Distribution Approximation for  $r_s$  707

Summary of Parametric and Nonparametric Tests 708

Synopsis of Introductory Case 709

#### 20.5 The Sign Test 711

#### 20.6 Tests Based on Runs 715

The Method of Runs Above and Below the Median 716

Using the Computer for the Runs Test 718

Writing with Statistics 719

**Conceptual Review** 721

**Additional Exercises and Case Studies** 722

Exercises 722

Case Studies 725

#### Appendix 20.1: Guidelines for Other Software Packages 726

## APPENDICES

**APPENDIX A** Tables 730

**APPENDIX B** Answers to Selected Even-Numbered Exercises 743

Glossary G-1

Photo Credits PC-1

Index I-1

# BUSINESS STATISTICS

# 1

## LEARNING OBJECTIVES

**After reading this chapter  
you should be able to:**

- LO 1.1 Describe the importance of statistics.**
- LO 1.2 Differentiate between descriptive statistics and inferential statistics.**
- LO 1.3 Explain the need for sampling and discuss various data types.**
- LO 1.4 Describe variables and various types of measurement scales.**

# Statistics and Data

Every day we are bombarded with data and claims. The analysis of data and the conclusions made from data are part of the field of statistics. A proper understanding of statistics is essential in understanding more of the real world around us, including business, sports, politics, health, social interactions—just about any area of contemporary human activity. In this first chapter, we will differentiate between sound statistical conclusions and questionable conclusions. We will also introduce some important terms, which are referenced throughout the text, that will help us describe different aspects of statistics and their practical importance. You are probably familiar with some of these terms already, from reading or hearing about opinion polls, surveys, and the all-pervasive product ads. Our goal is to place what you already know about these uses of statistics within a framework that we then use for explaining where they came from and what they really mean. A major portion of this chapter is also devoted to a discussion of variables and various types of measurement scales. As we will see in later chapters, we need to distinguish between different variables and measurement scales in order to choose the appropriate statistical methods for analyzing data.





## INTRODUCTORY CASE

### Tween Survey

Luke McCaffrey owns a ski resort two hours outside Boston, Massachusetts, and is in need of a new marketing manager. He is a fairly tough interviewer and believes that the person in this position should have a basic understanding of data fundamentals, including some background with statistical methods. Luke is particularly interested in serving the needs of the “tween” population (children aged 8 to 12 years old). He believes that tween spending power has grown over the past few years, and he wants their skiing experience to be memorable so that they want to return. At the end of last year’s ski season, Luke asked 20 tweens four specific questions.

Q1. On your car drive to the resort, which radio station was playing?

Q2. On a scale of 1 to 4, rate the quality of the food at the resort (where 1 is poor, 2 is fair, 3 is good, and 4 is excellent).

Q3. Presently, the main dining area closes at 3:00 pm. What time do you think it should close?

Q4. How much of your own money did you spend at the lodge today?

The responses to these questions are shown in Table 1.1

**TABLE 1.1** Tween Responses to Skylark Valley Resort Survey

Tween	Q1	Q2	Q3	Q4	Tween	Q1	Q2	Q3	Q4
1	JAMN94.5	4	5:00 pm	20	11	JAMN94.5	3	3:00 pm	0
2	MIX104.1	2	5:00 pm	10	12	JAMN94.5	4	4:00 pm	5
3	KISS108	2	4:30 pm	10	13	KISS108	2	4:30 pm	5
4	JAMN94.5	3	4:00 pm	0	14	KISS108	2	5:00 pm	10
5	KISS108	1	3:30 pm	0	15	KISS108	3	4:00 pm	5
6	JAMN94.5	1	6:00 pm	25	16	JAMN94.5	3	6:00 pm	20
7	KISS108	2	6:00 pm	15	17	KISS108	2	5:00 pm	15
8	KISS108	3	5:00 pm	10	18	MIX104.1	4	6:00 pm	15
9	KISS108	2	4:30 pm	10	19	KISS108	1	5:00 pm	25
10	KISS108	3	4:30 pm	20	20	KISS108	2	4:30 pm	10

Luke asks each job applicant to use the information to:

1. Summarize the results of the survey.
2. Provide management with suggestions for improvement.

A synopsis from the job applicant with the best answers is provided at the end of Section 1.3.

Describe the importance of statistics.

In order to make intelligent decisions in a world full of uncertainty, we all have to understand statistics—the language of data. Unfortunately, many people avoid learning statistics because they believe (incorrectly!) that statistics simply deals with incomprehensible formulas and tedious calculations, and that it has no use in real life. This type of thinking is far from the truth because we encounter statistics *every day* in real life. We must understand statistics or risk making uninformed decisions and costly mistakes. While it is true that statistics incorporates formulas and calculations, it is logical reasoning that dictates how the data are collected, the calculations implemented, and the results communicated. A knowledge of statistics also provides the necessary tools to differentiate between sound statistical conclusions and questionable conclusions drawn from an insufficient number of data points, “bad” data points, incomplete data points, or just misinformation. Consider the following examples.

**Example 1.** After Washington, DC, had record amounts of snow in the winter of 2010, the headline of a newspaper stated, “What global warming?”

**Problem with conclusion:** The existence or nonexistence of climate change cannot be based on one year’s worth of data. Instead, we must examine long-term trends and analyze decades’ worth of data.

**Example 2.** A gambler predicts that his next roll of the dice will be a lucky 7 because he did not get that outcome on the last three rolls.

**Problem with conclusion:** As we will see later in the text when we discuss probability, the probability of rolling a 7 stays constant with each roll of the dice. It does not become more likely if it did not appear on the last roll or, in fact, any number of preceding rolls.

**Example 3.** On January 10, 2010, nine days prior to a special election to fill the U.S. Senate seat that was vacated due to the death of Ted Kennedy, a *Boston Globe* poll gave the Democratic candidate, Martha Coakley, a 15-point lead over the Republican candidate, Scott Brown. On January 19, 2010, Brown won 52% of the vote, compared to Coakley’s 47%, and became a U.S. senator for Massachusetts.

**Problem with conclusion:** Critics accused the *Globe*, which had endorsed Coakley, of purposely running a bad poll to discourage voters from coming out for Brown. In reality, by the time the *Globe* released the poll, it contained old information from January 2–6, 2010. Even more problematic was that the poll included people who said that they were unlikely to vote!

**Example 4.** Starbucks Corp., the world’s largest coffee-shop operator, reported that sales at stores open at least a year climbed 4% at home and abroad in the quarter ended December 27, 2009. Chief Financial Officer Troy Alstead said that “the U.S. is back in a good track and the international business has similarly picked up. . . . Traffic is really coming back. It’s a good sign for what we’re going to see for the rest of the year” (www.bloomberg.com, January 20, 2010).

**Problem with conclusion:** In order to calculate same-store sales growth, which compares how much each store in the chain is selling compared with a year ago, we remove stores that have closed. Given that Starbucks closed more than 800 stores over the past few years to counter large sales declines, it is likely that the sales increases in many of the stores were caused by traffic from nearby, recently closed stores. In this case, same-store sales growth may overstate the overall health of Starbucks.

**Example 5.** Researchers at the University of Pennsylvania Medical Center found that infants who sleep with a nightlight are much more likely to develop myopia later in life (*Nature*, May 1999).

**Problem with conclusion:** This example appears to commit the *correlation-to-causation fallacy*. Even if two variables are highly correlated, one does not necessarily cause the other. *Spurious correlation* can make two variables appear closely related when no causal relation exists. Spurious correlation between two variables is not based on any demonstrable relationship, but rather on a relation that arises in the data solely because each of those variables is related to some third variable. In a follow-up study, researchers at The Ohio State University found no link between infants who sleep with a nightlight and the development of myopia (*Nature*, March 2000). They did, however, find strong links between parental myopia and the development of child myopia, and between parental myopia and the parents' use of a nightlight in their children's room. So the cause of both conditions (the use of a nightlight and the development of child myopia) is parental myopia.

Note the diversity of the sources of these examples—the environment, psychology, polling, business, and health. We could easily include others, from sports, sociology, the physical sciences, and elsewhere. Data and data interpretation show up in virtually every facet of life, sometimes spuriously. All of the preceding examples basically misuse data to add credibility to an argument. A solid understanding of statistics provides you with tools to react intelligently to information that you read or hear.

## 1.2 WHAT IS STATISTICS?

### LO 1.2

In the broadest sense, we can define the study of statistics as the methodology of extracting useful information from a data set. Three steps are essential for doing good statistics. First, we have to find the right data, which are both complete and lacking any misrepresentation. Second, we must use the appropriate statistical tools, depending on the data at hand. Finally, an important ingredient of a well-executed statistical analysis is to clearly communicate numerical information into written language.

We generally divide the study of statistics into two branches: descriptive statistics and inferential statistics. **Descriptive statistics** refers to the summary of important aspects of a data set. This includes collecting data, organizing the data, and then presenting the data in the form of charts and tables. In addition, we often calculate numerical measures that summarize, for instance, the data's typical value and the data's variability. Today, the techniques encountered in descriptive statistics account for the most visible application of statistics—the abundance of quantitative information that is collected and published in our society every day. The unemployment rate, the president's approval rating, the Dow Jones Industrial Average, batting averages, the crime rate, and the divorce rate are but a few of the many “statistics” that can be found in a reputable newspaper on a frequent, if not daily, basis. Yet, despite the familiarity of descriptive statistics, these methods represent only a minor portion of the body of statistical applications.

The phenomenal growth in statistics is mainly in the field called inferential statistics. Generally, **inferential statistics** refers to drawing conclusions about a large set of data—called a **population**—based on a smaller set of **sample** data. A population is defined as all members of a specified group (not necessarily people), whereas a sample is a subset of that particular population. In most statistical applications, we must rely on sample data in order to make inferences about various characteristics of the population. For example, a 2010 survey of 1,208 registered voters by a USA TODAY/Gallup Poll found that President Obama's job performance was viewed favorably by only 41% of those polled, his lowest rating in a USA TODAY/Gallup Poll since he took office in January 2009 (*USA TODAY*, August 3, 2010). Researchers use this sample result, called a **sample statistic**, in an attempt to estimate the corresponding unknown **population parameter**. In this case, the parameter of interest is the percentage of *all* registered voters that view the president's job performance favorably. It is generally not feasible to obtain population data and calculate the relevant parameter directly due to prohibitive costs and/or practicality, as discussed next.

Differentiate between descriptive statistics and inferential statistics.

## POPULATION VERSUS SAMPLE

A **population** consists of all items of interest in a statistical problem. A **sample** is a subset of the population. We analyze sample data and calculate a **sample statistic** to make inferences about the unknown **population parameter**.

### LO 1.3

Explain the need for sampling and discuss various data types.

## The Need for Sampling

A major portion of inferential statistics is concerned with the problem of estimating population parameters or testing hypotheses about such parameters. If we have access to data that encompass the entire population, then we would know the values of the parameters. Generally, however, we are unable to use population data for two main reasons.

- **Obtaining information on the entire population is expensive.** Consider how the monthly unemployment rate in the United States is calculated by the Bureau of Labor Statistics (BLS). Is it reasonable to assume that the BLS counts every unemployed person each month? The answer is a resounding NO! In order to do this, every home in the country would have to be contacted. Given that there are over 150 million individuals in the labor force, not only would this process cost too much, it would take an inordinate amount of time. Instead, the BLS conducts a monthly sample survey of about 60,000 households to measure the extent of unemployment in the United States.
- **It is impossible to examine every member of the population.** Suppose we are interested in the average length of life of a Duracell AAA battery. If we tested the duration of each Duracell AAA battery, then in the end, all batteries would be dead and the answer to the original question would be useless.

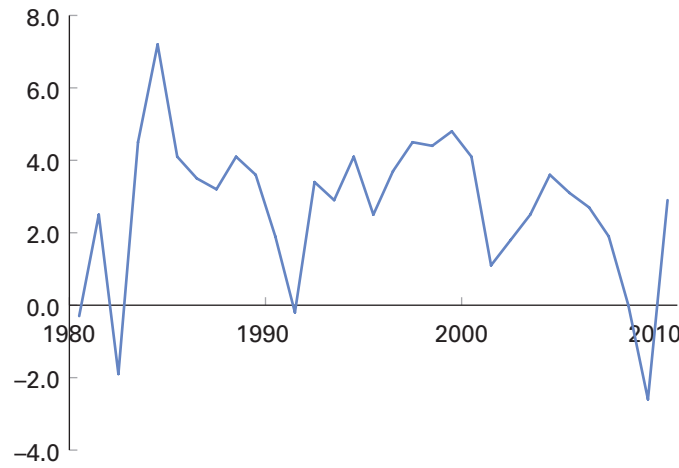
## Types of Data

Sample data are generally collected in one of two ways. **Cross-sectional data** refers to data collected by recording a characteristic of many subjects at the same point in time, or without regard to differences in time. Subjects might include individuals, households, firms, industries, regions, and countries. The tween data presented in Table 1.1 in the introductory case is an example of cross-sectional data because it contains tween responses to four questions at the end of the ski season. It is unlikely that all 20 tweens took the questionnaire at exactly the same time, but the differences in time are of no relevance in this example. Other examples of cross-sectional data include the recorded scores of students in a class, the sale prices of single-family homes sold last month, the current price of gasoline in different states in the United States, and the starting salaries of recent business graduates from The Ohio State University.

**Time series data** refers to data collected by recording a characteristic of a subject over several time periods. Time series can include hourly, daily, weekly, monthly, quarterly, or annual observations. Examples of time series data include the hourly body temperature of a patient in a hospital's intensive care unit, the daily price of IBM stock in the first quarter of 2015, the weekly exchange rate between the U.S. dollar and the euro, the monthly sales of cars at a dealership in 2014, and the annual growth rate of India in the last decade. Figure 1.1 shows a plot of the real (inflation-adjusted) GDP growth rate of the United States from 1980 through 2010. The average growth rate for this period is 2.7%, yet the plot indicates a great deal of variability in the series. It exhibits a wavelike movement, spiking downward in 2008 due to the economic recession before rebounding in 2010.



FILE  
GDP\_Growth



Source: Bureau of Economic Analysis.

**FIGURE 1.1** Real GDP growth rate from 1980 through 2010

**Cross-sectional data** contain values of a characteristic of many subjects at the same point or approximately the same point in time. **Time series data** contain values of a characteristic of a subject over time.

## Getting Started on the Web

As you can imagine, there is an abundance of data on the Internet. We accessed much of the data in this text by simply using a search engine like Google. These search engines often directed us to the same data-providing sites. For instance, the U.S. federal government publishes a great deal of economic and business data. The Bureau of Economic Analysis (BEA), the Bureau of Labor Statistics (BLS), the Federal Reserve Economic Data (FRED), and the U.S. Census Bureau provide data on inflation, unemployment, gross domestic product (GDP), and much more. Zillow.com is a real estate site that supplies data such as recent home sales, monthly rent, and mortgage rates. Finance.yahoo.com is a financial site that lists data such as stock prices, mutual fund performance, and international market data. *The Wall Street Journal*, *The New York Times*, *USA Today*, *The Economist*, and *Fortune* are all reputable publications that provide all sorts of data. Finally, espn.com offers comprehensive sports data on both professional and college teams. We list these sites in Table 1.2 and summarize *some* of the data that are available.



**TABLE 1.2** Select Internet Data Sites

Internet Site	Select Data Availability
Bureau of Economic Analysis (BEA)	National and regional data on gross domestic product (GDP) and personal income, international data on trade in goods and services.
Bureau of Labor Statistics (BLS)	Inflation rates, unemployment rates, employment, pay and benefits, spending and time use, productivity.
Federal Reserve Economic Data (FRED)	Banking, business/fiscal data, exchange rates, reserves, monetary base.
U.S. Census Bureau	Economic indicators, foreign trade, health insurance, housing, sector-specific data.
zillow.com	Recent home sales, home characteristics, monthly rent, mortgage rates.
finance.yahoo.com	Historical stock prices, mutual fund performance, international market data.
<i>The Wall Street Journal</i> , <i>The New York Times</i> , <i>USA Today</i> , <i>The Economist</i> , and <i>Fortune</i>	Poverty, crime, obesity, and plenty of business-related data.
espn.com	Professional and college teams' scores, rankings, standings, individual player statistics.

## EXERCISES 1.2

1. It came as a big surprise when Apple's touch screen iPhone 4, considered by many to be the best smartphone ever, was found to have a problem (*The New York Times*, June 24, 2010). Users complained of weak reception, and sometimes even dropped calls, when they cradled the phone in their hands in a particular way. A quick survey at a local store found that 2% of iPhone 4 users experienced this reception problem.
  - a. Describe the relevant population.
  - b. Does 2% denote the population parameter or the sample statistic?
2. Many people regard video games as an obsession for youngsters, but, in fact, the average age of a video game player is 35 years (Reuters.com, August 21, 2009). Is the value 35 likely the actual or the estimated average age of the population? Explain.
3. An accounting professor wants to know the average GPA of the students enrolled in her class. She looks up information on Blackboard about the students enrolled in her class and computes the average GPA as 3.29.
  - a. Describe the relevant population.
  - b. Does the value 3.29 represent the population parameter or the sample statistic?
4. Business graduates in the United States with a marketing concentration earn high salaries. According to the Bureau of Labor Statistics, the average annual salary for marketing managers was \$104,400 in 2007.
  - a. What is the relevant population?
  - b. Do you think the average salary of \$104,400 was computed from the population? Explain.
5. Recent research suggests that depression significantly increases the risk of developing dementia later in life (*BBC News*, July 6, 2010). In a study involving 949 elderly persons, it was reported that 22% of those who had depression went on to develop dementia, compared to only 17% of those who did not have depression.
  - a. Describe the relevant population and the sample.
  - b. Do the numbers 22% and 17% represent population parameters or sample statistics?
6. Go to [www.finance.yahoo.com/](http://www.finance.yahoo.com/) to get a current stock quote for Google, Inc. (ticker symbol = GOOG). Then, click on historical prices to record the monthly adjusted close price of Google stock in 2010. Create a table that uses this information. What type of data do these numbers represent? Comment on the data.
7. Ask 20 of your friends whether they live in a dormitory, a rental unit, or other form of accommodation. Also find out their approximate monthly lodging expenses. Create a table that uses this information. What type of data do these numbers represent? Comment on the data.
8. Go to [www.zillow.com/](http://www.zillow.com/) and find the sale price data of 20 single-family homes sold in Las Vegas, Nevada, in the last 30 days. In the data set, include the sale price, the number of bedrooms, the square footage, and the age of the house. What type of data do these numbers represent? Comment on the data.
9. The Federal Reserve Bank of St. Louis is a good source for downloading economic data. Go to [research.stlouisfed.org/fred2/](http://research.stlouisfed.org/fred2/) to extract quarterly data on gross private saving (GPSAVE) from 2008 to 2011 (16 observations). Create a table that uses this information. Plot the data over time and comment on the savings trend in the United States.
10. Another good source of data is the U.S. Census Bureau. Go to [www.census.gov/](http://www.census.gov/) and extract the most recent median household income for Alabama, Arizona, California, Florida, Georgia, Indiana, Iowa, Maine, Massachusetts, Minnesota, Mississippi, New Mexico, North Dakota, and Washington. What type of data do these numbers represent? Comment on the regional differences in income.

### LO 1.4

## 1.3 VARIABLES AND SCALES OF MEASUREMENT

Describe variables and various types of measurement scales.

When we conduct a statistical investigation, we invariably focus on people, objects, or events with particular characteristics. When a characteristic of interest differs in kind or degree among various observations, then the characteristic can be termed a **variable**. We further categorize a variable as either qualitative or quantitative. For a **qualitative variable**, we use labels or names to identify the distinguishing characteristic of each observation. For instance, the 2010 Census asked each respondent to indicate gender on the form. Each respondent chose either male or female. Gender is a qualitative variable. Other examples of qualitative variables include race, profession, type of business, the manufacturer of a car, and so on.

A variable that assumes meaningful numerical values is called a **quantitative variable**. Quantitative variables, in turn, are either discrete or continuous. A **discrete variable** assumes a countable number of values. Consider the number of children in a

family or the number of points scored in a basketball game. We may observe values such as 3 children in a family or 90 points being scored in a basketball game, but we will not observe 1.3 children or 92.5 scored points. The values that a discrete variable assumes need not be whole numbers. For example, the price of a stock for a particular firm is a discrete variable. The stock price may take on a value of \$20.37 or \$20.38, but it cannot take on a value between these two points. Finally, a discrete variable may assume an infinite number of values, but these values are countable; that is, they can be presented as a sequence  $x_1, x_2, x_3$ , and so on. The number of cars that cross the Golden Gate Bridge on a Saturday is a discrete variable. Theoretically, this variable assumes the values 0, 1, 2, . . .

A **continuous variable** is characterized by uncountable values within an interval. Weight, height, time, and investment return are all examples of continuous variables. For example, an unlimited number of values occur between the weights of 100 and 101 pounds, such as 100.3, 100.625, 100.8342, and so on. In practice, however, continuous variables may be measured in discrete values. We may report a newborn's weight (a continuous variable) in discrete terms as 6 pounds 10 ounces and another newborn's weight in similar discrete terms as 6 pounds 11 ounces.

#### QUALITATIVE VARIABLES VERSUS QUANTITATIVE VARIABLES

A **variable** is the general characteristic being observed on a set of people, objects, or events, where each observation varies in kind or degree. Labels or names are used to categorize the distinguishing characteristics of a **qualitative variable**; eventually, these attributes may be coded into numbers for purposes of data processing. A **quantitative variable** assumes meaningful numerical values, and can be further categorized as either **discrete** or **continuous**. A discrete variable assumes a countable number of values, whereas a continuous variable is characterized by uncountable values within an interval.

In order to choose the appropriate statistical methods for summarizing and analyzing data, we need to distinguish between different measurement scales. All data measurements can be classified into one of four major categories: nominal, ordinal, interval, and ratio. Nominal and ordinal scales are used for qualitative variables, whereas interval and ratio scales are used for quantitative variables. We discuss these scales in ascending order of sophistication.

## The Nominal Scale

The **nominal scale** represents the least sophisticated level of measurement. If we are presented with nominal data, all we can do is categorize or group the data. The values in the data set differ merely by name or label. Consider the following example.

Each company listed in Table 1.3 is a member of the Dow Jones Industrial Average (DJIA). The DJIA is a stock market index that shows how 30 large, publicly owned companies based in the United States have traded during a standard trading session in the stock market. Table 1.3 also shows where stocks of these companies are traded: on either the National Association of Securities Dealers Automated Quotations (Nasdaq) or the New York Stock Exchange (NYSE). These data are classified as nominal scale since we are simply able to group or categorize them. Specifically, only four stocks are traded on Nasdaq, whereas the remaining 26 are traded on the NYSE.

Often we substitute *numbers* for the particular qualitative characteristic or trait that we are grouping. One reason why we do this is for ease of exposition; always referring to the National Association of Securities Dealers Automated Quotations, or even Nasdaq, becomes awkward and unwieldy. In addition, as we will see later in the text, statistical analysis is greatly facilitated by using numbers instead of names.





**TABLE 1.3** Companies of the DJIA and Exchange Where Stock Is Traded

Company	Exchange	Company	Exchange
3M (MMM)	NYSE	Intel (INTC)	Nasdaq
American Express (AXP)	NYSE	Johnson & Johnson (JNJ)	NYSE
Apple (AAPL)	Nasdaq	JPMorgan Chase (JPM)	NYSE
Boeing (BA)	NYSE	McDonald's (MCD)	NYSE
Caterpillar (CAT)	NYSE	Merck (MRK)	NYSE
Chevron (CVX)	NYSE	Microsoft (MFST)	Nasdaq
Cisco (CSCO)	Nasdaq	Nike (NKE)	NYSE
Coca-Cola (KO)	NYSE	Pfizer (PFE)	NYSE
Disney (DIS)	NYSE	Procter & Gamble (PG)	NYSE
Dupont (DD)	NYSE	Travelers (TRV)	NYSE
ExxonMobil (XOM)	NYSE	United Health (UNH)	NYSE
General Electric (GE)	NYSE	United Tech. Corp. (UTX)	NYSE
Goldman Sachs (GS)	NYSE	Verizon (VZ)	NYSE
Home Depot (HD)	NYSE	Visa (V)	NYSE
IBM (IBM)	NYSE	Walmart (WMT)	NYSE

SOURCE: [www.money.cnn.com/data/dow30/](http://www.money.cnn.com/data/dow30/), information retrieved March 21, 2015.

For example, we might use the number 0 to show that a company's stock is traded on Nasdaq and the number 1 to show that a company's stock is traded on the NYSE. In tabular form:

Exchange	Number of Companies Trading on Exchange
0	4
1	26

## The Ordinal Scale

Compared to the nominal scale, the **ordinal scale** reflects a stronger level of measurement. With ordinal data we are able to both *categorize* and *rank* the data with respect to some characteristic or trait. The weakness with ordinal data is that we cannot interpret the difference between the ranked values because the actual numbers used are arbitrary. For example, suppose you are asked to classify the service at a particular hotel as excellent, good, fair, or poor. A standard way to record the ratings is

Excellent	4	Fair	2
Good	3	Poor	1

Here the value attached to excellent (4) is higher than the value attached to good (3), indicating that the response of excellent is preferred to good. However, another representation of the ratings might be

Excellent	100	Fair	70
Good	80	Poor	40

Excellent still receives a higher value than good, but now the difference between the two categories is 20 ( $100 - 80$ ), as compared to a difference of 1 ( $4 - 3$ ) when we use the first classification. In other words, *differences between categories are meaningless with ordinal data*. (We also should note that we could reverse the ordering so that, for instance, excellent equals 40 and poor equals 100; this renumbering would not change the nature of the data.)

### EXAMPLE 1.1

In the introductory case, four questions were posed to tweens. The first question (Q1) asked tweens to name the radio station that they listened to on the ride to the resort, and the second question (Q2) asked tweens to rate the food quality at the resort on a scale of 1 to 4. The tweens' responses to these questions are shown in Table 1.1 in the introductory case.

- What is the scale of measurement of the radio station data?
- How are the data based on the ratings of the food quality similar to the radio station data? How are the data different?
- Summarize the tweens' responses to Q1 and Q2 in tabular form. How can the resort use the information from these responses?

#### SOLUTION:

- When asked which radio station played on the car ride to the resort, tweens responded with one of the following answers: JAMN94.5, MIX104.1, or KISS108. These are nominal data—the values in the data differ merely in name or label.
- Since we can both categorize and rank the food quality data, we classify these responses as ordinal data. Ordinal data are similar to nominal data in the sense that we can categorize the data. The main difference between ordinal and nominal data is that the categories of ordinal data are ranked. A rating of 4 is better than a rating of 3. With the radio station data, we cannot say that KISS108 is ranked higher than MIX104.1; some tweens may argue otherwise, but we simply categorize nominal data without ranking.
- With respect to the radio station data (Q1), we can assign 1 to JAMN94.5, 2 to MIX104.1, and 3 to KISS108. Counting the responses that fall into each category, we find that six tweens listened to 1, two listened to 2, and 12 listened to 3, or in tabular form:

Radio Station	Number of Tweens Listening to Radio Station
1	6
2	2
3	12

Twelve of the 20 tweens, or 60%, listened to KISS108. This information could prove useful to the management of the resort as they make decisions as to where to allocate their advertising dollars. If the resort could only choose to advertise at one radio station, it would appear that KISS108 would be the wise choice.

Given the food quality responses (Q2), we find that three of the tweens rated food quality with a 4, six tweens rated food quality with a 3, eight tweens rated food quality with a 2, and three tweens rated food quality with a 1. In tabular form:

Rating	Number of Tweens
4	3
3	6
2	8
1	3

The food quality results may be of concern to management. Just as many tweens rated the food quality as excellent as compared to poor. Moreover, the majority  $[(8 + 3)/20 = 55\%]$  felt that the food was, at best, fair. Perhaps a more extensive survey that focuses solely on food quality would reveal the reason for their apparent dissatisfaction.

As mentioned earlier, nominal and ordinal scales are used for *qualitative variables*. Values corresponding to a qualitative variable are typically expressed in words but are coded into numbers for purposes of data processing. When summarizing the results of a qualitative variable, we typically count the number or calculate the percentage of persons or objects that fall into each possible category. With a qualitative variable, we are unable to perform meaningful arithmetic operations, such as adding and subtracting.

## The Interval Scale

With data that are measured on an **interval scale**, not only can we categorize and rank the data, we are also assured that the differences between scale values are meaningful. Thus, the arithmetic operations of addition and subtraction are meaningful. The Fahrenheit scale for temperatures is an example of an interval scale. Not only is 60 degrees Fahrenheit hotter than 50 degrees Fahrenheit, the same difference of 10 degrees also exists between 90 and 80 degrees Fahrenheit.

The main drawback of data on an interval scale is that the value of zero is arbitrarily chosen; the zero point of an interval scale does not reflect a complete absence of what is being measured. No specific meaning is attached to zero degrees Fahrenheit other than to say it is 10 degrees colder than 10 degrees Fahrenheit. With an arbitrary zero point, meaningful ratios cannot be constructed. For instance, it is senseless to say that 80 degrees is twice as hot as 40 degrees; in other words, the ratio 80/40 has no meaning.

## The Ratio Scale

The **ratio scale** represents the strongest level of measurement. Ratio data have all the characteristics of interval data as well as a *true zero* point, which allows us to interpret the ratios of values. A ratio scale is used to measure many types of data in business analysis. Variables such as sales, profits, and inventory levels are expressed as ratio data. A meaningful zero allows us to state, for example, that profits for firm A are double those of firm B. Measurements such as weight, time, and distance are also measured on a ratio scale since zero is meaningful.

Unlike qualitative data, arithmetic operations are valid on interval- and ratio-scaled values. In later chapters, we will calculate summary measures for the typical value and the variability of quantitative variables; we cannot calculate these measures if the variable is qualitative in nature.

### EXAMPLE 1.2

In the last two questions from the introductory case's survey (Q3 and Q4), the 20 tweens were asked: "What time should the main dining area close?" and "How much of your *own* money did you spend at the lodge today?" Their responses appear in Table 1.1 in the introductory case.

- a. How are the time data classified? In what ways do the time data differ from ordinal data? What is a potential weakness of this measurement scale?
- b. What is the measurement scale of the money data? Why is it considered the strongest form of data?
- c. In what ways is the information from Q3 and Q4 useful for the resort?

#### SOLUTION:

- a. Clock time responses, such as 3:00 pm and 3:30 pm, or 5:30 pm and 6:00 pm, are on an interval scale. Interval data are a stronger measurement scale than ordinal data because differences between interval-scaled values are meaningful. In this particular example, we can say that 3:30 pm is 30 minutes later than 3:00 pm and 6:00 pm is 30 minutes later than 5:30 pm. The weakness with interval data is that the value of zero is arbitrary. Here, with the clock time responses, we have no apparent zero point; however, we could always arbitrarily define a zero point, say, at 12:00 am. Thus, although differences are comparable with interval data, ratios are meaningless due to the arbitrariness

of the zero point. In other words, it is senseless to form the ratio 6:00 pm/3:00 pm and conclude that 6:00 pm is twice as long a time period as 3:00 pm.

- b. Since the tweens' responses are in dollar amounts, this is ratio data. The ratio scale is the strongest form of data because we can categorize and rank values as well as calculate meaningful differences. Moreover, since there is a natural zero point, valid ratios can also be calculated. For example, the data show that three tweens spent \$20. These tweens spent four times as much as the three tweens that spent \$5 ( $\$20/\$5 = 4$ ).
- c. A review of the clock time responses (Q3) in Table 1.1 shows that the vast majority of the tweens would like the dining area to remain open later. In fact, only one tween feels that the dining area should close at 3:00 pm. An inspection of the money responses (Q4) in Table 1.1 indicates that only three of the 20 tweens did not spend any of his/her own money. This is very important information. It does appear that the discretionary spending of this age group is significant. The resort would be wise to cater to some of their preferences.

## SYNOPSIS OF INTRODUCTORY CASE

A preliminary survey of tween preferences conducted by the management of a ski resort two hours outside Boston, Massachusetts, revealed some interesting information.

- Tweens were first asked to name the radio station that they listened to on the way to the resort. The responses show that 60% of the tweens listened to KISS108. If the resort wishes to contact tweens using this medium, it may want to direct its advertising dollars to this station.
- Next, the tweens were asked to rate the food quality at the resort on a scale of 1 to 4 (where 1 is poor, 2 is fair, 3 is good, and 4 is excellent). The survey results with respect to food quality are disturbing. The majority of the tweens, 55% (11/20), felt that the food was, at best, fair. A more extensive study focusing on food quality appears necessary.
- Tweens were then asked what time the main dining area should close, given a present closing time of 3:00 pm. The data suggest that the vast majority of the tweens (19 out of 20) would like the dining area to remain open later.
- Finally, the tweens were asked to report the amount of their *own* money they spent at the lodge. The resort is likely pleased with the responses to this question since 17 of the 20 tweens spent their own money at the lodge. This finding appears consistent with the belief that tween spending is growing.



## EXERCISES 1.3

11. Which of the following variables are qualitative and which are quantitative? If the variable is quantitative, then specify whether the variable is discrete or continuous.
  - a. Points scored in a football game.
  - b. Racial composition of a high school classroom.
  - c. Heights of 15-year-olds.
12. Which of the following variables are qualitative and which are quantitative? If the variable is quantitative, then specify whether the variable is discrete or continuous.
  - a. Colors of cars in a mall parking lot.
  - b. Time it takes each student to complete a final exam.
  - c. The number of patrons who frequent a restaurant.

13. In each of the following scenarios, define the type of measurement scale.
  - a. A kindergarten teacher marks whether each student is a boy or a girl.
  - b. A ski resort records the daily temperature during the month of January.
  - c. A restaurant surveys its customers about the quality of its waiting staff on a scale of 1 to 4, where 1 is poor and 4 is excellent.
14. In each of the following scenarios, define the type of measurement scale.
  - a. An investor collects data on the weekly closing price of gold throughout a year.
  - b. An analyst assigns a sample of bond issues to one of the following credit ratings, given in descending order of credit quality (increasing probability of default): AAA, AA, BBB, BB, CC, D.
  - c. The dean of the business school at a local university categorizes students by major (i.e., accounting, finance, marketing, etc.) to help in determining class offerings in the future.
15. In each of the following scenarios, define the type of measurement scale.
  - a. A meteorologist records the amount of monthly rainfall over the past year.
  - b. A sociologist notes the birth year of 50 individuals.
  - c. An investor monitors the daily stock price of BP following the 2010 oil disaster in the Gulf of Mexico.
16. A professor records the majors of her 30 students as follows:

Accounting	Economics	Undecided	Finance	Management
Management	Finance	Marketing	Economics	Management
Marketing	Finance	Marketing	Accounting	Finance
Finance	Undecided	Management	Undecided	Economics
Economics	Accounting	Management	Undecided	Economics
Accounting	Economics	Management	Accounting	Economics

- a. What is the measurement scale of these data?
  - b. Summarize the results in tabular form.
  - c. What information can be extracted from the data?
17. **FILE DOW\_Characteristics.** The accompanying table shows a portion of the 30 companies that comprise the Dow Jones Industrial Average (DJIA). The second column shows the year that the company joined the DJIA (Year). The third column shows each company's Morningstar rating (Rating). (Five stars is the best rating that a company can receive, indicating that the company's stock price is undervalued and thus a very good buy. One star is the worst rating a company can be given, implying that the stock price is overvalued and a bad buy.) Finally, the fourth column shows each company's stock price as of June 30, 2010 (Stock Price).

Company	Year	Rating	Stock Price
3M (MMM)	1976	*****	\$78.99
Alcoa (AA)	1959	****	10.03
:	:	:	:
Walt Disney (DIS)	1991	***	31.50

SOURCE: Morningstar ratings retrieved from [www.morningstar.com](http://www.morningstar.com) on June 30, 2010; stock prices retrieved from [www.finance.yahoo.com](http://www.finance.yahoo.com).

- a. What is the measurement scale of the Year data? What are the strengths of this type of data? What are the weaknesses?
  - b. What is the measurement scale of Morningstar's star-based rating system? Summarize Morningstar's star-based rating system for the companies in tabular form. Let 5 denote \*\*\*\*\*, 4 denote \*\*\*\*, and so on. What information can be extracted from these data?
  - c. What is the measurement scale of the Stock Price data? What are its strengths?

## CONCEPTUAL REVIEW

### LO 1.1 Describe the importance of statistics.

A proper understanding of statistical ideas and concepts helps us understand more of the real world around us, including issues in business, sports, politics, health, and social interactions. We must understand statistics or risk making bad decisions and costly mistakes. A knowledge of statistics also provides the necessary tools to differentiate between sound statistical conclusions and questionable conclusions drawn from an insufficient number of data points, "bad" data points, incomplete data points, or just misinformation.

### LO 1.2 Differentiate between descriptive statistics and inferential statistics.

The study of statistics is generally divided into two branches: descriptive statistics and inferential statistics. **Descriptive statistics** refers to the summary of a data set in the

form of tables, graphs, and/or the calculation of numerical measures. **Inferential statistics** refers to extracting useful information from a **sample** to draw conclusions about a **population**. A **population** consists of all items of interest in a statistical problem; a **sample** is a subset of that population.

**LO 1.3 Explain the need for sampling and discuss various data types.**

In general, we use sample data rather than population data for two main reasons: (1) obtaining information on the entire population is expensive and/or (2) it is impossible to examine every item of the population.

**Cross-sectional data** contain values of a characteristic of many subjects at the same point in time or without regard to differences in time. **Time series data** contain values of a characteristic of a subject over time.

**LO 1.4 Describe variables and various types of measurement scales.**

A variable is categorized as either qualitative or quantitative. For a **qualitative variable**, we use labels or names to identify the distinguishing characteristic of each observation. A **quantitative variable** assumes meaningful numerical values and can be further categorized as either **discrete** or **continuous**. A discrete variable assumes a countable number of values, whereas a continuous variable is characterized by uncountable values within an interval.

All data measurements can be classified into one of four major categories.

- The **nominal scale** represents the least sophisticated level of measurement. The values on a nominal scale differ merely by name or label. These values are then simply categorized or grouped by name.
- The values on an **ordinal scale** can be categorized *and* ranked; however, differences between the ranked values are meaningless.
- The **interval scale** is a stronger measurement scale as compared to nominal and ordinal scales. Values on the interval scale can be categorized and ranked, and differences between values are meaningful. The main drawback of the interval scale is that the value of zero is arbitrarily chosen; this implies that ratios constructed from interval-scaled values bear no significance.
- The **ratio scale** represents the strongest level of measurement. Ratio data have all the characteristics of interval data as well as a true zero point; thus, as its name implies, meaningful ratios can be calculated with values on the ratio scale.

Nominal and ordinal scales are used for qualitative variables. When summarizing the results of qualitative data, we typically count the number or calculate the percentage of persons or objects that fall into each possible category. Interval and ratio scales are used for quantitative variables. Unlike qualitative variables, arithmetic operations are valid on quantitative variables.



# 2

## LEARNING OBJECTIVES

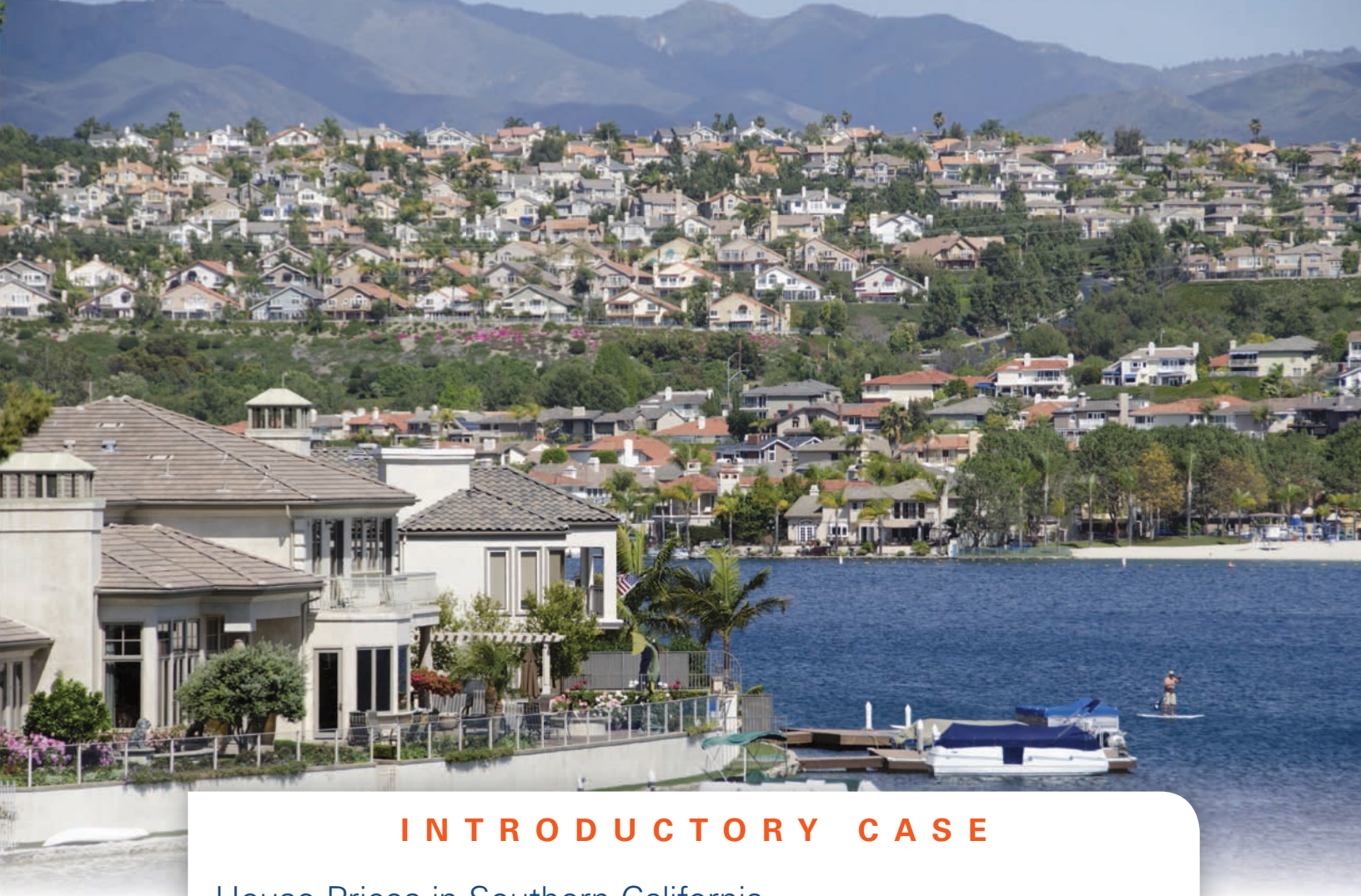
After reading this chapter  
you should be able to:

- LO 2.1 Summarize qualitative data by forming frequency distributions.
- LO 2.2 Construct and interpret pie charts and bar charts.
- LO 2.3 Summarize quantitative data by forming frequency distributions.
- LO 2.4 Construct and interpret histograms, polygons, and ogives.
- LO 2.5 Construct and interpret a stem-and-leaf diagram.
- LO 2.6 Construct and interpret a scatterplot.

# Tabular and Graphical Methods

People often have difficulty processing information provided by data in its raw form. A useful way of interpreting data effectively is to condense the data with some kind of visual or numerical summary. In this chapter, we present several tabular and graphical tools that can help us organize and present data. We first construct frequency distributions using qualitative data. We can visualize these frequency distributions by constructing pie charts and bar charts. For quantitative data, we again make frequency distributions. In addition to giving us an overall picture of where the data tend to cluster, frequency distributions using quantitative data also show us how the data are spread out from the lowest value to the highest value. For visual representations of quantitative data, we examine histograms, polygons, ogives, and stem-and-leaf diagrams. Finally, we show how to construct a scatterplot, which graphically depicts the relationship between two quantitative variables. We will find that a scatterplot is a very useful tool when conducting correlation and regression analysis, topics discussed in depth later in the text.





## INTRODUCTORY CASE

### House Prices in Southern California

Mission Viejo, a city located in Southern California, was named the safest city in California and the third-safest city in the nation (CQPress.com, November 23, 2009). Matthew Edwards, a relocation specialist for a real estate firm in Mission Viejo, often relays this piece of information to clients unfamiliar with the many benefits that the city offers. Recently, a client from Seattle, Washington, asked Matthew for a summary of recent sales. The client is particularly interested in the availability of houses in the \$500,000 range. Table 2.1 shows the sale price for 36 single-family houses in Mission Viejo during June 2010.

**TABLE 2.1** Recent Sale Price of Houses in Mission Viejo, CA, for June 2010 (data in \$1,000s)

\$430	670	530	521	669	445
520	417	525	350	660	412
460	533	430	399	702	735
475	525	330	560	540	537
670	538	575	440	460	630
521	370	555	425	588	430

Source: [www.zillow.com](http://www.zillow.com).

Matthew wants to use the sample information to:

1. Make summary statements concerning the range of house prices.
2. Comment on where house prices tend to cluster.
3. Calculate appropriate percentages in order to compare house prices in Mission Viejo, California, to those in Seattle, Washington.

A synopsis of this case is provided at the end of Section 2.2.

Summarize qualitative data by forming frequency distributions.

As we discussed in Chapter 1, nominal and ordinal data are types of qualitative data. Nominal data typically consist of observations that represent labels or names; information related to gender or race are examples. Nominal data are considered the least sophisticated form of data since all we can do with the data is categorize it. Ordinal data are stronger in the sense that we can categorize and order the data. Examples of ordinal data include the ratings of a product or a professor, where 1 represents the worst and 4 represents the best. In order to organize qualitative data, it is often useful to construct a frequency distribution.

### FREQUENCY DISTRIBUTION FOR QUALITATIVE DATA

A **frequency distribution** for qualitative data groups data into categories and records the number of observations that fall into each category.

To illustrate the construction of a frequency distribution with nominal data, Table 2.2 shows the weather for the month of February (2010) in Seattle, Washington.

**TABLE 2.2** Seattle Weather, February 2010

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
	1 Rainy	2 Rainy	3 Rainy	4 Rainy	5 Rainy	6 Rainy
7 Rainy	8 Rainy	9 Cloudy	10 Rainy	11 Rainy	12 Rainy	13 Rainy
14 Rainy	15 Rainy	16 Rainy	17 Sunny	18 Sunny	19 Sunny	20 Sunny
21 Sunny	22 Sunny	23 Rainy	24 Rainy	25 Rainy	26 Rainy	27 Rainy
28 Sunny						

Source: [www.wunderground.com](http://www.wunderground.com).



We first note that the weather in Seattle is categorized as cloudy, rainy, or sunny. The first column in Table 2.3 lists these categories. Initially, we use a “tally” column to record the number of days that fall into each category. Since the first eight days of February were rainy days, we place the first eight tally marks in the rainy category; the ninth day of February was cloudy, so we place one tally mark in the cloudy category, and so on. Finally, we convert each category’s total tally count into its respective numerical value in the frequency column. Since only one tally mark appears in the cloudy category, we record the value 1 as its frequency. Note that if we sum the frequency column, we obtain the sample size. A frequency distribution in its final form does not include the tally column.

**TABLE 2.3** Frequency Distribution for Seattle Weather, February 2010

Weather	Tally	Frequency
Cloudy	I	1
Rainy	HH HH HH HH	20
Sunny	HH II	7
		Total = 28 days

From the frequency distribution, we can now readily observe that the most common type of day in February was rainy since this type of day occurs with the highest frequency. In many applications, we want to compare data sets that differ in size. For example, we might

want to compare the weather in February to the weather in March. However, February has 28 days (except during a leap year) and March has 31 days. In this instance, we would convert the frequency distribution to a **relative frequency distribution**. We calculate each category's relative frequency by dividing the respective category's frequency by the total number of observations. The sum of the relative frequencies should equal one, or a value very close to one due to rounding.

In Table 2.4, we convert the frequency distribution from Table 2.3 into a relative frequency distribution. Similarly, we obtain the relative frequency distribution for the month of March; the raw data for March are not shown. March had 3 cloudy days, 18 rainy days, and 10 sunny days. Each of these frequencies was then divided by 31, the number of days in the month of March.

**TABLE 2.4** Relative Frequency Distribution for Seattle Weather

Weather	February 2010: Relative Frequency	March 2010: Relative Frequency
Cloudy	$1/28 = 0.036$	$3/31 = 0.097$
Rainy	$20/28 = 0.714$	$18/31 = 0.581$
Sunny	$7/28 = 0.250$	$10/31 = 0.323$
	Total = 1	Total = 1 (subject to rounding)

SOURCE: [www.wunderground.com](http://www.wunderground.com).

We can easily convert relative frequencies into percentages by multiplying by 100. For instance, the percent of cloudy days in February and March equals 3.6% and 9.7%, respectively. From the relative frequency distribution, we can now conclude that the weather in Seattle in both February and March was predominantly rainy. However, the weather in March was a bit nicer in that approximately 32% of the days were sunny, as opposed to only 25% of the days in February.

#### CALCULATING RELATIVE AND PERCENT FREQUENCIES

The **relative frequency** for each category in a frequency distribution equals the proportion (fraction) of observations in each category. A category's relative frequency is calculated by dividing the frequency by the total number of observations. The sum of the relative frequencies should equal one.

The **percent frequency** for each category in a frequency distribution equals the percent (%) of observations in each category; it equals the relative frequency of the category multiplied by 100.

## Visualizing Frequency Distributions for Qualitative Data

We can visualize the information found in frequency distributions by constructing various graphs. Graphical representations often portray the data more dramatically, as well as simplify interpretation. A **pie chart** and a **bar chart** are two widely used graphical representations of qualitative data.

#### LO 2.2

Construct and interpret pie charts and bar charts.

#### GRAPHICAL DISPLAY OF QUALITATIVE DATA: PIE CHART

A **pie chart** is a segmented circle whose segments portray the relative frequencies of the categories of some qualitative variable.

A pie chart is best explained by using an example. Consider Example 2.1.



## EXAMPLE 2.1

Is America having a “marriage crisis?” The answer depends on whom you ask, but nearly every study focuses on the women’s liberation movement of the late 1960s and 1970s. As more and more women earned college degrees, they entered the workforce and delayed motherhood. Marriage became less necessary for their economic survival. No matter what the reason for the decline in marriage, here are some facts. In 1960, 71% of all adults in the United States were married. Today, barely half of all adults are married, just 52%. Table 2.5 shows the proportions of all adults who were married, widowed, divorced, or single in 1960 compared to those same proportions in 2010. Construct pie charts to graphically depict marital status in the United States in these two time periods.

**TABLE 2.5** Marital Status, 1960 versus 2010

Marital Status	1960	2010
Married	0.71	0.52
Single	0.15	0.28
Divorced	0.05	0.14
Widowed	0.09	0.06

NOTE: Proportions for each year rounded so that they summed to one.

SOURCE: Pew Research Center analysis of Decennial Census (1960–2000) and American Community Survey data (2008, 2010).

**SOLUTION:** In order to construct a pie chart, we first draw a circle. We then cut the circle into slices, or sectors such that each sector is proportional to the size of the category we wish to display. For instance, Table 2.5 shows that married adults accounted for 0.71 in 1960. Since a circle contains 360 degrees, the portion of the circle representing married adults encompasses  $0.71 \times 360 = 255.6$  degrees. Similar calculations for the other three categories yield:

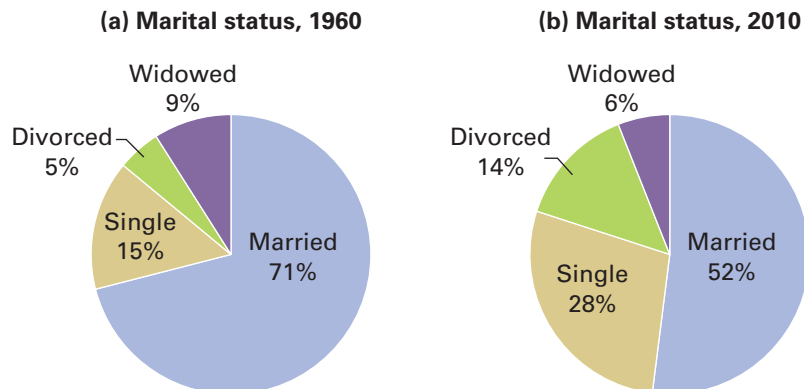
Single:  $0.15 \times 360 = 54$  degrees

Divorced:  $0.05 \times 360 = 18$  degrees

Widowed:  $0.09 \times 360 = 32.4$  degrees

The same methodology can be used to construct a pie chart for marital status in 2010. Figure 2.1 shows the resulting pie charts.

**FIGURE 2.1** Pie charts for marital status



## Using Excel to Construct a Pie Chart

Excel offers various options for displaying a pie chart. To replicate the pie chart in Figure 2.1(a), follow these steps:

- A. **FILE** Open *Marital\_Status* (Table 2.5) and select an empty cell.
- B. Select the category names and respective relative frequencies from the year 1960. Leave out the heading (top row).
- C. From the menu choose **Insert > Pie > 2-D Pie** and choose the graph on the top left.
- D. In order to give the pie chart category names and their respective percentages, from the menu choose **Layout > Data Labels > More Data Label Options**. Under *Label Options*, deselect “Value” and select “Category Name” and “Percentage.”

**FILE**  
*Marital\_Status*

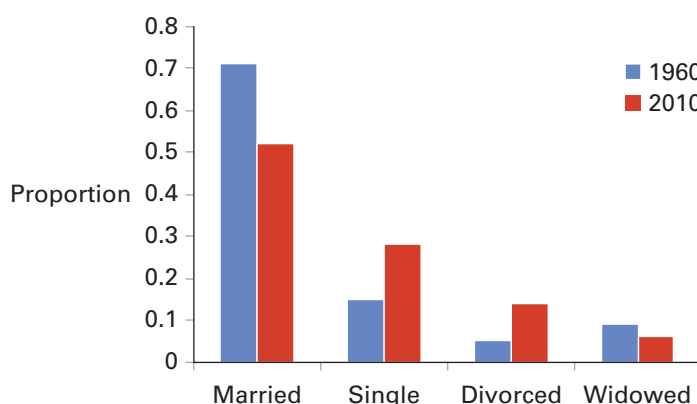
Another way to graphically depict qualitative data is to construct a **bar chart**.

### GRAPHICAL DISPLAY OF QUALITATIVE DATA: BAR CHART

A **bar chart** depicts the frequency or the relative frequency for each category of the qualitative variable as a series of horizontal or vertical bars, the lengths of which are proportional to the values that are to be depicted.

We first discuss a vertical bar chart, sometimes referred to as a column chart. Here, we place each category on the horizontal axis and mark the vertical axis with an appropriate range of values for either frequency or relative frequency. The height of each bar is equal to the frequency or the relative frequency of the corresponding category. Typically, we leave space between categories to improve clarity.

Figure 2.2 shows a relative frequency bar chart for the marital status example. It is particularly useful because we can group marital status by year, emphasizing the decline in the proportion of U.S. adults who are married and the rise in the proportion of U.S. adults who are single.



**FIGURE 2.2**  
Marital status of U.S. adults, 1960 versus 2010

## Using Excel to Construct a Bar Chart

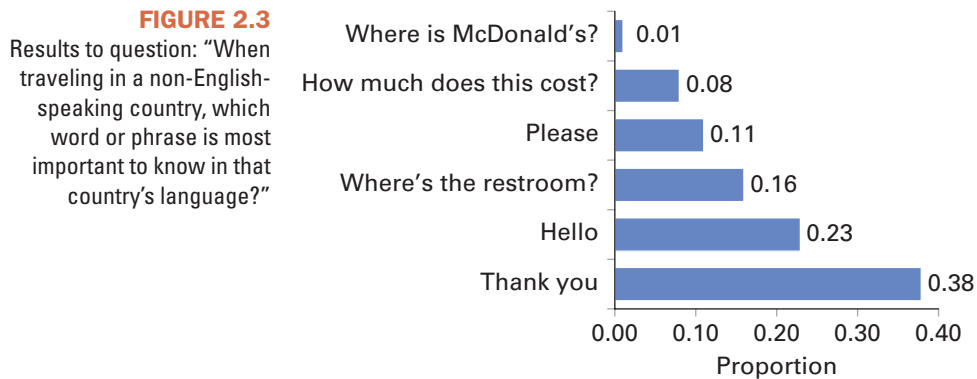
Excel provides many options for showing a bar chart. To replicate the bar chart in Figure 2.2, follow these steps:

- A. **FILE** Open *Marital\_Status* (Table 2.5) and select an empty cell.
- B. Select the category names and respective relative frequencies for the years 1960 and 2010. Leave out the heading (top row).

**FILE**  
*Marital\_Status*

- C. Choose **Insert > Column > 2-D Column**. From the options given, choose the graph on the top left. (This will create a vertical bar chart. If you want to construct a horizontal bar chart, choose **Insert > Bar > 2-D Bar**.)
- D. In the legend to the right of the bar chart, Excel labels the data for the year 1960 as “Series 1” and the data for the year 2010 as “Series 2” by default. In order to edit the legend, select the legend and choose **Design > Select Data**. From the *Legend Entries*, select “Series 1,” then select *Edit*, and under *Series Name*, type the new name of 1960. Follow the same steps to rename “Series 2” to 2010.

For a horizontal bar chart, we simply place each category on the vertical axis and mark the horizontal axis with an appropriate range of values for either frequency or relative frequency. For example, a recent poll asked more than 1,000 Americans: “When traveling in a non-English-speaking country, which word or phrase is most important to know in that country’s language?” (Source: *Vanity Fair*, January 2, 2012). Figure 2.3 shows the results of the poll. The phrase “Thank you” earned the largest percentage of votes (38%). Fortunately, only 1% of Americans believed that the phrase “Where is McDonald’s?” was of vital importance. The proportions in Figure 2.3 do not sum to one because we exclude those that responded with uncommon words or phrases.



## Cautionary Comments When Constructing or Interpreting Charts or Graphs

As with many of the statistical methods that we examine throughout this text, the possibility exists for unintentional, as well as purposeful, distortions of graphical information. As a careful researcher, you should follow these basic guidelines:

- The simplest graph should be used for a given set of data. Strive for clarity and avoid unnecessary adornments.
- Axes should be clearly marked with the numbers of their respective scales; each axis should be labeled.
- When creating a bar chart, each bar should be of the same width. Differing bar widths create distortions. The same principle holds in the next section when we discuss histograms.
- The vertical axis should not be given a very high value as an upper limit. In these instances, the data may appear compressed so that an increase (or decrease) of the data is not as apparent as it perhaps should be. For example, Figure 2.4(a) plots the daily price for a barrel of crude oil for the first quarter of 2011. Due to Middle East unrest, the price of crude oil rose from a low of \$83.13 per barrel to a high of \$106.19 per barrel, or approximately 28%  $\left( = \frac{106.19 - 83.13}{83.13} \right)$ . However, since Figure 2.4(a) uses a high value as an upper limit on the vertical axis (\$325), the rise in price appears dampened.
- The vertical axis should not be stretched so that an increase (or decrease) of the data appears more pronounced than warranted. For example, Figure 2.4(b) charts the daily

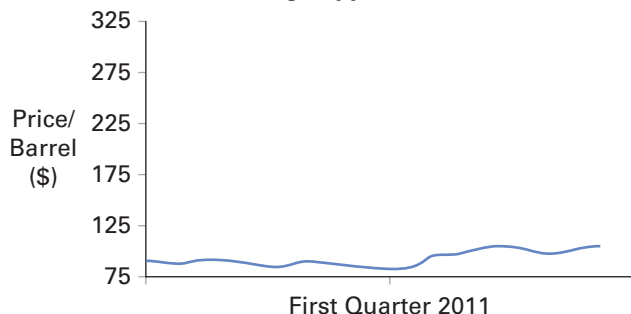
**FILE**  
*Crude\_Oil*

closing stock price for Johnson & Johnson (JNJ) for the week of April 4, 2011. It is true that the stock price declined over the week from a high of \$60.15 to a low of \$59.46; this amounts to a \$0.69 decrease or an approximate 1% decline. However, since the vertical axis is stretched, the drop in stock price appears more dramatic.

**FILE**  
JNJ

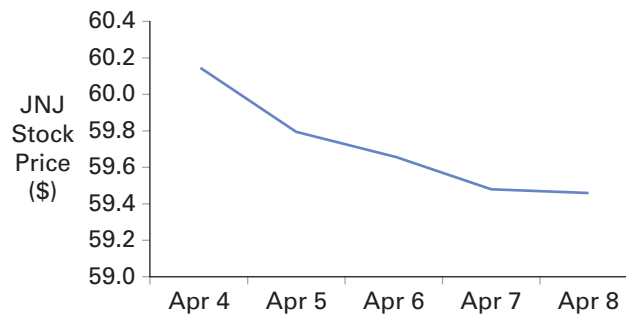
**FIGURE 2.4** Misleading scales on vertical axes

**(a) Vertical axis with high upper limit**



SOURCE: U.S. Energy Information Administration.

**(b) Stretched vertical axis**



SOURCE: www.finance.yahoo.com.

## EXERCISES 2.1

- A local restaurant is committed to providing its patrons with the best dining experience possible. On a recent survey, the restaurant asked patrons to rate the quality of their entrées. The responses ranged from 1 to 5, where 1 indicated a disappointing entrée and 5 indicated an exceptional entrée. The results of the survey are as follows:

3	5	4	4	3	2	3	3	2	5	5	5
5	3	3	2	1	4	5	5	4	2	5	5
5	4	4	3	1	5	2	1	5	4	4	4

- Construct frequency and relative frequency distributions that summarize the survey's results.
  - Are patrons generally satisfied with the quality of their entrées? Explain.
- First-time patients at North Shore Family Practice are required to fill out a questionnaire that gives the doctor an overall idea of each patient's health. The first question is: "In general, what is the quality of your health?" The patient chooses Excellent, Good, Fair, or Poor. Over the past month, the responses to this question from first-time patients were:

Fair	Good	Fair	Excellent
Good	Good	Good	Poor
Excellent	Excellent	Poor	Good
Fair	Good	Good	Good
Good	Poor	Fair	Excellent
Excellent	Good	Good	Good

- Construct frequency and relative frequency distributions that summarize the responses to the questionnaire.
- What is the most common response to the questionnaire? How would you characterize the health of first-time patients at this medical practice?

- A survey asked chief executives at leading U.S. firms the following question: "Where do you expect the U.S. economy to be 12 months from now?" A representative sample of their responses appears below:

Same	Same	Same	Better	Worse
Same	Same	Better	Same	Worse
Same	Better	Same	Better	Same
Worse	Same	Same	Same	Worse
Same	Same	Same	Better	Same

- Construct frequency and relative frequency distributions that summarize the responses to the survey. Where did most chief executives expect the U.S. economy to be in 12 months?
  - Use Excel to construct a pie chart and a bar chart to summarize your results.
- AccuWeather.com reported the following weather delays at these major U.S. airline hubs for July 21, 2010:

City	Delay	City	Delay
Atlanta	PM Delays	Mpls./St. Paul	None
Chicago	None	New York	All Day Delays
Dallas/Ft. Worth	None	Orlando	None
Denver	All Day Delays	Philadelphia	All Day Delays
Detroit	AM Delays	Phoenix	None
Houston	All Day Delays	Salt Lake City	None
Las Vegas	All Day Delays	San Francisco	AM Delays
Los Angeles	AM Delays	Seattle	None
Miami	AM Delays	Washington	All Day Delays

- Construct frequency and relative frequency distributions that summarize the delays at major U.S. hubs. What was the most common type of delay? Explain.
- Use Excel to construct a pie chart and a bar chart to summarize your results.



5. Fifty pro-football rookies were rated on a scale of 1 to 5, based on performance at a training camp as well as on past performance. A ranking of 1 indicated a poor prospect whereas a ranking of 5 indicated an excellent prospect. The following frequency distribution was constructed.

Rating	Frequency
1	4
2	10
3	14
4	18
5	4

- How many of the rookies received a rating of 4 or better? How many of the rookies received a rating of 2 or worse?
  - Construct the corresponding relative frequency distribution. What percent received a rating of 5?
  - Construct a bar chart for these data.
6. A recent survey asked 5,324 individuals: "What's most important to you when choosing where to live?" The responses are shown in the following relative frequency distribution.

Response	Relative Frequency
Good jobs	0.37
Affordable homes	0.15
Top schools	0.11
Low crime	0.23
Things to do	0.14

Copyright © 2010 Turner, Inc. Used with permission.

- Construct the corresponding frequency distribution. How many of the respondents chose "low crime" as the most important criterion when choosing where to live?
  - Construct a bar chart for the frequency distribution found in part a.
7. What is the perfect summer trip? A National Geographic Kids survey (*AAA Horizons*, April 2007) asked this question to 316 children ages 8 to 14. Their responses are given in the following frequency distribution.

Top Vacation Choice	Frequency
Cruises	140
Beaches	68
Amusement Parks	68
Big Cities	20
Lakes	12
Summer Camp	8

- Construct a relative frequency distribution. What percentage of the responses cited "Cruises" as the perfect summer trip?
  - Construct a bar chart for these data.
8. The following table lists U.S. revenue (in \$ billions) of the major car-rental companies.

Car-Rental Company	Revenue in 2009
Enterprise	\$10.7
Hertz	4.7
Avis Budget	4.0
Dollar Thrifty	1.5
Other	1.0

SOURCE: *The Wall Street Journal*, July 30, 2010.

- Compute the relative market share of the car-rental companies.
  - Hertz accounted for what percentage of sales?
  - Use Excel to construct a pie chart for these data.
9. A survey conducted by CBS News asked 829 respondents which of the following events will happen first. The responses are summarized in the following table:

Cure for cancer found	40%
End of dependence on oil	27%
Signs of life in outer space	12%
Peace in Middle East	8%
Other	6%
None will happen	7%

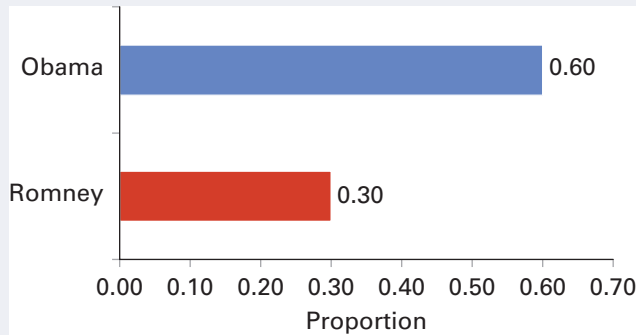
SOURCE: *Vanity Fair*, December 2009.

- Use Excel to construct a pie chart and a bar chart for these data.
  - How many people think that a cure for cancer will be found first?
10. A 2010 poll conducted by NBC asked respondents who would win Super Bowl XLV in 2011. The responses by 20,825 people are summarized in the following table.

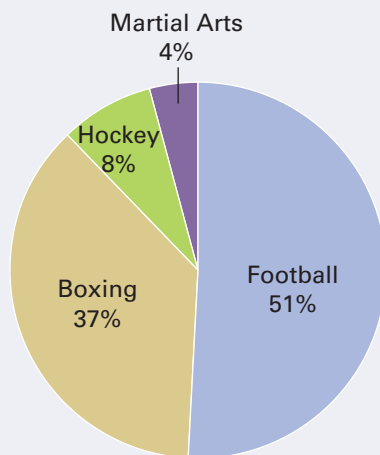
Team	Number of Votes
Atlanta Falcons	4,040
New Orleans Saints	1,880
Houston Texans	1,791
Dallas Cowboys	1,631
Minnesota Vikings	1,438
Indianapolis Colts	1,149
Pittsburgh Steelers	1,141
New England Patriots	1,095
Green Bay Packers	1,076
Others	

- How many responses were for "Others"?
  - The Green Bay Packers won Super Bowl XLV, defeating the Pittsburgh Steelers by the score of 31–25. What proportion of respondents felt that the Green Bay Packers would win?
  - Construct a bar chart for these data using relative frequencies.
11. In a USA TODAY/Gallup Poll, respondents favored Barack Obama over Mitt Romney in terms of likeability, 60% to

30% (*Los Angeles Times*, July 28, 2012). The following bar chart summarizes the responses.



- What percentage of respondents favored neither Obama nor Romney in terms of likeability?
  - Suppose this survey was based on 500 respondents. How many respondents favored Obama over Romney?
12. A recent survey of 992 people asked: In which professional sport—football, boxing, hockey, or martial arts—is an athlete most likely to sustain an injury that will affect the athlete after he or she retires? (*Vanity Fair*, January 29, 2012.) The following pie chart summarizes the responses.



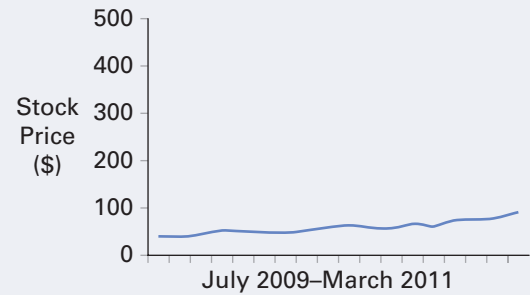
Copyright © 2012 Conde Nast. Used with permission.

- According to this survey, in which sport was an athlete most likely to sustain an injury with lifelong

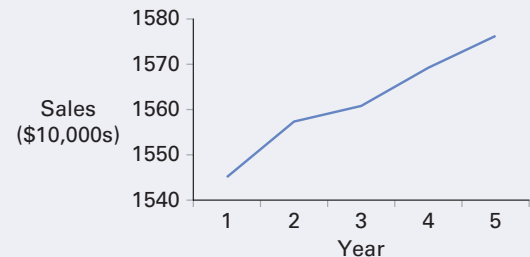
consequences? In which sport was an athlete least likely to sustain an injury with lifelong consequences?

- How many respondents believed that professional hockey players were most likely to sustain an injury with lifelong consequences?

13. The accompanying figure plots the monthly stock price of Caterpillar, Inc., from July 2009 through March 2011. The stock has experienced tremendous growth over this time period, almost tripling in price. Does the figure reflect this growth? If not, why not?



14. Annual sales at a small pharmaceutical firm have been rather stagnant over the most recent five-year period, exhibiting only 1.2% growth over this time frame. A research analyst prepares the accompanying graph for inclusion in a sales report.



Does this graph accurately reflect what has happened to sales over the last five years? If not, why not?

## 2.2 SUMMARIZING QUANTITATIVE DATA

### LO 2.3

With quantitative data, each observation represents a meaningful amount or count. The number of patents held by pharmaceutical firms (count) and household incomes (amount) are examples of quantitative data. Although different in nature from qualitative data, we still use frequency distributions to summarize quantitative data.

Before discussing the mechanics of constructing a frequency distribution, we find it useful to first examine one in its final form, using the house-price data from the introductory case to this chapter. We converted the raw data (the actual values) from Table 2.1 into a frequency distribution with five intervals or **classes**, each of width 100, as shown in

Summarize quantitative data by forming frequency distributions.

Table 2.6. We see, for instance, that four houses sold in the first class, where prices ranged from \$300,000 up to \$400,000. The data are more manageable using a frequency distribution, but some detail is lost because we no longer see the actual values.

**TABLE 2.6** Frequency Distribution for House-Price Data

Class (in \$1000s)	Frequency
300 up to 400	4
400 up to 500	11
500 up to 600	14
600 up to 700	5
700 up to 800	2
	Total = 36

### EXAMPLE 2.2

Based on the frequency distribution in Table 2.6, what is the price range over this time period? Over what price range did the majority of the houses sell?

**SOLUTION:** The frequency distribution shows that house prices ranged from \$300,000 up to \$800,000 over this time period. The most houses (14) sold in the \$500,000 up to \$600,000 range. Note that only four houses sold in the lowest price range and only two houses sold at the highest price range.

It turns out that reading and understanding a frequency distribution is actually easier than forming one. When we constructed a frequency distribution with qualitative data, the raw data could be categorized in a well-defined way. With quantitative data, we must make certain decisions about the number of classes, as well as the width of each class. We do not apply concrete rules when we define the classes in Table 2.6; however, we are able to follow several guidelines.

## Guidelines for Constructing a Frequency Distribution

- *Classes are mutually exclusive.* In other words, classes do not overlap. Each observation falls into one, and only one, class. For instance, suppose a value of 400 appeared in Table 2.1. Given the class divisions in Table 2.6, we would have included this observation in the second class interval. Mathematically, the second class interval is expressed as  $400 \leq \text{Price} < 500$ . Alternatively, we can define the second interval as  $400 < \text{Price} \leq 500$ , in which case the value 400 is included in the previous class interval. In short, no matter the specification of the classes, the observation is included in only one of the classes.
- *Classes are exhaustive.* The total number of classes covers the entire sample (or population). In Table 2.6, if we had left off the last class, 700 up to 800, then we would be omitting two observations from the sample.
- *The total number of classes in a frequency distribution usually ranges from 5 to 20.* Smaller data sets tend to have fewer classes than larger data sets. Recall that the goal of constructing a frequency distribution is to summarize the data in a form that accurately depicts the group as a whole. If we have too many classes, then this advantage

of the frequency distribution is lost. For instance, suppose we create a frequency distribution for the house-price data with 17 classes, each of width 25, as shown in Table 2.7. Technically, this is a valid frequency distribution, but the summarization advantage of the frequency distribution is lost because there are too many class intervals. Similarly, if the frequency distribution has too few classes, then considerable accuracy and detail are lost. Consider a frequency distribution of the house-price data with three classes, each of width 150, as shown in Table 2.8.

**TABLE 2.8** Too Few Classes in a Distribution

Class (in \$1000s)	Frequency
300 up to 450	12
450 up to 600	17
600 up to 750	7
	Total = 36

Again, this is a valid frequency distribution. However, we cannot tell whether the 17 houses that sold for \$450,000 up to \$600,000 fall closer to the price of \$450,000, fall closer to the price of \$600,000, or are evenly spread within the interval. With only three classes in the frequency distribution, too much detail is lost.

- Once we choose the number of classes for a raw data set, we can then *approximate the width of each class* by using the formula

$$\frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}.$$

Generally, the width of each class is the same for each class interval. If the class width varied, comparisons between the numbers of observations in different intervals would be misleading.

- It is preferable to *define class limits that are easy to recognize and interpret*. Suppose we conclude, as we do in Table 2.6, that we should have five classes in the frequency distribution for the house-price data. Applying the class-width formula with the largest value of 735 and the smallest value of 330 (from Table 2.1) yields  $\frac{735 - 330}{5} = 81$ . Table 2.9 shows the frequency distribution with five classes and a class width of 81.

**TABLE 2.9** Cumbersome Class Width in a Distribution

Class (in \$1000s)	Frequency
330 up to 411	4
411 up to 492	11
492 up to 573	12
573 up to 654	3
654 up to 735	6
	Total = 36

Again, this is a valid frequency distribution, but it proves unwieldy. Recall that one major goal in forming a frequency distribution is to provide more clarity in

**TABLE 2.7** Too Many Classes in a Distribution

Class (in \$1000s)	Frequency
325 up to 350	2
350 up to 375	1
375 up to 400	1
400 up to 425	3
425 up to 450	5
450 up to 475	3
475 up to 500	0
500 up to 525	5
525 up to 550	5
550 up to 575	3
575 up to 600	1
600 up to 625	0
625 up to 650	1
650 up to 675	4
675 up to 700	0
700 up to 725	1
725 up to 750	1
	Total = 36

interpreting the data. Grouping the data in this manner actually makes analyzing the data more difficult. In order to facilitate interpretation of the frequency distribution, it is best to define class limits with ease of recognition in mind. To this end, and as initially shown in Table 2.6, we set the lower limit of the first class at 300 (rather than 330) and obtain the remaining class limits by successively adding 100 (rather than 81).

Once we have clearly defined the classes for a particular data set, the next step is to count and record the number of data points that fall into each class. As we did with the construction of a qualitative frequency distribution, we usually include a tally column to aid in counting (see Table 2.10), but then we remove this column in the final presentation of the frequency distribution. For instance, in Table 2.1, the first data point, 430, falls in the second class, so we place a tally mark in the second class; the next value of 520 falls in the third class, so we place a tally mark in the third class, and so on. The frequency column shows the numerical value of the respective tally count. Since four tally marks appear in the first class, we record the value 4 as its frequency—the number of observations that fall into the first class. One way to ensure that we have included all the data points in the frequency distribution is to sum the frequency column. This sum should always equal the population or sample size.

**TABLE 2.10** Constructing Frequency Distributions for the House-Price Data

Class (in \$1,000s)	Tally	Frequency	Cumulative Frequency
300 up to 400	IIII	4	4
400 up to 500	HH III I	11	$4 + 11 = 15$
500 up to 600	HH III III	14	$4 + 11 + 14 = 29$
600 up to 700	HH	5	$4 + 11 + 14 + 5 = 34$
700 up to 800	II	2	$4 + 11 + 14 + 5 + 2 = 36$
		Total = 36	

A frequency distribution indicates how many observations (in this case, house prices) fall within some range. However, we might want to know how many observations fall below the upper limit of a particular class. In these cases, our needs are better served with a cumulative frequency distribution.

The last column of Table 2.10 shows values for cumulative frequency. The cumulative frequency of the first class is the same as the frequency of the first class—that is, the value 4. However, the interpretation is different. With respect to the frequency column, the value 4 tells us that four of the houses sold in the \$300,000 up to \$400,000 range. For the cumulative frequency column, the value 4 tells us that four of the houses sold for less than \$400,000. To obtain the cumulative frequency for the second class—we add its frequency, 11, with the preceding frequency, 4, and obtain 15. This tells us that 15 of the houses sold for less than \$500,000. We solve for the cumulative frequencies of the remaining classes in a like manner. Note that the cumulative frequency of the last class is equal to the sample size of 36. This indicates that all 36 houses sold for less than \$800,000.

#### FREQUENCY AND CUMULATIVE FREQUENCY DISTRIBUTIONS FOR QUANTITATIVE DATA

For quantitative data, a **frequency distribution** groups data into intervals called **classes** and records the number of observations that falls into each class.

A **cumulative frequency distribution** records the number of observations that fall below the upper limit of each class.

### EXAMPLE 2.3

Using Table 2.10, how many of the houses sold in the \$500,000 up to \$600,000 range? How many of the houses sold for less than \$600,000?

**SOLUTION:** From the frequency distribution, we find that 14 houses sold in the \$500,000 up to \$600,000 range. In order to find the number of houses that sold for less than \$600,000, we use the cumulative frequency distribution. We readily observe that 29 of the houses sold for less than \$600,000.

Suppose we want to compare house prices in Mission Viejo, California, to house prices in another region of the United States. Just as for qualitative data, when making comparisons between two quantitative data sets—especially if the data sets differ in size—a relative frequency distribution tends to provide more meaningful information than a frequency distribution.

The second column of Table 2.11 shows the construction of a relative frequency distribution from the frequency distribution in Table 2.10. We take each class's frequency and divide by the total number of observations. For instance, we observed four houses that sold in the lowest range of \$300,000 up to \$400,000. We take the class frequency of 4 and divide by the sample size, 36, and obtain 0.11. Equivalently, we can say 11% of the houses sold in this price range. We make similar calculations for each class and note that when we sum the column of relative frequencies, we should get a value of one (or, due to rounding, a number very close to one).

**TABLE 2.11** Constructing Relative Frequency Distributions for House-Price Data

Class (in \$1,000s)	Relative Frequency	Cumulative Relative Frequency
300 up to 400	$4/36 = 0.11$	0.11
400 up to 500	$11/36 = 0.31$	$0.11 + 0.31 = 0.42$
500 up to 600	$14/36 = 0.39$	$0.11 + 0.31 + 0.39 = 0.81$
600 up to 700	$5/36 = 0.14$	$0.11 + 0.31 + 0.39 + 0.14 = 0.95$
700 up to 800	$2/36 = 0.06$	$0.11 + 0.31 + 0.39 + 0.17 + 0.06 \approx 1$
	Total = 1 (subject to rounding)	

The last column of Table 2.11 shows the cumulative relative frequency distribution. The cumulative relative frequency for a particular class indicates the proportion (fraction) of the observations that falls below the upper limit of that particular class. We can calculate the cumulative relative frequency of each class in one of two ways: (1) We can sum successive relative frequencies or (2) we can divide each class's cumulative frequency by the sample size. In Table 2.11 we show the first way. The value for the first class is the same as the value for its relative frequency—that is, 0.11. For the second class, we add 0.31 to 0.11 and obtain 0.42; this value indicates that 42% of the house prices were less than \$500,000. We continue calculating cumulative relative frequencies in this manner until we reach the last class. Here, we get the value one, which means that 100% of the houses sold for less than \$800,000.

### RELATIVE AND CUMULATIVE RELATIVE FREQUENCY DISTRIBUTIONS

For quantitative data, a **relative frequency distribution** identifies the proportion (or the fraction) of observations that falls into each class—that is,

$$\text{Class relative frequency} = \frac{\text{Class frequency}}{\text{Total number of observations}}$$

A **cumulative relative frequency distribution** records the proportion (or the fraction) of observations that fall below the upper limit of each class.

### EXAMPLE 2.4

Using Table 2.11, what percent of the houses sold for at least \$500,000 but not more than \$600,000? What percent of the houses sold for less than \$600,000? What percent of the houses sold for \$600,000 or more?

**SOLUTION:** The relative frequency distribution indicates that 39% of the houses sold for at least \$500,000 but not more than \$600,000. Further, the cumulative relative frequency distribution indicates that 81% of the houses sold for less than \$600,000. This result implies that 19% sold for \$600,000 or more.

### LO 2.4

Construct and interpret histograms, polygons, and ogives.

## Visualizing Frequency Distributions for Quantitative Data

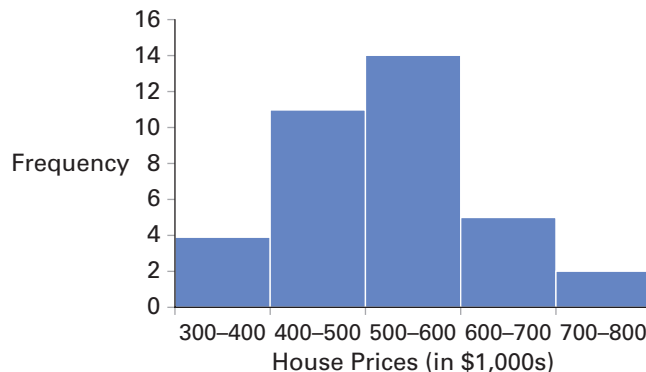
**Histograms** and **polygons** are graphical depictions of frequency and relative frequency distributions. The advantage of a visual display is that we can quickly see where most of the observations tend to cluster, as well as the spread and shape of the data. For instance, histograms and polygons may reveal whether or not the distribution is symmetrically shaped.

### GRAPHICAL DISPLAY OF QUANTITATIVE DATA: HISTOGRAM

A **histogram** is a series of rectangles where the width and height of each rectangle represent the class width and frequency (or relative frequency) of the respective class.

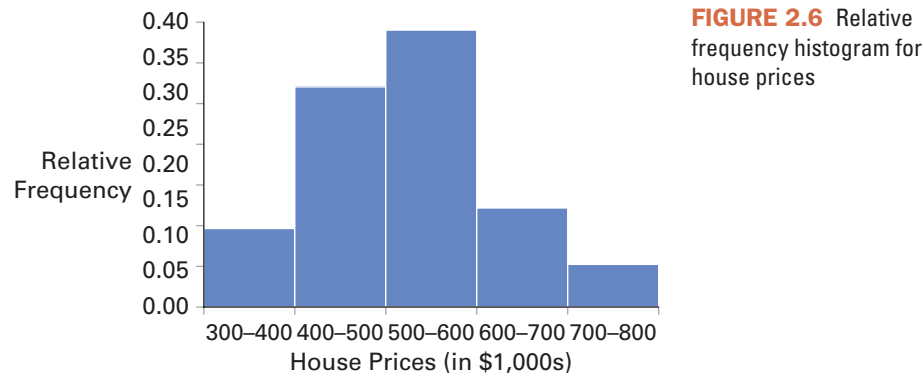
For quantitative data, a histogram is essentially the counterpart to the vertical bar chart we use for qualitative data. When constructing a histogram, we mark off the class limits along the horizontal axis. The height of each bar represents either the frequency or the relative frequency for each class. No gaps appear between the interval limits. Figure 2.5 shows a histogram for the frequency distribution of house prices shown in Table 2.6. A casual inspection of the histogram reveals that the selling price of houses in this sample ranged from \$300,000 to \$800,000; however, most house prices fell in the \$500,000 to \$600,000 range.

**FIGURE 2.5**  
Frequency histogram  
for house prices



The only difference between a frequency histogram and a relative frequency histogram is the unit of measurement on the vertical axis. For the frequency histogram, we use the frequency of each class to represent the height; for the relative frequency histogram we use the proportion (or the fraction) of each class to represent the height. In a relative frequency histogram, the area of any rectangle is proportional to the relative frequency of observations falling into that class. Figure 2.6 shows the relative frequency histogram for house prices.

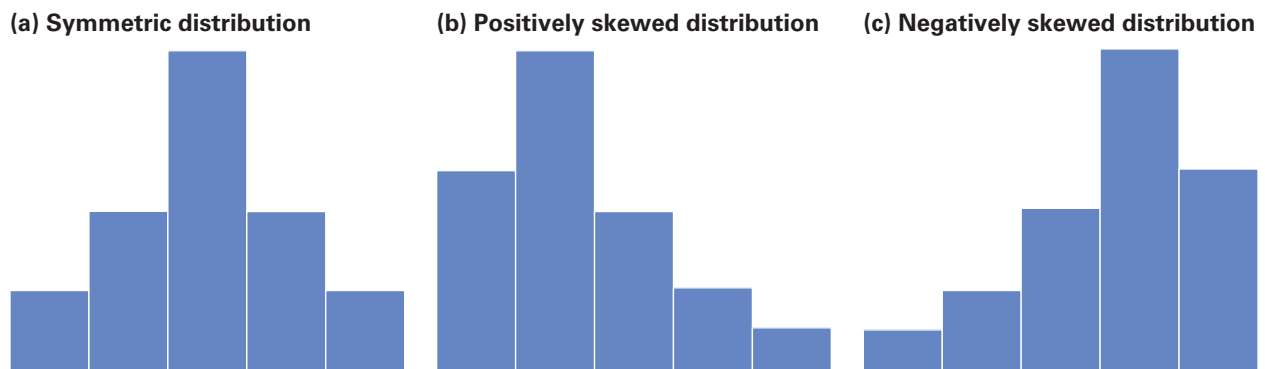




**FIGURE 2.6** Relative frequency histogram for house prices

In general, the shape of most data distributions can be categorized as either symmetric or skewed. A symmetric distribution is one that is a mirror image of itself on both sides of its center. That is, the location of values below the center correspond to those above the center. As we will see in later chapters, the smoothed histogram for many data sets approximates a bell-shaped curve, which is indicative of the well-known normal distribution. If the distribution is not symmetric, then it is either positively skewed or negatively skewed.

**FIGURE 2.7** Histograms with differing shapes



The histogram in Figure 2.7(a) shows a symmetric distribution. If the edges were smoothed, this histogram would look somewhat bell-shaped. In Figure 2.7(b), the histogram shows a positively skewed, or skewed to the right, distribution with a long tail extending to the right. This attribute reflects the presence of a small number of relatively large values. Finally, the histogram in Figure 2.7(c) indicates a negatively skewed, or skewed to the left, distribution since it has a long tail extending off to the left. Data that follow a negatively skewed distribution have a small number of relatively small values.

Though not nearly as skewed as the data exhibited in Figure 2.7(b), the house-price data in Figure 2.6 exhibit slight positive skew. This is the result of a few, relatively expensive homes in the city. It is common for distributions of house prices and incomes to exhibit positive skewness.

## Using Excel to Construct a Histogram

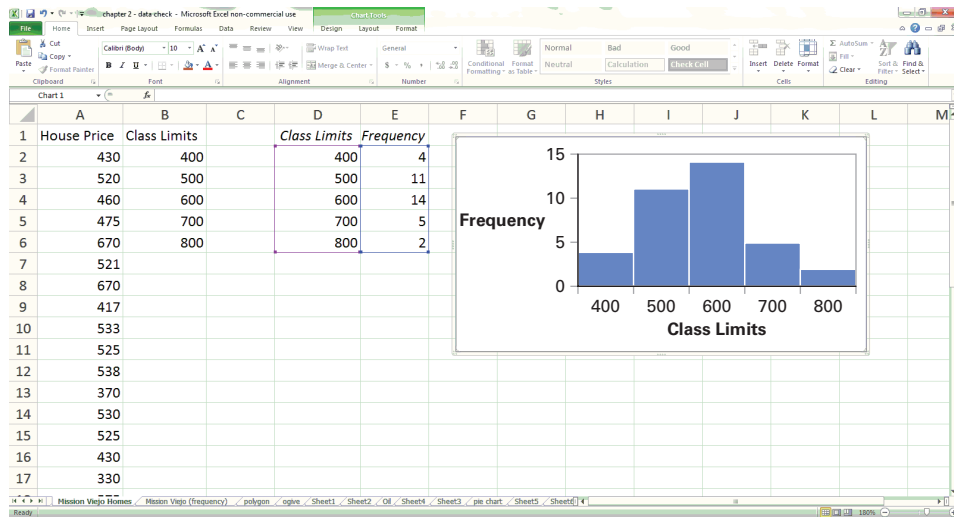
In general, Excel offers two different ways to construct a histogram, depending on whether we have access to the raw data or the frequency distribution. In either case, we need to have the classes clearly defined. We will first construct a histogram for house prices using the raw data from Table 2.1, and then show a histogram for the house prices from the frequency distribution from Table 2.6.

## Constructing a Histogram from a Set of Raw Data

**FILE**  
MV\_Houses

- A. **FILE** Open *MV\_Houses* (Table 2.1).
- B. In a column next to the data, enter the values of the upper limits of each class, or in this example, 400, 500, 600, 700, and 800; label this column “Class Limits.” The reason for these entries is explained in step D. The house-price data and the class limits (as well as the resulting frequency distribution and histogram) are shown in Figure 2.8.

**FIGURE 2.8** Constructing a histogram from raw data with Excel



- C. From the menu choose **Data > Data Analysis > Histogram > OK**. (Note: If you do not see the **Data Analysis** option under **Data**, you must *add in* this option. From the menu choose **File > Options > Add-Ins** and choose **Go** at the bottom of the dialog box. Select the box to the left of **Analysis Toolpak**, and then click **OK**. If you have installed this option properly, you should now see **Data Analysis** under **Data**.)
- D. In the *Histogram* dialog box (see Figure 2.9), under *Input Range*, select the data. Excel uses the term “bins” for the class limits. If we leave the *Bin Range* box empty, Excel creates evenly distributed intervals using the minimum and maximum values of the input range as end points. This methodology is rarely satisfactory. In order to construct a histogram that is more informative, we use the upper limit of each class as the bin values. Under *Bin Range*, we select the *Class Limits* data. (Check the *Labels* box if you have included the names House Price and Class Limits as part of the selection.) Under *Output Options*, we choose **Chart Output**, then click **OK**.

**FIGURE 2.9**  
Excel’s dialog box for a histogram

**Histogram**

Input  
Input Range: \$A\$1:\$A\$37  
Bin Range: \$B\$1:\$B\$6  
☒ Labels

Output options  
☐ Output Range:  
☒ New Worksheet Ply:  
☐ New Workbook  
☐ Pareto (sorted histogram)  
☐ Cumulative Percentage  
☒ Chart Output

OK Cancel Help

- E. Since Excel leaves spaces between the rectangles, we right-click on any of the rectangles, choose **Format Data Series** and change the *Gap Width* to 0, then choose **Close**. In the event that the given class limits do not include all the data points, Excel automatically adds another interval labeled “More” to the resulting frequency distribution and histogram. Since we observe zero observations in this interval for this example, we delete this interval for expositional purposes. Excel also defines its classes by excluding the value of the lower limit and including the value of the upper class limit for each interval. For example, if the value 400 appeared in the house-price data, Excel would have accounted for this observation in the first class. If any upper-limit value appeared in the house-price data, we would have adjusted the class limits in the *Bin Range* to 399, 499, etc., so that Excel’s frequency distribution and histogram would be consistent with those that we constructed in Table 2.10 and Figure 2.5. Further formatting regarding colors, axes, grids, etc. can be done by selecting **Layout** from the menu.

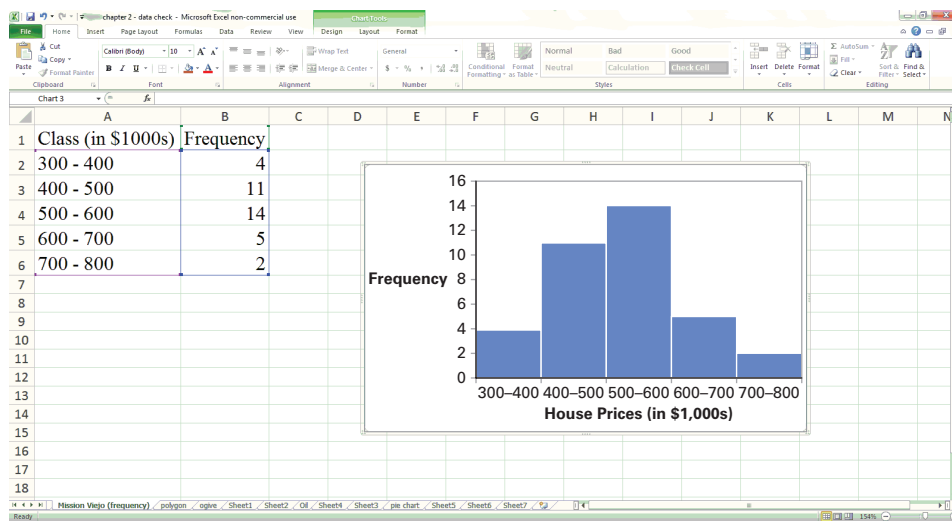
## Constructing a Histogram from a Frequency Distribution

Suppose we do not have the raw data for house prices, but we have the frequency distribution reported in Table 2.6.

- FILE** Open *MV\_Frequency* (Table 2.6).
- Select the classes and respective frequencies. See Figure 2.10 below.
- From the menu choose **Insert** > **Column** > **2-D Column** and choose the graph on the top left.
- In order to remove the spaces between the rectangles, right-click on any of the rectangles, choose **Format Data Series** and change the *Gap Width* to 0, then choose **Close**.
- Further formatting regarding colors, axes, grids, etc. can be done by selecting **Layout** from the menu.

**FILE**  
*MV\_Frequency*

**FIGURE 2.10** Constructing a histogram from a frequency distribution with Excel



A **polygon** provides another convenient way of depicting a frequency distribution. It too gives a general idea of the shape of a distribution. Like the histogram, we place either the frequency or the relative frequency of the distribution on the y-axis, and the upper and lower limits of each class on the x-axis. We plot the midpoint of each class with its corresponding frequency or relative frequency. We then connect neighboring points with a straight line.

### GRAPHICAL DISPLAY OF QUANTITATIVE DATA: POLYGON

A **polygon** connects a series of neighboring points where each point represents the midpoint of a particular class and its associated frequency or relative frequency.

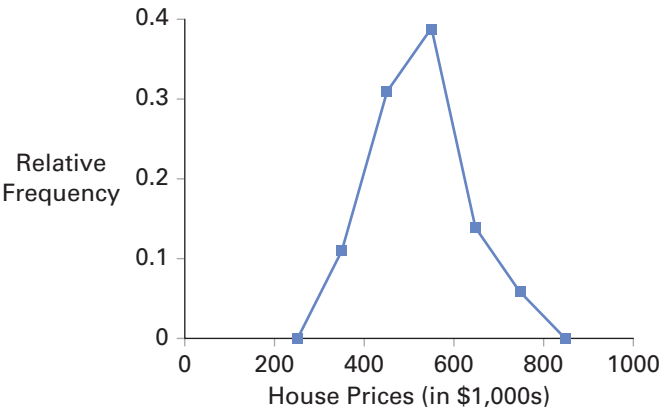
If we choose to construct a polygon for the house-price data, we first calculate the mid-point of each interval; thus, the midpoint for the first interval is  $\frac{300+400}{2} = 350$  and similarly, the midpoints for the remaining intervals are 450, 550, 650, and 750. We treat each midpoint as the  $x$ -coordinate and the respective frequency (or relative frequency) as the  $y$ -coordinate. After plotting the points, we connect neighboring points. In order to close off the graph at each end, we add one interval below the lowest interval (so, 200 up to 300 with midpoint 250) and one interval above the highest interval (so, 800 up to 900 with midpoint 850) and assign each of these classes zero frequencies. Table 2.12 shows the relevant coordinates for plotting a polygon using the house-price data. We chose to use relative frequency to represent the  $y$ -coordinate.

**TABLE 2.12** Coordinates for Plotting Relative Frequency Polygon

Classes	x-coordinate (midpoint)	y-coordinate (relative frequency)
(Lower end)	250	0
300–400	350	0.11
400–500	450	0.31
500–600	550	0.39
600–700	650	0.14
700–800	750	0.06
(Upper end)	850	0

Figure 2.11 plots a relative frequency polygon for the house-price data. Here the distribution appears to approximate the bell-shaped distribution discussed earlier. Only a careful inspection of the right tail suggests that the data are slightly positively skewed.

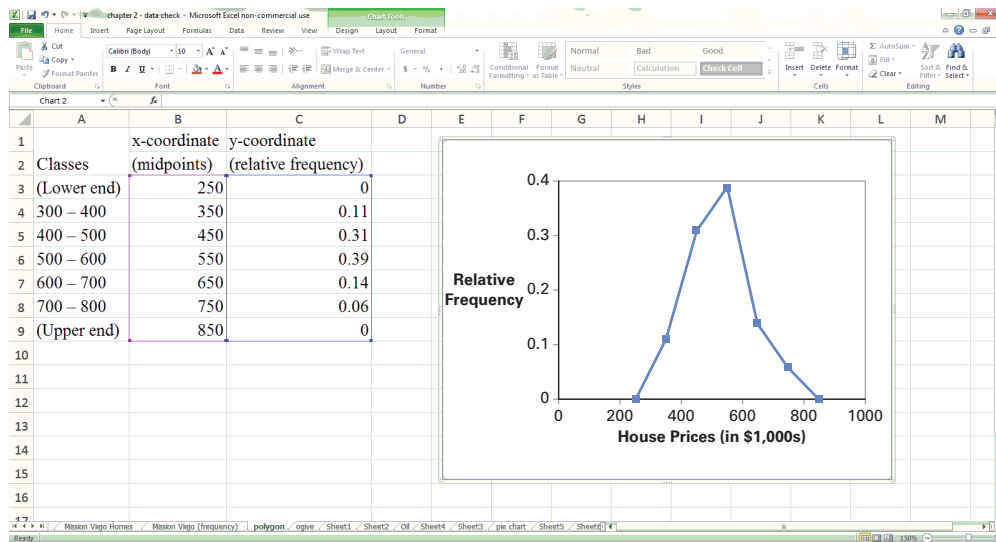
**FIGURE 2.11**  
Polygon for the  
house-price data



Using Excel to Construct a Polygon

- A. To construct a polygon, input the appropriate  $x$ - and  $y$ -coordinates into an Excel spreadsheet. We use the data from Table 2.12.
- B. Select the  $x$ - and the  $y$ -coordinates (as shown in Figure 2.12) and choose **Insert > Scatter**. Select the box at the middle right.
- C. Further formatting regarding colors, axes, grids, etc. can be done by selecting **Layout** from the menu.

**FIGURE 2.12** Constructing a polygon with Excel



In many instances, we might want to convey information by plotting an ogive (pronounced “ojive”).

#### GRAPHICAL DISPLAY OF QUANTITATIVE DATA: OGIVE

An **ogive** is a graph that plots the cumulative frequency or the cumulative relative frequency of each class against the upper limit of the corresponding class.

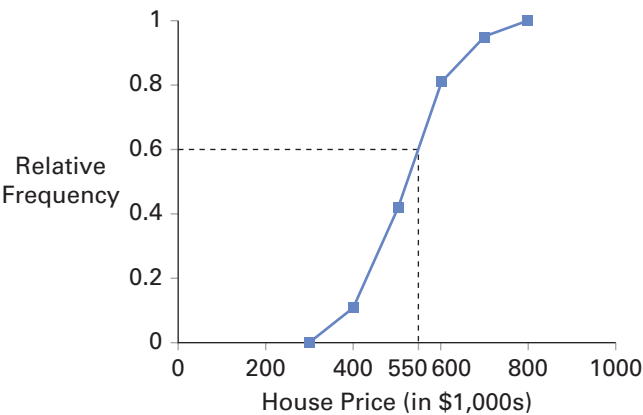
An ogive differs from a polygon in that we use the upper limit of each class as the  $x$ -coordinate and the cumulative frequency or cumulative relative frequency of the corresponding class as the  $y$ -coordinate. After plotting the points, we connect neighboring points. Lastly, we close the ogive only at the lower end by intersecting the  $x$ -axis at the lower limit of the first class. Table 2.13 shows the relevant coordinates for plotting an ogive using the house-price data. We choose to use cumulative relative frequency as the  $y$ -coordinate. The use of cumulative frequency would not change the shape of the ogive, just the unit of measurement on the  $y$ -axis.

**TABLE 2.13** Coordinates for the ogive for the house-price data

Classes	x-coordinate (upper limit)	y-coordinate (cumulative relative frequency)
(Lower end)	300	0
300–400	400	0.11
400–500	500	0.42
500–600	600	0.81
600–700	700	0.95
700–800	800	1

Figure 2.13 plots the ogive for the house-price data. In general, we can use an ogive to approximate the proportion of values that are less than a specified value on the horizontal axis. Consider an application to the house-price data in Example 2.5.

**FIGURE 2.13**  
Ogive for the  
house-price data



**EXAMPLE 2.5**

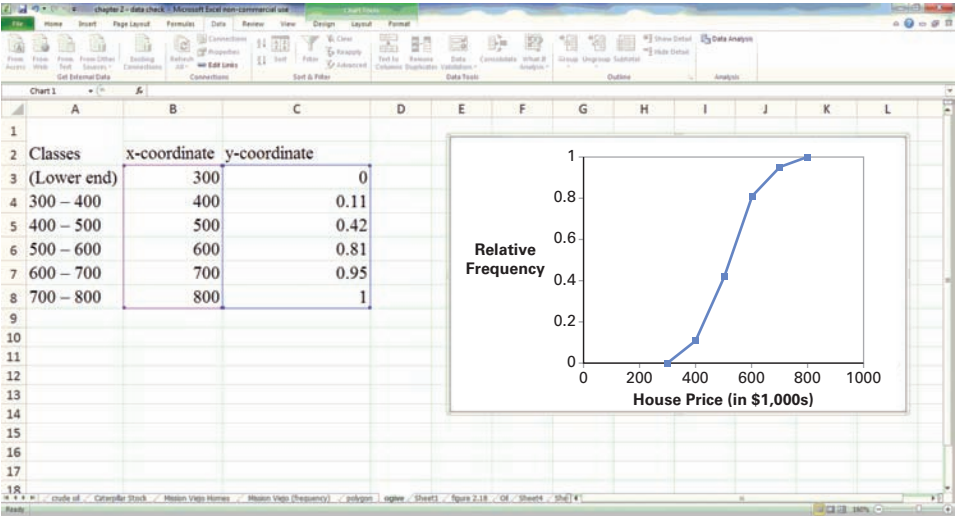
Using Figure 2.13, approximate the percentage of houses that sold for less than \$550,000.

**SOLUTION:** Draw a vertical line that starts at 550 and intersects the ogive. Then follow the line to the vertical axis and read the value. You can conclude that approximately 60% of the houses sold for less than \$550,000.

Using Excel to Construct an Ogive

- A. To construct an ogive, input the appropriate  $x$ - and  $y$ -coordinates into an Excel spreadsheet. We use the data from Table 2.13.
- B. Select the  $x$ - and the  $y$ -coordinates (as shown in Figure 2.14) and choose **Insert > Scatter**. Select the box at the middle right.
- C. Further formatting regarding colors, axes, grids, etc. can be done by selecting **Layout** from the menu.

**FIGURE 2.14** Constructing an ogive with Excel





## SYNOPSIS OF INTRODUCTORY CASE

During June 2010, Matthew Edwards reviewed the selling prices of 36 house sales in Mission Viejo, California, for a client from Seattle, Washington. After constructing various frequency distributions, he is able to make the following summary conclusions. House prices ranged from \$300,000 up to \$800,000 over this time period. Most of the houses (14) sold in the \$500,000 up to \$600,000 range, which is, more or less, the client's price range. Twenty-nine of the houses sold for less than \$600,000. Converting the data into percentages so the client can make comparisons with home sales in the Seattle area, Matthew found that 39% of the houses sold for \$500,000 up to \$600,000. Moreover, 81% of the houses sold for less than \$600,000, which implies that 19% sold for \$600,000 or more.



## EXERCISES 2.2

### Mechanics

15. Consider the following data set:

4	10	8	7	6	10	11	14	13	14
3	9	8	5	7	6	10	3	11	11
8	8	4	5	5	12	12	3	8	8

- Construct a frequency distribution using classes of 3 up to 5, 5 up to 7, etc.
- Construct relative frequency, cumulative frequency, and cumulative relative frequency distributions.
- How many of the observations are at least 7 but less than 9? How many of the observations are less than 9?
- What percentage of the observations are at least 7 but less than 9? What percentage of the observations are less than 9?
- Graph a relative frequency histogram.
- Graph an ogive.

16. Consider the following data set:

4	10	8	7	6	10	11	14	13	14
3	9	8	5	7	6	10	3	11	11
8	8	4	5	5	12	12	3	8	8
10	-9	28	14	-5	9	11	5	8	-3
33	-4	2	3	22	25	5	29	26	0
-8	-5	0	15	-4	35	21	15	19	23
4	6	-2	12	24	36	15	3	-5	2

- Construct a frequency distribution using classes of -10 up to 0, 0 up to 10, etc. How many of the observations are at least 10 but less than 20?
- Construct a relative frequency distribution and a cumulative relative frequency distribution. What

percent of the observations are at least 10 but less than 20? What percent of the observations are less than 20?

- Graph a relative frequency polygon. Is the distribution symmetric? If not, then how is it skewed?

17. Consider the following frequency distribution:

Class	Frequency
10 up to 20	12
20 up to 30	15
30 up to 40	25
40 up to 50	4

- Construct a relative frequency distribution. Graph a relative frequency histogram.
- Construct a cumulative frequency distribution and a cumulative relative frequency distribution.
- What percent of the observations are at least 30 but less than 40? What percent of the observations are less than 40?

18. Consider the following frequency distribution:

Class	Frequency
1000 up to 1100	2
1100 up to 1200	7
1200 up to 1300	3
1300 up to 1400	4

- Construct a relative frequency distribution. What percent of the observations are at least 1100 but less than 1200?
- Construct a cumulative frequency distribution and a cumulative relative frequency distribution. How many of the observations are less than 1300?
- Graph a frequency histogram.

19. Consider the following cumulative frequency distribution:

Class	Cumulative Frequency
15 up to 25	30
25 up to 35	50
35 up to 45	120
45 up to 55	130

- Construct a frequency distribution. How many observations are at least 35 but less than 45?
- Graph a frequency histogram.
- What percent of the observations are less than 45?

20. Consider the following relative frequency distribution:

Class	Relative Frequency
-20 up to -10	0.04
-10 up to 0	0.28
0 up to 10	0.26
10 up to 20	0.22
20 up to 30	0.20

- Suppose this relative frequency distribution is based on a sample of 50 observations. Construct a frequency distribution. How many of the observations are at least -10 but less than 0?
- Construct a cumulative frequency distribution. How many of the observations are less than 20?
- Graph a relative frequency polygon.

21. Consider the following cumulative relative frequency distribution.

Class	Cumulative Relative Frequency
150 up to 200	0.10
200 up to 250	0.35
250 up to 300	0.70
300 up to 350	1

- Construct a relative frequency distribution. What percent of the observations are at least 250 but less than 300?
- Graph an ogive.

## Applications

22. *Kiplinger's* (August 2007) lists the assets (in billions of \$) for the 20 largest stock mutual funds (ranked by size) as follows:

\$99.8	49.7	86.3	109.2	56.9
88.2	44.1	58.8	176.7	49.9
61.4	128.8	53.6	95.2	92.5
55.0	96.5	45.3	73.0	70.9

- Construct a frequency distribution using classes of 40 up to 70, 70 up to 100, etc.

- Construct the relative frequency distribution, the cumulative frequency distribution, and the cumulative relative frequency distribution.
- How many of the funds had assets of at least \$100 but less than \$130 (in billions)? How many of the funds had assets less than \$160 (in billions)?
- What percent of the funds had assets of at least \$70 but less than \$100 (in billions)? What percent of the funds had assets less than \$130 (in billions)?
- Construct a histogram. Comment on the shape of the distribution.

23. The number of text messages sent by 25 13-year-olds over the past month are as follows:

630	516	892	643	627	510	937	909	654
817	760	715	605	975	888	912	952	701
744	793	852	504	562	670	685		

- Construct a frequency distribution using classes of 500 up to 600, 600 up to 700, etc.
- Construct the relative frequency distribution, the cumulative frequency distribution, and the cumulative relative frequency distribution.
- How many of the 13-year-olds sent at least 600 but less than 700 text messages? How many sent less than 800 text messages?
- What percent of the 13-year-olds sent at least 500 but less than 600 text messages? What percent of the 13-year-olds sent less than 700 text messages?
- Construct a polygon. Comment on the shape of the distribution.

24. AccuWeather.com listed the following high temperatures (in degrees Fahrenheit) for 33 European cities on July 21, 2010.

75	92	81	85	90	73	94	95	81	64	85
62	84	85	81	86	90	79	74	90	91	95
88	87	81	73	76	86	90	83	75	92	83

- Construct a frequency distribution using classes of 60 up to 70, 70 up to 80, etc.
- Construct the relative frequency, the cumulative frequency, and the cumulative relative frequency distributions.
- How many of the cities had high temperatures less than 80°?
- What percent of the cities had high temperatures of at least 80° but less than 90°? What percent of the cities had high temperatures less than 90°?
- Construct a polygon. Comment on the shape of the distribution.

25. Fifty cities provided information on vacancy rates (in percent) in local apartments in the following frequency distribution.

Vacancy Rate (in percent)	Frequency
0 up to 3	5
3 up to 6	10
6 up to 9	20
9 up to 12	10
12 up to 15	5

- Construct the corresponding relative frequency distribution, cumulative frequency distribution, and cumulative relative frequency distribution.
  - How many of the cities had a vacancy rate less than 12%? What percent of the cities had a vacancy rate of at least 6% but less than 9%? What percent of the cities had a vacancy rate of less than 9%?
  - Construct a histogram. Comment on the shape of the distribution.
26. The following relative frequency distribution summarizes the ages of women who had a child in the last year.

Ages	Relative Frequency
15 up to 20	0.10
20 up to 25	0.25
25 up to 30	0.28
30 up to 35	0.24
35 up to 40	0.11
40 up to 45	0.02

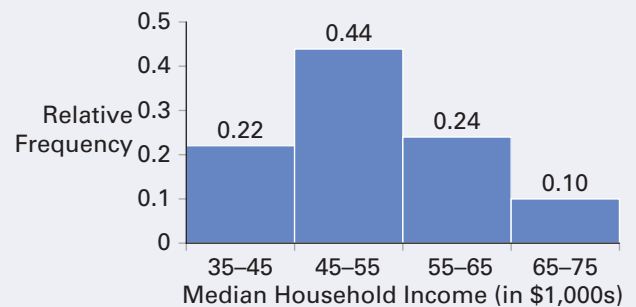
SOURCE: *The Statistical Abstract of the United States, 2010.*

- Assume the relative frequency distribution is based on a sample of 2,000 women. Construct the corresponding frequency distribution, cumulative frequency distribution, and cumulative relative frequency distribution.
  - What percent of the women were at least 25 but less than 30 years old? What percent of the women were younger than 35 years old?
  - Construct a relative frequency polygon. Comment on the shape of the distribution.
  - Construct an ogive. Using the graph, approximate the age of the middle 50% of the distribution.
27. The manager of a nightclub near a local university recorded the ages of the last 100 guests in the following cumulative frequency distribution.

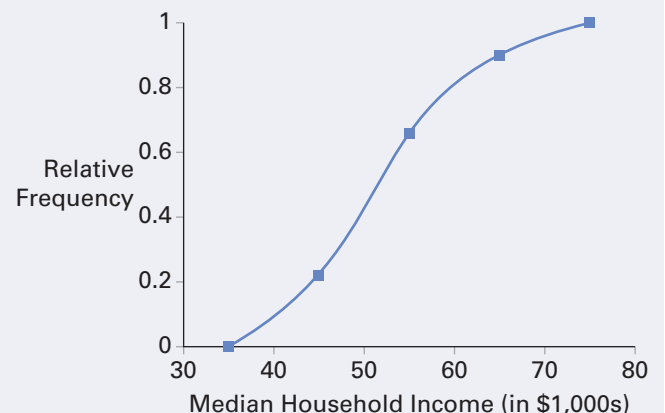
Ages	Cumulative Frequency
18 up to 22	45
22 up to 26	70
26 up to 30	85
30 up to 34	96
34 up to 38	100

- Construct the corresponding frequency, relative frequency, and cumulative relative frequency distributions.
- How many of the guests were at least 26 but less than 30 years old? What percent of the guests were at least 22 but less than 26 years old? What percent of the guests were younger than 34 years old? What percent were 34 years or older?
- Construct a histogram. Comment on the shape of the distribution.

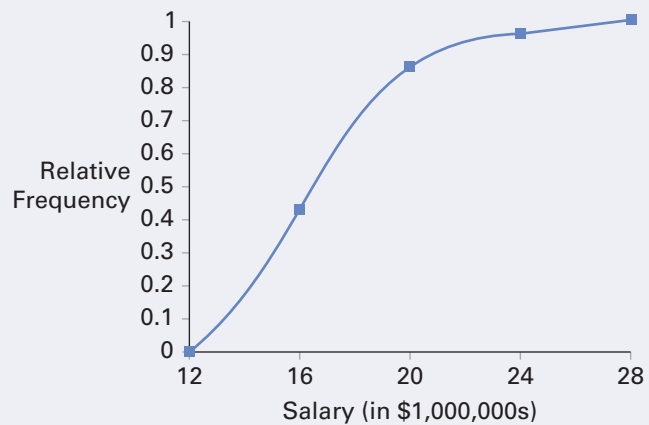
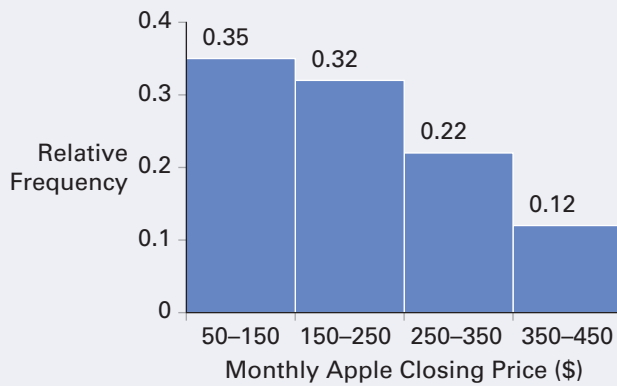
28. The following relative frequency histogram summarizes the median household income for the 50 states in the United States (*U.S. Census, 2010*).



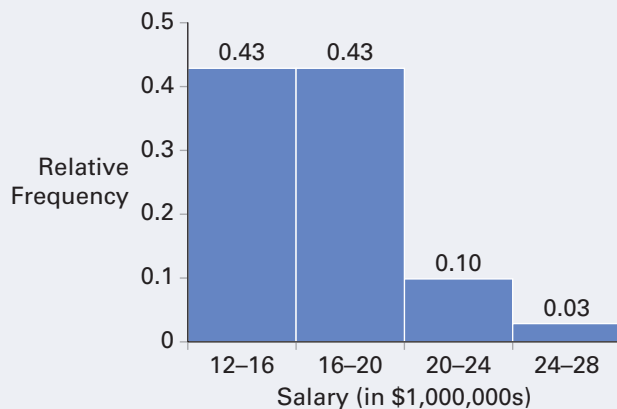
- Is the distribution symmetric? If not, is it positively or negatively skewed?
  - What percentage of the states had median household income between \$45,000 and \$55,000?
  - What percentage of the states had median household income between \$35,000 and \$55,000?
29. The following ogive summarizes the median household income for the 50 states in the United States (*U.S. Census, 2010*).



- Approximate the percentage of states with median household income less than \$50,000.
  - Approximate the percentage of states with median household income more than \$60,000.
30. The following histogram summarizes Apple Inc.'s monthly stock price for the years 2007 through 2011 (<http://finance.yahoo.com>, data retrieved April 20, 2012).



- a. Is the distribution symmetric? If not, is it positively or negatively skewed?
- b. Over this five-year period, approximate the minimum monthly stock price and the maximum monthly stock price.
- c. Over this five-year period, which class had the highest relative frequency.
31. The following histogram summarizes the salaries (in \$1,000,000s) for the 30 highest-paid players in the National Basketball Association (NBA) for the 2012 season (www.nba.com, data retrieved March 2012).



- a. Is the distribution symmetric? If not, is it positively or negatively skewed?
- b. How many NBA players earned between \$20,000,000 and \$24,000,000?
- c. Approximately how many NBA players earned between \$12,000,000 and \$20,000,000?
32. The following ogive summarizes the salary (in \$1,000,000s) for the 30 highest-paid players in the National Basketball Association (NBA) for the 2012 season (www.nba.com, data retrieved March 2012).

- a. Approximate the percentage of salaries that were less than \$18,000,000.
- b. Approximate the number of salaries that were more than \$14,000,000.
33. **FILE Math SAT.** The following table lists a portion of the average math SAT scores for each state for the year 2009.

State	SAT
Alabama	552
Alaska	516
⋮	⋮
Wyoming	568

SOURCE: www.collegeboard.com.

- a. Construct a frequency distribution and histogram using classes of 450 to 500, 501 to 550, etc. Comment on the shape of the distribution. How many of the states had scores between 551 and 600?
- b. Construct the relative frequency, the cumulative frequency, and the cumulative relative frequency distributions.
- c. How many of the states had math SAT scores of 550 or less?
- d. What percent of the states had math SAT scores between 551 and 600? What percent of the states had mean SAT scores of 550 or less?
34. **FILE Census.** The accompanying table shows a portion of median house values for the 50 states as reported by the U.S. Census Bureau in 2010.

State	House Value
Alabama	\$117,600
Alaska	229,100
⋮	⋮
Wyoming	174,000

- a. Construct a frequency distribution and a histogram for the median house values. Use six classes with upper limits of \$100,000, \$200,000, etc.

- b. Is the distribution symmetric? If not, is it positively or negatively skewed?
- c. Which class interval had the highest frequency?
- d. What percentage of the states had median house values between \$300,000 and \$400,000?
- e. How many of the states had median house values less than \$300,000?

35. **FILE Gas Prices.** The accompanying table shows a portion of the average price for a gallon of gas for the 50 states during April 2012.

State	Price per Gallon
Alabama	\$4.36
Alaska	3.79
⋮	⋮
Wyoming	3.63

SOURCE: www.AAA.com, data retrieved April 16, 2012.

- a. Construct a frequency distribution and histogram for the average gas prices. Use six classes with upper limits of \$3.70, \$3.90, etc.
- b. Is the distribution symmetric? If not, is it positively or negatively skewed?
- c. Which class interval had the highest frequency?
- d. Construct an ogive. Approximate the percentage of states that had average gas prices of \$3.90 or less.

Approximate the number of states that had average gas prices greater than \$3.90.

36. **FILE DJIA 2012.** For the first three months of 2012, the stock market put up its best first-quarter performance in over a decade (Money.cnn.com, April 9, 2012). The accompanying table shows a portion of the daily price index for the Dow Jones Industrial Average (DJIA) over this period.

Day	DJIA Price Index
January 3, 2012	12,397
January 4, 2012	12,418
⋮	⋮
March 31, 2012	13,212

SOURCE: Finance.yahoo.com, data retrieved April 20, 2012.

- a. Construct a frequency distribution and histogram for the DJIA price index. Use five classes with upper limits of 12,500, 12,750, etc. On how many days during this quarter was the DJIA less than 12,500?
- b. Construct a relative frequency polygon. Is the distribution symmetric? If not, is it positively or negatively skewed?
- c. Construct an ogive. Approximate the percentage of days that the DJIA was less than 13,000.

## 2.3 STEM-AND-LEAF DIAGRAMS

LO 2.5

John Tukey (1915–2000), a well-known statistician, provided another visual method for displaying quantitative data. A **stem-and-leaf diagram** is often a preliminary step when analyzing a data set. It is useful in that it gives an overall picture of where the data are centered and how the data are dispersed from the center.

Construct and interpret a stem-and-leaf diagram.

### GRAPHICAL DISPLAY OF QUANTITATIVE DATA: STEM-AND-LEAF DIAGRAMS

A **stem-and-leaf diagram** is constructed by separating each value of a data set into two parts: a *stem*, which consists of the leftmost digits, and a *leaf*, which consists of the last digit.

The best way to explain a stem-and-leaf diagram is to show an example.

### EXAMPLE 2.6

Table 2.14 shows the ages of the 25 wealthiest people in the world in 2010. Construct and interpret a stem-and-leaf diagram.

**TABLE 2.14** Wealthiest People in the World, 2010

Name	Age	Name	Age
Carlos Slim Helu	70	Li Ka-shing	81
William Gates III	54	Jim Walton	62
Warren Buffet	79	Alice Walton	60
Mukesh Ambani	52	Liliane Bettencourt	87
Lakshmi Mittal	59	S. Robson Walton	66
Lawrence Ellison	65	Prince Alwaleed Al Saud	54
Bernard Arnault	61	David Thomson	52
Eike Batista	53	Michael Otto	66
Amancio Ortega	74	Lee Shau Kee	82
Karl Albrecht	90	Michael Bloomberg	68
Ingvar Kamprad	83	Sergey Brin	36
Christy Walton	55	Charles Koch	74
Stefan Persson	62		

Reprinted by permission of Forbes Media LLC © 2011.

**SOLUTION:** For each age, we first decide that the number in the tens spot will denote the stem, thus leaving the number in the ones spot as the leaf. We then identify the lowest and highest values in the data set. Sergey Brin is the youngest member of this group at 36 years of age (stem: 3, leaf: 6) and Karl Albrecht is the oldest at 90 years of age (stem: 9, leaf: 0). These values give us the first and last values in the stem. This means our stems will be 3, 4, 5, 6, 7, 8, and 9, as shown in Panel A of Table 2.15.

**TABLE 2.15** Constructing a Stem-and-Leaf Diagram for Example 2.6

Panel A		Panel B		Panel C	
Stem	Leaf	Stem	Leaf	Stem	Leaf
3		3	6	3	6
4		4		4	
5		5	4 2 9 3 5 4 2	5	2 2 3 4 4 5 9
6		6	5 1 2 2 0 6 6 8	6	0 1 2 2 5 6 6 8
7	0	7	0 9 4 4	7	0 4 4 9
8		8	3 1 7 2	8	1 2 3 7
9		9	0	9	0

We then begin with the wealthiest man in the world, Carlos Slim Helu, whose age of 70 gives us a stem of 7 and a leaf of 0. We place a 0 in the row corresponding to a stem of 7, as shown in Panel A of Table 2.15. We continue this process with all the other ages and obtain the values in Panel B. Finally, in Panel C we arrange each individual leaf row in ascending order; this is the stem-and-leaf diagram in its final form.

The stem-and-leaf diagram (Panel C) presents the original 25 values in a more organized form. From the diagram we can readily observe that the ages range from 36 to 90. Wealthy individuals in their sixties make up the greatest group in the sample with eight members, while those in their fifties place a close second, accounting for seven members. We also note that the distribution is not perfectly symmetric. A stem-and-leaf diagram is similar to a histogram turned on its side with the added benefit of retaining the original values.



## EXERCISES 2.3

### Mechanics

37. Consider the following data set:

5.4	4.6	3.5	2.8	2.6	5.5	5.5	2.3	3.2	4.2
4.0	3.0	3.6	4.5	4.7	4.2	3.3	3.2	4.2	3.4

Construct a stem-and-leaf diagram. Is the distribution symmetric? Explain.

38. Consider the following data set:

-64	-52	-73	-82	-85	-80	-79	-65	-50	-71
-80	-85	-75	-65	-77	-87	-72	-83	-73	-80

Construct a stem-and-leaf diagram. Is the distribution symmetric? Explain.

### Applications

39. A sample of patients arriving at Overbrook Hospital's emergency room recorded the following body temperature readings over the weekend:

100.4	99.6	101.5	99.8	102.1	101.2	102.3	101.2	102.2	102.4
101.6	101.5	99.7	102.0	101.0	102.5	100.5	101.3	101.2	102.2

Construct and interpret a stem-and-leaf diagram.

40. Suppose the following high temperatures were recorded for major cities in the contiguous United States for a day in July.

84	92	96	91	96	94	93	82	81	76
90	95	84	90	84	98	94	90	83	78
88	96	106	78	92	98	91	84	80	94
94	93	107	87	77	99	94	73	74	92

Construct and interpret a stem-and-leaf diagram.

41. A police officer is concerned with excessive speeds on a portion of Interstate 90 with a posted speed limit of 65 miles per hour. Using his radar gun, he records the following speeds for 25 cars and trucks:

66	72	73	82	80	81	79	65	70	71
80	75	75	65	67	67	72	73	73	80
81	78	71	70	70					

Construct a stem-and-leaf diagram. Are the officer's concerns warranted?

42. Spain was the winner of the 2010 World Cup, beating the Netherlands by a score of 1–0. The ages of the players from both teams were as follows:

Spain									
29	25	23	30	32	25	29	30	26	29
21	28	24	21	27	22	25	21	23	24
Netherlands									
27	22	26	30	35	33	29	25	27	25
35	27	27	26	23	25	23	24	26	39

Construct a stem-and-leaf diagram for each country. Comment on similarities and differences between the two data sets.

## 2.4 SCATTERPLOTS

LO 2.6

All of the tabular and graphical tools presented thus far have focused on describing one variable. However, in many instances we are interested in the relationship between two variables. People in virtually every discipline examine how one variable may systematically influence another variable. Consider, for instance, how

- Incomes vary with education.
- Sales vary with advertising expenditures.
- Stock prices vary with corporate profits.
- Crop yields vary with the use of fertilizer.
- Cholesterol levels vary with dietary intake.
- Price varies with reliability.

Construct and interpret a scatterplot.

## SCATTERPLOT

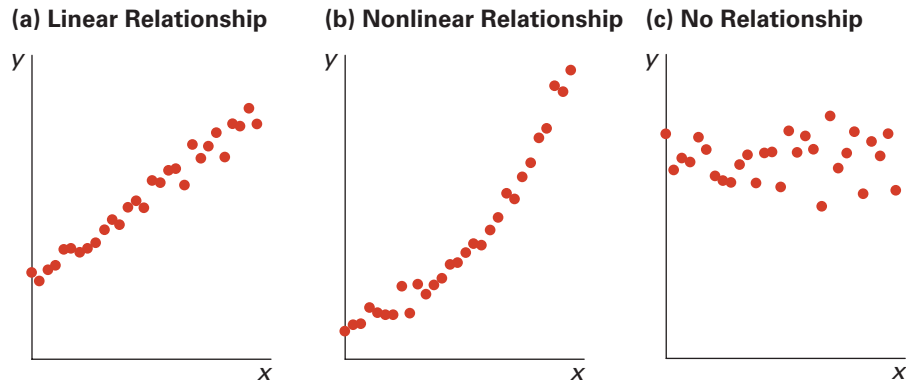
A **scatterplot** is a graphical tool that helps in determining whether or not two quantitative variables are related in some systematic way. Each point in the diagram represents a pair of observed values of the two variables.

When constructing a scatterplot, we generally refer to one of the variables as  $x$  and represent it on the horizontal axis and the other variable as  $y$  and represent it on the vertical axis. We then plot each pairing:  $(x_1, y_1)$ ,  $(x_2, y_2)$ , etc. Once the data are plotted, the graph may reveal that

- A linear relationship exists between the two variables;
- A nonlinear relationship exists between the two variables; or
- No relationship exists between the two variables.

For example, Figure 2.15(a) shows points on a scatterplot clustered together along a line with a positive slope; we infer that the two variables have a positive linear relationship. Part (b) depicts a positive nonlinear relationship; as  $x$  increases,  $y$  tends to increase at an increasing rate. The points in part (c) are scattered with no apparent pattern; thus, there is no relationship between the two variables.

**FIGURE 2.15** Scatterplots depicting relationships between two variables



In order to illustrate a scatterplot, consider the following example.

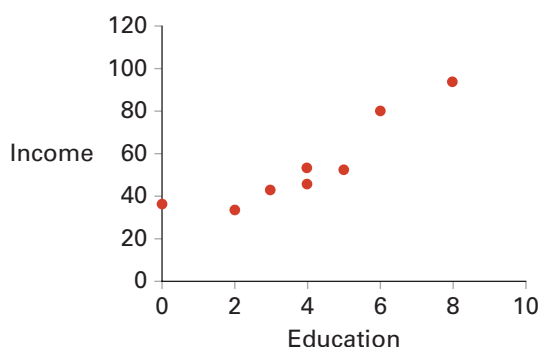
### EXAMPLE 2.7

A social scientist wants to analyze the relationship between educational attainment and income. He collects the data shown in Table 2.16, where Education refers to years of higher education and Income is the individual's annual income in thousands of dollars. Construct and interpret a scatterplot.

**TABLE 2.16** Education and Salary for Eight Individuals

Individual	Education	Income
1	3	45
2	4	56
3	6	85
4	2	35
5	5	55
6	4	48
7	8	100
8	0	38

**SOLUTION:** We let  $x$  and  $y$  denote Education and Income, respectively. We plot the first individual's pairing as (3, 45), the second individual's pairing as (4, 56), and so on. The graph should resemble Figure 2.16.

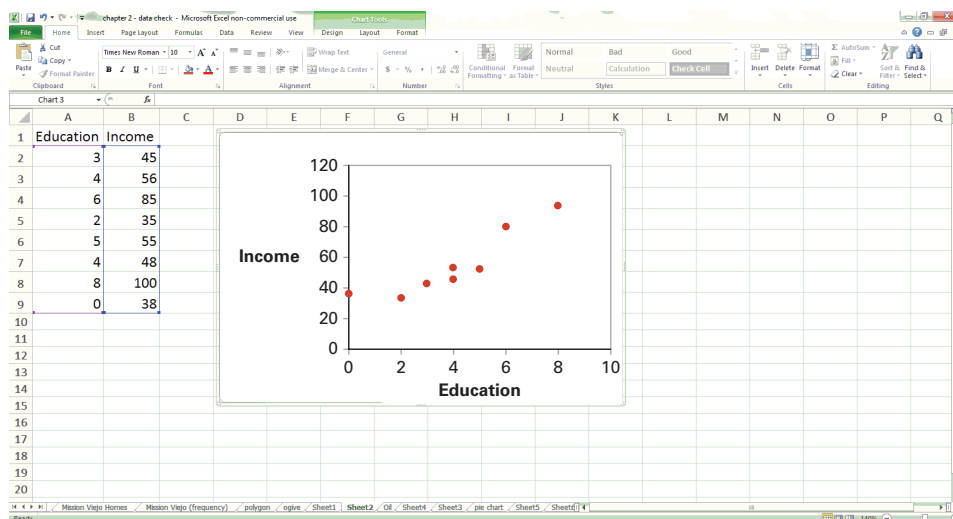


**FIGURE 2.16**  
Scatterplot of Income  
versus Education

As expected, we observe a positive relationship between the two variables; that is, when Education increases, Income tends to increase.

## Using Excel to Construct a Scatterplot

- To construct a scatterplot, input the appropriate  $x$ - and  $y$ -coordinates into an Excel spreadsheet. Here we use the data from Example 2.7.
- As shown in Figure 2.17, select the  $x$ - and  $y$ -coordinates and choose **Insert** > **Scatter**. Select the graph at the top left.



**FIGURE 2.17**  
Constructing a  
scatterplot with Excel

## EXERCISES 2.4

### Mechanics

43. Construct a scatterplot with the following data. Categorize the relationship between  $x$  and  $y$ .

$x$	3	7	12	5	6
$y$	22	10	5	14	12

44. Construct a scatterplot with the following data. Does a linear relationship exist between  $x$  and  $y$ ?

$x$	10	4	6	3	7
$y$	3	2	6	6	4

45. Construct a scatterplot with the following data. Categorize the relationship between  $x$  and  $y$ .

<i>x</i>	1	2	3	4	5	6	7	8
<i>y</i>	22	20	18	10	5	4	3	2

## Applications

46. A statistics instructor wants to examine whether a relationship exists between the hours a student spends studying for the final exam (Hours) and a student's grade on the final exam (Grade). She takes a sample of eight students.

Hours	8	2	3	8	10	15	25	5
Grade	75	47	50	80	85	88	93	55

Construct a scatterplot. What conclusions can you draw from the scatterplot?

47. A recent study offers evidence that the more weight a woman gains during pregnancy, the higher the risk of having a high-birth-weight baby, defined as at least 8 pounds, 13 ounces, or 4 kilograms (*The Wall Street Journal*, August 5, 2010). High-birth-weight babies are more likely to be obese in adulthood. The weight gain (in kilograms) of eight mothers and the birth weight of their newborns (in kilograms) are recorded in the accompanying table.

Mother's Weight Gain	Newborn's Birth Weight
18	4.0
7	2.5
8	3.0
22	4.5
21	4.0
9	3.5
8	3.0
10	3.5

Construct a scatterplot. Do the results support the findings of the study?

48. In order to diversify risk, investors are often encouraged to invest in assets whose returns have either a negative relationship or no relationship. The annual return data on two assets is shown in the accompanying table.

Return A	Return B
−20%	8%
−5	5
18	−1
15	−2
−12	2

Construct a scatterplot. For diversity purposes, would the investor be wise to include both of these assets in her portfolio? Explain.

49. In an attempt to determine whether a relationship exists between the price of a home and the number of days it takes to sell the home, a real estate agent collects data on the recent sales of eight homes.

Price (in \$1,000s)	Days to Sell Home
265	136
225	125
160	120
325	140
430	145
515	150
180	122
423	145

Construct a scatterplot. What can the realtor conclude?

## WRITING WITH STATISTICS

The tabular and graphical tools introduced in this chapter are the starting point for most studies and reports that involve statistics. They can help you organize data so you can see patterns and trends in the data, which can then be analyzed by the methods described in later chapters of this text. In this section, we present an example of using tabular and graphical methods in a sample report. Each of the remaining chapters contains a sample report incorporating the concepts developed in that respective chapter.

Camilla Walford is a newly hired journalist for a national newspaper. One of her first tasks is to analyze gas prices in the United States during the week of the Fourth of July holiday. She collects average gas prices for the 48 contiguous states and the District of Columbia (DC), a portion of which is shown in Table 2.17.

**FILE**  
Gas\_Prices\_2010

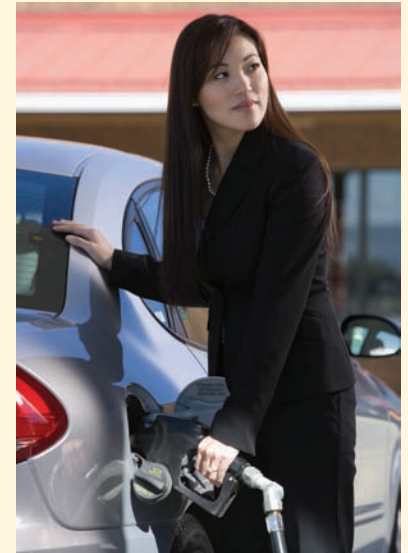
**TABLE 2.17** U.S. Gas Prices, July 2, 2010

State	Average Price (\$ per gallon)
Alabama	\$2.59
Arkansas	2.60
⋮	⋮
Wyoming	2.77

SOURCE: AAA's Daily Fuel Gauge Report, July 2, 2010.

Camilla wants to use the sample information to:

1. Construct frequency distributions to summarize the data.
2. Make summary statements concerning gas prices.
3. Convey the information from the distributions into graphical form.



Historically, in the United States, many people choose to take some time off during the Fourth of July holiday period and travel to the beach, the lake, or the mountains. The roads tend to be heavily traveled, making the cost of gas a concern. The following report provides an analysis of gas prices across the nation over this holiday period.

The analysis focuses on the average gas price for the 48 contiguous states and the District of Columbia (henceforth, referenced as 49 states for ease of exposition). The range of gas prices is from a low of \$2.52 per gallon (South Carolina) to a high of \$3.15 per gallon (California). To find out how gas prices are distributed between these extremes, the data have been organized into several frequency distributions as shown in Table 2.A. For instance, most states (17 of the 49) have an average gas price between \$2.70 and \$2.80 per gallon.

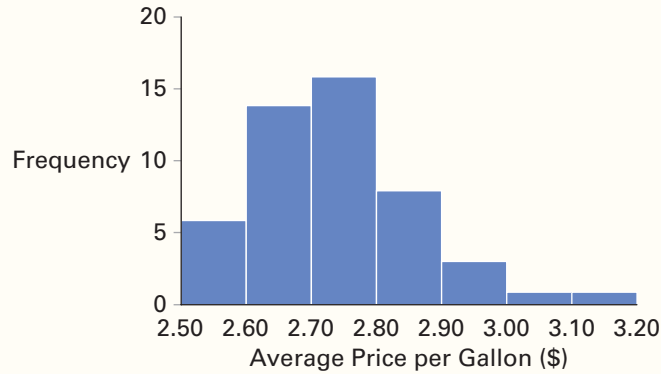
## Sample Report—Gas Prices across the United States

**TABLE 2.A** Frequency Distributions for Gas Prices in the United States, July 2, 2010

Average Price (\$ per gallon)	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
2.50 up to 2.60	5	0.10	5	0.10
2.60 up to 2.70	13	0.27	18	0.37
2.70 up to 2.80	17	0.35	35	0.72
2.80 up to 2.90	8	0.16	43	0.88
2.90 up to 3.00	4	0.08	47	0.96
3.00 up to 3.10	1	0.02	48	0.98
3.10 up to 3.20	1	0.02	49	1.00
Sample Size = 49				

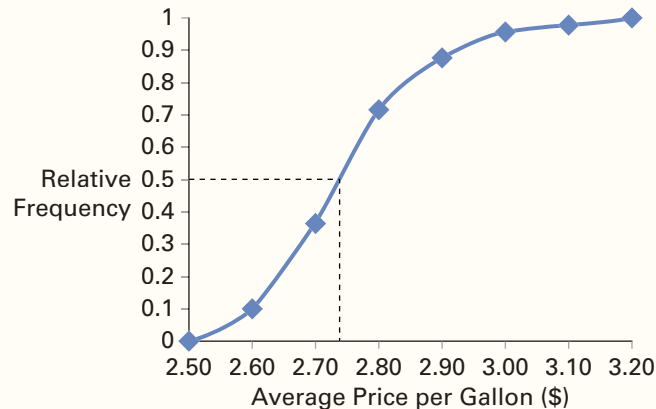
Equivalently, looking at the relative frequency column, 35% of the states have an average price in this range. The cumulative frequency column indicates that 35 states have an average price less than \$2.80 per gallon. Finally, the last column shows that the average price in 72% of the states (approximately three-quarters of the sample) is less than \$2.80 per gallon. Figure 2.A shows a histogram for gas prices, which graphs the frequency distribution from Table 2.A. This graph reinforces the fact that the average price of gas nationwide is between \$2.50 and \$3.20 per gallon. Moreover, gas prices are positively skewed since the distribution runs off to the right; only two states (California and Washington) have gas prices that are more than \$3.00 per gallon.

**FIGURE 2.A** Histogram of average gas prices nationwide



Another useful visual representation of the data is an ogive, shown in Figure 2.B. The ogive graphs the cumulative relative frequency distribution from Table 2.A. The ogive is useful for approximating the “middle” price. If we draw a horizontal line to the ogive at the 0.5 relative frequency mark, it intersects the plot at a point corresponding on the horizontal axis to a “middle price” of approximately \$2.75. This indicates that gas stations in approximately half of the states charged below this price and half charged above it.

**FIGURE 2.B** Ogive of average gas prices nationwide



## CONCEPTUAL REVIEW

### **LO 2.1** Summarize qualitative data by forming frequency distributions.

For **qualitative data**, a **frequency distribution** groups data into categories and records the number of observations that fall into each category. A **relative frequency distribution** shows the proportion (or the fraction) of observations in each category.

### **LO 2.2** Construct and interpret pie charts and bar charts.

Graphically, we can show a frequency distribution for qualitative data by constructing a **pie chart** or a **bar chart**. A pie chart is a segmented circle that clearly portrays the categories of some qualitative variable. A **bar chart** depicts the frequency or the relative frequency for each category of the qualitative variable as a series of horizontal or vertical bars, the lengths of which are proportional to the values that are to be depicted.



### LO 2.3 Summarize quantitative data by forming frequency distributions.

For quantitative data, a **frequency distribution** groups data into intervals called **classes**, and records the number of observations that falls into each class. A **cumulative frequency distribution** records the number of observations that falls below the upper limit of each class. A **relative frequency distribution** identifies the proportion (or the fraction) of observations that falls into each class. A **cumulative relative frequency distribution** shows the proportion (or the fraction) of observations that falls below the upper limit of each class.

### LO 2.4 Construct and interpret histograms, polygons, and ogives.

**Histograms** and **polygons** are graphical representations of frequency or relative frequency distributions for quantitative data. A casual inspection of these graphs reveals where most of the observations tend to cluster, as well as the general shape and spread of the data. An **ogive** is a graphical representation of a cumulative frequency or cumulative relative frequency distribution.

### LO 2.5 Construct and interpret a stem-and-leaf diagram.

A **stem-and-leaf diagram** is another visual method of displaying quantitative data. It is constructed by separating each value of a data set into a *stem*, which consists of the leftmost digits, and a *leaf*, which consists of the last digit. Like histograms and polygons, stem-and-leaf diagrams give an overall picture of where the data are centered and how the data are dispersed from the center.

### LO 2.6 Construct and interpret a scatterplot.

A **scatterplot** is a graphical tool that helps in determining whether or not two quantitative variables are related in some systematic way. Each point in the diagram represents a pair of observed values of the two variables.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

50. A 2003 survey by the Centers for Disease Control and Prevention concluded that smoking is forbidden in nearly 75% of U.S. households (*The Boston Globe*, May 25, 2007). The survey gathered responses from at least 900 households in each state. When residents of Utah were asked whether or not smoking was allowed in their households, a representative sample of responses was as follows:

No	No	No	No	No	No	Yes	No	No	No
No	Yes	No	No	No	No	No	No	No	No

When a similar survey was taken in Kentucky, a representative sample of responses was as follows:

No	No	Yes	No	Yes	No	Yes	Yes	No	No
No	Yes	Yes	No	Yes	No	No	Yes	Yes	No

- Construct a relative frequency distribution that summarizes the responses of residents from Utah and Kentucky. Comment on the results.
- Construct a bar chart that summarizes the results for each state.

51. Patrons at a local restaurant were asked to rate their recent experience at the restaurant with respect to its advertised atmosphere of upbeat, comfortable, and clean. Possible responses included Outstanding, Good, OK, and Horrible. The following table shows the responses of 28 patrons:

Horrible	OK	Horrible	Horrible
OK	OK	Horrible	Horrible
Horrible	OK	Horrible	Good
Horrible	Good	Good	Good
Horrible	OK	Horrible	OK
Good	Good	Horrible	Good
Horrible	OK	Horrible	Good

- Construct a relative frequency distribution that summarizes the responses of the patrons. Briefly summarize your findings. What recommendations would you make to the owner of the restaurant?
  - Use Excel to construct a pie chart and a bar chart for these data.
52. A survey conducted by CBS News asked parents about the professions they would want their children to pursue. The results are summarized in the following table.

Profession	Parents' Preference
Doctor, banker, lawyer, or president	65%
Internet mogul	13
Humanitarian-aid worker	6
Athlete	9
Movie star, rock star	2
Other	5

SOURCE: *Vanity Fair*, December 2009.

- Use Excel to construct a pie chart and a bar chart for these data.
  - How many parents wanted their children to become athletes if the results were based on 550 responses?
53. The one-year return (in %) for 24 mutual funds is as follows:

-14.5	-5.0	-3.7	2.5	-79	-11.2
4.8	-16.8	9.0	6.5	8.2	5.3
-12.2	15.9	18.2	25.4	3.4	-1.4
5.5	-4.2	-0.5	6.0	-2.4	10.5

- Construct a frequency distribution using classes of -20 up to -10, -10 up to 0, etc.
  - Construct the relative frequency, the cumulative frequency, and the cumulative relative frequency distributions.
  - How many of the funds had returns of at least 0% but less than 10%? How many of the funds had returns of 10% or more?
  - What percentage of the funds had returns of at least 10% but less than 20%? What percent of the funds had returns less than 20%?
54. *The Statistical Abstract of the United States, 2010* provided the following frequency distribution of the number of people who live below the poverty level by region.

Region	Number of People (in 1,000s)
Northeast	6,166
Midwest	7,237
South	15,501
West	8,372

- Construct a relative frequency distribution. What percentage of people who live below the poverty level live in the Midwest?
  - Use Excel to construct a pie chart and a bar chart for these data.
55. *Money* magazine (January 2007) reported that an average of 77 million adults in the United States make financial resolutions at the beginning of a new year. Consider the following frequency distribution, which reports the top financial resolutions of 1,026 Americans (MONEY/ICR poll conducted November 8–12, 2006).

Financial Resolution	Frequency
Saving more	328
Paying down debt	257
Making more income	154
Spending less	133
Investing more	103
Saving for a large purchase	41
Don't know	10

- Construct a relative frequency distribution for these data. What percentage of the respondents indicated that paying down debt was their top financial resolution?
  - Construct a bar chart.
56. A recent poll of 3,057 individuals asked: “What’s the longest vacation you plan to take this summer?” The following relative frequency distribution summarizes the results.

Response	Relative Frequency
A few days	0.21
A few long weekends	0.18
One week	0.36
Two weeks	0.25

- Construct a frequency distribution of these data. How many people are going to take a one-week vacation this summer?
  - Use Excel to construct a pie chart.
57. A survey conducted by CBS News asked 1,026 respondents: “What would you do with an unexpected tax refund?” The responses are summarized in the following table.

Pay off debts	47%
Put it in the bank	30%
Spend it	11%
I never get a refund	10%
Other	2%

Copyright © CBS News Archives. Used with permission.

- Construct a bar chart for these data.
  - How many people will spend the tax refund?
58. The following table reports the number of people as well as the number of people living below the poverty level across regions in the U.S. for the year 2013. (All numbers are in 1,000s.)

Region	Total	Below Poverty Level
Northeast	55,478	7,046
Midwest	66,785	8,590
South	116,961	18,870
West	73,742	10,812

SOURCE: [www.census.gov/hhes/www/poverty/data/incpovhlth/2013/table3.pdf](http://www.census.gov/hhes/www/poverty/data/incpovhlth/2013/table3.pdf), data retrieved March 23, 2015.

- a. Construct and interpret a pie chart that summarizes the proportion of people living in each region.
  - b. Construct and interpret a pie chart that summarizes the proportion of people living below the poverty level in each region. Is this pie chart consistent with the one you constructed in part (a); that is, in those regions that are relatively less populated, is the proportion of people living below the poverty level less?
59. The manager at a water park constructed the following frequency distribution to summarize attendance in July and August.

Attendance	Frequency
1,000 up to 1,250	5
1,250 up to 1,500	6
1,500 up to 1,750	10
1,750 up to 2,000	20
2,000 up to 2,250	15
2,250 up to 2,500	4

- a. Construct the corresponding relative frequency, cumulative frequency, and cumulative relative frequency distributions.
  - b. What is the most likely attendance range? How many times was attendance less than 2,000 people?
  - c. What percentage of the time was attendance at least 1,750 but less than 2,000 people? What percentage of the time was attendance less than 1,750 people? What percentage of the time was attendance 1,750 or more?
  - d. Construct a histogram. Comment on the shape of the distribution.
60. A researcher conducts a mileage economy test involving 80 cars. The frequency distribution describing average miles per gallon (mpg) appears in the following table.

Average mpg	Frequency
15 up to 20	15
20 up to 25	30
25 up to 30	15
30 up to 35	10
35 up to 40	7
40 up to 45	3

- a. Construct the corresponding relative frequency, cumulative frequency, and cumulative relative frequency distributions.
- b. How many of the cars got less than 30 mpg? What percentage of the cars got at least 20 but less than 25 mpg? What percentage of the cars got less than 35 mpg? What percent got 35 mpg or more?
- c. Construct a histogram. Comment on the shape of the distribution.

61. *The Wall Street Journal* (August 28, 2006) asked its readers: “Ideally, how many days a week, if any, would you work from home?” The following relative frequency distribution summarizes the responses from 3,478 readers.

Days Working from Home	Relative Frequency
0	0.12
1	0.18
2	0.30
3	0.15
4	0.07
5	0.19

Use Excel to construct a pie chart and a bar chart to summarize the data.

62. **FILE** *Wealthiest Americans*. The accompanying table lists a portion of the ages and net worth of the wealthiest people in America.

Name	Age	Net Worth (\$ billions)
William Gates III	53	50.0
Warren Buffet	79	40.0
⋮	⋮	⋮
Philip Knight	71	9.5

Source: *Forbes*, Special Report, September 2009.

- a. What percentage of the wealthiest people in America had net worth more than \$20 billion?
  - b. What percentage of the wealthiest people in America had net worth between \$10 billion and \$20 billion?
  - c. Construct a stem-and-leaf diagram on age. Comment on the shape of the distribution and how it compares with Table 2.15.
63. **FILE** *DOW PEG*. The price-to-earnings growth ratio, or PEG ratio, is the market’s valuation of a company relative to its earnings prospects. A PEG ratio of 1 indicates that the stock’s price is in line with growth expectations. A PEG ratio less than 1 suggests that the stock of the company is undervalued (typical of value stocks), whereas a PEG ratio greater than 1 suggests the stock is overvalued (typical of growth stocks). The accompanying table shows a portion of PEG ratios of companies listed on the Dow Jones Industrial Average.

Company	PEG Ratio
3M (MMM)	1.4
Alcoa (AA)	0.9
⋮	⋮
Walt Disney (DIS)	1.2

Source: [www.finance.yahoo](http://www.finance.yahoo), data retrieved April 13, 2011.

Construct a stem-and-leaf diagram on the PEG ratio. Interpret your findings.

64. The following table lists the sale price and type of 20 recently sold houses in New Jersey.

Price	Type	Price	Type
\$305,000	Ranch	\$568,000	Colonial
\$450,000	Colonial	\$385,000	Other
\$389,000	Contemporary	\$310,000	Contemporary
\$525,000	Other	\$450,000	Colonial
\$300,000	Ranch	\$400,000	Other
\$330,000	Contemporary	\$359,000	Ranch
\$355,000	Contemporary	\$379,000	Ranch
\$405,000	Colonial	\$509,000	Colonial
\$365,000	Ranch	\$435,000	Colonial
\$415,000	Ranch	\$510,000	Other

- Construct a frequency distribution on types of houses sold in New Jersey. Interpret your findings.
  - Construct a frequency distribution for house price using classes of \$300,000 up to \$350,000, \$350,000 up to \$400,000, etc.
  - Use a histogram and an ogive to summarize the data in part b.
65. A manager of a local retail store analyzes the relationship between Advertising (in \$100s) and Sales (in \$1,000s) by reviewing the store's data for the previous six months. Construct a scatterplot and comment on whether or not a relationship exists.

Advertising (in \$100s)	Sales (in \$1,000s)
20	15
25	18
30	20
22	16
27	19
26	20

66. The following table lists the National Basketball Association's (NBA's) leading scorers, their average minutes per game (MPG), and their average points per game (PPG) for 2008:

Player	MPG	PPG
D. Wade	38.6	30.2
L. James	37.7	28.4
K. Bryant	36.1	26.8
D. Nowitzki	37.3	25.9
D. Granger	36.2	25.8
K. Durant	39.0	25.3
C. Paul	38.5	22.8
C. Anthony	34.5	22.8
C. Bosh	38.0	22.7
B. Roy	37.2	22.6

SOURCE: [www.espn.com](http://www.espn.com).

Construct and interpret a scatterplot of PPG against MPG. Does a relationship exist between the two variables?

## CASE STUDIES

**CASE STUDY 2.1** There are six broad sectors that comprise the Dow Jones Industrial Average (DJIA). These are the broad areas in which the company conducts business. The following table shows a list of the 30 companies that comprise the DJIA and each company's sector.

**FILE**  
DJIA\_Sector

**Data for Case Study 2.1** Companies and Sectors of the DJIA

Company	Sector
3M (MMM)	Manufacturing
American Express (AXP)	Finance
⋮	⋮
Walmart (WMT)	Consumer

SOURCE: [www.money.cnn.com/data/dow30/](http://www.money.cnn.com/data/dow30/), information retrieved March 21, 2015.

In a report, use the sample information to:

- Construct a frequency distribution and a relative frequency distribution for the sectors that comprise the DJIA. Use pie charts for data visualization.
- Discuss how the various sectors are represented in the DJIA.

**CASE STUDY 2.2** When reviewing the overall strength of a particular firm, financial analysts typically examine the net profit margin. This statistic is generally calculated as the ratio of a firm's net profit after taxes (net income) to its revenue, expressed as a percentage. For example, a 20% net profit margin means that a firm has a net income of \$0.20 for each dollar of sales. A net profit margin can even be negative if the firm has a negative net income. In general, the higher the net profit margin, the more effective the firm is at converting revenue into actual profit. The net profit margin serves as a good way of comparing firms in the same industry, since such firms generally are subject to the same business conditions. However, financial analysts also use the net profit margin to compare firms in different industries in order to gauge which firms are relatively more profitable. The accompanying table shows a portion of net profit margins for a sample of clothing retailers.

**Data for Case Study 2.2** Net Profit Margin for Clothing Retailers

Firm	Net Profit Margin (in percent)
Abercrombie & Fitch	1.58
Aéropostale	10.64
⋮	⋮
Wet Seal	16.15

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com), data retrieved July 2010.

**FILE**  
*Net\_Profit\_Margins*

In a report, use the sample information to:

1. Provide a brief definition of net profit margin and explain why it is an important statistic.
2. Construct appropriate tables (frequency distribution, relative frequency distribution, etc.) and graphs that summarize the clothing industry's net profit margin. Use -5, 0, 5, etc. for the upper limits of the classes for the distributions.
3. Discuss where the data tend to cluster and how the data are spread from the lowest value to the highest value.
4. Comment on the net profit margin of the clothing industry, as compared to the beverage industry's net profit margin of approximately 10.9% (Source: [biz.yahoo.com](http://biz.yahoo.com), July 2010).

**CASE STUDY 2.3** The following table lists a portion of U.S. life expectancy (in years) for the 50 states.

**Data for Case Study 2.3** Life Expectancy by State, 2010–2011

Rank	State	Life Expectancy (in years)
1	Hawaii	81.5
2	Minnesota	80.9
⋮	⋮	⋮
50	Mississippi	74.8

SOURCE: [en.wikipedia.org/wiki/List\\_of\\_U.S.\\_states\\_by\\_life\\_expectancy](http://en.wikipedia.org/wiki/List_of_U.S._states_by_life_expectancy), data retrieved April 25, 2012.

**FILE**  
*Life\_Expectancy*

In a report, use the sample information to:

1. Construct appropriate tables (frequency distribution, relative frequency distribution, etc.) and graphs to summarize life expectancy in the United States. Use 75, 76.5, 78, etc. for the upper limits of the classes for the distributions.
2. Discuss where the data tend to cluster and how the data are spread from the lowest value to the highest value.
3. Comment on the shape of the distribution.

## APPENDIX 2.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### Pie Chart

##### FILE

*Marital\_Status*

- A. (Replicating Figure 2.1) From the menu choose **Graph > Pie Chart**. Select **Chart values from a table**, select Marital Status as the **Categorical variable**, and 1960 and 2010 as the **Summary variables**.
- B. Choose **Labels**. Select **Titles/Footnotes** and enter Marital Status, 1960 versus 2010. Then select **Slice Labels** and select **Category name** and **Percent**.
- C. Choose **Multiple Graphs**, and then select **On the same graph**.

#### Bar Chart

##### FILE

*Marital\_Status*

- A. (Replicating Figure 2.2) From the menu choose **Graph > Bar Chart**. From **Bars Represent** select **Values from a Table**, and from **Two-way Table** select **Cluster**.
- B. In the *Bar Chart—Two-Way Table—Cluster* dialog box, select 1960 and 2010 as **Graph variables**. Select Marital Status as **Row labels**. Under **Table Arrangement**, choose **Rows are outermost categories and columns are innermost**.

#### Histogram

*From Raw Data:*

##### FILE

*MV\_Houses*

- A. (Replicating Figure 2.5) From the menu choose **Graph > Histogram > Simple**. Click **OK**.
- B. Select House Price as **Graph Variables**. Click **OK**.
- C. Double-click *x*-axis and select **Edit Scale**. Under **Major Tick Positions**, choose **Position of Ticks** and enter 300 400 500 600 700 800. Under **Scale Range**, unclick **Auto** for *Minimum* and enter 300. Then unclick **Auto** for *Maximum* and enter 800. Select the **Binning** tab. Under **Interval Type**, select **Cut-point**. Under **Interval Definition**, select **Midpoint/Cutpoint Definitions** and enter 300 400 500 600 700 800.

*From a Frequency Distribution:*

##### FILE

*MV\_Frequency*

- A. (Replicating Figure 2.5) From the menu choose **Graph > Bar Chart**. From **Bars Represent** select **A function of a variable**, and from **One Y** select **Simple**. Click **OK**.
- B. Under **Function** select **Sum**. Under **Graph variables** select **Frequency**, and under **Categorical variable** select **Class (in \$1,000s)**.
- C. Double-click *x*-axis. Under **Space Between Scale Categories**, uncheck **Gap between Cluster** and enter 0.

#### Polygon

- A. (Replicating Figure 2.11) Input the *x*- and *y*-coordinates from Table 2.12 into a Minitab spreadsheet.
- B. From the menu choose **Graph > Scatterplot > With Connect Line**.
- C. Under **Y variables** select *y*-coordinate, and under **X variables** select *x*-coordinate.



## Ogive

- A. (Replicating Figure 2.13) Input the  $x$ - and  $y$ -coordinates from Table 2.13 into a Minitab spreadsheet.
- B. From the menu choose **Graph > Scatterplot > With Connect Line**.
- C. Under **Y variables** select  $y$ -coordinate, and under **X variables** select  $x$ -coordinate.

## Scatterplot

- A. (Replicating Figure 2.16) Input the Education and Income data from Example 2.7 into a Minitab spreadsheet.
- B. From the menu choose **Graph > Scatterplot > Simple**.
- C. Under **Y variables** select Income, and under **X variables** select Education.

## SPSS

### Pie Chart

- A. (Replicating Figure 2.1) From the menu choose **Graphs > Legacy Dialogs > Pie**. Under **Data in Chart Are**, select **Values of individual cases**. Click **Define**.
- B. Under **Slices Represent**, select 1960. Under **Slices Labels**, select **Variable**, then select Marital Status.
- C. Double-click on the graph to open **Chart Editor**, and then choose **Elements > Show Data Labels**. In the **Properties** dialog box, under **Displayed** select Percent and Marital Status.

**FILE**  
*Marital\_Status*

### Bar Chart

- A. (Replicating Figure 2.2) From the menu choose **Graphs > Legacy Dialogs > Bar**. Choose **Clustered**. Under **Data in Chart Are**, select **Values of individual cases**. Click **Define**.
- B. Under **Bars Represent**, select 1960 and 2010. Under **Category Labels**, select **Variable**, then select Marital Status.

**FILE**  
*Marital\_Status*

### Histogram

- A. (Replicating Figure 2.5) From the menu choose **Graphs > Legacy Dialogs > Histogram**. Under **Variable**, select HousePrice.
- B. In the Output window, double-click on Frequency ( $y$ -axis title), choose the **Scale** tab, and under **Range**, enter 0 as **Minimum**, 15 as **Maximum**, and 5 as **Major Increment**. Then click **Apply**.
- C. Double-click on the bars. Choose the **Binning** tab, and under **X Axis**, select **Custom** and **Interval width**, and enter 100 for the interval width. Then click **Apply**.

**FILE**  
*MV\_Houses*

### Polygon

- A. (Replicating Figure 2.11) Input the  $x$ - and  $y$ -coordinates from Table 2.12 into an SPSS spreadsheet.
- B. From the menu choose **Graphs > Legacy Dialogs > Scatter/Dot**. Choose **Simple Scatter** and then click **Define**.
- C. Under **Y Axis** select  $y$ , and under **X Axis** select  $x$ . Click **OK**.
- D. In the Output window, double-click on the graph to open the **Chart Editor**, then from the menu choose **Elements > Interpolation Line**.

## Ogive

- A. (Replicating Figure 2.13) Input the  $x$ - and  $y$ -coordinates from Table 2.13 into an SPSS spreadsheet.

- B. From the menu choose **Graphs > Legacy Dialogs > Scatter/Dot**. Choose **Simple Scatter**. Then click **Define**.
- C. Under **Y Axis**, select *y*, and under **X Axis**, select *x*. Click **OK**.
- D. In the Output window, double-click on the graph to open the **Chart Editor**, then from the menu choose **Elements > Interpolation Line**. Then click **Apply**. From the menu choose **Edit > Select X Axis**. Choose the **Scale** tab, and under **Range**, enter 300 as **Minimum**, 800 as **Maximum**, 100 as **Major Increment**, and 300 as **Origin**. Then click **Apply**.

### Scatterplot

- A. (Replicating Figure 2.16) Input the Education and Income data from Example 2.7 into an SPSS spreadsheet.
- B. From the menu choose **Graphs > Legacy Dialogs > Scatter/Dot**. Choose **Simple Scatter**. Then click **Define**.
- C. Under **Y Axis**, select *Income*, and under **X Axis**, select *Education*. Click **OK**.
- D. In the Output window, double-click on the graph to open the **Chart Editor**. From the menu, choose **Edit > Select Y Axis**. Under **Range**, enter 0 as **Minimum**, 120 as **Maximum**, and 20 as **Major Increment**. Then click **Apply**. From the menu choose **Edit > Select X Axis**. Choose the **Scale** tab, and under **Range**, enter 0 as **Minimum**, 10 as **Maximum**, and 2 as **Major Increment**. Then click **Apply**.

## JMP

### Pie Chart

(Replicating Figure 2.1) From the menu choose **Graph > Chart**. Under **Select Columns**, select *Marital Status*, and then under **Cast Selected Columns into Roles**, select **Categories, X, Levels**. Under **Select Columns**, select 1960 and 2010, and then select **Statistics > % of Total**. Under **Options**, choose **Pie Chart**.

### Bar Chart

(Replicating Figure 2.2) From the menu choose **Graph > Chart**. Under **Select Columns**, select *Marital Status*, and then under **Cast Selected Columns into Roles**, select **Categories, X, Levels**. Under **Select Columns**, select 1960 and 2010, and select **Statistics > Data**. Under **Options**, select **Overlay** and **Bar Chart**.

### Histogram

- A. (Replicating Figure 2.5) From the menu choose **Analyze > Distribution**. Under **Select Columns**, select *House Price*, then under **Cast Selected Columns into Roles**, select **Y, columns**.
- B. Right-click on the *y*-axis and select **Axis Settings**. For **Minimum**, enter 300; for **Maximum** enter 800; and for **Increment**, enter 100.

### Polygon

- A. (Replicating Figure 2.11) Input the *x*- and *y*-coordinates from Table 2.12 into a JMP spreadsheet.
- B. From the menu choose **Graph > Overlay Plot**. Under **Select Columns**, select *x*-coordinate, and then under **Cast Selected Columns into Roles**, select **X**. Under **Select Columns**, select *y*-coordinate, and then under **Cast Selected Columns into Roles**, select **Y**.
- C. Click on the red triangle next to the title **Overlay Plot**. Select **Y Options > Connect Points**.

**FILE**

*Marital\_Status*

**FILE**

*Marital\_Status*

**FILE**

*MV\_Houses*

## Ogive

- A. (Replicating Figure 2.13) Input the  $x$ - and  $y$ -coordinates from Table 2.13 into a JMP spreadsheet.
- B. From the menu choose **Graph > Overlay Plot**. Under **Select Columns**, select  $x$ -coordinate, and then under **Cast Selected Columns into Roles**, select **X**. Under **Select Columns**, select  $y$ -coordinate, and then under **Cast Selected Columns into Roles**, select **Y**.
- C. Click on the red triangle next to the title **Overlay Plot**. Select **Y Options > Connect Points**.

## Scatterplot

- A. (Replicating Figure 2.16) Input the Education and Income data from Example 2.7 into a JMP spreadsheet.
- B. From the menu choose **Graph > Overlay Plot**. Under **Select Columns**, select Education, and then under **Cast Selected Columns into Roles**, select **X**. Under **Select Columns**, select Income, and then under **Cast Selected Columns into Roles**, select **Y**.

# 3

## LEARNING OBJECTIVES

**After reading this chapter  
you should be able to:**

- LO 3.1 Calculate and interpret the mean, the median, and the mode.**
- LO 3.2 Calculate and interpret percentiles and a box plot.**
- LO 3.3 Calculate and interpret a geometric mean return and an average growth rate.**
- LO 3.4 Calculate and interpret the range, the mean absolute deviation, the variance, the standard deviation, and the coefficient of variation.**
- LO 3.5 Explain mean-variance analysis and the Sharpe ratio.**
- LO 3.6 Apply Chebyshev's theorem, the empirical rule, and z-scores.**
- LO 3.7 Calculate the mean and the variance for grouped data.**
- LO 3.8 Calculate and interpret the covariance and the correlation coefficient.**

# Numerical Descriptive Measures

In Chapter 2, we learned how to summarize data by using tables and graphs so that we can extract meaningful information. In this chapter, we focus on numerical descriptive measures. These measures provide precise, objectively determined values that are easy to calculate, interpret, and compare with one another. We first calculate several measures of central location, which attempt to find a typical or central value for the data. In addition to analyzing the center, we need to know how the data vary around the center. Measures of dispersion gauge the underlying variability of the data. We use measures of central location and dispersion to introduce some popular applications, including the Sharpe ratio and the empirical rule. Finally, we discuss measures that examine the linear relationship between two variables. These measures assess whether two variables have a positive linear relationship, a negative linear relationship, or no linear relationship.



## INTRODUCTORY CASE

### Investment Decision

Rebecca Johnson works as an investment counselor at a large bank. Recently, an inexperienced investor asked Rebecca about clarifying some differences between two top-performing mutual funds from the last decade: Vanguard's Precious Metals and Mining fund (henceforth, Metals) and Fidelity's Strategic Income fund (henceforth, Income). The investor shows Rebecca the return data that he has accessed over the Internet, but he acknowledges that he has trouble interpreting it. Table 3.1 shows the return data for these two mutual funds for the years 2000–2009.

**TABLE 3.1** Returns (in percent) for the Metals and the Income Funds, 2000–2009

Year	Metals	Income	Year	Metals	Income
2000	–7.34	4.07	2005	43.79	3.12
2001	18.33	6.52	2006	34.30	8.15
2002	33.35	9.38	2007	36.13	5.44
2003	59.45	18.62	2008	–56.02	–11.37
2004	8.09	9.44	2009	76.46	31.77

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com).

Rebecca will use the above sample information to:

1. Determine the typical return of the mutual funds.
2. Evaluate the investment risk of the mutual funds.

A synopsis of this case is provided at the end of Section 3.4.

Calculate and interpret the mean, the median, and the mode.

The term *central location* relates to the way quantitative data tend to cluster around some middle or central value. Measures of central location attempt to find a typical or central value that describes the data. Examples include finding a typical value that describes the return on an investment, the number of defects in a production process, the salary of a business graduate, the rental price in a neighborhood, the number of customers at a local convenience store, and so on.

## The Mean

The **arithmetic mean** is the primary measure of central location. Generally, we refer to the arithmetic mean as simply the **mean** or the **average**. In order to calculate the mean of a data set, we simply add up the values of all the data points and divide by the number of data points in the population or sample.

### EXAMPLE 3.1

Let's use the data in Table 3.1 in the introductory case to calculate and interpret the mean return of the Metals fund and the mean return of the Income fund.

**SOLUTION:** Let's start with the mean return for the Metals fund. We first add all the returns and then divide by the number of returns as follows:

$$\text{Metals fund mean return} = \frac{-7.34 + 18.33 + \cdots + 76.46}{10} = \frac{246.54}{10} = 24.65\%.$$

Similarly, we calculate the mean return for the Income fund as:

$$\text{Income fund mean return} = \frac{4.07 + 6.52 + \cdots + 31.77}{10} = \frac{85.14}{10} = 8.51\%.$$

Thus, over the 10-year period 2000–2009, the mean return for the Metals fund was greater than the mean return for the Income fund, or equivalently,  $24.65\% > 8.51\%$ . These means represent typical annual returns resulting from a one-year investment.

All of us have calculated a mean before. What might be new for some of us is the notation used to express the mean as a formula. For instance, when calculating the mean return for the Metals fund, we let  $x_1 = -7.34$ ,  $x_2 = 18.33$ , and so on, and let  $n$  represent the number of observations in the sample. So our calculation for the mean can be written as

$$\text{Mean} = \frac{x_1 + x_2 + \cdots + x_{10}}{n}.$$

The mean of the sample is referred to as  $\bar{x}$  (pronounced x-bar). Also, we can denote the numerator of this formula using summation notation, which yields the following compact formula for the **sample mean**:  $\bar{x} = \frac{\sum x_i}{n}$ . We should also point out that if we had all the return data for this mutual fund, instead of just the data for the past 10 years, then we would have been able to calculate the **population mean**  $\mu$  as  $\mu = \frac{\sum x_i}{N}$ , where  $\mu$  is the Greek letter mu (pronounced as “mew”) and  $N$  is the number of observations in the population.

### THE MEAN

For sample values  $x_1, x_2, \dots, x_n$ , the **sample mean**  $\bar{x}$  is computed as

$$\bar{x} = \frac{\sum x_i}{n}.$$

For population values  $x_1, x_2, \dots, x_N$ , the **population mean**  $\mu$  is computed as

$$\mu = \frac{\sum x_i}{N}.$$



The calculation method is identical for the sample mean and the population mean except that the sample mean uses  $n$  observations and the population mean uses  $N$  observations, where  $n < N$ . We refer to the population mean as a **parameter** and the sample mean as a **statistic**. Since the population mean is generally unknown, we often use the sample mean to estimate the population mean.

The mean is used extensively in statistics. However, it can give a misleading description of the center of the distribution in the presence of extremely small or large values.

The mean is the most commonly used measure of central location. One weakness of this measure is that it is unduly influenced by **outliers** — that is, extremely small or large values.

Example 3.2 highlights the main weakness of the mean.

### EXAMPLE 3.2

Seven people work at Acetech, a small technology firm in Seattle. Their salaries over the past year are listed in Table 3.2. Compute the mean salary for this firm and discuss whether it accurately indicates a typical value.

**TABLE 3.2** Salaries of Employees at Acetech

Title	Salary
Administrative Assistant	\$ 40,000
Research Assistant	40,000
Computer Programmer	65,000
Senior Research Associate	90,000
Senior Sales Associate	145,000
Chief Financial Officer	150,000
President (and owner)	550,000

**SOLUTION:** Since the salaries of all employees of Acetech are included in Table 3.2, we calculate the population mean salary as:

$$\mu = \frac{\sum x_i}{N} = \frac{40,000 + 40,000 + \dots + 550,000}{7} = \$154,286.$$

It is true that the mean salary for this firm is \$154,286, but this value does not reflect the typical salary at this firm. In fact, six of the seven employees earn less than \$154,286. This example highlights the main weakness of the mean — that is, it is very sensitive to extreme observations (extremely large or extremely small values), or outliers.

## The Median

Since the mean can be affected by outliers, we often also calculate the **median** as a measure of central location. The median is the middle value of a data set. It divides the data in half; an equal number of observations lie above and below the median. Many government publications and other data sources publish both the mean and the median in order to accurately portray a data set's typical value. If the values of the mean and the median differ significantly, then it is likely that the data set contains outliers. For instance, in 2007 the U.S. Census Bureau determined that the median income

for American households was \$46,326, whereas the mean income was \$63,344. It is well documented that a small number of households in the United States have income considerably higher than the typical American household income. As a result, these top-earning households influence the mean by pushing its value significantly above the value of the median.

### THE MEDIAN

The **median** is the middle value of a data set. We arrange the data in ascending (smallest to largest) order and calculate the median as

- The middle value if the number of observations is odd, or
- The average of the two middle values if the number of observations is even.

The median is especially useful when outliers are present.

### EXAMPLE 3.3

Use the data in Table 3.2 to calculate the median salary of employees at Acetech.

**SOLUTION:** In Table 3.2, the data are already arranged in ascending order. We reproduce the salaries along with their relative positions.

Position:	1	2	3	4	5	6	7
Value:	\$40,000	40,000	65,000	90,000	145,000	150,000	550,000

Given seven salaries, the median occupies the 4th position. Thus, the median is \$90,000. Three salaries are less than \$90,000 and three salaries are greater than \$90,000. As compared to the mean income of \$154,286, the median in this case better reflects the typical salary.

### EXAMPLE 3.4

Use the data in Table 3.1 in the introductory case to calculate and interpret the median returns for the Metals and the Income funds.

**SOLUTION:** Let's start with the median return for the Metals fund. We first arrange the data in ascending order:

Position:	1	2	3	4	5	6	7	8	9	10
Value:	-56.02	-7.34	8.09	18.33	33.35	34.30	36.13	43.79	59.45	76.46

Given 10 observations, the median is the average of the values in the 5th and 6th positions. These values are 33.35 and 34.30, so the median is  $\frac{33.35 + 34.30}{2} = 33.83\%$ . Over the period 2000–2009, the Metals fund had a median return of 33.83%, which indicates that 5 years had returns less than 33.83% and 5 years had returns greater than 33.83%. A comparison of the median return (33.83%) and the mean return

(24.65%) reveals a mean that is less than the median by almost 10 percentage points, which indicates that the Metals data may possibly be affected by extremely small values; we will discuss the detection of outliers later. Thus, in order to give a more transparent description of a data's center, it is wise to report both the mean and the median.

Similarly, we can find the median for the Income fund as 7.34%. In this case, the median return of 7.34% does not appear to deviate drastically from the mean return of 8.51%. This is not surprising since a casual inspection reveals that the relative magnitude of very small or large values is weaker in the Income fund data.

Note that the mean and the median suggest that a typical annual return for the Metals fund is much higher than that for the Income fund. Then why would anyone want to invest in the Income fund? We will come back to this question later in this chapter, when we explore the risk associated with investing in these funds.

## The Mode

The **mode** of a data set is the value that occurs most frequently. A data set can have more than one mode, or even no mode. For instance, if we try to calculate the mode return for either the Metals fund or the Income fund in Table 3.1, we see that no value in either fund occurs more than once. Thus, there is no mode value for either fund. If a data set has one mode, then we say it is unimodal. If two or more modes exist, then the data set is multimodal; it is common to call it bimodal in the case of two modes. Generally, the mode's value as a measure of central location tends to diminish with data sets that have more than three modes.

### THE MODE

The **mode** is the most frequently occurring value in a data set. A data set may have no mode or more than one mode.

### EXAMPLE 3.5

Use the data in Table 3.2 to calculate the modal salary of employees at Acetech.

**SOLUTION:** The salary \$40,000 is earned by two employees. Every other salary occurs just once. So \$40,000 is the modal salary. Just because a value occurs with the most frequency does not guarantee that it best reflects the center of the data. It is true that the modal salary at Acetech is \$40,000, but most employees earn considerably more than this amount.

In the preceding examples, we used measures of central location to describe quantitative data. However, in many instances we want to summarize qualitative data, where the mode is the only meaningful measure of central location.

### EXAMPLE 3.6

Kenneth Forbes is a manager at the University of Wisconsin campus bookstore. There has been a recent surge in the sale of women's sweatshirts, which are available in three sizes: Small (S), Medium (M), and Large (L). Kenneth notes that the campus bookstore sold 10 sweatshirts over the weekend in the following sizes:



S	L	L	M	S	L	M	L	L	M
---	---	---	---	---	---	---	---	---	---

Comment on the data set and use the appropriate measure of central location that best reflects the typical size of a sweatshirt.

**SOLUTION:** This data set is an example of qualitative data. Here, the mode is the only relevant measure of central location. The modal size is L since it appears 5 times as compared to S and M, which appear 2 and 3 times, respectively. Often, when examining issues relating to the demand for a product, such as replenishing stock, the mode tends to be the most relevant measure of central location.

## Using Excel to Calculate Measures of Central Location

In general, Excel offers a couple of ways to calculate most of the descriptive measures that we discuss in this chapter.

### Excel's Formula Option

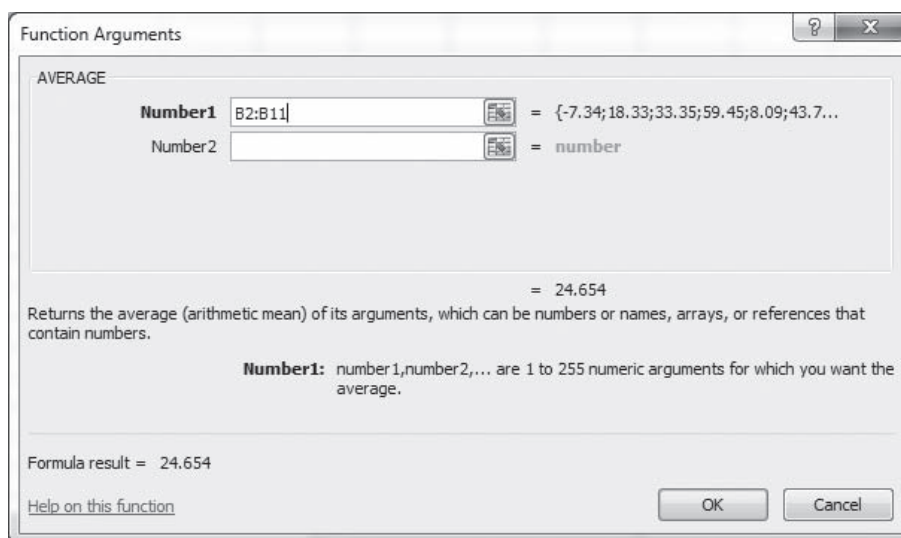
Excel provides built-in formulas for virtually every summary measure that we may need. To illustrate, we follow these steps to calculate the mean for the Metals fund.

#### FILE Fund\_Returns

- Open the *Fund\_Returns* data (Table 3.1) and select an empty cell.
- From the menu choose **Formulas > Insert Function**. In the *Insert Function* dialog box, choose **Statistical** under *Select a Category*. Here you will see a list of all the relevant summary measures that Excel calculates.
- Since we want to calculate the mean return for the Metals fund, under *Select a Function*, choose **AVERAGE**. Click **OK**.
- See Figure 3.1. In the *Average* dialog box, click on the box to the right of *Number 1* and then select the Metals data. Click **OK**. You should see the value 24.65, which equals the value that we calculated manually. In order to calculate the median and the mode, we repeat these steps, but we choose **MEDIAN** and **MODE** as the functions instead of **AVERAGE**.

Once you get familiar with Excel's function names, an easier way to perform these calculations is to select an empty cell in the spreadsheet and input "**=Function Name(array)**", where you replace Function Name with Excel's syntax for that particular function and select the relevant data for the array or input the cell designations. For example, when calculating the mean for the Metals fund, we input "**=AVERAGE(B2:B11)**"; the data for the Metals return data are occupying cells B2

through B11 on the spreadsheet. After choosing <Enter>, Excel returns the function result in the cell. When introducing new functions later in this chapter and other chapters, we will follow this format.



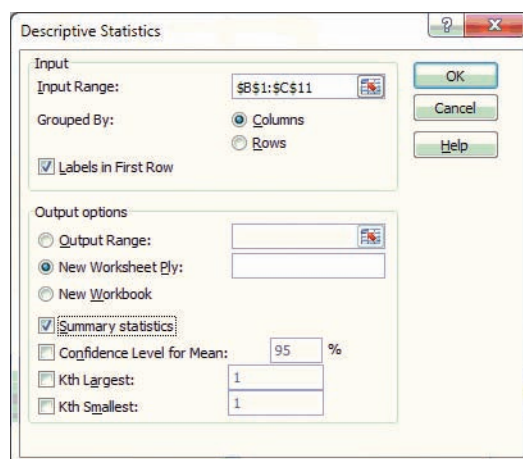
**FIGURE 3.1** Excel's AVERAGE dialog box

## Excel's Data Analysis Toolpak Option

Another way to obtain values for the mean, the median, and the mode is to use Excel's Data Analysis Toolpak option. One advantage of this option is that it provides numerous summary measures using a single command. Again, we illustrate this option using the data from the introductory case.

- A. Open the *Fund>Returns* data (Table 3.1).
- B. From the menu choose **Data > Data Analysis > Descriptive Statistics > OK**. (Note: As mentioned in Chapter 2, if you do not see **Data Analysis** under **Data**, you must *Add-in* the Analysis Toolpak option.)
- C. See Figure 3.2. In the *Descriptive Statistics* dialog box, click on the box next to *Input Range*, then select the data. If you included the fund names when you highlighted the data, make sure you click on the option next to *Labels in First Row*. Click the box in front of *Summary Statistics*. Then click **OK**.

**FILE**  
*Fund>Returns*



**FIGURE 3.2** Excel's Descriptive Statistics dialog box

- D. Table 3.3 presents the Excel output. If the output is difficult to read, highlight the data and choose **Home > Format > Column > Autofit Selection**. As noted earlier, Excel provides numerous summary measures; we have put the measures of central location in boldface. (Measures of dispersion are also in boldface; we analyze these measures in more detail shortly.) Note that Excel reports the mode as #N/A, which means “no value is available”; this is consistent with our finding that no value in the data appeared more than once.

**TABLE 3.3** Excel Output Using Descriptive Statistics Dialog Box

Metals		Income	
<b>Mean</b>	<b>24.654</b>	<b>Mean</b>	<b>8.514</b>
Standard Error	11.7414004	Standard Error	3.4997715
<b>Median</b>	<b>33.825</b>	<b>Median</b>	<b>7.335</b>
<b>Mode</b>	<b>#N/A</b>	<b>Mode</b>	<b>#N/A</b>
<b>Standard Deviation</b>	<b>37.1295681</b>	<b>Standard Deviation</b>	<b>11.067249</b>
<b>Sample Variance</b>	<b>1378.60483</b>	<b>Sample Variance</b>	<b>122.484</b>
Kurtosis	1.668701	Kurtosis	2.3615757
Skewness	−1.0076169	Skewness	0.5602496
<b>Range</b>	<b>132.48</b>	<b>Range</b>	<b>43.14</b>
Minimum	−56.02	Minimum	−11.37
Maximum	76.46	Maximum	31.77
Sum	246.54	Sum	85.14
Count	10	Count	10

In Chapter 2, we used histograms to discuss **symmetry** and **skewness**. Recall that the distribution is symmetric if one side of the histogram is a mirror image of the other side. For a symmetric and unimodal distribution, the mean, the median, and the mode are equal. In business applications, it is common to encounter data that are skewed. The mean is usually greater than the median when the data are positively skewed and less than the median, when the data are negatively skewed. We would also like to comment on the numerical measure of skewness that Excel reports, even though we will not discuss its calculation. A skewness coefficient of zero indicates the data values are evenly distributed on both sides of the mean. A positive skewness coefficient implies that extreme values are concentrated in the right tail of the distribution, pulling the mean up, relative to the median, and the bulk of values lie to the left of the mean. Similarly, a negative skewness coefficient implies that extreme values are concentrated in the left tail of the distribution, pulling the mean down, relative to the median, and the bulk of values lie to the right of the mean. We find that the returns are negatively skewed (Skewness = −1.0076) for the Metals fund and positively skewed (Skewness = 0.5602) for the Income fund.

### The Weighted Mean

So far we have focused on applications where each observation in the data contributed equally to the mean. The **weighted mean** is relevant when some observations contribute more than others. For example, a student is often evaluated on the basis of the weighted mean since the score on the final exam is typically worth more than the score on the midterm.

#### THE WEIGHTED MEAN

Let  $w_1, w_2, \dots, w_n$  denote the weights of the sample observations  $x_1, x_2, \dots, x_n$  such that  $w_1 + w_2 + \dots + w_n = 1$ . The **weighted mean** for the sample is computed as

$$\bar{x} = \sum w_i x_i.$$

The weighted mean for the population is computed similarly.



### EXAMPLE 3.7

A student scores 60 on Exam 1, 70 on Exam 2, and 80 on Exam 3. What is the student's average score for the course if Exams 1, 2, and 3 are worth 25%, 25%, and 50% of the grade, respectively?

**SOLUTION:** We define the weights as  $w_1 = 0.25$ ,  $w_2 = 0.25$ , and  $w_3 = 0.50$ . We compute the average score as  $\bar{x} = \sum w_i x_i = 0.25(60) + 0.25(70) + 0.50(80) = 72.50$ . Note that the unweighted mean is only 70 because it does not incorporate the higher weight given to the score on Exam 3.

## EXERCISES 3.1

### Mechanics

1. Given the following observations from a sample, calculate the mean, the median, and the mode.

8	10	9	12	12
---	----	---	----	----

2. Given the following observations from a sample, calculate the mean, the median, and the mode.

-4	0	-6	1	-3	-4
----	---	----	---	----	----

3. Given the following observations from a population, calculate the mean, the median, and the mode.

150	257	55	110	110	43	201	125	55
-----	-----	----	-----	-----	----	-----	-----	----

4. Given the following observations from a population, calculate the mean, the median, and the mode.

20	15	25	20	10	15	25	20	15
----	----	----	----	----	----	----	----	----

### Applications

5. At a small firm in Boston, seven employees were asked to report their one-way commute time (in minutes) into the city. Their responses were the following.

20	35	90	45	40	35	50
----	----	----	----	----	----	----

- a. How long was the shortest commute? The longest commute?
- b. Calculate the mean, the median, and the mode.
6. In order to get an idea on current buying trends, a real estate agent collects data on 10 recent house sales in the area. Specifically, she notes the number of bedrooms in each house as follows:

3	4	3	3	5	2	4	2	5	6
---	---	---	---	---	---	---	---	---	---

- a. Calculate the mean, the median, and the mode.

- b. Which measure of central location best reflects the typical value with respect to the number of bedrooms in recent house sales?

7. The following table shows the 10 highest-paid chief executive officers of the last decade.

Name	Firm	Compensation (in millions)
Lawrence Ellison	Oracle	\$1,835.7
Barry Diller	IAC, Expedia	1,142.9
Ray Irani	Occidental Petroleum	857.1
Steve Jobs	Apple	748.8
Richard Fairbank	Capital One	568.5
Angelo Mozilo	Countrywide	528.6
Eugene Isenberg	Nabors Industries	518.0
Terry Semel	Yahoo	489.6
Henry Silverman	Cendant	481.2
William McGuire	UnitedHealth Group	469.3

SOURCE: *The Wall Street Journal*, July 27, 2010.

- a. Calculate the mean compensation for the 10 highest-paid chief executive officers.
- b. Does the mean accurately reflect the center of the data? Explain.
8. An investor bought common stock of Microsoft Corporation on three occasions at the following prices.

Date	Price Per Share	Number of Shares
January 2009	\$19.58	70
July 2009	\$24.06	80
December 2009	\$29.54	50

Calculate the average price per share at which the investor bought these shares.

9. You score 90 on the midterm, 60 on the final, and 80 on the class project. What is your average score if the midterm is worth 30%, the final is worth 50%, and the class project is worth 20%?

10. An investor bought common stock of Dell Inc. on three occasions at the following prices.

Date	Price Per Share
January 2009	\$10.34
July 2009	\$13.98
December 2009	\$14.02

- a. What is the average price per share if the investor had bought 100 shares in January, 60 in July, and 40 in December?
- b. What is the average price per share if the investor had bought 40 shares in January, 60 in July, and 100 in December?
11. **FILE ERA.** One important statistic in baseball is a pitcher's earned run average, or ERA. This number represents the average number of earned runs given up by the pitcher per nine innings. The following table lists a portion of the ERAs for pitchers playing for the New York Yankees and the Baltimore Orioles as of July 22, 2010.

New York Yankees	ERA	Baltimore Orioles	ERA
Sabathia	3.13	Guthrie	4.58
Pettitte	2.88	Millwood	5.77
:	:	:	:

SOURCE: [www.mlb.com](http://www.mlb.com).

- a. Calculate the mean and the median ERAs for the New York Yankees.
- b. Calculate the mean and the median ERAs for the Baltimore Orioles.
- c. Based solely on your calculations above, which team is likely to have the better winning record? Explain.
12. **FILE Largest Corporations.** The following table shows Fortune 500's rankings of America's 10 largest corporations for 2010. Next to each corporation is its market capitalization (in billions of dollars as of March 26, 2010) and its total return to investors for the year 2009.

Company	Mkt. Cap. (in \$ billions)	Total Return
Walmart	\$209	-2.7%
Exxon Mobil	314	-12.6
Chevron	149	8.1
General Electric	196	-0.4
Bank of America	180	7.3
ConocoPhillips	78	2.9
AT&T	155	4.8
Ford Motor	47	336.7
JP Morgan Chase	188	19.9
Hewlett-Packard	125	43.1

SOURCE: <http://money.cnn.com>, May 3, 2010.

- a. Calculate the mean and the median for market capitalization.
- b. Calculate the mean and the median for total return.
- c. For each variable (market capitalization and total return), comment on which measure better reflects central location.

13. **FILE MV\_Houses.** The following table shows a portion of the sale price (in \$1,000s) for 36 homes sold in Mission Viejo, CA, during June 2010.

Number	Sale Price (in \$1,000s)
1	\$430
2	520
:	:
36	430

Calculate the mean, the median, and the mode.

14. **FILE Gas\_Prices\_2012.** The accompanying table shows a portion of the average price for a gallon of gas for the 50 states in the U.S.

State	Price per Gallon
Alabama	\$4.36
Alaska	3.79
:	:
Wyoming	3.63

SOURCE: <http://AAA.com>, data retrieved April 16, 2012.

Find the mean, the median, and the mode for the price per gallon.

15. **FILE Life\_Expectancy.** The following table lists a portion of U.S. life expectancy (in years) for the 50 states.

Rank	State	Life Expectancy (in years)
1	Hawaii	81.5
2	Alaska	80.9
:	:	:
50	Mississippi	74.8

SOURCE: [http://en.wikipedia.org/wiki/List\\_of\\_U.S.\\_states\\_by\\_life\\_expectancy](http://en.wikipedia.org/wiki/List_of_U.S._states_by_life_expectancy), data retrieved April 25, 2012.

Find the mean, the median, and the mode of life expectancy.

As discussed earlier, the median is a measure of central location that divides the data in half; that is, half of the data points fall below the median and half fall above the median. The median is also called the 50th percentile. In many instances, we are interested in a **percentile** other than the 50th percentile. Here we discuss calculating and interpreting percentiles. Generally, percentiles are calculated for large data sets; for ease of exposition, we show their use with small data sets. In addition, we construct a box plot, which is, more or less, a visual representation of particular percentiles. It also helps us identify outliers and skewness in the data.

Percentiles provide detailed information about how data are spread over the interval from the smallest value to the largest value. You have probably been exposed to percentiles. For example, the SAT is the most widely used test in the undergraduate admissions process. Scores on the math portion of the SAT range from 200 to 800. Suppose you obtained a raw score of 650 on this section of the test. It may not be readily apparent how you did relative to other students that took the same test. However, if you know that the raw score corresponds to the 75th percentile, then you know that approximately 75% of students had scores lower than your score and approximately 25% of students had scores higher than your score.

Calculate and interpret percentiles and a box plot.

### PERCENTILES

In general, the  $p$ th **percentile** divides a data set into two parts:

- Approximately  $p$  percent of the observations have values less than the  $p$ th percentile;
- Approximately  $(100 - p)$  percent of the observations have values greater than the  $p$ th percentile.

### Calculating the $p$ th Percentile

- First arrange the data in ascending (smallest to largest) order.
- Locate the approximate position of the percentile by calculating  $L_p$ :

$$L_p = (n + 1) \frac{p}{100},$$

where  $L_p$  indicates the location of the desired  $p$ th percentile and  $n$  is the sample size; for the population percentile, replace  $n$  with  $N$ . For example, we set  $p = 50$  for the median because it is the 50th percentile.

- Once you find the value for  $L_p$ , observe whether or not  $L_p$  is an integer:
  - If  $L_p$  is an integer, then  $L_p$  denotes the location of the  $p$ th percentile. For instance, if  $L_{20}$  is equal to 2, then the 20th percentile is equal to the second observation in the ordered data set.
  - If  $L_p$  is not an integer, we need to interpolate between two observations to approximate the desired percentile. So if  $L_{20}$  is equal to 2.25, then we need to interpolate 25% of the distance between the second and third observations in order to find the 20th percentile.

#### EXAMPLE 3.8

Consider the information presented in the introductory case of this chapter. Calculate and interpret the 25th and the 75th percentiles for the Metals fund.

**SOLUTION:** The first step is to arrange the data in ascending order:

Position:	1	2	3	4	5	6	7	8	9	10
Value:	-56.02	-7.34	8.09	18.33	33.35	34.30	36.13	43.79	59.45	76.46

For the 25th percentile:  $L_{25} = (n + 1) \frac{p}{100} = (10 + 1) \frac{25}{100} = 2.75$ . So, the 25th percentile is located 75% of the distance between the second and third observations; it is calculated as

$$-7.34 + 0.75(8.09 - (-7.34)) = -7.34 + 11.57 = 4.23.$$

Thus, 25% of the returns were less than 4.23%, and 75% of the returns were greater than 4.23%.

For the 75th percentile:  $L_{75} = (n + 1) \frac{p}{100} = (10 + 1) \frac{75}{100} = 8.25$ . So, the 75th percentile is located 25% of the distance between the eighth and ninth observations; it is calculated as

$$43.79 + 0.25(59.45 - 43.79) = 43.79 + 3.92 = 47.71.$$

Thus, 75% of the returns were less than 47.71%, and 25% of the returns were greater than 47.71%.

Earlier, we calculated the median or the 50th percentile for the Metals fund and obtained a value of 33.83%. When we calculate the 25th, the 50th, and the 75th percentiles for a data set, we have effectively divided the data into four equal parts, or quarters. Thus, the 25th percentile is also referred to as the first quartile (Q1), the 50th percentile is referred to as the second quartile (Q2), and the 75th percentile is referred to as the third quartile (Q3).

## Constructing and Interpreting a Box Plot

A **box plot**, also referred to as a box-and-whisker plot, is a convenient way to graphically display the minimum value (Min), the quartiles (Q1, Q2, and Q3), and the maximum value (Max) of a data set. Using our results from the Metals fund, Table 3.4 summarizes the five values that we will plot:

**TABLE 3.4** Summary Values for the Metals Fund

Min	Q1	Q2	Q3	Max
-56.02%	4.23%	33.83%	47.71%	76.46%

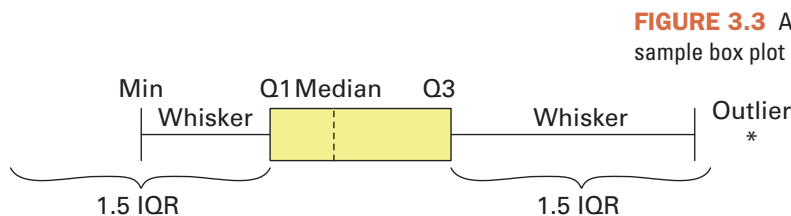
The values in Table 3.4 are often referred to as the five-number summary for the data set. Box plots are particularly useful when comparing similar information gathered at another place or time. They also are used as an effective tool for identifying outliers and skewness. In Section 3.1, we discussed that the mean is unduly influenced by outliers. Sometimes outliers may indicate bad data due to incorrectly recorded observations or incorrectly included observations in the data set. In such cases, the relevant observations should be corrected or simply deleted from the data set. Alternatively, outliers may just be due to random variations, in which case the relevant observations should remain in the data set. In any event, it is important to be able to identify potential outliers so that one can take corrective actions, if needed.

In order to construct a box plot, we follow these steps.

- A. Plot the five-number summary values in ascending order on the horizontal axis.
- B. Draw a box encompassing the first and third quartiles.
- C. Draw a dashed vertical line in the box at the median.
- D. To determine if a given observation is an outlier, first calculate the difference between Q3 and Q1. This difference is called the **interquartile range** or IQR. Therefore, the length of the box is equal to the IQR and the span of the box contains the middle half of the data. Draw a line (“whisker”) that extends from Q1 to the minimum data value that is not farther than  $1.5 \times \text{IQR}$  from Q1. Similarly, draw a line that extends from Q3 to the maximum data value that is not farther than  $1.5 \times \text{IQR}$  from Q3.
- E. Use an asterisk to indicate points that are farther than  $1.5 \times \text{IQR}$  from the box. These points are considered outliers.

Consider the box plot in Figure 3.3 for illustration. In the figure, the left whisker extends from Q1 to Min since Min is not farther than  $1.5 \times \text{IQR}$  from Q1. The right whisker, on the other hand, does not extend from Q3 to Max since there is an observation that is farther than  $1.5 \times \text{IQR}$  from Q3. The asterisk on the right indicates this observation is considered an outlier.

Box plots are also used to informally gauge the shape of the distribution. Symmetry is implied if the median is in the center of the box and the left and right whiskers are equidistant from their respective quartiles. If the median is left of center and the right whisker is longer than the left whisker, then the distribution is positively skewed. Similarly, if the median is right of center and the left whisker is longer than the right whisker, then the distribution is negatively skewed. From Figure 3.3, we note that the median is located to the left of center and the right whisker is longer than the left whisker. This indicates that the underlying distribution is positively skewed.

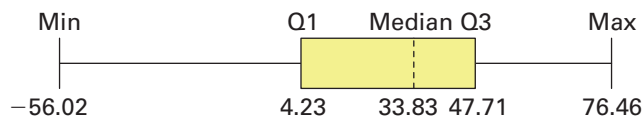


### EXAMPLE 3.9

Use the information presented in the introductory case of this chapter to construct and interpret the box plot for the Metals fund.

**SOLUTION:** Based on the information in Table 3.4, we calculate the IQR as the difference between Q3 and Q1, or  $47.71\% - 4.23\% = 43.48\%$ . We then calculate  $1.5 \times \text{IQR} = 1.5 \times 43.48\% = 65.22\%$ . The distance between Q1 and the smallest value,  $4.23 - (-56.02\%) = 60.25\%$ , is within the limit of  $65.22\%$ ; thus, the line will extend to the minimum value of  $-56.02\%$  on the left side of the box plot (see Figure 3.4). Similarly, the distance between the largest value and Q3,  $76.46\% - 47.71\% = 28.75$ , is also well within the limit of  $65.22\%$ ; here the line will extend to the right up to the maximum value of  $76.46\%$ . Given the criteria for constructing a box plot, there are no outliers in this data set.

**FIGURE 3.4** Box plot for the Metals Fund



From this box plot we can quickly grasp several points concerning the distribution of returns for the Metals fund. First, returns range from  $-56.02\%$  to  $76.46\%$ , with about half being less than  $33.83\%$  and half being greater than  $33.83\%$ . We make two further observations: (1) the median is off-center within the box, being located to the right of center, and (2) the left whisker is longer than the right whisker. This suggests that the distribution is negatively skewed.

## EXERCISES 3.2

### Mechanics

16. Calculate the 20th, 50th, and 80th percentiles for the following data set:

120	215	187	343	268	196	312
-----	-----	-----	-----	-----	-----	-----

17. Calculate the 20th, 40th, and 70th percentiles for the following data set:

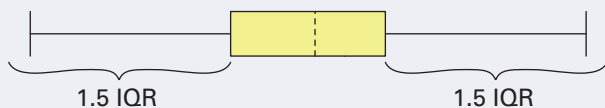
-300	-257	-325	-234	-297	-362	-255
------	------	------	------	------	------	------

18. Consider the following box plot.



- Does the above box plot indicate possible outliers in the data?
- Comment on the skewness of the underlying distribution.

19. Consider the following box plot.



- Does the above box plot indicate possible outliers in the data?
- Comment on the skewness of the underlying distribution.

20. Consider the following data set:

0.04	0.10	-0.05	-0.02	0.08	0.15	-0.09
------	------	-------	-------	------	------	-------

- Calculate and interpret the 25th, 50th, and 75th percentiles.
- Construct a box plot. Are there any outliers?

21. Consider the following data set:

12	9	27	15	58	35	21	32	22
----	---	----	----	----	----	----	----	----

- Calculate and interpret the 25th, 50th, and 75th percentiles.
- Construct a box plot. Are there any outliers?

### Applications

22. Scores on the final in a statistics class are as follows.

75	25	75	62	80	85	80	99	90	60
86	92	40	74	72	65	87	70	85	70

- Calculate and interpret the 25th, 50th, and 75th percentiles.
  - Construct a box plot. Are there any outliers? Is the distribution symmetric? If not, comment on its skewness.
23. Consider the return data (in percent) for the Income fund in Table 3.1.
- Calculate and interpret the 25th, 50th, and 75th percentiles.
  - Construct a box plot. Are there any outliers?
  - Is the distribution symmetric? If not, comment on its skewness.
24. **FILE Census.** The accompanying table shows a portion of median household income (Income) and median house value (House Value) for the 50 states in 2010.

State	Income	House Value
Alabama	\$42,081	\$117,600
Alaska	66,521	229,100
⋮	⋮	⋮
Wyoming	53,802	174,000

SOURCE: 2010 U.S. Census.

- Construct a box plot for household income and use it to identify outliers, if any, and comment on skewness.
- Construct a box plot for median house value and use it to identify outliers, if any, and comment on skewness.
- Are you surprised by the above results?



25. **FILE PE\_Ratio.** A price-earnings ratio or P/E ratio is calculated as a firm's share price compared to the income or profit earned by the firm per share. Generally, a high P/E ratio suggests that investors are expecting higher earnings growth in the future compared to companies with a lower P/E ratio. The accompanying table shows a portion of companies that comprise the Dow Jones Industrial Average (DJIA) and their P/E ratios as of May 17, 2012 (at the time data were retrieved, the P/E ratio for one firm on the DJIA, Bank of America, was not available).

Company	P/E Ratio
3M (MMM)	14
Alcoa (AA)	24
:	:
Walt Disney (DIS)	14

- Calculate and interpret the 25th, 50th, and 75th percentiles.
- Construct a box plot. Are there any outliers? Is the distribution symmetric? If not, comment on its skewness.

### 3.3 THE GEOMETRIC MEAN

LO 3.3

The geometric mean is a multiplicative average, as opposed to an additive average (the arithmetic mean). It is the relevant measure when evaluating investment returns over several years. It is also the relevant measure when calculating average growth rates.

Calculate and interpret a geometric mean return and an average growth rate.

#### The Geometric Mean Return

Suppose you invested \$1,000 in a stock that had a 10% return in 2009 and a -10% return in 2010. The arithmetic mean suggests that by the end of year 2010, you would be right back where you started with \$1,000 worth of stock. It is true that the arithmetic mean return over the two-year period is 0% ( $\bar{x} = \frac{0.10 + (-0.10)}{2} = 0$ ); however, the arithmetic mean ignores the effects of compounding. As shown in Table 3.5, the value of your investment at the end of two years is \$990, a loss of \$10. The geometric mean accurately captures a negative annual return from the two-year investment period.

**TABLE 3.5** End of Year Holdings Given an Initial Investment of \$1,000

Year	Return	Value at the End of Year
2009	10%	\$1,000 + 1,000(0.10) = \$1,100
2010	-10%	\$1,100 + 1,100(-0.10) = \$990

#### THE GEOMETRIC MEAN RETURN

For multiperiod returns  $R_1, R_2, \dots, R_n$ , the **geometric mean return**  $G_R$  is computed as

$$G_R = \sqrt[n]{(1 + R_1)(1 + R_2) \cdots (1 + R_n)} - 1,$$

where  $n$  is the number of multiperiod returns.

Let us revisit the above case where you invested \$1,000 in a stock that had a 10% return in 2009 and a -10% return in 2010. The geometric mean is computed as

$$G_R = \sqrt[2]{(1 + 0.10)(1 + (-0.10))} - 1 = ((1.10)(0.90))^{1/2} - 1 = -0.005, \text{ or } -0.5\%.$$

We interpret the geometric mean return as the **annualized return**, that you will earn from a two-year investment period. Table 3.6 shows that with the computed annualized return of -0.5%, the end investment value is the same as shown in Table 3.5.

**TABLE 3.6** End-of-Year Holdings Given an Initial Investment of \$1,000

Year	Annualized Return	Value at the End of Year
2009	-0.5%	\$1,000 + 1,000(-0.005) = \$995
2010	-0.5%	995 + 995(-0.005) = \$990

**EXAMPLE 3.10**

Use the data in Table 3.1 to calculate the geometric mean for the Metals and the Income funds.

**SOLUTION:**

$$\begin{aligned}\text{Metals Fund: } G_R &= \sqrt[10]{(1 - 0.0734)(1 + 0.1833) \cdots (1 + 0.7646)} - 1 \\ &= (5.1410)^{1/10} - 1 = 0.1779, \text{ or } 17.79\%.\end{aligned}$$

$$\begin{aligned}\text{Income Fund: } G_R &= \sqrt[10]{(1 + 0.0407)(1 + 0.0652) \cdots (1 + 0.3177)} - 1 \\ &= (2.1617)^{1/10} - 1 = 0.0801, \text{ or } 8.01\%.\end{aligned}$$

Therefore, for the 10-year period, the annualized return for the Metals fund is higher than that of the Income fund,  $17.79\% > 8.01\%$ . However, the magnitude of the difference is relatively smaller than that of the arithmetic means, which for the Metals and Income funds are 24.65% and 8.51%, respectively. This shows that the geometric mean is not as sensitive to extreme values as is the arithmetic mean. As discussed earlier, the arithmetic mean for the Metals fund is unduly influenced by the extreme return of 76.46% in 2009.

**Arithmetic Mean versus Geometric Mean**

An issue that begs for explanation is the relevance of the arithmetic mean and the geometric mean as summary measures for financial returns. Both means are relevant descriptive measures for annual return; however, each has a different interpretation. The arithmetic mean is appropriate for analyzing a one-year investment, whereas the geometric mean is appropriate for analyzing a multi-year investment. In Example 3.10, the arithmetic mean of 24.65% is the average annual return for summarizing returns with an investment horizon of one year. The geometric mean of 17.79% is the average annual return when the investment horizon is 10 years. For illustration, we can think of the arithmetic mean as the relevant metric for an investor who is saving/investing to buy a house in about a year's time. The geometric mean is the relevant metric for an investor who is saving for retirement.

**The Average Growth Rate**

We also use the geometric mean when we calculate average growth rates.

**THE AVERAGE GROWTH RATE**

For growth rates  $g_1, g_2, \dots, g_n$ , the **average growth rate**  $G_g$  is computed as:

$$G_g = \sqrt[n]{(1 + g_1)(1 + g_2) \cdots (1 + g_n)} - 1$$

where  $n$  is the number of multiperiod growth rates.

### EXAMPLE 3.11

Table 3.7 shows sales for Adidas (in millions of €) for the years 2005 through 2009.

**TABLE 3.7** Sales for Adidas (in millions of €), 2005–2009

Year	2005	2006	2007	2008	2009
Sales	6,636	10,084	10,299	10,799	10,381

Calculate the growth rates for 2005–2006, 2006–2007, 2007–2008, and 2008–2009 and use them to compute the average growth rate.

**SOLUTION:** The growth rates for Adidas for four years are computed as:

- 2005–2006:  $\frac{10,084 - 6,636}{6,636} = 0.5196$
- 2006–2007:  $\frac{10,299 - 10,084}{10,084} = 0.0213$
- 2007–2008:  $\frac{10,799 - 10,299}{10,299} = 0.0485$
- 2008–2009:  $\frac{10,381 - 10,799}{10,799} = -0.0387$

Therefore,

$$\begin{aligned} G_g &= \sqrt[4]{(1 + 0.5196)(1 + 0.0213)(1 + 0.0485)(1 - 0.0387)} - 1 \\ &= \sqrt[4]{(1.5196)(1.0213)(1.0485)(0.9613)} = 1.5643^{1/4} - 1 = 0.1184, \text{ or } 11.84\%. \end{aligned}$$

Sales for Adidas from 2005 to 2009 had an average growth rate of 11.84% per year.

There is a simpler way to compute the average growth rate when the underlying values of the series are given. In the above example, it is cumbersome to first calculate the relevant growth rates and then use them to compute the average growth rate.

#### AN ALTERNATIVE FORMULA FOR THE AVERAGE GROWTH RATE

For observations  $x_1, x_2, \dots, x_n$ , the **average growth rate**  $G_g$  is computed as:

$$G_g = \sqrt[n-1]{\frac{x_n}{x_{n-1}} \frac{x_{n-1}}{x_{n-2}} \frac{x_{n-2}}{x_{n-3}} \dots \frac{x_2}{x_1}} - 1 = \sqrt[n-1]{\frac{x_n}{x_1}} - 1$$

where  $n - 1$  is the number of distinct growth rates. Note that only the first and last observations are needed in the time series due to cancellations in the formula.

### EXAMPLE 3.12

Calculate the average growth rate for Adidas directly from the sales data in Table 3.7.

**SOLUTION:** Using the first and last observations from the time series consisting of five observations, we calculate

$$G_g = \sqrt[n-1]{\frac{x_n}{x_1}} - 1 = \sqrt[5-1]{\frac{10,381}{6,636}} - 1 = 1.5643^{1/4} - 1 = 0.1184, \text{ or } 11.84\%$$

which is the same as in Example 3.11.

## EXERCISES 3.3

### Mechanics

26. Calculate the average growth rate return for the following data set:

4%	8%	-5%	6%
----	----	-----	----

27. Calculate the geometric mean return for the following data set:

-3%	2%	-5%	2.7%	3.1%
-----	----	-----	------	------

28. The returns for a pharmaceutical firm are 10% in Year 1, 5% in Year 2, and -15% in Year 3. What is the annualized return for the period?
29. The returns from an investment are 2% in Year 1, 5% in Year 2, and 1.8% in the first half of Year 3. Calculate the annualized return for the entire period.
30. The returns for an auto firm are 5% in Year 1 and 3% in the first quarter of Year 2. Calculate the annualized return for the period.

31. Consider the following observations for a series:

Year 1	Year 2	Year 3	Year 4
90	110	150	160

- a. Calculate the growth rates for Year 1–Year 2, Year 2–Year 3, and Year 3–Year 4.
- b. Calculate the average growth rate.
32. Consider the following observations for a time series:

Year 1	Year 2	Year 3	Year 4
1,200	1,280	1,380	1,520

- a. Calculate the growth rates for Year 1–Year 2, Year 2–Year 3, and Year 3–Year 4.
- b. Calculate the average growth rate.
33. Calculate the average growth rate from the following growth rates.

2.5%	3.6%	1.8%	2.2%	5.2%
------	------	------	------	------

### Applications

34. Suppose at the beginning of 2006 you decide to invest \$1,000 in Vanguard's European Stock Index mutual fund. The following table shows the returns for the years 2006–2009.

Year	Annual Return
2006	33.42%
2007	13.82%
2008	-44.73%
2009	31.91%

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com).

- a. Calculate and interpret the arithmetic mean return.
- b. Calculate and interpret the geometric mean return.
- c. How much money would you have accumulated by the end of 2009?

35. Home Depot and Lowe's are the two largest home improvement retailers in the U.S. The following table shows the total revenue (in billions) for each retailer for the years 2008–2010.

Year	Home Depot	Lowe's
2008	\$77.35	\$48.28
2009	71.29	48.23
2010	66.18	47.22

SOURCE: Annual Reports of Home Depot, Inc., and Lowe's Companies Inc.

- a. Calculate the growth rate for 2008–2009 and 2009–2010 for each retailer.
- b. Calculate the average growth rate for each retailer.
36. Suppose at the beginning of 2005 you decide to invest \$20,000 in Driehaus' Emerging Markets Growth mutual fund. The following table shows the returns for the years 2005–2009.

Year	Annual Return
2005	25.85%
2006	27.55%
2007	27.47%
2008	-47.02%
2009	75.75%

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com).

- a. Calculate and interpret the arithmetic mean return.
- b. Calculate and interpret the geometric mean return.
- c. How much money would you have accumulated by the end of 2009?
37. The following table shows the total revenue (in billions of \$) for Walmart Stores, Inc. and Target Corp. for the years 2008–2010.

Year	2008	2009	2010
Walmart	379.8	404.3	408.2
Target	63.4	65.0	65.3

SOURCE: Annual Reports of Walmart Stores, Inc., and Target Corp.

- a. Calculate the average growth rate for each firm.
- b. Which firm had the higher growth rate over the 2008–2010 period?
38. The following table shows sales for Nike (in millions of \$) for the years 2005 through 2009.

Year	2005	2006	2007	2008	2009
Sales	13,740	14,955	16,326	18,627	19,176

SOURCE: Annual Reports of Nike, Inc.

- a. Use the growth rates for 2005–2006, 2006–2007, 2007–2008, and 2008–2009 to calculate the average growth rate.
- b. Calculate the average growth rate directly from sales.

In Section 3.1, we focused on measures of central location in an attempt to find a typical or central value that describes the data. It is also important to analyze how the data vary around the center. Recall that over the 10-year period 2000–2009, the average returns for the Metals and Income funds were 24.65% and 8.51%, respectively. As an investor, you might ask why anyone would put money in the Income fund when, on average, this fund has a lower return. The answer to this question will become readily apparent once we analyze measures of variability or dispersion.

Table 3.8 shows each fund's minimum and maximum returns, as well as each fund's average return, over this time period. Note that the average return for the Income fund is relatively closer to its minimum and maximum returns as compared to the Metals fund. The comparison of the funds illustrates that the average is not sufficient when summarizing a data set; that is, it fails to describe the underlying variability of the data.

Calculate and interpret the range, the mean absolute deviation, the variance, the standard deviation, and the coefficient of variation.

**TABLE 3.8** Select Measures for the Metal and Income Funds, 2000–2009

	Minimum Return	Average Return	Maximum Return
Metals fund	−56.02%	24.65%	76.46%
Income fund	−11.37%	8.51%	31.77%

We now discuss several measures of dispersion that gauge the variability of a data set. Each measure is a numerical value that equals zero if all data values are identical, and increases as data values become more diverse.

## Range

The **range** is the simplest measure of dispersion; it is the difference between the maximum (Max) and the minimum (Min) values in a data set.

$$\text{Range} = \text{Max} - \text{Min}$$

### EXAMPLE 3.13

Use the data in Table 3.8 to calculate the range for the Metals and the Income funds.

**SOLUTION:** Metals fund:  $76.46\% - (-56.02\%) = 132.48\%$   
Income fund:  $31.77\% - (-11.37\%) = 43.14\%$

The Metals fund has the higher value for the range, indicating that it has more dispersion with respect to its minimum and maximum values.

The range is not considered a good measure of dispersion because it focuses solely on the extreme values and ignores every other observation in the data set. While the interquartile range,  $IQR = Q3 - Q1$ , discussed in Section 3.2, does not depend on extreme values, this measure still does not incorporate all the data.

## The Mean Absolute Deviation

A good measure of dispersion should consider differences of all observations from the mean. If we simply average all differences from the mean, the positives and the negatives will cancel out, even though they both contribute to dispersion, and the resulting average

will equal zero. The **mean absolute deviation** (MAD) is an average of the absolute differences between the observations and the mean.

#### THE MEAN ABSOLUTE DEVIATION (MAD)

For sample values,  $x_1, x_2, \dots, x_n$ , the **sample MAD** is computed as

$$\text{Sample MAD} = \frac{\sum |x_i - \bar{x}|}{n}.$$

For population values,  $x_1, x_2, \dots, x_N$ , the **population MAD** is computed as

$$\text{Population MAD} = \frac{\sum |x_i - \mu|}{N}.$$

#### EXAMPLE 3.14

Use the data in Table 3.1 to calculate MAD for the Metals and the Income funds.

**SOLUTION:** We first compute MAD for the Metals fund. The second column in Table 3.9 shows differences from the sample mean,  $\bar{x} = 24.65$ . As mentioned earlier, the sum of these differences equals zero (or a number very close to zero due to rounding). The third column shows the absolute value of each deviation from the mean. Summing these values yields the numerator for the MAD formula.

**TABLE 3.9** MAD Calculations for the Metals Fund

$x_i$	$x_i - \bar{x}$	$ x_i - \bar{x} $
-7.34	$-7.34 - 24.65 = -31.99$	31.99
18.33	$18.33 - 24.65 = -6.32$	6.32
$\vdots$	$\vdots$	$\vdots$
76.46	$76.46 - 24.65 = 51.81$	51.81
Total = 0 (subject to rounding)		Total = 271.12

For the Metals fund:  $\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n} = \frac{271.12}{10} = 27.11$ .

Similar calculations for the Income fund yield:  $\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n} = \frac{70.30}{10} = 7.03$ .

The Income fund has a smaller value for MAD than the Metals fund, again indicating a less dispersed data set.

## The Variance and the Standard Deviation

The **variance** and the **standard deviation** are the two most widely used measures of dispersion. Instead of calculating the average of the absolute differences from the mean, as in MAD, we calculate the average of the squared differences from the mean. The squaring of differences from the mean emphasizes larger differences more than smaller ones; MAD weighs large and small differences equally.

The variance is defined as the average of the squared differences between the observations and the mean. The formula for the variance differs depending on whether we have a sample or a population. We also note that variance squares the original units of measurement. In order to return to the original units of measurement, we take the positive square root of variance, which gives us the standard deviation.



### THE VARIANCE AND THE STANDARD DEVIATION

For sample values  $x_1, x_2, \dots, x_n$ , the **sample variance**  $s^2$  and the **sample standard deviation**  $s$  are computed as

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} \quad \text{and} \quad s = \sqrt{s^2}.$$

For population values  $x_1, x_2, \dots, x_N$ , the **population variance**  $\sigma^2$  (the Greek letter sigma, squared) and the **population standard deviation**  $\sigma$  are computed as

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad \text{and} \quad \sigma = \sqrt{\sigma^2}.$$

*Note:* The sample variance uses  $n - 1$  rather than  $n$  in the denominator; the reason is discussed in Chapter 8.

### EXAMPLE 3.15

Use the data in Table 3.1 to calculate the sample variance and the sample standard deviation for the Metals and the Income funds. Express the answers in the correct units of measurement.

**SOLUTION:** We will show the calculations for the Metals fund with the mean return of 24.65%. The second column in Table 3.10 shows each return less the mean. The third column shows the square of each deviation from the mean. Summing these values yields the numerator for the sample variance formula.

**TABLE 3.10** Sample Variance Calculation for the Metals Fund

$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
-7.34	$-7.34 - 24.65 = -31.99$	$(-31.99)^2 = 1,023.36$
18.33	$18.33 - 24.65 = -6.32$	$(-6.32)^2 = 39.94$
$\vdots$	$\vdots$	$\vdots$
76.46	$76.46 - 24.65 = 51.81$	$(51.81)^2 = 2,684.28$
	Total = 0 (subject to rounding)	Total = 12,407.44

For the Metals fund:  $s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{12,407.44}{10 - 1} = 1,378.60(\%)^2$ . Note that the units of measurement are squared. The sample standard deviation is  $s = \sqrt{1,378.60} = 37.13(\%)$ .

Similar calculations for the Income fund yield

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{1,102.34}{10 - 1} = 122.48(\%)^2 \text{ and } s = \sqrt{122.48} = 11.07(\%).$$

Based on all measures of dispersion discussed thus far, we can conclude that the Income fund is less dispersed than the Metals fund. With financial data, standard deviation tends to be the most common measure of risk. Therefore, the investment risk of the Income fund is lower than that of the Metals fund.

## The Coefficient of Variation

In some instances, analysis entails comparing the variability of two or more data sets that have different means or units of measurement. The **coefficient of variation (CV)** serves as a relative measure of dispersion and adjusts for differences in the magnitudes of the means. Calculated by dividing a data set's standard deviation by its mean, CV is a unitless measure that allows for direct comparisons of mean-adjusted dispersion across different data sets.

### THE COEFFICIENT OF VARIATION (CV)

$$\text{Sample CV} = \frac{s}{\bar{x}}$$

$$\text{Population CV} = \frac{\sigma}{\mu}$$

### EXAMPLE 3.16

Calculate and interpret the coefficient of variation for the Metals and Income funds.

**SOLUTION:** We use the sample means and the sample standard deviations computed earlier.

$$\text{For the Metals fund: CV} = \frac{s}{\bar{x}} = \frac{37.13\%}{24.65\%} = 1.51.$$

$$\text{For the Income fund: CV} = \frac{s}{\bar{x}} = \frac{11.07\%}{8.51\%} = 1.30.$$

Since 1.51 is greater than 1.30, we can conclude that the data for the Metals fund have more relative dispersion than the Income fund.

## Using Excel to Calculate Measures of Dispersion

### Excel's Formula Option

**FILE**  
*Fund\_Returns*

As discussed in Section 3.1, Excel provides built-in formulas for most summary measures. Table 3.11 shows each measure of dispersion that we discussed and its corresponding Function Name in Excel. For example, in order to calculate the standard deviation for the Metals fund, we open *Fund\_Returns*. We find an empty cell and insert “=STDEV.S(B2:B11)” and then choose <Enter>. Excel returns a value of 37.13, which matches the value that we calculated by hand.

**TABLE 3.11** Excel's Functions for Measures of Dispersion

Measure of Dispersion	Excel Function Name
Range	=MAX(array) – MIN(array)
Mean Absolute Deviation	=AVEDEV(array)
Sample Variance	=VAR.S(array)
Sample Standard Deviation	=STDEV.S(array)
Population Variance	=VAR.P(array)
Population Standard Deviation	=STDEV.P(array)

## Excel's Data Analysis Toolpak Option

In Section 3.1, we also discussed using Excel's Data Analysis Toolpak option, **Data > Data Analysis > Descriptive Statistics**, for calculating summary measures. For measures of dispersion, Excel treats the data as a sample and calculates the range, the sample variance, and the sample standard deviation. These values for the Metals and Income funds are shown in boldface in Table 3.3.

## SYNOPSIS OF INTRODUCTORY CASE

Vanguard's Precious Metals and Mining fund (Metals) and Fidelity's Strategic Income fund (Income) were two top-performing mutual funds for the years 2000 through 2009. An analysis of annual return data for these two funds provides important information for any type of investor. Over the past 10 years, the Metals fund posts the higher values for both the mean return and the median return, with values of 24.65% and 33.83%, respectively. When the mean differs dramatically from the median, it is often indicative of extreme values or outliers. Although the mean and the median for the Metals fund do differ by almost 10 percentage points, a boxplot analysis reveals no outliers. The mean return and the median return for the Income fund, on the other hand, are quite comparable at 8.51% and 7.34%, respectively.



While measures of central location typically represent the reward of investing, these measures do not incorporate the risk of investing. Standard deviation tends to be the most common measure of risk with financial data. Since the standard deviation for the Metals fund is substantially greater than the standard deviation for the Income fund ( $37.13\% > 11.07\%$ ), the Metals fund is likelier to have returns far above as well as far below its mean. Also, the coefficient of variation—a relative measure of dispersion—for the Metals fund is greater than the coefficient of variation for the Income fund. These two measures of dispersion indicate that the Metals fund is the riskier investment. These funds provide credence to the theory that funds with higher average returns often carry higher risk.

## EXERCISES 3.4

### Mechanics

39. Consider the following population data:

34	42	12	10	22
----	----	----	----	----

- Calculate the range.
- Calculate MAD.
- Calculate the population variance.
- Calculate the population standard deviation.

40. Consider the following population data:

0	-4	2	-8	10
---	----	---	----	----

- Calculate the range.
- Calculate MAD.
- Calculate the population variance.
- Calculate the population standard deviation.

41. Consider the following sample data:

40	48	32	52	38	42
----	----	----	----	----	----

- Calculate the range.
- Calculate MAD.
- Calculate the sample variance.
- Calculate the sample standard deviation.

42. Consider the following sample data:

-10	12	-8	-2	-6	8
-----	----	----	----	----	---

- Calculate the range.
- Calculate MAD.
- Calculate the sample variance and the sample standard deviation.

## Applications

43. The Department of Transportation (DOT) fields thousands of complaints about airlines each year. The DOT categorizes and tallies complaints, and then periodically publishes rankings of airline performance. The following table presents the 2006 results for the 10 largest U.S. airlines.

Airline	Complaints*	Airline	Complaints*
Southwest Airlines	1.82	Northwest Airlines	8.84
JetBlue Airways	3.98	Delta Airlines	10.35
Alaska Airlines	5.24	American Airlines	10.87
AirTran Airways	6.24	US Airways	13.59
Continental Airlines	8.83	United Airlines	13.60

SOURCE: Department of Transportation; \*per million passengers.

- Which airline fielded the least amount of complaints? Which airline fielded the most? Calculate the range.
  - Calculate the mean and the median number of complaints for this sample.
  - Calculate the variance and the standard deviation.
44. The monthly closing stock prices (rounded to the nearest dollar) for Starbucks Corp. and Panera Bread Co. for the first six months of 2010 are reported in the following table.

Month	Starbucks Corp.	Panera Bread Co.
January 2010	\$22	\$71
February 2010	23	73
March 2010	24	76
April 2010	26	78
May 2010	26	81
June 2010	24	75

SOURCE: www.finance.yahoo.com.

- Calculate the sample variance and the sample standard deviation for each firm's stock price.
  - Which firm's stock price had greater variability as measured by the standard deviation?
  - Which firm's stock price had the greater relative dispersion?
45. **FILE AnnArbor\_Rental.** Real estate investment in college towns continues to promise good returns (*The Wall Street Journal*, September 24, 2010). Marcela

Treisman works for an investment firm in Michigan. Her assignment is to analyze the rental market in Ann Arbor, which is home to the University of Michigan. She gathers data on monthly rent for 2011 along with the square footage of 40 homes. A portion of the data is shown in the accompanying table.

Monthly Rent	Square Footage
645	500
675	648
:	:
2400	2700

SOURCE: www.zillow.com.

- Calculate the mean and the standard deviation for monthly rent.
  - Calculate the mean and the standard deviation for square footage.
  - Which sample data exhibit greater relative dispersion?
46. **FILE Largest\_Corporations.** The accompanying data file shows the Fortune 500 rankings of America's largest corporations for 2010. Next to each corporation are its market capitalization (in billions of dollars as of March 26, 2010) and its total return to investors for the year 2009.
- Calculate the coefficient of variation for market capitalization.
  - Calculate the coefficient of variation for total return.
  - Which sample data exhibit greater relative dispersion?
47. **FILE Census.** The accompanying data file shows, among other variables, median household income and median house value for the 50 states.
- Compute and discuss the range of household income and house value.
  - Compute the sample MAD and the sample standard deviation of household income and house value.
  - Discuss why we cannot directly compare the sample MAD and the standard deviations of the two data sets.

## 3.5 MEAN-VARIANCE ANALYSIS AND THE SHARPE RATIO

LO 3.5

In the introduction to Section 3.4, we asked why any rational investor would invest in the Income fund over the Metals fund since the average return for the Income fund over the 2000–2009 period was approximately 9%, whereas the average return for the Metals fund was close to 25%. It turns out that investments with higher returns also carry higher risk. Investments include financial assets such as stocks, bonds, and mutual funds. The average return represents an investor's reward, whereas variance, or equivalently standard deviation, corresponds to risk.

According to mean-variance analysis, we can measure performance of any risky asset solely on the basis of the average and the variance of its returns.

Explain mean-variance analysis and the Sharpe ratio.

### MEAN-VARIANCE ANALYSIS

**Mean-variance analysis** postulates that the performance of an asset is measured by its rate of return, and this rate of return is evaluated in terms of its reward (mean) and risk (variance). In general, investments with higher average returns are also associated with higher risk.

Consider Table 3.12, which summarizes the mean and variance for the Metals and Income funds.

**TABLE 3.12** Mean-Variance Analysis of Two Mutual Funds, 2000–2009

Fund	Mean Return	Variance
Metals fund	24.65%	1,378.61(%) <sup>2</sup>
Income fund	8.51%	122.48(%) <sup>2</sup>

It is true that the Metals fund provided an investor with a higher reward over the 10-year period, but this same investor encountered considerable risk compared to an investor who invested in the Income fund. Table 3.12 shows that the variance of the Metals fund (1,378.61(%)<sup>2</sup>) is significantly greater than the variance of the Income fund (122.48(%)<sup>2</sup>). If we look back at Table 3.1 and focus on the Metals fund, we see returns far above the average return of 24.65% (for example, 59.45% and 76.46%), but also returns far below the average return of 24.65% (for example, −7.34% and −56.02%). Repeating this same analysis for the Income fund, the returns are far closer to the average return of 8.51%; thus, the Income fund provided a lower return, but also far less risk.

A discussion of mean-variance analysis seems almost incomplete without mention of the **Sharpe ratio**. Nobel Laureate William Sharpe developed what he originally referred to as the “reward-to-variability” ratio. However, academics and finance professionals prefer to call it the “Sharpe ratio.” The Sharpe ratio is used to characterize how well the return of an asset compensates for the risk that the investor takes. Investors are often advised to pick investments that have high Sharpe ratios.

The Sharpe ratio is defined with the reward specified in terms of the population mean and the variability specified in terms of the population standard deviation. However, we often compute the Sharpe ratio in terms of the sample mean and the sample standard deviation, where the return is usually expressed as a percent and not a decimal.

### THE SHARPE RATIO

The **Sharpe ratio** measures the extra reward per unit of risk. The Sharpe ratio for an investment  $I$  is computed as:

$$\frac{\bar{x}_I - \bar{R}_f}{s_I},$$

where  $\bar{x}_I$  is the mean return for the investment,  $\bar{R}_f$  is the mean return for a risk-free asset such as a Treasury bill (T-bill), and  $s_I$  is the standard deviation for the investment.

The numerator of the Sharpe ratio measures the extra reward that investors receive for the added risk taken—this difference is often called excess return. The higher the Sharpe ratio, the better the investment compensates its investors for risk.

### EXAMPLE 3.17

Calculate and interpret the Sharpe ratios for the Metals and Income funds given that the return on a 1-year T-bill is 2%.

**SOLUTION:** Since the return on a 1-year T-bill is 2%,  $\bar{R}_f = 2$ . Plugging in the values of the relevant means and standard deviations into the Sharpe ratio yields:

$$\text{Sharpe ratio for the Metals fund: } \frac{\bar{x}_I - \bar{R}_f}{s_I} = \frac{24.65 - 2}{37.13} = 0.61.$$

$$\text{Sharpe ratio for the Income fund: } \frac{\bar{x}_I - \bar{R}_f}{s_I} = \frac{8.51 - 2}{11.07} = 0.59.$$

We had earlier shown that the Metals fund had a higher return, which is good, along with a higher variance, which is bad. We can use the Sharpe ratio to make a valid comparison between the funds. The Metals fund provides the higher Sharpe ratio than the Income fund ( $0.61 > 0.59$ ); therefore, the Metals fund offered more reward per unit of risk compared to the Income fund.

## EXERCISES 3.5

### Mechanics

48. Consider the following data for two investments, A and B:

Investment A: $\bar{x} = 10$ and $s = 5$
--

Investment B: $\bar{x} = 15$ and $s = 10$
---

- Which investment provides the higher return? Which investment provides less risk? Explain.
  - Given a risk-free rate of 1.4%, calculate the Sharpe ratio for each investment. Which investment provides the higher reward per unit of risk? Explain.
49. Consider the following data for two investments, A and B:

Investment A: $\bar{x} = 8$ and $s = 5$
---

Investment B: $\bar{x} = 10$ and $s = 7$
--

- Which investment provides the higher return? Which investment provides less risk? Explain.
  - Given a risk-free rate of 2%, calculate the Sharpe ratio for each investment. Which investment provides the higher reward per unit of risk? Explain.
50. Consider the following returns for two investments, A and B, over the past four years:

Investment 1:	2%	8%	−4%	6%
Investment 2:	6%	12%	−8%	10%

- Which investment provides the higher return?
- Which investment provides less risk?
- Given a risk-free rate of 1.2%, calculate the Sharpe ratio for each investment. Which investment has performed better? Explain.

### Applications

51. The following table shows the annual returns (in percent) and summary measures for the Vanguard Energy Fund and the Vanguard Health Care Fund from 2005 through 2009.

Year	Energy	Health Care
2005	44.60	15.41
2006	19.68	10.87
2007	37.00	4.43
2008	−42.87	−18.45
2009	38.36	20.96
	$\bar{x}_{\text{Energy}} = 19.35$ $s_{\text{Energy}} = 35.99$	$\bar{x}_{\text{Health}} = 6.64$ $s_{\text{Health}} = 15.28$

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com).

- Which fund had the higher average return?
  - Which fund was riskier over this time period?
  - Given a risk-free rate of 3%, which fund has the higher Sharpe ratio? What does this ratio imply?
52. The following table shows the annual returns (in percent) for the Fidelity Latin America Fund and the Fidelity Canada Fund from 2005 through 2009.



Year	Latin America	Canada
2005	55.17	27.89
2006	44.33	15.04
2007	43.71	35.02
2008	-54.64	-42.64
2009	91.60	39.63

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com).

- Which fund had the higher average return?
  - Which fund was riskier over this time period?
  - Given a risk-free rate of 3%, which fund has the higher Sharpe ratio? What does this ratio imply?
53. **FILE Fidelity\_Select.** The accompanying table shows a portion of the annual return (in percent) for the Fidelity

Select Technology Fund and Fidelity Select Energy Fund from 2000 through 2011.

Year	Technology	Energy
2000	-24.31	30.47
2001	-38.55	-12.49
⋮	⋮	⋮
2011	-12.21	-8.76

SOURCE: [www.finance.com](http://www.finance.com).

- Compare the sample mean and the sample standard deviation of the two fund returns.
- Use a risk-free rate of 2% to compare the Sharpe ratios of the two funds.

## 3.6 ANALYSIS OF RELATIVE LOCATION

LO 3.6

The mean and the standard deviation are the most extensively used measures of central location and dispersion, respectively. Unlike the mean, it is not easy to interpret the standard deviation intuitively. All we can say is that a low value for the standard deviation indicates that the data points are close to the mean, while a high value for the standard deviation indicates that the data are spread out. In this section, we will use Chebyshev's theorem and the empirical rule to make precise statements regarding the percentage of data values that fall within a specified number of standard deviations from the mean. We will also use the mean and the standard deviation to compute  $z$ -scores that measure the relative location of a value within a data set;  $z$ -scores are also used to detect outliers.

Apply Chebyshev's theorem, the empirical rule, and  $z$ -scores.

### Chebyshev's Theorem

As we will see in more detail in later chapters, it is important to be able to use the standard deviation to make statements about the proportion of observations that fall within certain intervals. Fortunately, a Russian mathematician named Pavroty Chebyshev (1821–1894) found bounds for the proportion of the data that lie within a specified number of standard deviations from the mean.

#### CHEBYSHEV'S THEOREM

For any data set, the proportion of observations that lie within  $k$  standard deviations from the mean is at least  $1 - 1/k^2$ , where  $k$  is any number greater than 1.

This theorem holds both for a sample and for a population. For example, it implies that at least 0.75, or 75%, of the observations fall within  $k = 2$  standard deviations from the mean. Similarly, at least 0.89, or 89%, of the observations fall within  $k = 3$  standard deviations from the mean.

#### EXAMPLE 3.18

A large lecture class has 280 students. The professor has announced that the mean score on an exam is 74 with a standard deviation of 8. At least how many students scored within 58 and 90?

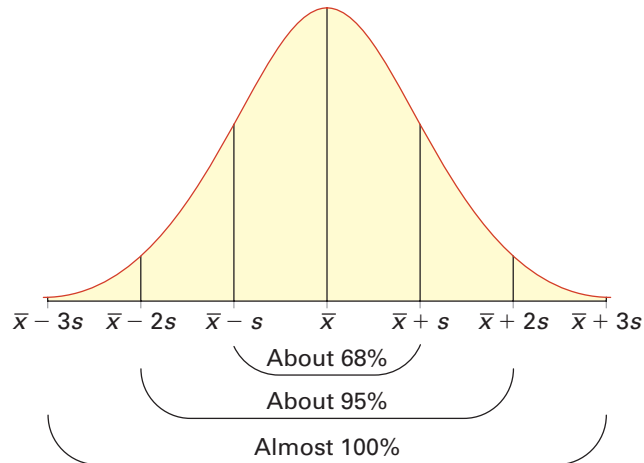
**SOLUTION:** The score 58 is two standard deviations below the mean ( $\bar{x} - 2s = 74 - (2 \times 8) = 58$ ), while the score 90 is two standard deviations above the mean ( $\bar{x} + 2s = 74 + (2 \times 8) = 90$ ). Using Chebyshev's theorem and  $k = 2$ , we have  $1 - 1/2^2 = 0.75$ . In other words, Chebyshev's theorem asserts that at least 75% of the scores will fall within 58 and 90. Therefore, at least 75% of 280 students, or  $0.75(280) = 210$  students, scored within 58 and 90.

The main advantage of Chebyshev's theorem is that it applies to all data sets, regardless of the shape of the distribution. However, it results in conservative bounds for the percentage of observations falling in a particular interval. The actual percentage of observations lying in the interval may in fact be much larger.

## The Empirical Rule

If we know that our data are drawn from a relatively symmetric and bell-shaped distribution—perhaps by a visual inspection of its histogram—then we can make more precise statements about the percentage of observations that fall within certain intervals. Symmetry and bell-shape are characteristics of the normal distribution, a topic that we discuss in Chapter 6. The normal distribution is often used as an approximation for many real-world applications. The **empirical rule** is illustrated in Figure 3.5. It provides the approximate percentage of observations that fall within 1, 2, or 3 standard deviations from the mean.

**FIGURE 3.5**  
Graphical description  
of the empirical rule



### THE EMPIRICAL RULE

Given a sample mean  $\bar{x}$ , a sample standard deviation  $s$ , and a relatively symmetric and bell-shaped distribution:

- Approximately 68% of all observations fall in the interval  $\bar{x} \pm s$ ,
- Approximately 95% of all observations fall in the interval  $\bar{x} \pm 2s$ , and
- Almost all observations fall in the interval  $\bar{x} \pm 3s$ .

### EXAMPLE 3.19

Let's revisit Example 3.18 regarding a large lecture class with 280 students with a mean score of 74 and a standard deviation of 8. Assume that the distribution is symmetric and bell-shaped.

- Approximately how many students scored within 58 and 90?
- Approximately how many students scored more than 90?

#### SOLUTION:

- a. As shown in Example 3.18, the score 58 is two standard deviations below the mean while the score 90 is two standard deviations above the mean. The empirical rule states that approximately 95% of the observations fall within two standard deviations of the mean. Therefore, about 95% of 280 students, or  $0.95(280) = 266$  students, scored within 58 and 90.
- b. We know that the score 90 is two standard deviations above the mean. Since approximately 95% of the observations fall within two standard deviations of the mean, we can infer that 5% of the observations fall outside the interval. Therefore, about half of 5%, or 2.5%, of 280 students scored above 90. Equivalently, about 7 students ( $0.025 \times 280$ ) scored above 90 on the exam. If the professor uses a cutoff score above 90 for an A, then only seven students in the class are expected to get an A.

The main difference between Chebyshev's theorem and the empirical rule is that Chebyshev's theorem applies to all data sets whereas the empirical rule is appropriate when the distribution is symmetric and bell-shaped. In the preceding two examples, while Chebyshev's theorem asserts that at least 75% of the students scored between 58 and 90, we are able to make a more precise statement with the empirical rule that suggests that about 95% of the students scored between 58 and 90. It is preferable to use the empirical rule if the histogram or other visual and numerical measures suggest a symmetric and bell-shaped distribution.

## z-Scores

It is often instructive to use the mean and the standard deviation to find the relative location of values within a data set. Suppose a student gets a score of 90 on her accounting exam and 90 on her marketing exam. While the student's scores are identical in both classes, her relative position in these classes may be quite different. What if the mean score was different in the classes? Even with the same mean scores, what if the standard deviation was different in the classes? Both the mean and the standard deviation are needed to find the relative position of this student in both classes.

We use the **z-score** to find the relative position of a sample value within the data set by dividing the deviation of the sample value from the mean by the standard deviation.

#### z-SCORE

A z-score is computed as

$$z = \frac{x - \bar{x}}{s},$$

where  $x$  is a sample value and  $\bar{x}$  and  $s$  are the sample mean and the sample standard deviation, respectively.

A z-score is a unitless measure since its numerator and the denominator have the same units, which cancel out with each other. It measures the distance of a given sample value from the mean in terms of standard deviations. For example, a z-score of 2 implies that the given sample value is 2 standard deviations above the mean. Similarly, a z-score of  $-1.5$  implies that the given sample value is 1.5 standard deviations below the mean. Converting sample data into z-scores is also called **standardizing** the data.

#### EXAMPLE 3.20

The mean and the standard deviation of scores on an accounting exam are 74 and 8, respectively. The mean and standard deviation of scores on a marketing exam are 78 and 10, respectively. Find the z-scores for a student who scores 90 in both classes.

**SOLUTION:** The  $z$ -score in the accounting class is  $z = \frac{90 - 74}{8} = 2$ . Similarly, the  $z$ -score in the marketing class is  $z = \frac{90 - 78}{10} = 1.2$ . Therefore, the student has fared relatively better in accounting since she is two standard deviations above the mean as compared to marketing where she is only 1.2 standard deviations above the mean.

In Section 3.2, we used box plots as an effective tool to identify outliers. If the data are relatively symmetric and bell-shaped, we can also use  $z$ -scores to detect outliers. Since almost all observations fall within three standard deviations of the mean, it is common to treat an observation as an outlier if its  $z$ -score is more than 3 or less than  $-3$ . Such observations must be reviewed to determine if they should remain in the data set.

### EXAMPLE 3.21

Consider the information presented in the introductory case of this chapter. Use  $z$ -scores to determine if there are outliers in the Metals fund data.

**SOLUTION:** The smallest and the largest observations in the data set are  $-56.02$  and  $76.46$ , respectively. The  $z$ -score for the smallest observation is  $z = \frac{-56.02 - 24.65}{37.13} = -2.17$  and the  $z$ -score for the largest observation is  $z = \frac{76.46 - 24.65}{37.13} = 1.40$ . Since the absolute value of both  $z$ -scores is less than 3, we conclude that there are no outliers in the Metals fund data, assuming that the distribution is relatively symmetric and bell-shaped. This result is consistent with our earlier analysis with the box plot.

## EXERCISES 3.6

### Mechanics

54. A data set has a mean of 80 and a standard deviation of 5.
  - a. Using Chebyshev's theorem, what percentage of the observations fall between 70 and 90?
  - b. Using Chebyshev's theorem, what percentage of the observations fall between 65 and 95?
55. A data set has a mean of 1500 and a standard deviation of 100.
  - a. Using Chebyshev's theorem, what percentage of the observations fall between 1300 and 1700?
  - b. Using Chebyshev's theorem, what percentage of the observations fall between 1100 and 1900?
56. A data set has a mean of 500 and a standard deviation of 25.
  - a. Using Chebyshev's theorem, find the interval that encompasses at least 75% of the data.
  - b. Using Chebyshev's theorem, find the interval that encompasses at least 89% of the data.
57. Data are drawn from a bell-shaped distribution with a mean of 20 and a standard deviation of 2.
  - a. Approximately what percentage of the observations fall between 18 and 22?
  - b. Approximately what percentage of the observations fall between 16 and 24?
  - c. Approximately what percentage of the observations are less than 16?
58. Consider a bell-shaped distribution with a mean of 750 and a standard deviation of 50. There are 500 observations in the data set.
  - a. Approximately what percentage of the observations are less than 700?
  - b. Approximately how many observations are less than 700?
59. Data are drawn from a bell-shaped distribution with a mean of 25 and a standard deviation of 4. There are 1,000 observations in the data set.
  - a. Approximately what percentage of the observations are less than 33?
  - b. Approximately how many observations are less than 33?
60. Data are drawn from a bell-shaped distribution with a mean of 5 and a standard deviation of 2.5.
  - a. Approximately what percentage of the observations are positive?
  - b. Approximately what percentage of the observations are not positive?

61. Data with 250 observations are drawn from a bell-shaped distribution with a mean of 50 and a standard deviation of 12. Approximately how many observations are more than 74?
62. Consider a sample with six observations of 6, 9, 12, 10, 9, and 8. Compute the z-scores for each sample observation.
63. Consider a sample with 10 observations of -3, 8, 4, 2, -4, 15, 6, 0, -4, and 5. Use z-scores to determine if there are any outliers in the data; assume a bell-shaped distribution.
67. An investment strategy has an expected return of 8% and a standard deviation of 6%. Assume investment returns are bell shaped.
  - a. How likely is it to earn a return between 2% and 14%?
  - b. How likely is it to earn a return greater than 14%?
  - c. How likely is it to earn a return below -4%?
68. Average talk time between charges of a given cell phone is advertised as 4 hours. Let the standard deviation be 0.8 hour.
  - a. Use Chebyshev's theorem to approximate the proportion of cell phones that will have talk time between 2.4 hours and 5.6 hours.
  - b. Assume a bell-shaped distribution to approximate the proportion of cell phones that will have talk time between 2.4 hours and 5.6 hours.

### Applications

64. A sample of the salaries of assistant professors on the business faculty at a local university revealed a mean income of \$72,000 with a standard deviation of \$3,000.
  - a. Using Chebyshev's theorem, what percentage of the faculty earns at least \$66,000 but no more than \$78,000?
  - b. Using Chebyshev's theorem, what percentage of the faculty earns at least \$63,000 but no more than \$81,000?
65. The historical returns on a portfolio had an average return of 8% and a standard deviation of 12%. Assume that returns on this portfolio follow a bell-shaped distribution.
  - a. Approximately what percentage of returns were greater than 20%?
  - b. Approximately what percentage of returns were below -16%?
66. It is often assumed that IQ scores follow a bell-shaped distribution with a mean of 100 and a standard deviation of 16.
  - a. Approximately what percentage of scores are between 84 and 116?
  - b. Approximately what percentage of scores are less than 68?
  - c. Approximately what percentage of scores are more than 116?
69. **FILE Census.** The accompanying data file shows, among other variables, median household income and median house value for the 50 states in 2010. Assume that income and house value data are bell-shaped.
  - a. Use z-scores to determine if there are any outliers in the household income data.
  - b. Use z-scores to determine if there are any outliers in the house value data.
70. **FILE Fidelity Select.** The accompanying data file shows the annual return (in percent) for the Fidelity Select Technology Fund and the Fidelity Select Energy Fund from 2000 through 2011. Assume that the return data are bell-shaped.
  - a. Use z-scores to determine if there are any outliers in the technology return data.
  - b. Use z-scores to determine if there are any outliers in the energy return data.

## 3.7 SUMMARIZING GROUPED DATA

### LO 3.7

The mean and the variance are the most widely used descriptive measures in statistics. However, the formulas in Sections 3.1 and 3.4 apply to ungrouped or raw data. In many instances, we access data that are in the form of a frequency distribution or grouped data. This is especially true of secondary data, such as data we obtain from government publications. When data are grouped or aggregated, the formulas for the mean and the variance must be modified.

Calculate the mean and the variance for grouped data.

### THE MEAN AND THE VARIANCE FOR A FREQUENCY DISTRIBUTION

#### Sample:

$$\text{Mean: } \bar{x} = \frac{\sum m_i f_i}{n}$$

$$\text{Variance: } s^2 = \frac{\sum (m_i - \bar{x})^2 f_i}{n - 1}$$

#### Population:

$$\text{Mean: } \mu = \frac{\sum m_i f_i}{N}$$

$$\text{Variance: } \sigma^2 = \frac{\sum (m_i - \mu)^2 f_i}{N}$$

where  $m_i$  and  $f_i$  are the midpoint and the frequency of the  $i$ th class, respectively. The standard deviation is the positive square root of the variance.

We note that by aggregating, some of the data information is lost. Therefore, unlike in the case of raw data, we can only compute approximate values for the summary measures with grouped data.

### EXAMPLE 3.22

Recall the frequency distribution of house prices that we constructed in Chapter 2.

Class (in \$1000s)	Frequency
300 up to 400	4
400 up to 500	11
500 up to 600	14
600 up to 700	5
700 up to 800	2

- Calculate the average house price.
- Calculate the sample variance and the sample standard deviation.

**SOLUTION:** Table 3.13 shows the frequency  $f_i$  and the midpoint  $m_i$  for each class in the second and third columns, respectively.

**TABLE 3.13** The Sample Mean and the Sample Variance Calculation for Grouped Data

Class (in \$1,000s)	$f_i$	$m_i$	$m_i f_i$	$(m_i - \bar{x})^2 f_i$
300 up to 400	4	350	1,400	$(350 - 522)^2 \times 4 = 118,336$
400 up to 500	11	450	4,950	$(450 - 522)^2 \times 11 = 57,024$
500 up to 600	14	550	7,700	$(550 - 522)^2 \times 14 = 10,976$
600 up to 700	5	650	3,250	$(650 - 522)^2 \times 5 = 81,920$
700 up to 800	2	750	1,500	$(750 - 522)^2 \times 2 = 103,968$
Total	36		18,800	372,224

- For the mean, we multiply each class's midpoint by its respective frequency, as shown in the fourth column of Table 3.13. Finally, we sum the fourth column and divide by the sample size. Or,

$$\bar{x} = \frac{\sum m_i f_i}{n} = \frac{18,800}{36} = 522. \text{ The average house price is thus \$522,000.}$$

- For the sample variance, we first calculate the sum of the weighted squared differences from the mean. The fifth column in Table 3.13 shows the appropriate calculations for each class. Summing the values in the fifth column yields the numerator for the variance formula. Thus, we calculate the variance as:

$$s^2 = \frac{\sum (m_i - \bar{x})^2 f_i}{n - 1} = \frac{372,224}{36 - 1} = 10,635.$$

The standard deviation is simply the positive square root of the sample variance, or  $s = \sqrt{10,635} = 103.13$ . The standard deviation for house price is thus \$103,130.



Many times, the data from secondary sources are distributed in the form of a relative frequency distribution rather than a frequency distribution. In order to use the formulas for the mean and variance for grouped data, first convert the relative frequency distribution into a frequency distribution, as discussed in Section 2.2 of Chapter 2.

## EXERCISES 3.7

### Mechanics

71. Consider the following frequency distribution.

Class	Frequency
2 up to 4	20
4 up to 6	60
6 up to 8	80
8 up to 10	20

- Calculate the population mean.
- Calculate the population variance and the population standard deviation.

72. Consider the following frequency distribution.

Class	Frequency
50 up to 60	10
60 up to 70	15
70 up to 80	8
80 up to 100	2

- Calculate the sample mean.
- Calculate the sample variance and the sample standard deviation.

73. The following relative frequency distribution was constructed from a population of 200. Calculate the population mean, the population variance, and the population standard deviation.

Class	Relative Frequency
-20 up to -10	0.35
-10 up to 0	0.20
0 up to 10	0.40
10 up to 20	0.05

74. The following relative frequency distribution was constructed from a sample of 50. Calculate the sample mean, the sample variance, and the sample standard deviation.

Class	Relative Frequency
0 up to 2	0.34
2 up to 4	0.20
4 up to 6	0.40
6 up to 8	0.06

### Applications

75. Fifty cities provided information on vacancy rates (in percent) for local apartments in the following frequency distribution.

Vacancy Rate (in percent)	Frequency
0 up to 3	5
3 up to 6	5
6 up to 9	10
9 up to 12	20
12 up to 15	10

- Calculate the average vacancy rate.
- Calculate the variance and the standard deviation for this sample.

76. A local hospital provided the following frequency distribution summarizing the weights of babies delivered over the month of January.

Weight (in pounds)	Number of Babies
2 up to 4	3
4 up to 6	8
6 up to 8	25
8 up to 10	30
10 up to 12	4

- Calculate the mean weight.
- Calculate the variance and the standard deviation for this sample.

77. A researcher conducts a mileage economy test involving 80 cars. The frequency distribution describing average miles per gallon (mpg) appears in the accompanying frequency distribution.

Average MPG	Frequency
15 up to 20	15
20 up to 25	30
25 up to 30	15
30 up to 35	10
35 up to 40	7
40 up to 45	3

- Calculate the mean mpg.
- Calculate the variance and the standard deviation.

78. The Boston Security Analysts Society, Inc. (BSAS) is a nonprofit association that serves as a forum for the exchange of ideas for the investment community. Suppose the ages of its members are based on the following frequency distribution.

Age	Frequency
21–31	11
32–42	44
43–53	26
54–64	7

- a. Calculate the mean age.  
b. Calculate the sample variance and the sample standard deviation.
79. The National Sporting Goods Association (NSGA) conducted a survey of the ages of people who purchased

athletic footwear in 2009. The ages are summarized in the following percent frequency distribution.

Age of Purchaser	Percent
Under 14 years old	19
14 to 17 years old	6
18 to 24 years old	10
25 to 34 years old	13
35 to 44 years old	14
45 to 64 years old	25
65 years old and over	13

Suppose the survey was based on 100 individuals. Calculate the average age of this distribution. Calculate the sample standard deviation. Use 10 as the midpoint of the first class and 75 as the midpoint of the last class.

### LO 3.8

## 3.8 COVARIANCE AND CORRELATION

Calculate and interpret the covariance and the correlation coefficient.

In Chapter 2, we introduced the idea of a scatterplot to visually assess whether two variables had some type of linear relationship. In this section, we present two numerical measures that quantify the direction and strength of the linear relationship between two variables,  $x$  and  $y$ . It is important to point out that these measures are not appropriate when the underlying relationship between the variables is nonlinear.

An objective numerical measure that reveals the direction of the linear relationship between two variables is called the **covariance**. We use  $s_{xy}$  to refer to a sample covariance, and  $\sigma_{xy}$  to refer to a population covariance.

### THE COVARIANCE

For values  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , the **sample covariance**  $s_{xy}$  is computed as

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}.$$

For values  $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ , the **population covariance**  $\sigma_{xy}$  is computed as

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}.$$

*Note:* As in the case of the sample variance, the sample covariance uses  $n - 1$  rather than  $n$  in the denominator.

- A positive value for covariance indicates a positive linear relationship between the two variables; on average, if  $x$  is above its mean, then  $y$  tends to be above its mean, and vice versa.
- A negative value for covariance indicates a negative linear relationship between the two variables; on average, if  $x$  is above its mean, then  $y$  tends to be below its mean, and vice versa.
- The covariance is zero if  $y$  and  $x$  have no linear relationship.

The covariance is difficult to interpret because it is sensitive to the units of measurement. That is, the covariance between two variables might be 100 and the covariance between

another two variables might be 1,000; yet all we can conclude is that both sets of variables are positively related. We cannot comment on the strength of the relationships. An easier measure to interpret is the **correlation coefficient**; it describes both the direction and strength of the linear relationship between  $x$  and  $y$ . We use  $r_{xy}$  to refer to a sample correlation coefficient and  $\rho_{xy}$  (the Greek letter rho) to refer to a population correlation coefficient.

### THE CORRELATION COEFFICIENT

The **sample correlation coefficient** is computed as  $r_{xy} = \frac{s_{xy}}{s_x s_y}$ , and the **population correlation coefficient** is computed as  $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ .

The correlation coefficient is unit free since the units in the numerator cancel with those in the denominator. The value of the correlation coefficient falls between  $-1$  and  $1$ . A perfect positive relationship exists if it equals  $1$ , and a perfect negative relationship exists if it equals  $-1$ . Other values for the correlation coefficient must be interpreted with reference to  $-1$ ,  $0$ , or  $1$ . For instance, a correlation coefficient equal to  $-0.80$  indicates a strong negative relationship, whereas a correlation coefficient equal to  $0.12$  indicates a weak positive relationship.

### EXAMPLE 3.23

Calculate and interpret the covariance and the correlation coefficient for the Metals ( $x$ ) and Income ( $y$ ) funds. Recall that  $\bar{x} = 24.65$ ,  $s_x = 37.13$ ,  $\bar{y} = 8.51$ , and  $s_y = 11.07$ .

**SOLUTION:** As a first step, Figure 3.6 shows a scatterplot of the return data for the Metals and Income funds; scatterplots were introduced in Section 2.4. It appears that there is a positive linear relationship between the two fund returns.

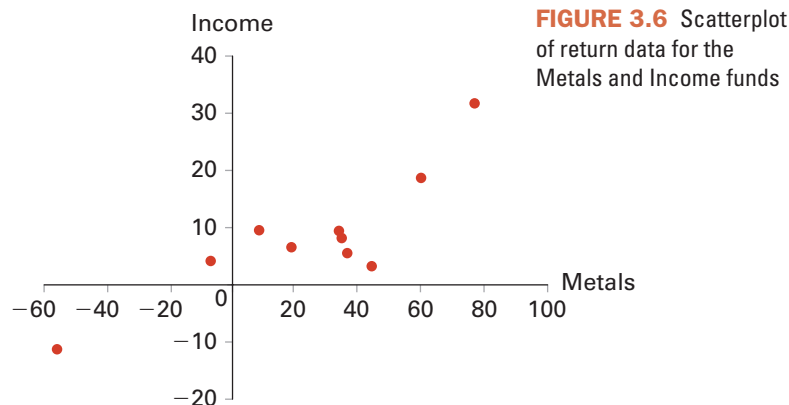


Table 3.14 shows the return data for each fund in the first two columns. The third column shows the product of differences from the mean.

Summing the values in the third column yields the numerator for the covariance formula. Thus, we calculate the covariance as:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{3,165.55}{10 - 1} = 351.73.$$

**TABLE 3.14** Covariance Calculation for the Metals and Income Funds

$x_i$	$y_i$	$(x_i - \bar{x})(y_i - \bar{y})$
-7.34	4.07	$(-7.34 - 24.65)(4.07 - 8.51) = 142.04$
18.33	6.52	$(18.33 - 24.65)(6.52 - 8.51) = 12.58$
$\vdots$	$\vdots$	$\vdots$
76.46	31.77	$(76.46 - 24.65)(31.77 - 8.51) = 1,205.10$
		Total = 3,165.55

The covariance of 351.73 indicates that the variables have a positive linear relationship. In other words, on average, when one fund's return is above its mean, the other fund's return is above its mean, and vice versa. The covariance is used to compute the correlation coefficient as:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{351.73}{(37.13)(11.07)} = 0.86.$$

The correlation coefficient of 0.86 indicates a strong positive linear relationship. In order to diversify the risk in an investor's portfolio, an investor is often advised to invest in assets (such as stocks, bonds, and mutual funds) whose returns are not strongly correlated. If asset returns are not strongly correlated, then if one investment does poorly, the other may still do well.

## Using Excel to Calculate the Covariance and the Correlation Coefficient

### FILE Fund\_Returns

Excel provides formulas for the covariance and the correlation coefficient. Table 3.15 shows Excel's function names for these descriptive measures. For example, in order to calculate the sample covariance between the Metals fund and the Income fund, we open *Fund\_Returns*. We find an empty cell and insert "`=COVARIANCE.S(B2:B11,C2:C11)`"; note that the data for the Metals fund are sitting in cells B2 through B11 (array1) and the data for the Income fund are sitting in cells C2 through C11 (array2). After choosing **<Enter>**, Excel returns a value of 351.73, which matches the value that we calculated by hand.

**TABLE 3.15** Excel's Functions for the Covariance and the Correlation Coefficient

Measure of Dispersion	Excel Function Name
Sample Covariance	<code>=COVARIANCE.S(array1,array2)</code>
Population Covariance	<code>=COVARIANCE.P(array1,array2)</code>
Correlation Coefficient	<code>=CORREL(array1,array2)</code>

## EXERCISES 3.8

### Mechanics

80. Consider the following sample data:

$x$	-2	0	3	4	7
$y$	-2	-3	-8	-9	-10

- Calculate the covariance between the variables.
- Calculate and interpret the correlation coefficient.

81. Consider the following sample data:

$x$	12	18	20	22	25
$y$	15	20	25	22	27

- Calculate the covariance between the variables.
- Calculate and interpret the correlation coefficient.

## Applications

82. In an attempt to determine whether a linear relationship exists between the price of a home and the number of days it takes to sell the home, a real estate agent collected the following data from recent sales in his city.

Price (in \$1,000s)	Days to Sell Home	Price (in \$1,000s)	Days to Sell Home
265	136	430	145
225	125	515	121
160	120	180	122
325	140	423	145

- a. Calculate the covariance. What kind of linear relationship exists?
- b. Calculate and interpret the correlation coefficient.
83. The following table shows the annual returns (in percent) for T. Rowe Price's Value and International Stock funds for the time period 2005–2009.

Year	Value Fund	International Fund
2005	6.30	16.27
2006	19.75	19.26
2007	0.75	13.43
2008	−39.76	−48.02
2009	37.15	52.20

- a. Calculate and interpret the covariance between the returns.
- b. Calculate and interpret the correlation coefficient.
84. A social scientist wants to analyze the relationship between educational attainment and salary. He interviews eight people. The accompanying table shows each person's years of higher education (Education) and corresponding salary (Salary, measured in \$1,000s).

Education	3	4	6	2	5	4	8	0
Salary	40	53	60	35	55	50	80	35

- a. Calculate the covariance. What kind of linear relationship exists?
- b. Calculate and interpret the correlation coefficient.
85. The director of graduate admissions at a local university is analyzing the relationship between scores on the Graduate Record Examination (GRE) and subsequent performance in graduate school, as measured by a student's grade point average (GPA). She uses a sample of 10 students who graduated within the past five years.

GRE	GPA
1500	3.4
1400	3.5
1000	3.0
1050	2.9
1100	3.0
1250	3.3
800	2.7
850	2.8
950	3.2
1350	3.3

- a. Calculate and interpret the covariance.
- b. Calculate and interpret the correlation coefficient. Does an applicant's GRE score seem to be a good indicator of subsequent performance in graduate school?
86. **FILE Census.** Access the data accompanying this exercise.
- a. Compute and interpret the correlation coefficient for household income and house value.
- b. Compute and interpret the correlation coefficient for household income and the percentage of the residents who are foreign born.
- c. Compute and interpret the correlation coefficient for household income and the percentage of the residents who are without a high school diploma.
87. **FILE Happiness\_Age.** Many attempts have been made to relate happiness with various factors. One such study relates happiness with age and finds that holding everything else constant, people are least happy when they are in their mid-40s (*The Economist*, December 16, 2010). Data are collected on a respondent's age and his/her perception of well-being on a scale from 0 to 100; a portion of the data is presented below.

Age	Happiness
49	62
51	66
⋮	⋮
69	72

- a. Calculate and interpret the sample correlation coefficient between age and happiness.
- b. Construct a scatterplot to point out a flaw with the above correlation analysis.

# WRITING WITH STATISTICS

Many environmental groups and politicians are suggesting a return to the federal 55-mile-per-hour (mph) speed limit on America’s highways. They argue that not only will a lower national speed limit reduce greenhouse emissions, it will also increase traffic safety.

Cameron Grinnell believes that a lower speed limit will not increase traffic safety. He believes that traffic safety is based on the variability of the speeds with which people are driving, rather than the average speed. The person who drives 20 mph below the pace of traffic is often as much a safety menace as the speeder. Cameron gathers the speeds of 40 cars from a highway with a speed limit of 55 mph (Highway 1) and the speeds of 40 cars from a highway with a speed limit of 65 mph (Highway 2). A portion of the data is shown in Table 3.16.



**FILE**  
*Highway\_Speeds*

**TABLE 3.16** Speed of Cars from Highway 1 and Highway 2

Highway 1 (55-mph limit)	Highway 2 (65-mph limit)
60	70
55	65
⋮	⋮
52	65

Cameron would like to use the above sample information to:

1. Compute and interpret the typical speed on these highways.
2. Compute and interpret the variability of speed on these highways.
3. Discuss if the reduction in the speed limit to 55 mph would increase safety on the highways.

## Sample Report—Analyzing Speed Limits

Recently, many concerned citizens have lobbied for a return to the federal 55-mile-per-hour (mph) speed limit on America’s highways. The reduction may lower gas emissions and save consumers on gasoline costs, but whether it will increase traffic safety is not clear. Many researchers believe that traffic safety is based on the variability of the speed rather than the average speed with which people are driving—the more variability in speed, the more dangerous the roads. Is there less variability in speed on a highway with a 55-mph speed limit as opposed to a 65-mph speed limit?

To compare average speeds, as well as the variability of speeds on highways, the speeds of 40 cars were recorded on a highway with a 55-mph speed limit (Highway 1) and on a highway with a 65-mph speed limit (Highway 2). Table 3.A shows the most relevant descriptive measures for the analysis.

**TABLE 3.A** Summary Measures for Highway 1 and Highway 2

	Highway 1 (55-mph speed limit)	Highway 2 (65-mph speed limit)
Mean	57	66
Median	56	66
Mode	50	70
Minimum	45	60
Maximum	74	70
Standard deviation	7.0	3.0
Coefficient of variation	0.12	0.05
Number of cars	40	40



The average speed of a car on Highway 1 was 57 mph, as opposed to 66 mph on Highway 2. On Highway 1, half of the 40 cars drove faster than 56 mph and half drove slower than 56 mph, as measured by the median; the median for Highway 2 was 66 mph. The mode shows that the most common speeds on Highway 1 and Highway 2 were 50 mph and 70 mph, respectively. Based on each measure of central location, Highway 2 experiences higher speeds as compared to Highway 1.

While measures of central location typically represent where the data cluster, these measures do not relay information about the variability in the data. The range of speeds is 29 mph for Highway 1 as compared to a range of just 10 mph for Highway 2. Generally, standard deviation is a more credible measure of dispersion, since range is based entirely on the minimum and the maximum values. The standard deviation for Highway 1 is substantially greater than the standard deviation for Highway 2 (7.0 mph > 3.0 mph). Therefore, the speeds on Highway 1 are more variable than the speeds on Highway 2. Even adjusting for differences in the magnitudes of the means by calculating the coefficient of variation, the speeds on Highway 1 are still more dispersed than on Highway 2 ( $0.12 > 0.05$ ).

On average, it is true that the speeds on Highway 2 are higher than the speeds on Highway 1; however, the variability of speeds is greater on Highway 1. If traffic safety improves when the variability of speeds declines, then the data suggest that a return to a federal 55-mph speed limit may not enhance the well-being of highway travelers.

## CONCEPTUAL REVIEW

### LO 3.1 Calculate and interpret the mean, the median, and the mode.

The mean (average) is the most widely used measure of central location. The **sample mean** and the **population mean** are computed as  $\bar{x} = \frac{\sum x_i}{n}$  and  $\mu = \frac{\sum x_i}{N}$ , respectively. One weakness of the mean is that it is unduly influenced by **outliers**—extremely small or large values.

The **median** is the middle value of a data set and is especially useful when outliers are present. We arrange the data in ascending (smallest to largest) order and find the median as the middle value if the number of observations is odd, or the average of the two middle values if the number of observations is even.

The **mode** is the value in the data set that occurs with the most frequency. A data set may have no mode or more than one mode. If the data are qualitative, then the mode is the only meaningful measure of central location.

### LO 3.2 Calculate and interpret percentiles and a box plot.

**Percentiles** provide detailed information about how the data are spread over the interval from the smallest value to the largest value. In general, the  $p$ th percentile divides the data set into two parts, where approximately  $p$  percent of the observations have values less than the  $p$ th percentile and the rest have values greater than the  $p$ th percentile. The 25th percentile is also referred to as the first quartile (Q1), the 50th percentile is referred to as the second quartile (Q2), and the 75th percentile is referred to as the third quartile (Q3).

A **box plot** displays the five-number summary (the minimum value, Q1, Q2, Q3, and the maximum value) for the data set. Box plots are particularly useful when comparing similar information gathered at another place or time. They are also used as an effective tool for identifying outliers and skewness.

### LO 3.3 Calculate and interpret a geometric mean return and an average growth rate.

The **geometric mean** is the multiplicative average of a data set. In general, the geometric mean is smaller than the arithmetic mean and is less sensitive to outliers. The geometric mean is relevant when summarizing financial returns over several years. For multiperiod returns  $R_1, R_2, \dots, R_n$ , the **geometric mean return** is computed as  $G_R = \sqrt[n]{(1 + R_1)(1 + R_2) \cdots (1 + R_n)} - 1$ , where  $n$  is the number of multiperiod returns.

The geometric mean is also used when summarizing **average growth rates**. For growth rates  $g_1, g_2, \dots, g_n$ , the average growth rate is computed as  $G_g = \sqrt[n]{(1 + g_1)(1 + g_2) \cdots (1 + g_n)} - 1$ , where  $n$  is the number of multiperiod growth rates. When the underlying values of the series are given, there is a simpler way to compute the average growth rate. For observations  $x_1, x_2, \dots, x_n$ , the average growth rate is computed as  $G_g = \sqrt[n-1]{\frac{x_n}{x_1}} - 1$ , where  $n - 1$  is the number of distinct growth rates.

### LO 3.4 Calculate and interpret the range, the mean absolute deviation, the variance, the standard deviation, and the coefficient of variation.

The **range** is the difference between the maximum and the minimum values in a data set.

The **mean absolute deviation** (MAD) is an average of the absolute differences between the observations and the mean of a data set. The sample MAD and the population MAD are computed as  $\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$  and  $\text{MAD} = \frac{\sum |x_i - \mu|}{N}$ , respectively.

The **variance** and the **standard deviation**, which are based on squared differences from the mean, are the two most widely used measures of dispersion. The sample variance  $s^2$  and the sample standard deviation  $s$  are computed as  $s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$  and  $s = \sqrt{s^2}$ , respectively. The population variance  $\sigma^2$  and the population standard deviation  $\sigma$  are computed as  $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$  and  $\sigma = \sqrt{\sigma^2}$ , respectively. Variance squares the original units of measurement; by calculating the standard deviation, we return to the original units of measurement.

The **coefficient of variation** CV is a relative measure of dispersion. The CV allows comparisons of variability between data sets with different means or different units of measurement. The sample CV and the population CV are computed as  $\text{CV} = \frac{s}{\bar{x}}$  and  $\text{CV} = \frac{\sigma}{\mu}$ , respectively.

### LO 3.5 Explain mean-variance analysis and the Sharpe ratio.

**Mean-variance analysis** postulates that we measure the performance of an asset by its rate of return and evaluate this rate of return in terms of its reward (mean) and risk (variance). In general, investments with higher average returns are also associated with higher risk.

The **Sharpe ratio** measures extra reward per unit of risk. The Sharpe ratio for an investment  $I$  is computed as  $\frac{\bar{x}_I - \bar{R}_f}{s_I}$ , where  $\bar{R}_f$  denotes the mean return on a risk-free asset. The higher the Sharpe ratio, the better the investment compensates its investors for risk.

### LO 3.6 Apply Chebyshev's theorem, the empirical rule, and z-scores.

**Chebyshev's theorem** dictates that for any data set, the proportion of observations that lie within  $k$  standard deviations from the mean will be at least  $1 - 1/k^2$ , where  $k$  is any number greater than 1.

Given a sample mean  $\bar{x}$ , a sample standard deviation  $s$ , and a bell-shaped distribution, the **empirical rule** dictates that

- Approximately 68% of all observations fall in the interval  $\bar{x} \pm s$ ,
- Approximately 95% of all observations fall in the interval  $\bar{x} \pm 2s$ , and
- Almost all observations fall in the interval  $\bar{x} \pm 3s$ .

A **z-score**, calculated as  $(x - \bar{x})/s$ , measures the relative location of the sample value  $x$ ; it is also used to detect outliers.

### LO 3.7 Calculate the mean and the variance for grouped data.

When analyzing **grouped data**, the formulas for the mean and the variance are modified as follows:

- The sample mean and the population mean are computed as  $\bar{x} = \frac{\sum m_i f_i}{n}$  and  $\mu = \frac{\sum m_i f_i}{N}$ , respectively.
- The sample variance and the population variance are computed as  $s^2 = \frac{\sum (m_i - \bar{x})^2 f_i}{n - 1}$  and  $\sigma^2 = \frac{\sum (m_i - \mu)^2 f_i}{N}$ , respectively. As always, the standard deviation is calculated as the positive square root of the variance.

### LO 3.8 Calculate and interpret the covariance and the correlation coefficient.

The **covariance** and the **correlation coefficient** are measures that assess the direction and strength of a linear relationship between two variables,  $x$  and  $y$ . The sample covariance  $s_{xy}$  and the population covariance  $\sigma_{xy}$  are computed as  $s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$  and  $\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$ , respectively. The sample correlation coefficient  $r_{xy}$  and the population correlation coefficient  $\rho_{xy}$  are computed as  $r_{xy} = \frac{s_{xy}}{s_x s_y}$  and  $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ , respectively.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

88. The following table lists the sales (in millions of dollars) of the top Italian restaurant chains in 2009.

Restaurant	Sales (millions)
Olive Garden	\$3,300
Carrabba's Italian Grill	629
Romano's Macaroni Grill	583
Maggiano's	366
Carino's Italian Grill	356
Buca di Beppo	220
Bertucci's	210

SOURCE: *The Boston Globe*, July 31, 2010.

Calculate the mean, the median, and the mode. Which measure of central tendency best reflects typical sales? Explain.

89. The following table shows the annual returns (in percent) for Fidelity's Electronic and Utilities funds.

Year	Electronic	Utilities
2005	13.23	9.36
2006	1.97	32.33
2007	2.77	21.03
2008	-50.00	-35.21
2009	81.65	14.71

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com).

- Calculate the sample mean, the sample variance, and the sample standard deviation for each fund.
- Which fund had the higher average return?
- Which fund was riskier over this time period? Use both the standard deviation and the coefficient of variation in your explanation.
- Given a risk-free rate of 4%, which fund has the higher Sharpe ratio? What does this ratio imply?

90. Monthly stock prices for two competing firms are as follows.

Month	Firm A	Firm B
January	\$28	\$21
February	31	24
March	32	24
April	35	27
May	34	25
June	28	20

- Calculate the sample mean, the sample variance, and the sample standard deviation for each firm's stock price.
  - Which firm had the higher stock price over the time period?
  - Which firm's stock price had greater variability as measured by the standard deviation? Which firm's stock price had the greater relative dispersion?
91. The manager at a water park constructed the following frequency distribution to summarize attendance for 60 days in July and August.

Attendance	Frequency
1,000 up to 1,250	5
1,250 up to 1,500	6
1,500 up to 1,750	10
1,750 up to 2,000	20
2,000 up to 2,250	15
2,250 up to 2,500	4

- Calculate the mean attendance.
  - Calculate the variance and the standard deviation.
92. The following table shows the revenues (in millions of dollars) for The Gap, Inc., and American Eagle Outfitters, Inc., for the years 2008–2010.

Year	Gap	American Eagle
2008	\$15.73	\$3.06
2009	14.53	2.99
2010	14.20	2.99

SOURCE: Annual Reports for Gap, Inc., and American Eagle Outfitters, Inc.

- Calculate the average growth rate for each firm.
  - Which firm had the higher growth rate over the 2008–2010 period?
93. Annual growth rates for individual firms in the toy industry tend to fluctuate dramatically, depending on consumers' tastes and current fads. Consider the following growth rates (in percent) for two companies in this industry, Hasbro and Mattel.

Year	2005	2006	2007	2008	2009
Hasbro	3.0	2.1	21.8	4.8	1.2
Mattel	1.5	9.1	5.7	−0.1	−8.2

SOURCE: Annual Reports for Hasbro, Inc., and Mattel Inc.

- Calculate the average growth rates for each firm.
- Use the standard deviation to evaluate the variability for each firm.
- Which company had the higher average growth rate? Which company's growth rate had greater variability?

94. The National Sporting Goods Association (NSGA) conducted a survey of the ages of individuals that purchased skateboarding footwear. The ages of this survey are summarized in the following percent frequency distribution.

Age of User	Percent
Under 14 years old	35
14 to 17 years old	41
18 to 24 years old	15
25 to 34 years old	4
35 to 44 years old	4
45 to 64 years old	1

Suppose the survey was based on a sample of 200 individuals. Calculate the mean and the standard deviation of the age of individuals that purchased skateboarding shoes. Use 10 as the midpoint of the first class.

95. A manager of a local retail store analyzes the relationship between advertising and sales by reviewing the store's data for the previous six months.

Advertising (in \$100s)	Sales (in \$1,000s)
20	15
25	18
30	20
22	16
27	19
26	20

- Calculate the mean of advertising and the mean of sales.
  - Calculate the standard deviation of advertising and the standard deviation of sales.
  - Calculate and interpret the covariance between advertising and sales.
  - Calculate and interpret the correlation coefficient.
96. The following table shows the annual returns (in percent) for two of Putnam's mutual funds: the Voyager Growth Fund and the George Putnam Fund of Boston.

Year	Growth Fund	Fund of Boston
2002	−26.43	−8.42
2003	24.71	17.40
2004	4.80	8.32
2005	5.50	4.04
2006	5.23	12.25

SOURCE: www.finance.yahoo.com.

- Calculate and interpret the covariance.
- Calculate the correlation coefficient and interpret.

97. **FILE Debt Payments.** An economist wishes to summarize sample data from 26 metropolitan areas in the United States. The following table lists a portion of each area's 2010–2011 median income as well as the monthly unemployment rate and average consumer debt for August 2010.

Metropolitan Area	Income (in \$1,000s)	Unemployment	Debt
Washington, D.C.	\$103.50	6.3%	\$1,285
Seattle	81.70	8.5	1,135
⋮	⋮	⋮	⋮
Pittsburgh	63.00	8.3	763

SOURCE: eFannieMae.com reports 2010–2011 area median incomes; www.bls.gov gives monthly unemployment rates for August 2010; Experian.com collected average monthly consumer debt payments in August 2010 and published the data in November 2010.

Use Excel to compute the summary measures for income, the monthly unemployment rate, and average consumer debt. Interpret these summary measures.

98. **FILE Car Theft.** The accompanying table shows a portion of the number of cases of car thefts for the 50 states during 2010.

State	Car Theft
Alabama	658
Alaska	280
⋮	⋮
Wyoming	84

SOURCE: www.fbi.gov.

- Calculate the mean, the median, and the mode for the number of car thefts.
  - Use z-scores to determine if there are any outliers in the data. Are you surprised by the result?
99. **FILE Quarterback Salaries.** American football is the highest paying sport on a per-game basis. Given that the quarterback is considered the most important player on an NFL team, he is typically well-compensated. Consider a portion of the following quarterback salary data in 2009.

Name	Salary (in \$ millions)
Philip Rivers	25.5566
Jay Cutler	22.0441
⋮	⋮
Tony Romo	0.6260

SOURCE: www.nfl.com.

- Use Excel to compute and interpret the mean and the median salary for a quarterback.
- Use Excel to compute and interpret the range and the standard deviation for quarterback salaries.

100. **FILE Gambling.** The accompanying table shows a portion of the number of cases of crime related to gambling (Gambling) and offenses against the family and children (Family Abuse) for the 50 states in the United States during 2010.

State	Gambling	Family Abuse
Alabama	47	1,022
Alaska	10	315
⋮	⋮	⋮
Wyoming	0	194

SOURCE: www.fbi.gov.

- Construct a box plot for gambling and use it to identify outliers, if any.
- Construct a box plot for abuse and use it to identify outliers, if any.
- Calculate and interpret the sample correlation coefficient between gambling and family abuse.

101. **FILE Gas Prices 2012.** The accompanying table shows a portion of the average price for a gallon of gas for the 50 states during April 2012.

State	Price per Gallon
Alabama	\$4.36
Alaska	3.79
⋮	⋮
Wyoming	3.63

SOURCE: AAA.com, data retrieved April 16, 2012.

- Construct a box plot for the gasoline price and use it to identify outliers, if any.
- Confirm your analysis by using z-scores to determine if there are any outliers in the gasoline price.

## CASE STUDIES

**CASE STUDY 3.1** An article in *The Wall Street Journal* (July 11, 2008) outlined a number of reasons as to why the 16 teams in Major League Baseball's National League (NL) are inferior to the 14 teams in the American League (AL). One reason for the imbalance pointed to the disparity in opening-day payrolls: the average AL payroll is greater than the NL average. A portion of the data showing opening-day payroll for each team is shown in the accompanying table.

**FILE**  
MLB\_Salaries

**Data for Case Study 3.1** Major League Baseball's Opening-Day Payrolls, 2010

American League	Payroll	National League	Payroll
New York Yankees	\$206,333,389	Chicago Cubs	\$146,609,000
Boston Red Sox	162,447,333	Philadelphia Phillies	141,928,379
⋮	⋮	⋮	⋮

SOURCE: [www.bizofbaseball.com](http://www.bizofbaseball.com).

In a report, use the sample information to:

1. Discuss the mean and the median of AL and NL opening-day salaries and comment on skewness.
2. Compare the range and the standard deviation of AL and NL opening-day salaries.
3. Use these summary measures to comment on the findings in *The Wall Street Journal*.

**CASE STUDY 3.2** Five years after graduating from college, Lucia Li feels that she is finally ready to invest some of her earnings. She has eliminated her credit card debt and has established an emergency fund. Her parents have been pleased with the performance of their mutual fund investments with Janus Capital Group. She has narrowed her search down to two mutual funds:

The Janus Balanced Fund: This “core” fund consists of stocks and bonds and its goal is diversification. It has historically produced solid long-term returns through different market cycles.

The Janus Overseas Fund: This fund invests in overseas companies based on their individual merits instead of their geography or industry sector.

The following table reports the annual returns (in percent) of these two funds over the past 10 years.

**FILE**  
Janus\_Funds

**Data for Case Study 3.2** Returns (in percent) for Janus Funds

Year	Janus Balanced Fund	Janus Overseas Fund	Year	Janus Balanced Fund	Janus Overseas Fund
2000	-2.16	-18.57	2005	7.75	32.39
2001	-5.04	-23.11	2006	10.56	47.21
2002	-6.56	-23.89	2007	10.15	27.76
2003	13.74	36.79	2008	-15.22	-52.75
2004	8.71	18.58	2009	24.28	78.12

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com).

In a report, use the sample information to:

1. Calculate measures of central location to describe the similarities and the differences in these two funds' returns.
2. Calculate measures of dispersion to assess the risk of each fund.
3. Calculate and interpret measures of correlation between the two funds.



**CASE STUDY 3.3** Due to a crisis in subprime lending, obtaining a mortgage has become difficult even for people with solid credit. In a report by the Associated Press (August 25, 2007), sales of existing homes fell for a 5th consecutive month, while home prices dropped for a record 12th month in July 2007. Mayan Horowitz, a research analyst for QuantExperts, wishes to study how the mortgage crunch has impacted the once-booming market of Florida. He collects data on the sale prices (in \$1,000s) of 25 single-family homes in Fort Myers, Florida, in January 2007 and collects another sample in July 2007. For a valid comparison, he samples only three-bedroom homes, each with 1,500 square feet or less of space on a lot size of 10,000 square feet or less. A portion of the data is shown in the accompanying table.

**Data for Case Study 3.3** Home Prices (in \$1,000s) in January 2007 and July 2007

Number	January	July
1	\$100	\$136
2	190	235
⋮	⋮	⋮
25	200	180

SOURCE: www.zillow.com.

**FILE**  
Fort\_Myers\_Sales

In a report, use the sample information to:

1. Compare the mean, the median, and the mode in each of the two sample periods.
2. Compare the standard deviation and the coefficient of variation in each of the two sample periods.
3. Discuss significant changes in the housing market in Fort Myers over the 6-month period.

**CASE STUDY 3.4** Nike's Online Annual Report provides total revenues (in millions of \$) for the Asian and Latin American regions for the years 2005 through 2009 as follows:

**Data for Case Study 3.4**

(a) Nike Revenues in Asia and Latin America (in millions of \$)

	2005	2006	2007	2008	2009
Asia	1,897	2,054	2,296	2,888	3,322
Latin America	696	905	967	1,165	1,285

Adidas' Online Annual Report provides total revenues (in millions of €) for the Asian and Latin American regions for the years 2005 through 2009 as follows:

**Data for Case Study 3.4**

(b) Nike Revenues in Asia and Latin America (in millions of \$)

	2005	2006	2007	2008	2009
Asia	1,523	2,020	2,254	2,662	2,614
Latin America	319	499	657	893	1,006

In a report, use the sample information to:

1. Summarize the growth rates in Asia and Latin America for Nike.
2. Summarize the growth rates in Asia and Latin America for Adidas.
3. Discuss the similarities and the differences of the growth rates in the two companies.

## APPENDIX 3.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### Calculating Summary Measures

**FILE**  
*Fund\_Returns*

- A. (Replicating Table 3.3) From the menu choose **Stat > Basic Statistics > Display Descriptive Statistics**. Then, under **Variables**, select Metals and Income. Click **Statistics**.
- B. Choose the summary measures that you wish to calculate, such as Mean, Standard deviation, etc.

#### Constructing a Box Plot

**FILE**  
*Fund\_Returns*

- A. (Replicating Figure 3.4) From the menu choose **Graph > Boxplot > One Y > Simple**.
- B. Under **Graph variables**, select Metals. Click on **Data View**. Choose **Interquartile range box**, **Outlier symbols**, **Individual symbols**, and **Median connect line**.
- C. Click on **Scale** and select the **Transpose value and category scales** box.

#### Calculating the Covariance and the Correlation Coefficient

**FILE**  
*Fund\_Returns*

(Replicating Example 3.23) From the menu choose **Stat > Basic Statistics > Covariance** (choose **Correlation** to calculate the correlation coefficient). Under **Variables**, select Metals and Income.

### SPSS

#### Calculating Summary Measures

**FILE**  
*Fund\_Returns*

- A. (Replicating Table 3.3) From the menu choose **Analyze > Descriptive Statistics > Descriptives**.
- B. Under **Variables**, select Metals and Income. Choose **Options**. Select the summary measures that you wish to calculate, such as Mean, Std. deviation, etc.

#### Calculating the Covariance and the Correlation Coefficient

**FILE**  
*Fund\_Returns*

- A. (Replicating Example 3.23) From the menu choose **Analyze > Correlate > Bivariate**.
- B. Under **Variables**, select Metals and Income. Under **Correlation Coefficients**, select **Pearson**. Choose **Options**. Under **Statistics**, select **Cross-product deviations and covariances**.

### JMP

#### Calculating Summary Measures and Constructing a Box Plot

**FILE**  
*Fund\_Returns*

(Replicating Table 3.3 and Figure 3.4) From the menu choose **Analyze > Distribution**. Under **Select Columns**, select Metals and Income, and under **Cast Selected Columns into Roles**, choose **Y, Columns**.

## Calculating the Covariance and the Correlation Coefficient

- A. (Replicating Example 3.23) From the menu choose **Analyze > Multivariate Methods > Multivariate**. Under **Select Columns**, select Metals and Income, and under **Cast Selected Columns into Roles**, select **Y, Columns**.
- B. Click the red triangle beside **Multivariate**. Select **Covariance Matrix**.

**FILE**  
*Fund\_Returns*

# 4

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 4.1 Describe fundamental probability concepts.
- LO 4.2 Formulate and explain subjective, empirical, and classical probabilities.
- LO 4.3 Calculate and interpret the probability of the complement of an event.
- LO 4.4 Calculate and interpret the probability that at least one of two events will occur.
- LO 4.5 Calculate and interpret a conditional probability and apply the multiplication rule.
- LO 4.6 Distinguish between independent and dependent events.
- LO 4.7 Calculate and interpret probabilities from a contingency table.
- LO 4.8 Apply the total probability rule.
- LO 4.9 Apply Bayes' theorem.
- LO 4.10 Use a counting rule to calculate the probability of an event.

# Introduction to Probability

Every day we make choices about issues in the presence of uncertainty. Uncertainty describes a situation where a variety of events are possible. Usually, we either implicitly or explicitly assign probabilities to these events and plan or act accordingly. For instance, we read the paper, watch the news, or check the Internet to determine the likelihood of rain and whether we should carry an umbrella. Retailers strengthen their sales force before the end-of-year holiday season in anticipation of an increase in shoppers. The Federal Reserve cuts interest rates when it believes the economy is at risk for weak growth and raises interest rates when it feels that inflation is the greater risk. By figuring out the chances of various events, we are better prepared to make the more desirable choices. This chapter presents the essential probability tools needed to frame and address many real-world issues involving uncertainty. Probability theory turns out to be the very foundation for statistical inference, and numerous concepts introduced in this chapter are essential for understanding later chapters.



## INTRODUCTORY CASE

### Sportswear Brands

Annabel Gonzalez is chief retail analyst at Longmeadow Consultants, a marketing firm. One aspect of her job is to track sports-apparel sales and uncover any particular trends that may be unfolding in the industry. Recently, she has been following Under Armour, Inc., the pioneer in the compression-gear market. Compression garments are meant to keep moisture away from a wearer's body during athletic activities in warm and cool weather. Under Armour has experienced exponential growth since the firm went public in November 2005. However, Nike, Inc., and Adidas Group, with 18% and 10% market shares, respectively, have aggressively entered the compression-gear market (*The Wall Street Journal*, October 23, 2007).

As part of her analysis, Annabel would first like to examine whether the age of the customer matters when buying compression clothing. Her initial feeling is that the Under Armour brand attracts a younger customer, whereas the more established companies, Nike and Adidas, draw an older clientele. She believes this information is relevant to advertisers and retailers in the sporting-goods industry, as well as to some in the financial community. She collects data on 600 recent purchases in the compression-gear market. She cross-classifies the data by age group and brand name, as shown in Table 4.1.

**TABLE 4.1** Purchases of Compression Garments Based on Age and Brand Name

Age Group	Brand Name		
	Under Armour	Nike	Adidas
Under 35 years	174	132	90
35 years and older	54	72	78

Annabel wants to use the sample information to:

1. Calculate and interpret relevant probabilities concerning brand name and age.
2. Determine whether the appeal of the Under Armour brand is mostly to younger customers.

A synopsis of this case is provided at the end of Section 4.3.

Describe fundamental probability concepts.

Since many choices we make involve some degree of uncertainty, we are better prepared for the eventual outcome if we can use probabilities to describe which events are likely and which are unlikely.

A **probability** is a numerical value that measures the likelihood that an event occurs. This value is between zero and one, where a value of zero indicates *impossible* events and a value of one indicates *definite* events.

In order to define an event and assign the appropriate probability to it, it is useful to first establish some terminology and impose some structure on the situation.

An **experiment** is a process that leads to one of several possible outcomes. The diversity of the outcomes of an experiment is due to the uncertainty of the real world. When you purchase a new computer, there is no guarantee as to how long it will last before any repair work is needed. It may need repair in the first year, in the second year, or after two years. You can think of this as an experiment because the actual outcome will be determined only over time. Other examples of an experiment include whether a roll of a fair die will result in a value of 1, 2, 3, 4, 5, or 6; whether the toss of a coin results in heads or tails; whether a project is finished early, on time, or late; whether the economy will improve, stay the same, or deteriorate; whether a ball game will end in a win, loss, or tie.

A **sample space**, denoted by  $S$ , of an experiment contains all possible outcomes of the experiment. For example, suppose the sample space representing the letter grade in a course is given by  $S = \{A, B, C, D, F\}$ . If the teacher also gives out an I (incomplete) grade, then  $S$  is not valid because all outcomes of the experiment are not included in  $S$ . The sample space for an experiment need not be unique. For example, in the above experiment, we can also define the sample space with just P (pass) and F (fail) outcomes; that is,  $S = \{P, F\}$ .

An **experiment** is a process that leads to one of several possible outcomes. A **sample space**, denoted  $S$ , of an experiment contains all possible outcomes of the experiment.

### EXAMPLE 4.1

A snowboarder competing in the Winter Olympic Games is trying to assess her probability of earning a medal in her event, the ladies' halfpipe. Construct the appropriate sample space.

**SOLUTION:** The athlete's attempt to predict her chances of earning a medal is an experiment because, until the Winter Games occur, the outcome is unknown. We formalize an experiment by constructing its sample space. The athlete's competition has four possible outcomes: gold medal, silver medal, bronze medal, and no medal. We formally write the sample space as  $S = \{\text{gold, silver, bronze, no medal}\}$ .

## Events

An **event** is a subset of the sample space. A simple event consists of just one of the possible outcomes of an experiment. Getting an A in a course is an example of a simple event. An event may also contain several outcomes of an experiment. For example, we can



define an event as getting a passing grade in a course; this event is formed by the subset of outcomes A, B, C, and D.

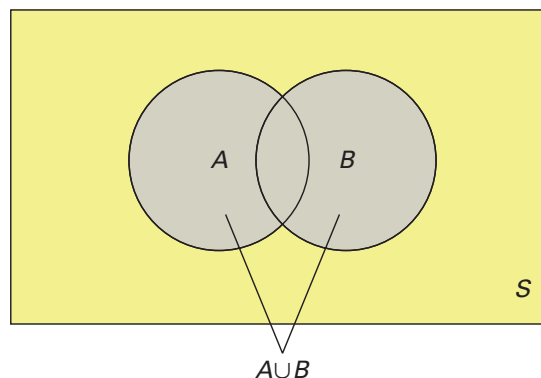
An **event** is any subset of outcomes of the experiment. It is called a simple event if it contains a single outcome.

Let us define two events from Example 4.1, where one event represents “earning a medal” and the other denotes “failing to earn a medal.” These events are **exhaustive** because they include all outcomes in the sample space. In the earlier grade-distribution example, the events of getting grades A and B are not exhaustive events because they do not include many feasible grades in the sample space. However, the events P and F, defined as pass and fail, respectively, are exhaustive.

Another important probability concept concerns **mutually exclusive** events. For two mutually exclusive events, the occurrence of one event precludes the occurrence of the other. Suppose we define the two events “at least earning a silver medal” (outcomes of gold and silver) and “at most earning a silver medal” (outcomes of silver, bronze, no medal). These two events are exhaustive because no outcome of the experiment is omitted. However, in this case, the events are not mutually exclusive because the outcome “silver” appears in both events. Going back to the grade-distribution example, while the events of getting grades A and B are not exhaustive, they are mutually exclusive, since you cannot possibly get an A as well as a B in the same course. However, getting grades P and F are mutually exclusive and exhaustive. Similarly, the events defined as “at least earning a silver medal” and “at most earning a bronze medal” are mutually exclusive and exhaustive.

Events are **exhaustive** if all possible outcomes of an experiment belong to the events. Events are **mutually exclusive** if they do not share any common outcome of an experiment.

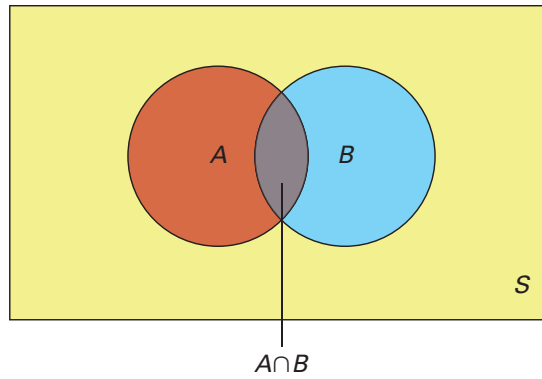
For any experiment, we can define events based on one or more outcomes of the experiment and also combine events to form new events. The **union** of two events, denoted  $A \cup B$ , is the event consisting of all outcomes in A or B. A useful way to illustrate these concepts is through the use of a Venn diagram, named after the British mathematician John Venn (1834–1923). Figure 4.1 shows a Venn diagram where the rectangle represents the sample space S and the two circles represent events A and B. The union  $A \cup B$  is the portion in the Venn diagram that is included in either A or B.



**FIGURE 4.1**  
The union of two events,  $A \cup B$

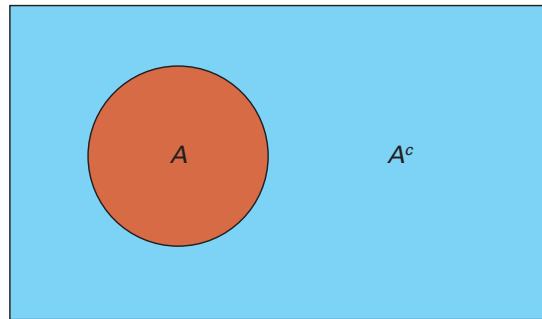
The **intersection** of two events, denoted  $A \cap B$ , is the event consisting of all outcomes in A and B. Figure 4.2 depicts the intersection of two events A and B. The intersection  $A \cap B$  is the portion in the Venn diagram that is included in both A and B.

**FIGURE 4.2**  
The intersection of two  
events,  $A \cap B$



The **complement** of event  $A$ , denoted  $A^c$ , is the event consisting of all outcomes in the sample space  $S$  that are not in  $A$ . In Figure 4.3,  $A^c$  is everything in  $S$  that is not included in  $A$ .

**FIGURE 4.3**  
The complement  
of an event,  $A^c$



#### COMBINING EVENTS

- The **union** of two events, denoted  $A \cup B$ , is the event consisting of all outcomes in  $A$  or  $B$ .
- The **intersection** of two events, denoted  $A \cap B$ , is the event consisting of all outcomes in  $A$  and  $B$ .
- The **complement** of event  $A$ , denoted  $A^c$ , is the event consisting of all outcomes in the sample space  $S$  that are not in  $A$ .

#### EXAMPLE 4.2

Recall that the snowboarder's sample space from Example 4.1 is defined as  $S = \{\text{gold, silver, bronze, no medal}\}$ . Now suppose the snowboarder defines the following three events:

- $A = \{\text{gold, silver, bronze}\}$ ; that is, event  $A$  denotes earning a medal;
  - $B = \{\text{silver, bronze, no medal}\}$ ; that is, event  $B$  denotes earning at most a silver medal; and
  - $C = \{\text{no medal}\}$ ; that is, event  $C$  denotes failing to earn a medal.
- Find  $A \cup B$  and  $B \cup C$ .
  - Find  $A \cap B$  and  $A \cap C$ .
  - Find  $B^c$ .

#### SOLUTION:

- The union of  $A$  and  $B$  denotes all outcomes common to  $A$  or  $B$ ; here, the event  $A \cup B = \{\text{gold, silver, bronze, no medal}\}$ . Note that there is no double counting of the outcomes "silver" or "bronze" in  $A \cup B$ . Similarly, we have the event  $B \cup C = \{\text{silver, bronze, no medal}\}$ .

- b. The intersection of  $A$  and  $B$  denotes all outcomes common to  $A$  and  $B$ ; here, the event  $A \cap B = \{\text{silver, bronze}\}$ . The event  $A \cap C = \emptyset$ , where  $\emptyset$  denotes the null (empty) set; no common outcomes appear in both  $A$  and  $C$ .
- c. The complement of  $B$  denotes all outcomes in  $S$  that are not in  $B$ ; here, the event  $B^c = \{\text{gold}\}$ .

## Assigning Probabilities

Now that we have described a valid sample space and the various ways in which we can define events from that sample space, we are ready to assign probabilities. When we arrive at a probability, we generally are able to categorize the probability as a *subjective probability*, an *empirical probability*, or a *classical probability*. Regardless of the method used, there are two defining properties of probability.

### LO 4.2

Formulate and explain subjective, empirical, and classical probabilities.

#### THE TWO DEFINING PROPERTIES OF PROBABILITY

1. The probability of any event  $A$  is a value between 0 and 1; that is,  $0 \leq P(A) \leq 1$ .
2. The sum of the probabilities of any list of mutually exclusive and exhaustive events equals 1.

Suppose the snowboarder from Example 4.1 believes that there is a 10% chance that she will earn a gold medal, a 15% chance that she will earn a silver medal, a 20% chance that she will earn a bronze medal, and a 55% chance that she will fail to earn a medal. She has assigned a **subjective probability** to each of the simple events. She made a personal assessment of these probabilities without referencing any data.

The snowboarder believes that the most likely outcome is failing to earn a medal since she gives that outcome the greatest chance of occurring at 55%. When formally writing out the probability that an event occurs, we generally construct a probability statement. Here, the probability statement might take the form:  $P(\{\text{no medal}\}) = 0.55$ , where  $P(\text{"event"})$  represents the probability that a given event occurs. Table 4.2 summarizes these events and their respective subjective probabilities. Note that here the events are mutually exclusive and exhaustive.

**TABLE 4.2** Snowboarder's Subjective Probabilities

Event	Probability
Gold	0.10
Silver	0.15
Bronze	0.20
No medal	0.55

Reading from the table we can readily see, for instance, that she assesses that there is a 15% chance that she will earn a silver medal, or  $P(\{\text{silver}\}) = 0.15$ . We should note that all the probabilities are between the values of zero and one, and they add up to one, thus meeting the defining properties of probability.

Suppose the snowboarder wants to calculate the probability of earning a medal. In Example 4.2, we defined “earning a medal” as event  $A$ , so the probability statement takes the form  $P(A)$ . We calculate this probability by summing the probabilities of the outcomes in  $A$ , or equivalently,

$$P(A) = P(\{\text{gold}\}) + P(\{\text{silver}\}) + P(\{\text{bronze}\}) = 0.10 + 0.15 + 0.20 = 0.45.$$

### EXAMPLE 4.3

Given the events in Example 4.2 and the probabilities in Table 4.2, calculate the following probabilities.

- a.  $P(B \cup C)$
- b.  $P(A \cap C)$
- c.  $P(B^c)$

**SOLUTION:**

- a. The probability that event  $B$  or event  $C$  occurs is

$$P(B \cup C) = P(\{\text{silver}\}) + P(\{\text{bronze}\}) + P(\{\text{no medal}\}) \\ = 0.15 + 0.20 + 0.55 = 0.90.$$

- b. The probability that event  $A$  and event  $C$  occur is

$$P(A \cap C) = 0; \text{ recall that there are no common outcomes in } A \text{ and } C.$$

- c. The probability that the complement of  $B$  occurs is

$$P(B^c) = P(\{\text{gold}\}) = 0.10.$$

In many instances, we calculate probabilities by referencing data based on the observed outcomes of an experiment. The **empirical probability** of an event is the observed relative frequency with which an event occurs. The experiment must be repeated a large number of times for empirical probabilities to be accurate.

**EXAMPLE 4.4**

The frequency distribution in Table 4.3 summarizes the ages of the richest 400 Americans. Suppose we randomly select one of these individuals.

- What is the probability that the individual is at least 50 but less than 60 years old?
- What is the probability that the individual is younger than 60 years old?
- What is the probability that the individual is at least 80 years old?

**TABLE 4.3** Frequency Distribution of Ages of 400 Richest Americans

Ages	Frequency
30 up to 40	7
40 up to 50	47
50 up to 60	90
60 up to 70	109
70 up to 80	93
80 up to 90	45
90 up to 100	9

SOURCE: www.forbes.com.

**SOLUTION:** In Table 4.3a, we first label each outcome with letter notation; for instance, the outcome “30 up to 40” is denoted as event  $A$ . Next we calculate the relative frequency of each event and use the relative frequency to denote the probability of the event.

**TABLE 4.3a** Relative Frequency Distribution of Ages of 400 Richest Americans

Ages	Event	Frequency	Relative Frequency
30 up to 40	$A$	7	$7/400 = 0.0175$
40 up to 50	$B$	47	0.1175
50 up to 60	$C$	90	0.2250
60 up to 70	$D$	109	0.2725
70 up to 80	$E$	93	0.2325
80 up to 90	$F$	45	0.1125
90 up to 100	$G$	9	0.0225

- a. The probability that an individual is at least 50 but less than 60 years old is

$$P(C) = \frac{90}{400} = 0.225.$$

- b. The probability that an individual is younger than 60 years old is

$$P(A \cup B \cup C) = \frac{7 + 47 + 90}{400} = 0.360.$$

- c. The probability that an individual is at least 80 years old is

$$P(F \cup G) = \frac{45 + 9}{400} = 0.135.$$

In a more narrow range of well-defined problems, we can sometimes deduce probabilities by reasoning about the problem. The resulting probability is a **classical probability**. Classical probabilities are often used in games of chance. They are based on the assumption that all outcomes of an experiment are equally likely. Therefore, the classical probability of an event is computed as the number of outcomes belonging to the event divided by the total number of outcomes.

### EXAMPLE 4.5

Suppose our experiment consists of rolling a six-sided die. Then we can define the appropriate sample space as  $S = \{1, 2, 3, 4, 5, 6\}$ .

- a. What is the probability that we roll a 2?
- b. What is the probability that we roll a 2 or 5?
- c. What is the probability that we roll an even number?

**SOLUTION:** Here we recognize that each outcome is equally likely. So with 6 possible outcomes, each outcome has a  $1/6$  chance of occurring.

- a. The probability that we roll a 2,  $P(\{2\})$ , is thus  $1/6$ .
- b. The probability that we roll a 2 or 5,  $P(\{2\}) + P(\{5\})$ , is  $1/6 + 1/6 = 1/3$ .
- c. The probability that we roll an even number,  $P(\{2\}) + P(\{4\}) + P(\{6\})$ , is  $1/6 + 1/6 + 1/6 = 1/2$ .

### CATEGORIZING PROBABILITIES

- A **subjective probability** is calculated by drawing on personal and subjective judgment.
- An **empirical probability** is calculated as a relative frequency of occurrence.
- A **classical probability** is based on logical analysis rather than on observation or personal judgment.

Since empirical and classical probabilities generally do not vary from person to person, they are often grouped as **objective probabilities**.

According to a famous **law of large numbers**, the empirical probability approaches the classical probability if the experiment is run a very large number of times. Consider, for example, flipping a fair coin 10 times. It is possible that heads may not show up exactly 5 times and, therefore, the relative frequency may not be 0.5. However, if we flip the fair coin a very large number of times, heads will show up approximately  $1/2$  of the time.

## Probabilities Expressed as Odds

Even though we tend to report the probability of an event occurring as a number between 0 and 1, alternative approaches to expressing probabilities include percentages and odds. Specifically, in wagering it is common to state probabilities in terms of odds. For instance,



at the start of the 2008–2009 football season, the Pittsburgh Steelers were not one of the strong favorites to win the Super Bowl, with odds for winning of 1:24 (*Betfair* website). In other words, an individual who bet \$1 on the Steelers winning the Super Bowl prior to the season would have won \$24 in profits. Since the bettor also receives the original stake back, for every \$1 staked in the wager, he/she would have gotten back \$25. We can convert the odds ratio into a probability by using the following generalization.

#### CONVERTING AN ODDS RATIO TO A PROBABILITY

Given odds *for* event  $A$  occurring of “ $a$  to  $b$ ,” the probability of  $A$  is  $\frac{a}{a+b}$ .

Given odds *against* event  $A$  occurring of “ $a$  to  $b$ ,” the probability of  $A$  is  $\frac{b}{a+b}$ .

Thus, with odds for winning the Super Bowl of 1:24, we can solve for the probability of the Steelers winning as:  $1/(1 + 24) = 1/25$  or 0.04. Moreover, the bet’s anticipated profit is \$0 because  $(0.04 \text{ probability of winning}) \times (\$24 \text{ profit if the wager is won}) + (0.96 \text{ probability of losing}) \times (-\$1 \text{ if the wager is lost}) = 0.96 + (-0.96) = 0$ .

This is an example of an expected value calculation, which we discuss further in Chapter 5. We would also like to point out that sports betting odds are usually displayed in various formats, including American, British, or European formats; the details are beyond the scope of this chapter.

#### EXAMPLE 4.6

Days prior to the 2009 Super Bowl, the Pittsburgh Steelers’ odds for beating the Arizona Cardinals increased to approximately 2:1. What was the probability of the Steelers winning just prior to the Super Bowl?

**SOLUTION:** The probability that the Steelers would win the Super Bowl rose to

$$\frac{a}{a+b} = \frac{2}{2+1} = 0.67.$$

(Note: The Steelers did win the Super Bowl, but just barely, scoring the winning touchdown with 35 seconds left in the game.)

Similarly, we can convert a probability to an odds ratio using the following generalization:

#### CONVERTING A PROBABILITY TO AN ODDS RATIO

If  $P(A)$  denotes the probability of an event  $A$  occurring, and  $P(A)$  does not equal zero or one, then:

The odds *for*  $A$  occurring equal  $\frac{P(A)}{1-P(A)}$ , and

The odds *against*  $A$  occurring equal  $\frac{1-P(A)}{P(A)}$ .

#### EXAMPLE 4.7

The summer of 2008 proved to be another difficult period for travelers. New York’s Kennedy Airport topped the list with the lowest on-time arrival rate: the likelihood that a plane arrived on-time occurred only 56% of the time (*The Wall Street*



*Journal*, September 9, 2008). Travelers at Atlanta's Airport fared a bit better, where the on-time arrival rate was 74%.

- a. Calculate the odds for a plane arriving on-time at New York's Kennedy Airport.
- b. Calculate the odds for a plane arriving on-time at Atlanta's Airport.

**SOLUTION:**

- a. With an on-time arrival probability of 0.56 for New York's Kennedy Airport we find

$$\frac{P(\{\text{on-time}\})}{1 - P(\{\text{on-time}\})} = \frac{0.56}{1 - 0.56} = \frac{0.56}{0.44} = 1.27.$$

Therefore, we would report the odds for arriving on-time in New York as 1.27 to 1. Note that given the odds for arriving on-time as 1.27:1, we can deduce  $P(\{\text{on-time}\})$  as

$$\frac{1.27}{2.27} = 0.56.$$

- b. With an on-time arrival probability of 0.74 for Atlanta's Airport, we find

$$\frac{P(\{\text{on-time}\})}{1 - P(\{\text{on-time}\})} = \frac{0.74}{1 - 0.74} = \frac{0.74}{0.26} = 2.85.$$

Therefore, we report the odds for arriving on-time in Atlanta as 2.85 to 1.

## EXERCISES 4.1

### Mechanics

1. Determine whether the following probabilities are best categorized as subjective, empirical, or classical probabilities.
  - a. Before flipping a fair coin, Sunil assesses that he has a 50% chance of obtaining tails.
  - b. At the beginning of the semester, John believes he has a 90% chance of receiving straight A's.
  - c. A political reporter announces that there is a 40% chance that the next person to come out of the conference room will be a Republican, since there are 60 Republicans and 90 Democrats in the room.
2. Express each of the probabilities in the preceding question as
  - a. odds assessed by Sunil for obtaining tails.
  - b. odds assessed by John for receiving straight A's.
  - c. odds assessed by the reporter for a Republican coming out of the room.
3. A sample space  $S$  yields five equally likely events,  $A$ ,  $B$ ,  $C$ ,  $D$ , and  $E$ .
  - a. Find  $P(D)$ .
  - b. Find  $P(B^c)$ .
  - c. Find  $P(A \cup C \cup E)$ .
4. You roll a die with the sample space  $S = \{1, 2, 3, 4, 5, 6\}$ . You define  $A$  as  $\{1, 2, 3\}$ ,  $B$  as  $\{1, 2, 3, 5, 6\}$ ,  $C$  as  $\{4, 6\}$ , and  $D$  as  $\{4, 5, 6\}$ . Determine which of the following events are exhaustive and/or mutually exclusive.
  - a.  $A$  and  $B$
  - b.  $A$  and  $C$

c.  $A$  and  $D$

d.  $B$  and  $C$

5. A sample space,  $S$ , yields four simple events,  $A$ ,  $B$ ,  $C$ , and  $D$ , such that  $P(A) = 0.35$ ,  $P(B) = 0.10$ , and  $P(C) = 0.25$ .
  - a. Find  $P(D)$ .
  - b. Find  $P(C^c)$ .
  - c. Find  $P(A \cup B)$ .

### Applications

6. Jane Peterson has taken Amtrak to travel from New York to Washington, DC, on six occasions, of which three times the train was late. Therefore, Jane tells her friends that the probability that this train will arrive on time is 0.50. Would you label this probability as empirical or classical? Why would this probability not be accurate?
7. Survey data, based on 65,000 mobile phone subscribers, shows that 44% of the subscribers use smartphones (*Forbes*, December 15, 2011). Based on this information, you infer that the probability that a mobile phone subscriber uses a smartphone is 0.44. Would you consider this probability estimate accurate? Is it a subjective, empirical, or classical probability?
8. Consider the following scenarios to determine if the mentioned combination of attributes represents a union or an intersection.
  - a. A marketing firm is looking for a candidate with a business degree and at least five years of work experience.
  - b. A family has decided to purchase Toyota or Honda.

9. Consider the following scenarios to determine if the mentioned combination of attributes represents a union or an intersection.
  - a. There are two courses that seem interesting to you, and you would be happy if you can take at least one of them.
  - b. There are two courses that seem interesting to you, and you would be happy if you can take both of them.
10. You apply for a position at two firms. Let event  $A$  represent the outcome of getting an offer from the first firm and event  $B$  represent the outcome of getting an offer from the second firm.
  - a. Explain why events  $A$  and  $B$  are not exhaustive.
  - b. Explain why events  $A$  and  $B$  are not mutually exclusive.
11. An alarming number of U.S. adults are either overweight or obese. The distinction between overweight and obese is made on the basis of body mass index (BMI), expressed as  $\text{weight/height}^2$ . An adult is considered overweight if the BMI is 25 or more but less than 30. An obese adult will have a BMI of 30 or greater. According to a January 2012 article in the *Journal of the American Medical Association*, 33.1% of the adult population in the United States is overweight and 35.7% is obese. Use this information to answer the following questions.
  - a. What is the probability that a randomly selected adult is either overweight or obese?
  - b. What is the probability that a randomly selected adult is neither overweight nor obese?
  - c. Are the events “overweight” and “obese” exhaustive?
  - d. Are the events “overweight” and “obese” mutually exclusive?
12. Many communities are finding it more and more difficult to fill municipal positions such as town administrators, finance directors, and treasurers. The following table shows the percentage of municipal managers by age group in the United States for the years 1971 and 2006.

Age	1971	2006
Under 30	26%	1%
30 to 40	45%	12%
41 to 50	21%	28%
51 to 60	5%	48%
Over 60	3%	11%

SOURCE: *The International City-County Management Association.*

- a. In 1971, what was the probability that a municipal manager was 40 years old or younger? In 2006, what was the probability that a municipal manager was 40 years old or younger?
- b. In 1971, what was the probability that a municipal manager was 51 years old or older? In 2006, what

was the probability that a municipal manager was 51 years old or older?

- c. What trends in ages can you detect from municipal managers in 1971 versus municipal managers in 2006?
13. At four community health centers on Cape Cod, Massachusetts, 15,164 patients were asked to respond to questions designed to detect depression (*The Boston Globe*, June 11, 2008). The survey produced the following results.

Diagnosis	Number
Mild	3,257
Moderate	1,546
Moderately Severe	975
Severe	773
No Depression	8,613

- a. What is the probability that a randomly selected patient suffered from mild depression?
  - b. What is the probability that a randomly selected patient did not suffer from depression?
  - c. What is the probability that a randomly selected patient suffered from moderately severe to severe depression?
  - d. Given that the national figure for moderately severe to severe depression is approximately 6.7%, does it appear that there is a higher rate of depression in this summer resort community? Explain.
14. On Sunday, July 11, 2010, Spain and the Netherlands played in the 2010 World Cup Final in Johannesburg. On the eve of the final, many betting lines were offering Spain's odds for winning at 15:8 (*Oddschecker* website).
- a. Spain won the World Cup. Suppose you had bet \$1,000 on Spain. What was your net gain? If Spain had lost, what would have been your net loss?
  - b. What was the implied probability of Spain winning the final?
15. Prior to the Academy Awards ceremony in 2009, the United Kingdom bookmaker Ladbrokes reported the following odds for winning an Oscar in the category of best actress (*The Wall Street Journal*, February 20, 2009).

Best Actress	Movie	Odds
Anne Hathaway	Rachel Getting Married	2:11
Angelina Jolie	Changeling	1:20
Melissa Leo	Frozen River	1:33
Meryl Streep	Doubt	3:10
Kate Winslet	The Reader	5:2

- a. Express the odds for each actress winning as a probability.
- b. According to your calculations, which actress was most likely to win an Oscar? Kate Winslet won her first Oscar on February 22, 2009. Was your prediction realized?

## 4.2 RULES OF PROBABILITY

In the previous section, we discussed how the probability of an event is assigned. Here we present various rules used to combine probabilities of events.

### The Complement Rule

The complement rule follows from one of the defining properties of probability: The sum of probabilities assigned to simple events in a sample space must equal one. Note that since  $S$  is a collection of all possible outcomes of the experiment (nothing else can happen),  $P(S) = 1$ . Let's revisit the sample space that we constructed when we rolled a six-sided die:  $S = \{1, 2, 3, 4, 5, 6\}$ . Suppose event  $A$  is defined as an even-numbered outcome or  $A = \{2, 4, 6\}$ . We then know that the complement of  $A$ ,  $A^c$ , is the set consisting of  $\{1, 3, 5\}$ . Moreover, we can deduce that  $P(A) = 1/2$  and  $P(A^c) = 1/2$ , so  $P(A) + P(A^c) = 1$ . Rearranging this equation, we obtain the complement rule:  $P(A^c) = 1 - P(A)$ .

#### LO 4.3

Calculate and interpret the probability of the complement of an event.

#### THE COMPLEMENT RULE

The **complement rule** states that the probability of the complement of an event,  $P(A^c)$ , is equal to one minus the probability of the event; that is,  $P(A^c) = 1 - P(A)$ .

The complement rule is quite straightforward and rather simple, but it is widely used and powerful.

#### EXAMPLE 4.8

According to the 2010 U.S. Census, 37% of women ages 25 to 34 have earned at least a college degree as compared with 30% of men in the same age group.

- What is the probability that a randomly selected woman between the ages of 25 to 34 does not have a college degree?
- What is the probability that a randomly selected man between the ages of 25 to 34 does not have a college degree?

#### SOLUTION:

- Let's define  $A$  as the event that a randomly selected woman between the ages of 25 and 34 has a college degree; thus,  $P(A) = 0.37$ . In this problem, we are interested in the complement of  $A$ . So  $P(A^c) = 1 - P(A) = 1 - 0.37 = 0.63$ .
- Similarly, we define  $B$  as the event that a randomly selected man between the ages of 25 to 34 has a college degree, so  $P(B) = 0.30$ . Thus,  $P(B^c) = 1 - P(B) = 1 - 0.30 = 0.70$ .

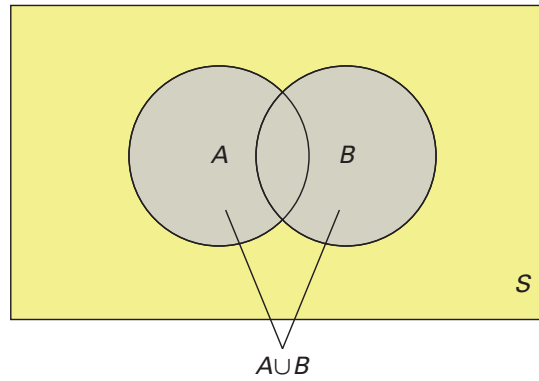
### The Addition Rule

The addition rule allows us to find the probability of the union of two events. Suppose we want to find the probability that either  $A$  occurs or  $B$  occurs, so in probability terms,  $P(A \cup B)$ . We reproduce the Venn diagram, used earlier in Figure 4.1, to help in exposition. Figure 4.4 shows a sample space  $S$  with the two events  $A$  and  $B$ . Recall that the union,  $A \cup B$ , is the portion in the Venn diagram that is included in either  $A$  or  $B$ . The intersection,  $A \cap B$ , is the portion in the Venn diagram that is included in both  $A$  and  $B$ .

#### LO 4.4

Calculate and interpret the probability that at least one of two events will occur.

**FIGURE 4.4**  
Finding the probability  
of the union of two  
events,  $P(A \cup B)$



If we try to obtain  $P(A \cup B)$  by simply summing  $P(A)$  with  $P(B)$ , then we overstate the probability because we double-count the probability of the intersection of A and B,  $P(A \cap B)$ . When implementing the addition rule, we sum  $P(A)$  and  $P(B)$  and then subtract  $P(A \cap B)$  from this sum.

#### THE ADDITION RULE

The **addition rule** states that the probability that A or B occurs, or that at least one of these events occurs, is equal to the probability that A occurs, plus the probability that B occurs, minus the probability that both A and B occur. Equivalently,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

#### EXAMPLE 4.9

Anthony feels that he has a 75% chance of getting an A in Statistics and a 55% chance of getting an A in Managerial Economics. He also believes he has a 40% chance of getting an A in both classes.

- What is the probability that he gets an A in at least one of these courses?
- What is the probability that he does not get an A in either of these courses?

#### SOLUTION:

- Let  $P(A_S)$  correspond to the probability of getting an A in Statistics and  $P(A_M)$  correspond to the probability of getting an A in Managerial Economics. Thus,  $P(A_S) = 0.75$  and  $P(A_M) = 0.55$ . In addition, there is a 40% chance that Anthony gets an A in both classes; that is,  $P(A_S \cap A_M) = 0.40$ . In order to find the probability that he receives an A in at least one of these courses, we calculate:

$$P(A_S \cup A_M) = P(A_S) + P(A_M) - P(A_S \cap A_M) = 0.75 + 0.55 - 0.40 = 0.90.$$

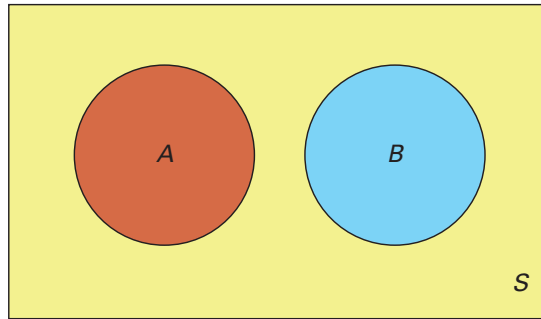
- The probability that he does not receive an A in either of these two courses is actually the complement of the union of the two events; that is,  $P((A_S \cup A_M)^c)$ . We calculated the union in part a, so using the complement rule we have

$$P((A_S \cup A_M)^c) = 1 - P(A_S \cup A_M) = 1 - 0.90 = 0.10.$$

An alternative expression that correctly captures the required probability is  $P((A_S \cup A_M)^c) = P(A_S^c \cap A_M^c)$ . A common mistake is to calculate the probability as  $P((A_S \cap A_M)^c) = 1 - P(A_S \cap A_M) = 1 - 0.40 = 0.60$ , which simply indicates that there is a 60% chance that Anthony will not get an A in both courses. This is clearly not the required probability that Anthony does not get an A in either course.

## The Addition Rule for Mutually Exclusive Events

As mentioned earlier, mutually exclusive events do not share any outcome of an experiment. Figure 4.5 shows the Venn diagram for two mutually exclusive events; note that the circles do not intersect.



**FIGURE 4.5**  
Mutually exclusive events

For mutually exclusive events  $A$  and  $B$ , the probability of their intersection is zero; that is,  $P(A \cap B) = 0$ . We need not concern ourselves with double-counting, and, therefore, the probability of the union is simply the sum of the two probabilities.

### THE ADDITION RULE FOR MUTUALLY EXCLUSIVE EVENTS

If  $A$  and  $B$  are mutually exclusive events, then  $P(A \cap B) = 0$  and, therefore, the addition rule simplifies to  $P(A \cup B) = P(A) + P(B)$ .

### EXAMPLE 4.10

Samantha Greene, a college senior, contemplates her future immediately after graduation. She thinks there is a 25% chance that she will join the Peace Corps and teach English in Madagascar for the next few years. Alternatively, she believes there is a 35% chance that she will enroll in a full-time law school program in the United States.

- a. What is the probability that she joins the Peace Corps or enrolls in law school?
- b. What is the probability that she does not choose either of these options?

#### SOLUTION:

- a. We can write the probability that Samantha joins the Peace Corps as  $P(A) = 0.25$  and the probability that she enrolls in law school as  $P(B) = 0.35$ . Immediately after college, Samantha cannot choose both of these options. This implies that these events are mutually exclusive, so  $P(A \cap B) = 0$ . Thus, when solving for the probability that Samantha joins the Peace Corps or enrolls in law school,  $P(A \cup B)$ , we can simply sum  $P(A)$  and  $P(B)$ :  $P(A \cup B) = P(A) + P(B) = 0.25 + 0.35 = 0.60$ .
- b. In order to find the probability that she does not choose either of these options, we need to recognize that this probability is the complement of the union of the two events; that is,  $P((A \cup B)^c)$ . Therefore, using the complement rule, we have

$$P((A \cup B)^c) = 1 - P(A \cup B) = 1 - 0.60 = 0.40.$$

## Conditional Probability

In business applications, the probability of interest is often a conditional probability. Examples include the probability that the housing market will improve conditional on the Federal Reserve taking remedial actions; the probability of making a six-figure salary

### LO 4.5

Calculate and interpret a conditional probability and apply the multiplication rule.

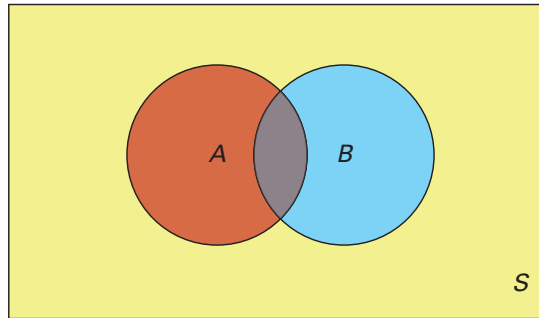
conditional on getting an MBA; the probability that a company's stock price will go up conditional on higher-than-expected profits; the probability that sales will improve conditional on the firm launching a new innovative product.

Let's use an example to illustrate the concept of conditional probability. Suppose the probability that a recent business college graduate finds a suitable job is 0.80. The probability of finding a suitable job is 0.90 if the recent business college graduate has prior work experience. This type of probability is called a **conditional probability**, where the probability of an event is conditional on the occurrence of another event. If  $A$  represents "finding a job" and  $B$  represents "prior work experience," then  $P(A) = 0.80$  and the conditional probability is denoted as  $P(A|B) = 0.90$ . The vertical mark  $|$  means "given that" and the conditional probability is typically read as "the probability of  $A$  given  $B$ ." In the above example, the probability of finding a suitable job increases from 0.80 to 0.90 when conditioned on prior work experience. In general, the conditional probability,  $P(A|B)$ , is greater than the **unconditional probability**,  $P(A)$ , if  $B$  exerts a positive influence on  $A$ . Similarly,  $P(A|B)$  is less than  $P(A)$  when  $B$  exerts a negative influence on  $A$ . Finally, if  $B$  exerts no influence on  $A$ , then  $P(A|B)$  equals  $P(A)$ . It is common to refer to "unconditional probability" simply as "probability."

As we will see later, it is important that we write the event that has already occurred after the vertical mark, since in most instances  $P(A|B) \neq P(B|A)$ . In the above example  $P(B|A)$  would represent the probability of prior work experience conditional on having found a job.

We again rely on the Venn diagram in Figure 4.6 to explain the conditional probability.

**FIGURE 4.6**  
Finding the conditional  
probability,  $P(A|B)$



Since  $P(A|B)$  represents the probability of  $A$  conditional on  $B$  ( $B$  has occurred), the original sample space  $S$  reduces to  $B$ . The conditional probability  $P(A|B)$  is based on the portion of  $A$  that is included in  $B$ . It is derived as the ratio of the probability of the intersection of  $A$  and  $B$  to the probability of  $B$ .

#### CONDITIONAL PROBABILITY

Given two events  $A$  and  $B$ , each with a positive probability of occurring, the probability that  $A$  occurs given that  $B$  has occurred ( $A$  conditioned on  $B$ ) is equal to  $P(A|B) = \frac{P(A \cap B)}{P(B)}$ . Similarly, the probability that  $B$  occurs given that  $A$  has occurred ( $B$  conditioned on  $A$ ) is equal to  $P(B|A) = \frac{P(A \cap B)}{P(A)}$ .

#### EXAMPLE 4.11

Economic globalization is defined as the integration of national economies into the international economy through trade, foreign direct investment, capital flows, migration, and the spread of technology. Although globalization is generally viewed favorably, it also increases the vulnerability of a country to economic conditions of the other country. An economist predicts a 60% chance that country A will perform



poorly and a 25% chance that country B will perform poorly. There is also a 16% chance that both countries will perform poorly.

- a. What is the probability that country A performs poorly given that country B performs poorly?
- b. What is the probability that country B performs poorly given that country A performs poorly?
- c. Interpret your findings.

**SOLUTION:** We first write down the available information in probability terms. Defining  $A$  as “country A performing poorly” and  $B$  as “country B performing poorly,” we have the following information:  $P(A) = 0.60$ ,  $P(B) = 0.25$ , and  $P(A \cap B) = 0.16$ .

a.  $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.16}{0.25} = 0.64$

b.  $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.16}{0.60} = 0.27$

- c. It appears that globalization has definitely made these countries vulnerable to the economic woes of the other country. The probability that country A performs poorly increases from 60% to 64% when country B has performed poorly. Similarly, the probability that country B performs poorly increases from 25% to 27% when conditioned on country A performing poorly.

## Independent and Dependent Events

Of particular interest to researchers is whether or not two events influence one another. Two events are **independent** if the occurrence of one event does not affect the probability of the occurrence of the other event. Let’s revisit the earlier example where the probability of finding a job is 0.80 and the probability of finding a job given prior work experience is 0.90. Prior work experience exerts a positive influence on finding a job because the conditional probability,  $P(A|B) = 0.90$ , exceeds the probability,  $P(A) = 0.80$ . Now consider the probability of finding a job given that your neighbor has bought a red car. Obviously, your neighbor’s decision to buy a red car has no influence on your probability of finding a job, which remains at 0.80.

Events are considered **dependent** if the occurrence of one is related to the probability of the occurrence of the other. We generally test for the independence of two events by comparing the conditional probability of one event, for instance  $P(A|B)$ , to the probability,  $P(A)$ . If these two probabilities are the same, we say that the two events,  $A$  and  $B$ , are independent; if the probabilities differ, the two events are dependent.

### LO 4.6

Distinguish between independent and dependent events.

#### INDEPENDENT VERSUS DEPENDENT EVENTS

Two events,  $A$  and  $B$ , are **independent** if  $P(A|B) = P(A)$  or, equivalently,  $P(B|A) = P(B)$ . Otherwise, the events are **dependent**.

#### EXAMPLE 4.12

Suppose that for a given year there is a 2% chance that your desktop computer will crash and a 6% chance that your laptop computer will crash. Moreover, there is a 0.12% chance that both computers will crash. Is the reliability of the two computers independent of each other?

**SOLUTION:** Let event  $D$  represent the outcome that your desktop crashes and event  $L$  represent the outcome that your laptop crashes. Therefore,  $P(D) = 0.02$ ,  $P(L) = 0.06$ , and  $P(D \cap L) = 0.0012$ . The reliability of the two computers is independent because

$$P(D|L) = \frac{P(D \cap L)}{P(L)} = \frac{0.0012}{0.06} = 0.02 = P(D).$$

In other words, if your laptop crashes, it does not alter the probability that your desktop also crashes. Equivalently,

$$P(L|D) = \frac{P(D \cap L)}{P(D)} = \frac{0.0012}{0.02} = 0.06 = P(L).$$

## The Multiplication Rule

In some situations, we are interested in finding the probability that two events,  $A$  and  $B$ , both occur; that is,  $P(A \cap B)$ . In order to obtain this probability, we can rearrange the formula for conditional probability to derive  $P(A \cap B)$ . For instance, from  $P(A|B) = \frac{P(A \cap B)}{P(B)}$ , we can easily derive  $P(A \cap B) = P(A|B)P(B)$ . Similarly, from  $P(B|A) = \frac{P(A \cap B)}{P(A)}$ , we derive  $P(A \cap B) = P(B|A)P(A)$ . Since we calculate the product of two probabilities to find  $P(A \cap B)$ , we refer to it as the **multiplication rule** for probabilities.

### THE MULTIPLICATION RULE

The **multiplication rule** states that the probability that  $A$  and  $B$  both occur is equal to the probability that  $A$  occurs given that  $B$  has occurred times the probability that  $B$  occurs; that is,  $P(A \cap B) = P(A|B)P(B)$ . Equivalently, we can also arrive at this probability as  $P(A \cap B) = P(B|A)P(A)$ .

### EXAMPLE 4.13

A stockbroker knows from past experience that the probability that a client owns stocks is 0.60 and the probability that a client owns bonds is 0.50. The probability that the client owns bonds if he/she already owns stocks is 0.55.

- What is the probability that the client owns both of these securities?
- Given that the client owns bonds, what is the probability that the client owns stocks?

#### SOLUTION:

- Let  $S$  correspond to the event that a client owns stocks and  $B$  correspond to the event that a client owns bonds. Thus, the probabilities that the client owns stocks and that the client owns bonds are  $P(S) = 0.60$  and  $P(B) = 0.50$ , respectively. The conditional probability that the client owns bonds given that he/she owns stocks is  $P(B|S) = 0.55$ . We calculate the probability that the client owns both of these securities as  $P(S \cap B) = P(B|S)P(S) = 0.55 \times 0.60 = 0.33$ .
- We need to calculate the conditional probability that the client owns stocks given that he/she owns bonds, or  $P(S|B)$ . Using the formula for conditional probability and the answer from part a, we find  $P(S|B) = \frac{P(S \cap B)}{P(B)} = \frac{0.33}{0.50} = 0.66$ .

## The Multiplication Rule for Independent Events

We know that two events,  $A$  and  $B$ , are independent if  $P(A|B) = P(A)$ . With independent events, the multiplication rule  $P(A \cap B) = P(A|B)P(B)$  simplifies to  $P(A \cap B) = P(A)P(B)$ . We can also use this rule to determine whether or not two

events are independent. That is, two events are independent if the probability  $P(A \cap B)$  equals the product of their probabilities,  $P(A)P(B)$ . In Example 4.12, we were given the probabilities  $P(D) = 0.02$ ,  $P(L) = 0.06$ , and  $P(D \cap L) = 0.0012$ . Consistent with the earlier result, events  $D$  and  $L$  are independent because  $P(D \cap L) = 0.0012$  equals  $P(D)P(L) = 0.02 \times 0.06 = 0.0012$ .

#### THE MULTIPLICATION RULE FOR INDEPENDENT EVENTS

If  $A$  and  $B$  are **independent events**, then the probability that  $A$  and  $B$  both occur equals the product of the probability of  $A$  and the probability of  $B$ ; that is,  $P(A \cap B) = P(A)P(B)$ .

#### EXAMPLE 4.14

The probability of passing the Level 1 CFA (Chartered Financial Analyst) exam is 0.50 for John Campbell and 0.80 for Linda Lee. The prospect of John's passing the exam is completely unrelated to Linda's success on the exam.

- What is the probability that both John and Linda pass the exam?
- What is the probability that at least one of them passes the exam?

**SOLUTION:** We can write the probabilities that John passes the exam and that Linda passes the exam as  $P(J) = 0.50$  and  $P(L) = 0.80$ , respectively.

- Since we are told that John's chances of passing the exam are not influenced by Linda's success at the exam, we can conclude that these events are independent, so  $P(J) = P(J|L) = 0.50$  and  $P(L) = P(L|J) = 0.80$ . Thus, when solving for the probability that both John and Linda pass the exam, we calculate the product of the probabilities, so  $P(J \cap L) = P(J) \times P(L) = 0.50 \times 0.80 = 0.40$ .
- We calculate the probability that at least one of them passes the exam as:  $P(J \cup L) = P(J) + P(L) - P(J \cap L) = 0.50 + 0.80 - 0.40 = 0.90$ .

## EXERCISES 4.2

### Mechanics

- Let  $P(A) = 0.65$ ,  $P(B) = 0.30$ , and  $P(A|B) = 0.45$ .
  - Calculate  $P(A \cap B)$ .
  - Calculate  $P(A \cup B)$ .
  - Calculate  $P(B|A)$ .
- Let  $P(A) = 0.55$ ,  $P(B) = 0.30$ , and  $P(A \cap B) = 0.10$ .
  - Calculate  $P(A|B)$ .
  - Calculate  $P(A \cup B)$ .
  - Calculate  $P((A \cup B)^c)$ .
- Let  $A$  and  $B$  be mutually exclusive events with  $P(A) = 0.25$  and  $P(B) = 0.30$ .
  - Calculate  $P(A \cap B)$ .
  - Calculate  $P(A \cup B)$ .
  - Calculate  $P(A|B)$ .
- Let  $A$  and  $B$  be independent events with  $P(A) = 0.40$  and  $P(B) = 0.50$ .
  - Calculate  $P(A \cap B)$ .
  - Calculate  $P((A \cup B)^c)$ .
  - Calculate  $P(A|B)$ .
- Let  $P(A) = 0.65$ ,  $P(B) = 0.30$ , and  $P(A|B) = 0.45$ .
  - Are  $A$  and  $B$  independent events? Explain.
  - Are  $A$  and  $B$  mutually exclusive events? Explain.
  - What is the probability that neither  $A$  nor  $B$  takes place?
- Let  $P(A) = 0.15$ ,  $P(B) = 0.10$ , and  $P(A \cap B) = 0.05$ .
  - Are  $A$  and  $B$  independent events? Explain.
  - Are  $A$  and  $B$  mutually exclusive events? Explain.
  - What is the probability that neither  $A$  nor  $B$  takes place?

22. Consider the following probabilities:  $P(A) = 0.25$ ,  $P(B^c) = 0.40$ , and  $P(A \cap B) = 0.08$ . Find:
  - a.  $P(B)$
  - b.  $P(A|B)$
  - c.  $P(B|A)$
23. Consider the following probabilities:  $P(A^c) = 0.30$ ,  $P(B) = 0.60$ , and  $P(A \cap B^c) = 0.24$ . Find:
  - a.  $P(A|B^c)$
  - b.  $P(B^c|A)$
  - c. Are  $A$  and  $B$  independent events? Explain.
24. Consider the following probabilities:  $P(A) = 0.40$ ,  $P(B) = 0.50$ , and  $P(A^c \cap B^c) = 0.24$ . Find:
  - a.  $P(A^c|B^c)$
  - b.  $P(A^c \cup B^c)$
  - c.  $P(A \cup B)$

## Applications

25. Survey data, based on 65,000 mobile phone subscribers, shows that 44% of the subscribers use smartphones (*Forbes*, December 15, 2011). Moreover, 51% of smartphone users are women.
  - a. Find the probability that a mobile phone subscriber is a woman who uses a smartphone.
  - b. Find the probability that a mobile phone subscriber is a man who uses a smartphone.
26. Only 20% of students in a college ever go to their professor during office hours. Of those who go, 30% seek minor clarification and 70% seek major clarification.
  - a. What is the probability that a student goes to the professor during her office hours for a minor clarification?
  - b. What is the probability that a student goes to the professor during her office hours for a major clarification?
27. The probabilities that stock  $A$  will rise in price is 0.40 and that stock  $B$  will rise in price is 0.60. Further, if stock  $B$  rises in price, the probability that stock  $A$  will also rise in price is 0.80.
  - a. What is the probability that at least one of the stocks will rise in price?
  - b. Are events  $A$  and  $B$  mutually exclusive? Explain.
  - c. Are events  $A$  and  $B$  independent? Explain.
28. Despite government bailouts and stimulus money, unemployment in the United States had not decreased significantly as economists had expected (*U.S. News & World Report*, July 2, 2010). Many analysts predicted only an 18% chance of a reduction in U.S. unemployment. However, if Europe slipped back into a recession, the probability of a reduction in U.S. unemployment would drop to 0.06.
  - a. What is the probability that there is not a reduction in U.S. unemployment?
  - b. Assume there is an 8% chance that Europe slips back into a recession. What is the probability that there is not a reduction in U.S. unemployment and that Europe slips into a recession?
29. Dr. Miriam Johnson has been teaching accounting for over 20 years. From her experience, she knows that 60% of her students do homework regularly. Moreover, 95% of the students who do their homework regularly generally pass the course. She also knows that 85% of her students pass the course.
  - a. What is the probability that a student will do homework regularly and also pass the course?
  - b. What is the probability that a student will neither do homework regularly nor will pass the course?
  - c. Are the events “pass the course” and “do homework regularly” mutually exclusive? Explain.
  - d. Are the events “pass the course” and “do homework regularly” independent? Explain.
30. Records show that 5% of all college students are foreign students who also smoke. It is also known that 50% of all foreign college students smoke. What percent of the students at this university are foreign?
31. An analyst estimates that the probability of default on a seven-year AA-rated bond is 0.06, while that on a seven-year A-rated bond is 0.13. The probability that they will both default is 0.04.
  - a. What is the probability that at least one of the bonds defaults?
  - b. What is the probability that neither the seven-year AA-rated bond nor the seven-year A-rated bond defaults?
  - c. Given that the seven-year AA-rated bond defaults, what is the probability that the seven-year A-rated bond also defaults?
32. Mike Danes has been delayed in going to the annual sales event at one of his favorite apparel stores. His friend has just texted him that there are only 20 shirts left, of which 8 are in size M, 10 in size L, and 2 in size XL. Also 3 of the shirts are white, 5 are blue, and the remaining are of mixed colors. Mike is interested in getting a white or a blue shirt in size L. Define the events  $A$  = Getting a white or a blue shirt and  $B$  = Getting a shirt in size L.
  - a. Find  $P(A)$ ,  $P(A^c)$ , and  $P(B)$ .
  - b. Are the events  $A$  and  $B$  mutually exclusive and exhaustive? Explain.
  - c. Would you describe Mike’s preference by the events  $A \cup B$  or  $A \cap B$ ?
33. In general, shopping online is supposed to be more convenient than going to stores. However, according to a recent Harris Interactive poll, 87% of people have experienced problems with an online transaction (*The Wall Street Journal*, October 2, 2007). Forty-two percent

of people who experienced a problem abandoned the transaction or switched to a competitor's website. Fifty-three percent of people who experienced problems contacted customer-service representatives.

- a. What percentage of people did not experience problems with an online transaction?
  - b. What percentage of people experienced problems with an online transaction and abandoned the transaction or switched to a competitor's website?
  - c. What percentage of people experienced problems with an online transaction and contacted customer-service representatives?
34. A manufacturing firm just received a shipment of 20 assembly parts, of slightly varied sizes, from a vendor. The manager knows that there are only 15 parts in the shipment that would be suitable. He examines these parts one at a time.
- a. Find the probability that the first part is suitable.
  - b. If the first part is suitable, find the probability that the second part is also suitable.
  - c. If the first part is suitable, find the probability that the second part is not suitable.
35. Apple products have become a household name in America with 51% of all households owning at least one Apple product (*CNN*, March 19, 2012). In the Midwest, the likelihood of owning an Apple product is 61% for households with kids and 48% for households without kids. Suppose there are 1,200 households in a representative community of which 820 are with kids and the rest are without kids.
- a. Are the events "household with kids" and "household without kids" mutually exclusive and exhaustive? Explain.
  - b. What is the probability that a household is without kids?
  - c. What is the probability that a household is with kids and owns an Apple product?
  - d. What is the probability that a household is without kids and does not own an Apple product?
36. Despite the repeated effort by the government to reform how Wall Street pays its executives, some of the nation's biggest banks are continuing to pay out bonuses nearly as large as those in the best years before the crisis (*The Washington Post*, January 15, 2010). It is known that 10 out of 15 members of the board of directors of a company were in favor of the bonus. Suppose two members were randomly selected by the media.
- a. What is the probability that both of them were in favor of the bonus?
  - b. What is the probability that neither of them was in favor of the bonus?
37. Christine Wong has asked Dave and Mike to help her move into a new apartment on Sunday morning. She has asked them both in case one of them does not show up. From past experience, Christine knows that there is a 40% chance that Dave will not show up and a 30% chance that Mike will not show up. Dave and Mike do not know each other and their decisions can be assumed to be independent.
- a. What is the probability that both Dave and Mike will show up?
  - b. What is the probability that at least one of them will show up?
  - c. What is the probability that neither Dave nor Mike will show up?
38. According to the Census's Population Survey, the percentage of children with two parents at home is the highest for Asians and lowest for blacks (*USA TODAY*, February 26, 2009). It is reported that 85% of Asian, 78% of white, 70% of Hispanic, and 38% of black children have two parents at home. Suppose there are 500 students in a representative school of which 280 are white, 50 are Asian, 100 are Hispanic, and 70 are black.
- a. Are the events "Asians" and "black" mutually exclusive and exhaustive? Explain.
  - b. What is the probability that a given child is not white?
  - c. What is the probability that a child is white and has both parents at home?
  - d. What is the probability that a child is Asian and does not have both parents at home?
39. Since the fall of 2008, millions of Americans have lost jobs due to the economic meltdown. A recent study shows that unemployment has not impacted white-collar and blue-collar workers equally (*Newsweek*, April 20, 2009). According to the Bureau of Labor Statistics report, while the national unemployment rate is 8.5%, it is only 4.3% for those with a college degree. It is fair to assume that 27% of people in the labor force are college educated. You have just heard that another worker in a large firm has been laid off. What is the probability that the worker is college educated?
40. According to a recent survey by two United Nations agencies and a nongovernmental organization, two in every three women in the Indian capital of New Delhi are likely to face some form of sexual harassment in a year (*BBC World News*, July 9, 2010). The study also reports that women who use public transportation are especially vulnerable. Suppose the corresponding probability of harassment for women who use public transportation is 0.82. It is also known that 28% of women use public transportation.
- a. What is the probability that a woman takes public transportation and also faces sexual harassment?
  - b. If a woman is sexually harassed, what is the probability that she had taken public transportation?
41. According to results from the Spine Patient Outcomes Research Trial, or SPORT, surgery for a painful, common back condition resulted in significantly reduced back pain

and better physical function than treatment with drugs and physical therapy (*The Wall Street Journal*, February 21, 2008). SPORT followed 803 patients, of whom 398 ended up getting surgery. After two years, of those who had surgery, 63% said they had a major improvement in their condition, compared with 29% among those who received nonsurgical treatment.

- a. What is the probability that a patient had surgery? What is the probability that a patient did not have surgery?
- b. What is the probability that a patient had surgery and experienced a major improvement in his or her condition?

- c. What is the probability that a patient received nonsurgical treatment and experienced a major improvement in his or her condition?

42. A recent study challenges the media narrative that foreclosures are dangerously widespread (*The New York Times*, March 2, 2009). According to this study, 62% of all foreclosures were centered in only four states, namely, Arizona, California, Florida, and Nevada. The national average rate of foreclosures in 2008 was 0.79%. What percent of the homes in the United States were foreclosed in 2008 and also centered in Arizona, California, Florida, or Nevada?

## LO 4.7

Calculate and interpret probabilities from a contingency table.

## 4.3 CONTINGENCY TABLES AND PROBABILITIES

We learned in Chapter 2 that, when organizing qualitative data, it is often useful to construct a frequency distribution. A frequency distribution is a useful tool when we want to sort one variable at a time. However, in many instances we want to examine or compare two qualitative variables. On these occasions, a **contingency table** proves very useful. Contingency tables are widely used in marketing and biomedical research, as well as in the social sciences.

### A CONTINGENCY TABLE

A **contingency table** generally shows frequencies for two qualitative (categorical) variables,  $x$  and  $y$ , where each cell represents a mutually exclusive combination of the pair of  $x$  and  $y$  values.

Table 4.4, first presented in the introductory case study of this chapter, is an example of a contingency table where the qualitative variables of interest,  $x$  and  $y$ , are Age Group and Brand Name, respectively. Age Group has two possible categories: (1) under 35 years and (2) 35 years and older; Brand Name, has three possible categories: (1) Under Armour, (2) Nike, and (3) Adidas.

**TABLE 4.4** Purchases of Compression Garments Based on Age and Brand Name

Age Group	Brand Name		
	Under Armour	Nike	Adidas
Under 35 years	174	132	90
35 years and older	54	72	78

Each cell in Table 4.4 represents a frequency; for example, there are 174 customers under the age of 35 who purchase an Under Armour product, whereas there are 54 customers at least 35 years old who purchase an Under Armour product. Recall that we estimate an empirical probability by calculating the relative frequency of the occurrence of the event. To make calculating these probabilities less cumbersome, it is often useful to denote each event with letter notation and calculate totals for each column and row as shown in Table 4.4a.

**TABLE 4.4a** A Contingency Table Labeled Using Event Notation

Age Group	Brand Name			Total
	$B_1$	$B_2$	$B_3$	
$A$	174	132	90	396
$A^c$	54	72	78	204
<b>Total</b>	228	204	168	600



Thus, let events  $A$  and  $A^c$  correspond to “under 35 years” and “35 years and older,” respectively; similarly, let events  $B_1$ ,  $B_2$ , and  $B_3$  correspond to “Under Armour,” “Nike,” and “Adidas,” respectively. In addition, after calculating row totals, it is now easier to recognize that 396 of the customers are under 35 years old and 204 of the customers are at least 35 years old. Similarly, column totals indicate that 228 customers purchase Under Armour, 204 purchase Nike, and 168 purchase Adidas. Finally, the frequency corresponding to the cell in the last column and the last row is 600. This value represents the sample size; that is, the total number of customers in the sample. We arrive at this value by either summing the values in the last column ( $396 + 204$ ) or summing the values in the last row ( $228 + 204 + 168$ ).

The following example illustrates how to calculate probabilities when the data are presented in the form of a contingency table.

### EXAMPLE 4.15

Using the information in Table 4.4a, answer the following questions.

- What is the probability that a randomly selected customer is younger than 35 years old?
- What is the probability that a randomly selected customer purchases an Under Armour garment?
- What is the probability that a customer is younger than 35 years old and purchases an Under Armour garment?
- What is the probability that a customer is either younger than 35 years old or purchases an Under Armour garment?
- What is the probability that a customer is under 35 years of age, given that the customer purchases an Under Armour garment?

#### SOLUTION:

- $P(A) = \frac{396}{600} = 0.66$ ; there is a 66% chance that a randomly selected customer is less than 35 years old.
- $P(B_1) = \frac{228}{600} = 0.38$ ; there is a 38% chance that a randomly selected customer purchases an Under Armour garment.
- $P(A \cap B_1) = \frac{174}{600} = 0.29$ ; there is a 29% chance that a randomly selected customer is younger than 35 years old and purchases an Under Armour garment.
- $P(A \cup B_1) = \frac{174 + 132 + 90 + 54}{600} = \frac{450}{600} = 0.75$ ; there is a 75% chance that a randomly selected customer is either younger than 35 years old or purchases an Under Armour garment. Alternatively, we can use the addition rule to solve this problem as  $P(A \cup B_1) = P(A) + P(B_1) - P(A \cap B_1) = 0.66 + 0.38 - 0.29 = 0.75$ .
- We wish to calculate the conditional probability,  $P(A|B_1)$ . When the information is in the form of a contingency table, calculating a conditional probability is rather straightforward. We are given the information that the customer purchases an Under Armour garment, so the sample space shrinks from 600 customers to 228 customers. We can ignore all customers that make Nike or Adidas purchases, or all outcomes in events  $B_2$  and  $B_3$ . Thus, of the 228 customers who make an Under Armour purchase, 174 of them are under 35 years of age. Therefore, the probability that a customer is under 35 years of age given that the customer makes an Under Armour purchase is calculated as  $P(A|B_1) = \frac{174}{228} = 0.76$ . Alternatively, we can use the conditional probability formula to solve the problem as  $P(A|B_1) = \frac{P(A \cap B_1)}{P(B_1)} = \frac{174/600}{228/600} = \frac{174}{228} = 0.76$ .

Arguably, a more convenient way of calculating relevant probabilities is to convert the contingency table to a **joint probability table**. The frequency in each cell is divided by the number of outcomes in the sample space, which in this example is 600 customers. Table 4.4b shows the results.

**TABLE 4.4b** Converting a Contingency Table to a Joint Probability Table

Age Group	Brand Name			Total
	$B_1$	$B_2$	$B_3$	
$A$	0.29	0.22	0.15	0.66
$A^c$	0.09	0.12	0.13	0.34
<b>Total</b>	0.38	0.34	0.28	1.00

The values in the interior of the table represent the probabilities of the intersection of two events, also referred to as **joint probabilities**. For instance, the probability that a randomly selected person is under 35 years of age and makes an Under Armour purchase, denoted  $P(A \cap B_1)$ , is 0.29. Similarly, we can readily read from this table that 12% of the customers purchase a Nike garment and are at least 35 years old, or  $P(A^c \cap B_2) = 0.12$ .

The values in the margins of Table 4.4b represent unconditional probabilities. These probabilities are also referred to as **marginal probabilities**. For example, the probability that a randomly selected customer is under 35 years of age,  $P(A)$ , is simply 0.66. Also, the probability of purchasing a Nike garment,  $P(B_2)$ , is 0.34.

Note that the conditional probability is basically the ratio of a joint probability to an unconditional probability. Since  $P(A|B_1) = \frac{P(A \cap B_1)}{P(B_1)}$ , the numerator is the joint probability,  $P(A \cap B_1)$ , and the denominator is the unconditional probability,  $P(B_1)$ . Let's refer back to the probability that we calculated earlier; that is, the probability that a customer is under 35 years of age, given that the customer purchases an Under Armour product. This

conditional probability is easily computed as  $P(A|B_1) = \frac{P(A \cap B_1)}{P(B_1)} = \frac{0.29}{0.38} = 0.76$ .

### EXAMPLE 4.16

Given the information in Table 4.4b, what is the probability that a customer purchases an Under Armour product, given that the customer is under 35 years of age?

**SOLUTION:** Now we are solving for  $P(B_1|A)$ . So

$$P(B_1|A) = \frac{P(A \cap B_1)}{P(A)} = \frac{0.29}{0.66} = 0.44.$$

Note that  $P(B_1|A) = 0.44 \neq P(A|B_1) = 0.76$ .

### EXAMPLE 4.17

Determine whether the events “under 35 years old” and “Under Armour” are independent.

**SOLUTION:** In order to determine whether two events are independent, we compare an event's conditional probability to its unconditional probability; that is, events  $A$  and  $B$  are independent if  $P(A|B) = P(A)$ . In the Under Armour example, we have already found that  $P(A|B_1) = 0.76$ . In other words, there is a 76% chance that a customer is under 35 years old given that the customer purchases an Under Armour product. We compare this conditional probability to its unconditional probability,  $P(A) = 0.66$ . Since these probabilities differ, the events “under 35 years old” and

“Under Armour” are not independent events. We could have compared  $P(B_1|A)$  to  $P(B_1)$  and found that  $0.44 \neq 0.38$ , which leads us to the same conclusion that the events are dependent. As discussed in the preceding section, an alternative approach is to compare the joint probability with the product of the two unconditional probabilities. Events are independent if  $P(A \cap B_1) = P(A)P(B_1)$ . In this example,  $P(A \cap B_1) = 0.29$  does not equal  $P(A)P(B_1) = 0.66 \times 0.38 = 0.25$ , so the two events are not independent.

It is important to note that the conclusions about independence, such as the one made in Example 4.17, are informal since they are based on empirical probabilities computed from given sample information. In the preceding example, these probabilities will change if a different sample of 600 customers is used. Formal tests of independence are discussed in Chapter 12.

## SYNOPSIS OF INTRODUCTORY CASE

After careful analysis of the contingency table representing customer purchases of compression garments based on age and brand name, several interesting remarks can be made. From a sample of 600 customers, it appears that the majority of the customers who purchase these products tend to be younger: 66% of the customers were younger than 35 years old, whereas 34% were at least 35 years old. It is true that more customers chose to purchase Under Armour garments (with 38% of purchases) as compared to Nike or Adidas garments (with 34% and 28% of purchases, respectively). However, given that Under Armour was the pioneer



in the compression-gear market, this company should be concerned with the competition posed by Nike and Adidas. Further inspection of the contingency table reveals that if a customer was under 35 years old, the chances of the customer purchasing an Under Armour garment rises to about 44%. This result indicates that the age of a customer seems to influence the brand name purchased. In other words, 38% of the customers choose to buy Under Armour products, but as soon as the attention is confined to those customers who are under 35 years old, the likelihood of a purchase from Under Armour rises to about 44%. The information that the Under Armour brand appeals to younger customers is relevant not only to Under Armour and how the firm may focus its advertising efforts, but also to competitors and retailers in the compression garment market.

## EXERCISES 4.3

### Mechanics

43. Consider the following contingency table.

	$B$	$B^c$
$A$	26	34
$A^c$	14	26

- Convert the contingency table into a joint probability table.
- What is the probability that  $A$  occurs?
- What is the probability that  $A$  and  $B$  occur?
- Given that  $B$  has occurred, what is the probability that  $A$  occurs?

- Given that  $A^c$  has occurred, what is the probability that  $B$  occurs?
- Are  $A$  and  $B$  mutually exclusive events? Explain.
- Are  $A$  and  $B$  independent events? Explain.

44. Consider the following joint probability table.

	$B_1$	$B_2$	$B_3$	$B_4$
$A$	0.09	0.22	0.15	0.20
$A^c$	0.03	0.10	0.09	0.12

- What is the probability that  $A$  occurs?
- What is the probability that  $B_2$  occurs?
- What is the probability that  $A^c$  and  $B_4$  occur?

- d. What is the probability that  $A$  or  $B_3$  occurs?
- e. Given that  $B_2$  has occurred, what is the probability that  $A$  occurs?
- f. Given that  $A$  has occurred, what is the probability that  $B_4$  occurs?

## Applications

45. According to an online survey by Harris Interactive for job site CareerBuilder.com, more than half of IT (information technology) workers say they have fallen asleep at work (*InformationWeek*, September 27, 2007). Sixty-four percent of government workers admitted to falling asleep on the job. Consider the following contingency table that is representative of the survey results.

Slept on the Job?	Job Category	
	IT Professional	Government Professional
Yes	155	256
No	145	144

- a. Convert the contingency table into a joint probability table.
  - b. What is the probability that a randomly selected worker is an IT professional?
  - c. What is the probability that a randomly selected worker slept on the job?
  - d. If a randomly selected worker slept on the job, what is the probability that he/she is an IT professional?
  - e. If a randomly selected worker is a government professional, what is the probability that he/she slept on the job?
  - f. Are the events "IT Professional" and "Slept on the Job" independent? Explain using probabilities.
46. A recent report suggests that business majors spend the least amount of time on course work than all other college students (*The New York Times*, November 17, 2011). A provost of a university decides to conduct a survey where students are asked if they study hard, defined by spending at least 20 hours per week on course work. Of 120 business majors included in the survey, 20 said that they studied hard as compared to 48 out of 150 nonbusiness majors who said that they studied hard.
    - a. Construct a contingency table that shows the frequencies for the qualitative variables Major (business or nonbusiness) and Study Hard (yes or no).
    - b. Find the probability that a business major spends less than 20 hours per week on course work.
    - c. What is the probability that a student studies hard?
    - d. If a student spends at least 20 hours on course work, what is the probability that he/she is a business major? What is the corresponding probability that he/she is a nonbusiness major?

47. A recent poll asked 16- to 21-year-olds whether or not they are likely to serve in the U.S. military. The following table, cross-classified by gender and race, reports the percentage of those polled who responded that they are likely or very likely to serve in the active-duty military.

Gender	Race		
	Hispanic	Black	White
Male	33.5%	20.5%	16.5%
Female	14.5%	10.5%	4.5%

SOURCE: Defense Human Resources Activity telephone poll of 3,228 Americans conducted October through December 2005.

- a. What is the probability that a randomly selected respondent is female?
  - b. What is the probability that a randomly selected respondent is Hispanic?
  - c. Given that a respondent is female, what is the probability that she is Hispanic?
  - d. Given that a respondent is white, what is the probability that the respondent is male?
  - e. Are the events "Male" and "White" independent? Explain using probabilities.
48. According to a Michigan State University researcher, Americans are becoming increasingly polarized on issues pertaining to the environment (<http://news.msu.edu>, April 19, 2011). It is reported that 70% of Democrats see signs of global warming as compared to only 30% of Republicans who feel the same. Suppose the survey was based on 400 Democrats and 400 Republicans.
    - a. Construct a contingency table that shows frequencies for the qualitative variables Political Affiliation (Democrat or Republican) and Global Warming (yes or no).
    - b. Find the probability that a Republican sees signs of global warming.
    - c. Find the probability that a person does not see signs of global warming.
    - d. If a person sees signs of global warming, what is the probability that this person is a Democrat?
  49. Merck & Co. conducted a study to test the promise of its experimental AIDS vaccine (*The Boston Globe*, September 22, 2007). Volunteers in the study were all free of the human immunodeficiency virus (HIV), which causes AIDS, at the start of the study, but all were at high risk for getting the virus. Volunteers were given either the vaccine or a dummy shot; 24 of 741 volunteers who got the vaccine became infected with HIV, whereas 21 of 762 volunteers who got the dummy shot became infected with HIV. The following table summarizes the results of the study.

	Vaccinated	Dummy Shot
Infected	24	21
Not Infected	717	741

- Convert the contingency table into a joint probability table.
- What is the probability that a randomly selected volunteer got vaccinated?
- What is the probability that a randomly selected volunteer became infected with the HIV virus?
- If the randomly selected volunteer was vaccinated, what is the probability that he/she got infected?
- Are the events “Vaccinated” and “Infected” independent? Explain using probabilities. Given your answer, is it surprising that Merck & Co. ended enrollment and vaccination of volunteers in the study? Explain.

50. More and more households are struggling to pay utility bills given a shaky economy and high heating costs (*The Wall Street Journal*, February 14, 2008). Particularly hard hit are households with homes heated with propane or heating oil. Many of these households are spending twice as much to stay warm this winter compared to those who heat with natural gas or electricity. A representative sample of 500 households was taken to investigate if the type of heating influences whether or not a household is delinquent in paying its utility bill. The following table reports the results.

Delinquent in Payment?	Type of Heating			
	Natural Gas	Electricity	Heating Oil	Propane
Yes	50	20	15	10
No	240	130	20	15

- What is the probability that a randomly selected household uses heating oil?
- What is the probability that a randomly selected household is delinquent in paying its utility bill?
- What is the probability that a randomly selected household uses heating oil and is delinquent in paying its utility bill?

- Given that a household uses heating oil, what is the probability that it is delinquent in paying its utility bill?
- Given that a household is delinquent in paying its utility bill, what is the probability that the household uses electricity?
- Are the events “Heating Oil” and “Delinquent in Payment” independent? Explain using probabilities.

51. The research team at a leading perfume company is trying to test the market for its newly introduced perfume. In particular the team wishes to look for gender and international differences in the preference for this perfume. They sample 2,500 people internationally and each person in the sample is asked to try the new perfume and list his/her preference. The following table reports the results.

Preference	Gender	America	Europe	Asia
Like it	Men	210	150	120
	Women	370	310	180
Don't like it	Men	290	150	80
	Women	330	190	120

- What is the probability that a randomly selected man likes the perfume?
- What is the probability that a randomly selected Asian likes the perfume?
- What is the probability that a randomly selected European woman does not like the perfume?
- What is the probability that a randomly selected American man does not like the perfume?
- Are the events “Men” and “Like Perfume” independent in (i) America, (ii) Europe, (iii) Asia? Explain using probabilities.
- Internationally, are the events “Men” and “Like Perfume” independent? Explain using probabilities.

## 4.4 THE TOTAL PROBABILITY RULE AND BAYES' THEOREM

In this section, we present two important rules in probability theory: the total probability rule and Bayes' theorem. The **total probability rule** is a useful tool for breaking the computation of a probability into distinct cases. **Bayes' theorem** uses this rule to update the probability of an event that has been affected by a new piece of evidence.

### The Total Probability Rule

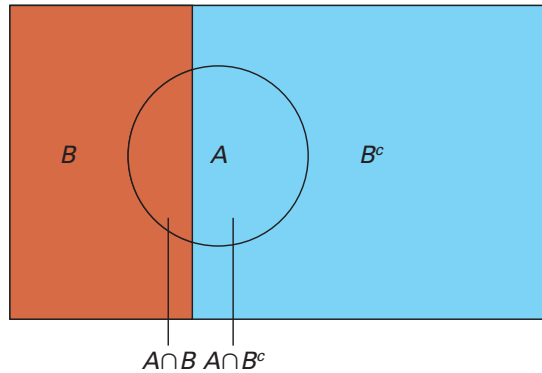
Sometimes the probability of an event is not readily apparent from the given information. The total probability rule expresses the probability of an event in terms of joint or conditional probabilities. Let  $P(A)$  denote the probability of an event of

#### LO 4.8

Apply the total probability rule.

interest. We can express  $P(A)$  as the sum of probabilities of the intersections of  $A$  with some mutually exclusive and exhaustive events corresponding to an experiment. For instance, consider event  $B$  and its complement  $B^c$ . Figure 4.7 shows the sample space partitioned entirely into these two mutually exclusive and exhaustive events. The circle, representing event  $A$ , consists entirely of its intersections with  $B$  and  $B^c$ . According to the total probability rule,  $P(A)$  equals the sum of  $P(A \cap B)$  and  $P(A \cap B^c)$ .

**FIGURE 4.7**  
The total probability rule:  
 $P(A) = P(A \cap B) + P(A \cap B^c)$



Oftentimes the joint probabilities needed to compute the total probability are not explicitly specified. Therefore, we use the multiplication rule to derive these probabilities from the conditional probabilities as  $P(A \cap B) = P(A|B)P(B)$  and  $P(A \cap B^c) = P(A|B^c)P(B^c)$ .

#### THE TOTAL PROBABILITY RULE CONDITIONAL ON TWO EVENTS

The **total probability rule** expresses the probability of an event,  $A$ , in terms of probabilities of the intersection of  $A$  with any mutually exclusive and exhaustive events. The total probability rule based on two events,  $B$  and  $B^c$ , is

$$P(A) = P(A \cap B) + P(A \cap B^c),$$

or equivalently,

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

An intuitive way to express the total probability rule is with the help of a **probability tree**. Whenever an experiment can be broken down into stages, with a different aspect of the result observed at each stage, we can use a probability tree to represent the various possible sequences of observations. We also use an alternative tabular method for computing the probability  $P(A)$ . The following example illustrates the mechanics of a probability tree and the tabular method.

#### EXAMPLE 4.18

Even though a certain statistics professor does not require attendance as part of a student's overall grade, she has noticed that those who regularly attend class have a higher tendency to get a final grade of A. The professor calculates that there is an 80% chance that a student attends class regularly. Moreover, given that a student attends class regularly, there is a 35% chance that the student receives an A grade; however, if a student does not attend class regularly, there is only a 5% chance of an A grade. Use this information to answer the following questions.

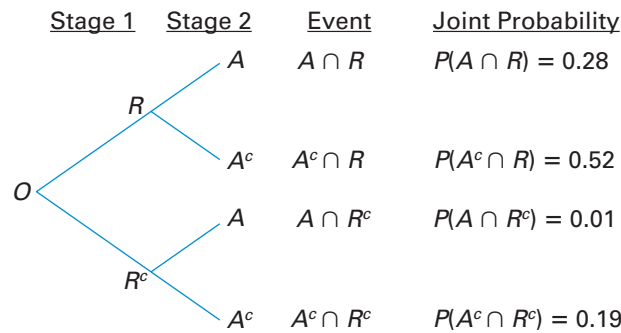
- What is the probability that a student does not attend class regularly?
- What is the probability that a student attends class regularly and receives an A grade?



- c. What is the probability that a student does not attend class regularly and receives an A grade?
- d. What is the probability that a student receives an A grade?

**SOLUTION:** We first let  $A$  correspond to the event that a student receives an A grade and  $R$  correspond to the event that a student attends class regularly. From the preceding information, we then have the following probabilities:  $P(R) = 0.80$ ,  $P(A|R) = 0.35$ , and  $P(A|R^c) = 0.05$ . Figure 4.8 shows a probability tree that consists of nodes (junctions) and branches (lines) where the initial node  $O$  is called the origin. The branches emanating from  $O$  represent the possible outcomes that may occur at the first stage. Thus, at stage 1 we have events  $R$  and  $R^c$  originating from  $O$ . These events become the nodes at the second stage. The sum of the probabilities coming from any particular node is equal to one.

**FIGURE 4.8** Probability tree for class attendance and final grade in statistics



- a. Using the complement rule, if we know that there is an 80% chance that a student attends class regularly,  $P(R) = 0.80$ , then the probability that a student does not attend class regularly is found as  $P(R^c) = 1 - P(R) = 1 - 0.80 = 0.20$ .
- In order to arrive at a subsequent stage, and deduce the corresponding probabilities, we use the information obtained from the previous stage. For instance, given that a student attends class regularly, there is a 35% chance that the student receives an A grade; that is,  $P(A|R) = 0.35$ . Given that a student regularly attends class, the likelihood of not receiving an A grade is 65% because  $P(A^c|R) = 1 - P(A|R) = 0.65$ . Similarly, given  $P(A|R^c) = 0.05$ , we compute  $P(A^c|R^c) = 1 - P(A|R^c) = 1 - 0.05 = 0.95$ . Any path through branches of the tree from the origin to a terminal node defines the intersection of the earlier two events. Thus, following the top branches, we arrive at the event  $A \cap R$ , meaning that a student attends class regularly and receives an A grade. The probability of this event is the product of the probabilities attached to the branches forming that path; here we are simply applying the multiplication rule. Now we are prepared to answer parts b and c.
- b. Multiplying the probabilities attached to the top branches we obtain  $P(A \cap R) = P(A|R)P(R) = 0.35 \times 0.80 = 0.28$ ; there is a 28% chance that a student attends class regularly and receives an A grade.
  - c. In order to find the probability that a student does not attend class regularly and receives an A grade, we compute  $P(A \cap R^c) = P(A|R^c)P(R^c) = 0.05 \times 0.20 = 0.01$ .
  - d. The probability that a student receives an A grade,  $P(A)$ , is not explicitly given in Example 4.18. However, we can sum the relevant joint probabilities in parts b and c to obtain this probability:

$$P(A) = P(A \cap R) + P(A \cap R^c) = 0.28 + 0.01 = 0.29.$$

An alternative method uses a tabular representation of probabilities. Table 4.5 contains all relevant probabilities that are directly or indirectly specified in Example 4.18.

**TABLE 4.5** Tabular Method for Computing  $P(A)$

Unconditional Probability	Conditional Probability	Joint Probability
$P(R) = 0.80$	$P(A R) = 0.35$	$P(A \cap R) = P(A R)P(R) = 0.28$
$P(R^c) = 0.20$	$P(A R^c) = 0.05$	$P(A \cap R^c) = P(A R^c)P(R^c) = 0.01$
$P(R) + P(R^c) = 1$		$P(A) = P(A \cap R) + P(A \cap R^c) = 0.29$

As we saw earlier, each joint probability is computed as a product of its conditional probability and the corresponding unconditional probability; that is,  $P(A \cap R) = P(A|R)P(R) = 0.35 \times 0.80 = 0.28$ . Similarly,  $P(A \cap R^c) = P(A|R^c)P(R^c) = 0.05 \times 0.20 = 0.01$ . Therefore,  $P(A) = P(A \cap R) + P(A \cap R^c) = 0.29$ .

#### LO 4.9

Apply Bayes' theorem.

## Bayes' Theorem

The total probability rule is also needed to derive Bayes' theorem, developed by the Reverend Thomas Bayes (1702–1761). Bayes' theorem is a procedure for updating probabilities based on new information. The original probability is an unconditional probability called a **prior probability** in the sense that it reflects only what we know now before the arrival of any new information. On the basis of new information, we update the prior probability to arrive at a conditional probability called a **posterior probability**.

Suppose we know that 99% of the individuals who take a lie detector test tell the truth. Therefore, the prior probability of telling the truth is 0.99. Suppose an individual takes the lie detector test and the results indicate that the individual lied. Bayes' theorem updates a prior probability to compute a posterior probability, which in the above example is essentially a conditional probability based on the information that the lie detector has detected a lie.

Let  $P(B)$  denote the prior probability and  $P(B|A)$  the posterior probability. Note that the posterior probability is conditional on event  $A$ , representing new information. Recall the conditional probability formula from Section 4.2:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

In some instances, we may have to evaluate  $P(B|A)$ , but we do not have explicit information on  $P(A \cap B)$  or  $P(A)$ . However, given information on  $P(B)$ ,  $P(A|B)$  and  $P(A|B^c)$ , we can use the total probability rule and the multiplication rule to find  $P(B|A)$  as follows:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

### BAYES' THEOREM

The posterior probability  $P(B|A)$  can be found using the information on the prior probability  $P(B)$  along with the conditional probabilities  $P(A|B)$  and  $P(A|B^c)$  as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

In the above formula, we have used Bayes' theorem to update the prior probability  $P(B)$  to the posterior probability  $P(B|A)$ . Equivalently, we can use Bayes' theorem to update the prior probability  $P(A)$  to derive the posterior probability  $P(A|B)$  by interchanging the events  $A$  and  $B$  in the above formula.

### EXAMPLE 4.19

In a lie-detector test, an individual is asked to answer a series of questions while connected to a polygraph (lie detector). This instrument measures and records several physiological responses of the individual on the basis that false answers will produce distinctive measurements. Assume that 99% of the individuals who go in for a polygraph test tell the truth. These tests are considered to be 95% reliable. In other words, there is a 95% chance that the test will detect a lie if an individual actually lies. Let there also be a 0.5% chance that the test erroneously detects a lie even when the individual is telling the truth. An individual has just taken a polygraph test and the test has detected a lie. What is the probability that the individual was actually telling the truth?

**SOLUTION:** First we define some events and their associated probabilities. Let  $D$  and  $T$  correspond to the events that the polygraph detects a lie and that an individual is telling the truth, respectively. We are given that  $P(T) = 0.99$ , implying that  $P(T^c) = 1 - 0.99 = 0.01$ . In addition, we formulate  $P(D|T^c) = 0.95$  and  $P(D|T) = 0.005$ . We need to find  $P(T|D)$  when we are not explicitly given  $P(D \cap T)$  and  $P(D)$ . We can use Bayes' theorem to find:

$$P(T|D) = \frac{P(D \cap T)}{P(D)} = \frac{P(D \cap T)}{P(D \cap T) + P(D \cap T^c)} = \frac{P(D|T)P(T)}{P(D|T)P(T) + P(D|T^c)P(T^c)}.$$

Although we can use this formula to solve the problem directly, it is often easier to solve it systematically with the help of the following table.

**TABLE 4.6** Computing Posterior Probabilities for Example 4.19

Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
$P(T) = 0.99$	$P(D T) = 0.005$	$P(D \cap T) = 0.00495$	$P(T D) = 0.34256$
$P(T^c) = 0.01$	$P(D T^c) = 0.95$	$P(D \cap T^c) = 0.00950$	$P(T^c D) = 0.65744$
$P(T) + P(T^c) = 1$		$P(D) = 0.01445$	$P(T D) + P(T^c D) = 1$

The first column presents prior probabilities and the second column shows related conditional probabilities. We first compute the denominator of Bayes' theorem by using the total probability rule,  $P(D) = P(D \cap T) + P(D \cap T^c)$ . Joint probabilities are calculated as products of conditional probabilities with their corresponding prior probabilities. For instance, in Table 4.6, in order to obtain  $P(D \cap T)$ , we multiply  $P(D|T)$  with  $P(T)$ , which yields  $P(D \cap T) = 0.005 \times 0.99 = 0.00495$ . Similarly, we find  $P(D \cap T^c) = 0.95 \times 0.01 = 0.00950$ . Thus, according to the total probability rule,  $P(D) = 0.00495 + 0.00950 = 0.01445$ . Finally,  $P(T|D) = \frac{P(D \cap T)}{P(D \cap T) + P(D \cap T^c)} = \frac{0.00495}{0.01445} = 0.34256$ . The prior probability of an individual telling the truth is 0.99. However, given the new information that the polygraph detected the individual telling a lie, the posterior probability of this individual telling the truth is now revised downward to 0.34256.

So far we have used the total probability rule as well as Bayes' theorem based on two mutually exclusive and exhaustive events, namely,  $B$  and  $B^c$ . We can easily extend the analysis to include  $n$  mutually exclusive and exhaustive events,  $B_1, B_2, \dots, B_n$ .

## EXTENSIONS OF THE TOTAL PROBABILITY RULE AND BAYES' THEOREM

If  $B_1, B_2, \dots, B_n$  represent  $n$  mutually exclusive and exhaustive events, then the **total probability rule** extends to:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n),$$

or equivalently,

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n).$$

Similarly, **Bayes' theorem**, for any  $i = 1, 2, \dots, n$ , extends to:

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n)},$$

or equivalently,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n)}.$$

### EXAMPLE 4.20

Scott Myers is a security analyst for a telecommunications firm called Webtalk. Although he is optimistic about the firm's future, he is concerned that its stock price will be considerably affected by the condition of credit flow in the economy. He believes that the probability is 0.20 that credit flow will improve significantly, 0.50 that it will improve only marginally, and 0.30 that it will not improve at all. He also estimates that the probability that the stock price of Webtalk will go up is 0.90 with significant improvement in credit flow in the economy, 0.40 with marginal improvement in credit flow in the economy, and 0.10 with no improvement in credit flow in the economy.

- Based on Scott's estimates, what is the probability that the stock price of Webtalk goes up?
- If we know that the stock price of Webtalk has gone up, what is the probability that credit flow in the economy has improved significantly?

**SOLUTION:** As always, we first define the relevant events and their associated probabilities. Let  $S$ ,  $M$ , and  $N$  denote significant, marginal, and no improvement in credit flow, respectively. Then  $P(S) = 0.20$ ,  $P(M) = 0.50$ , and  $P(N) = 0.30$ . In addition, if we allow  $G$  to denote an increase in stock price, we formulate  $P(G|S) = 0.90$ ,  $P(G|M) = 0.40$ , and  $P(G|N) = 0.10$ . We need to calculate  $P(G)$  in part a and  $P(S|G)$  in part b. Table 4.7 aids in assigning probabilities.

**TABLE 4.7** Computing Posterior Probabilities for Example 4.20

Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
$P(S) = 0.20$	$P(G S) = 0.90$	$P(G \cap S) = 0.18$	$P(S G) = 0.4390$
$P(M) = 0.50$	$P(G M) = 0.40$	$P(G \cap M) = 0.20$	$P(M G) = 0.4878$
$P(N) = 0.30$	$P(G N) = 0.10$	$P(G \cap N) = 0.03$	$P(N G) = 0.0732$
$P(S) + P(M) + P(N) = 1$		$P(G) = 0.41$	$P(S G) + P(M G) + P(N G) = 1$

- In order to calculate  $P(G)$ , we use the total probability rule,  $P(G) = P(G \cap S) + P(G \cap M) + P(G \cap N)$ . The joint probabilities are calculated as a product of conditional probabilities with their corresponding prior probabilities. For instance, in Table 4.7,  $P(G \cap S) = P(G|S)P(S) = 0.90 \times 0.20 = 0.18$ . Therefore, the probability that the stock price of Webtalk goes up equals  $P(G) = 0.18 + 0.20 + 0.03 = 0.41$ .

- b. According to Bayes' theorem,  $P(S|G) = \frac{P(G \cap S)}{P(G)} = \frac{P(G \cap S)}{P(G \cap S) + P(G \cap M) + P(G \cap N)}$ . We use the total probability rule in the denominator to find  $P(G) = 0.18 + 0.20 + 0.03 = 0.41$ . Therefore,  $P(S|G) = \frac{P(G \cap S)}{P(G)} = \frac{0.18}{0.41} = 0.4390$ . Note that the prior probability of a significant improvement in credit flow is revised upward from 0.20 to a posterior probability of 0.4390.

## EXERCISES 4.4

### Mechanics

52. Let  $P(B) = 0.60$ ,  $P(A|B) = 0.80$ , and  $P(A|B^c) = 0.10$ . Calculate the following probabilities:
- $P(B^c)$
  - $P(A \cap B)$  and  $P(A \cap B^c)$
  - $P(A)$
  - $P(B|A)$
53. Let  $P(A) = 0.70$ ,  $P(B|A) = 0.55$ , and  $P(B|A^c) = 0.10$ . Use a probability tree to calculate the following probabilities:
- $P(A^c)$
  - $P(A \cap B)$  and  $P(A^c \cap B)$
  - $P(B)$
  - $P(A|B)$
54. Complete the following probability table.

Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
$P(A) = 0.30$	$P(B A) = 0.25$	$P(A \cap B) =$	$P(A B) =$
$P(A^c) =$	$P(B A^c) = 0.80$	$P(A^c \cap B) =$	$P(A^c B) =$
Total =		$P(B) =$	Total =

55. Complete the following probability table.

Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
$P(B) = 0.85$	$P(A B) = 0.05$	$P(A \cap B) =$	$P(B A) =$
$P(B^c) =$	$P(A B^c) = 0.80$	$P(A \cap B^c) =$	$P(B^c A) =$
Total =		$P(A) =$	Total =

56. Let a sample space be partitioned into three mutually exclusive and exhaustive events,  $B_1$ ,  $B_2$ , and  $B_3$ . Complete the following probability table.

Prior Probabilities	Conditional Probabilities	Joint Probabilities	Posterior Probabilities
$P(B_1) = 0.10$	$P(A B_1) = 0.40$	$P(A \cap B_1) =$	$P(B_1 A) =$
$P(B_2) =$	$P(A B_2) = 0.60$	$P(A \cap B_2) =$	$P(B_2 A) =$
$P(B_3) = 0.30$	$P(A B_3) = 0.80$	$P(A \cap B_3) =$	$P(B_3 A) =$
Total =		$P(A) =$	Total =

### Applications

57. Christine has always been weak in mathematics. Based on her performance prior to the final exam in Calculus, there is a 40% chance that she will fail the course if she does not have a tutor. With a tutor, her probability of failing

decreases to 10%. There is only a 50% chance that she will find a tutor at such short notice.

- What is the probability that Christine fails the course?
  - Christine ends up failing the course. What is the probability that she had found a tutor?
58. An analyst expects that 20% of all publicly traded companies will experience a decline in earnings next year. The analyst has developed a ratio to help forecast this decline. If the company is headed for a decline, there is a 70% chance that this ratio will be negative. If the company is not headed for a decline, there is a 15% chance that the ratio will be negative. The analyst randomly selects a company and its ratio is negative. What is the posterior probability that the company will experience a decline?
59. The State Police are trying to crack down on speeding on a particular portion of the Massachusetts Turnpike. To aid in this pursuit, they have purchased a new radar gun that promises greater consistency and reliability. Specifically, the gun advertises  $\pm$  one-mile-per-hour accuracy 98% of the time; that is, there is a 0.98 probability that the gun will detect a speeder, if the driver is actually speeding. Assume there is a 1% chance that the gun erroneously detects a speeder even when the driver is below the speed limit. Suppose that 95% of the drivers drive below the speed limit on this stretch of the Massachusetts Turnpike.
- What is the probability that the gun detects speeding and the driver was speeding?
  - What is the probability that the gun detects speeding and the driver was not speeding?
  - Suppose the police stop a driver because the gun detects speeding. What is the probability that the driver was actually driving below the speed limit?
60. According to a recent study, cell phones are the main medium for teenagers to stay connected with friends and family (*CNN*, March 19, 2012). It is estimated that 90% of older teens (aged 14 to 17) and 60% of younger teens (aged 12 to 13) own a cell phone. Suppose 70% of all teens are older teens.
- What is the implied probability that a teen owns a cell phone?
  - Given that a teen owns a cell phone, what is the probability that he/she is an older teen?

- c. Given that the teen owns a cell phone, what is the probability that he/she is a younger teen?
61. According to data from the *National Health and Nutrition Examination Survey*, 33% of white, 49.6% of black, 43% of Hispanic, and 8.9% of Asian women are obese. In a representative town, 48% of women are white, 19% are black, 26% are Hispanic, and the remaining 7% are Asian.
- Find the probability that a randomly selected woman in this town is obese.
  - Given that a woman is obese, what is the probability that she is white?
  - Given that a woman is obese, what is the probability that she is black?
  - Given that a woman is obese, what is the probability that she is Asian?
62. A crucial game of the Los Angeles Lakers basketball team depends on the health of their key player. According to his doctor's report, there is a 40% chance that he will be fully fit to play, a 30% chance that he will be somewhat fit to play, and a 30% chance that he will not be able to play at all. The coach has estimated the chances of winning at 80% if the player is fully fit, 60% if he is somewhat fit, and 40% if he is unable to play.
- What is the probability that the Lakers will win the game?
  - You have just heard that the Lakers won the game. What is the probability that the key player had been fully fit to play in the game?

63. An analyst thinks that next year there is a 20% chance that the world economy will be good, a 50% chance that it will be neutral, and a 30% chance that it will be poor. She also predicts probabilities that the performance of a start-up firm, Creative Ideas, will be good, neutral, or poor for each of the economic states of the world economy. The following table presents probabilities for three states of the world economy and the corresponding conditional probabilities for Creative Ideas.

State of the World Economy	Probability of Economic State	Performance of Creative Ideas	Conditional Probability of Creative Ideas
Good	0.20	Good	0.60
		Neutral	0.30
		Poor	0.10
Neutral	0.50	Good	0.40
		Neutral	0.30
		Poor	0.30
Poor	0.30	Good	0.20
		Neutral	0.30
		Poor	0.50

- What is the probability that the performance of the world economy will be neutral and that of creative ideas will be poor?
- What is the probability that the performance of Creative Ideas will be poor?
- The performance of Creative Ideas was poor. What is the probability that the performance of the world economy had also been poor?

## LO 4.10

## 4.5 COUNTING RULES

Use a counting rule to calculate the probability of an event.

In several areas of statistics, including the binomial distribution discussed in the next chapter, the calculation of probabilities involves defining and counting outcomes. Here we discuss principles and shortcuts for counting. Specifically, we explore the factorial, combination, and permutation notations. We then find that in certain circumstances counting rules can aid in calculating the probability of an event.

When we are interested in counting the arrangements of a given set of  $n$  items, we calculate  **$n$  factorial**, denoted  $n!$ . In other words, given  $n$  items, there are  $n!$  ways of arranging them. We apply the factorial when there are no groups—we are only arranging a given set of  $n$  items.

### THE FACTORIAL FORMULA

The number of ways to assign every member of a group of size  $n$  to  $n$  slots is calculated using the **factorial formula**:

$$n! = n \times (n - 1) \times (n - 2) \times (n - 3) \times \cdots \times 1$$

By definition,  $0! = 1$ .



### EXAMPLE 4.21

A little-league coach has nine players on his team and he has to assign each of the players to one of nine positions (pitcher, catcher, first base, etc.). In how many ways can the assignments be made?

**SOLUTION:** The first player may be assigned to nine different positions. Then eight positions remain. The second player can be assigned to eight different positions. The third player can be assigned to seven different positions, and so on, until the ninth and last player can be assigned in only one way. The total number of different assignments is equal to  $9! = 9 \times 8 \times \cdots \times 1 = 362,880$ .

The **combination** and **permutation formulas** apply to two groups of predetermined size. We apply the combination formula when the order of the arrangement does not matter, whereas we use the permutation formula when the order is important. Generally, we look for a specific reference to “order” being important when employing the permutation formula.

#### THE COMBINATION FORMULA

The number of ways to choose  $x$  objects from a total of  $n$  objects, where the order in which the  $x$  objects are listed *does not matter*, is calculated using the **combination formula**:

$${}_nC_x = \binom{n}{x} = \frac{n!}{(n-x)!x!}$$

### EXAMPLE 4.22

The little-league coach from Example 4.21 recruits three more players so that his team has backups in case of injury. Now his team totals 12.

- a. How many ways can the coach select nine players from the 12-player roster?
- b. If each of the lineups from part a is equally likely, what is the probability that the coach selects a particular lineup?

**SOLUTION:**

- a. This is a combination problem because we are simply interested in placing 9 players on the field. We have no concern, for instance, as to whether a player pitches, catches, or plays first base. In other words, the order in which the players are selected is not important. We make use of the combination formula as follows:

$${}_{12}C_9 = \binom{12}{9} = \frac{12!}{(12-9)! \times 9!} = \frac{12 \times 11 \times \cdots \times 1}{(3 \times 2 \times 1) \times (9 \times 8 \times \cdots \times 1)} = 220.$$

- b. Given that each lineup in part a is equally likely, the probability that any one lineup occurs is  $1/220 = 0.0045$ .

#### THE PERMUTATION FORMULA

The number of ways to choose  $x$  objects from a total of  $n$  objects, where the order in which the  $x$  objects is listed *does matter*, is calculated using the **permutation formula**:

$${}_nP_x = \frac{n!}{(n-x)!}$$

### EXAMPLE 4.23

Now suppose the little league coach from Example 4.22 recognizes that the nine positions of baseball are quite different. It matters whether one player is pitching or whether that same player is in the outfield.

- a. In how many ways can the coach assign his 12-player roster to the nine different positions?
- b. If each of the lineups from part a are equally likely, what is the probability that the coach selects a particular lineup?

#### SOLUTION:

- a. This is a permutation problem because the order in which the coach assigns the positions matters. For example, a lineup that has one player playing in the outfield is different from a lineup that has that same player pitching. We calculate the answer as follows:

$${}_{12}P_9 = \frac{12!}{(12-9)!} = \frac{12 \times 11 \times \cdots \times 1}{3 \times 2 \times 1} = 79,833,600.$$

Comparing the answers we obtained from Examples 4.22 and 4.23, we see there is a big difference between the number of arrangements when the position of the player does not matter versus the number of arrangements when the position is important.

- b. Given that each lineup in part a is equally likely, the probability that any one lineup occurs is  $1/79,833,600 \approx 0.0000$ ; that is, the probability approaches zero.

## EXERCISES 4.5

### Mechanics

- 64. Calculate the following values.
  - a.  $8!$  and  $6!$
  - b.  ${}_8C_6$
  - c.  ${}_8P_6$
- 65. Calculate the following values.
  - a.  $7!$  and  $3!$
  - b.  ${}_7C_3$
  - c.  ${}_7P_3$

### Applications

- 66. At a local elementary school, a principal is making random class assignments for her 8 teachers. Each teacher must be assigned to exactly one job. In how many ways can the assignments be made?
- 67. Twenty cancer patients volunteer for a clinical trial. Ten of the patients will receive a placebo and 10 will receive the trial drug. In how many different ways can the researchers select 10 patients to receive the trial drug from the total of 20?
- 68. There are 10 players on the local basketball team. The coach decides to randomly pick 5 players for the game.
  - a. In how many different ways can the coach select 5 players to start the game if order does not matter?
  - b. In how many different ways can the coach select 5 players to start the game if order (the type of position, i.e., point guard, center, etc.) matters?
- 69. A horse-racing fan is contemplating the many different outcomes in an eight-horse race.
  - a. How many different combinations are possible if only the first three places (first, second, and third) are considered?
  - b. If each of the combinations in part a is equally likely, what is the probability of selecting the winning combination?
  - c. How many different *rankings* are possible if only the first three places are considered?
  - d. If each of the rankings in part c is equally likely, what is the probability of selecting the winning ranking?
- 70. Jacqueline Fibbe manages 10 employees at a small ice cream store in Beverly Farms, MA. She assigns three employees for each eight-hour shift.

- a. If order is not important, in how many different ways can she select three employees from the total of 10 for each eight-hour shift?
  - b. Megan H. is one of the 10 employees. If the assignment of employees is random, how many of the shifts in part a will include Megan H.?
71. David Barnes and his fiancée Valerie Shah are visiting Hawaii. At the Hawaiian Cultural Center in Honolulu, they are told that 2 out of a group of 8 people will be randomly picked for a free lesson of a Tahitian dance.
- a. What is the probability that both David and Valerie get picked for the Tahitian dance lesson?
  - b. What is the probability that Valerie gets picked before David for the Tahitian dance lesson?

## WRITING WITH STATISTICS

A University of Utah study examined 7,925 severely obese adults who had gastric bypass surgery and an identical number of people who did not have the surgery (*The Boston Globe*, August 23, 2007). The study wanted to investigate whether or not losing weight through stomach surgery prolonged the lives of severely obese patients, thereby reducing their deaths from heart disease, cancer, and diabetes.

Over the course of the study, 534 of the participants died. Of those who died, the cause of death was classified as either a disease death (such as heart disease, cancer, and diabetes) or a nondisease death (such as suicide or accident). Lawrence Plummer, a research analyst, is handed Table 4.8, which summarizes the study's findings:

**TABLE 4.8** Deaths Cross-Classified by Cause and Method of Losing Weight

Cause of Death	Method of Losing Weight	
	No Surgery	Surgery
Death from Disease	285	150
Death from Nondisease	36	63



Lawrence wants to use the sample information to:

1. Calculate and interpret relevant probabilities for the cause of death and the method of losing weight.
2. Determine whether the events "Death from Disease" and "No Surgery" are independent.

Numerous studies have documented the health risks posed to severely obese people—those people who are at least 100 pounds overweight. Severely obese people, for instance, typically suffer from high blood pressure and are more likely to develop diabetes. A University of Utah study examined whether the manner in which a severely obese person lost weight influenced a person's longevity. The study followed 7,925 patients who had stomach surgery and an identical number who did not have the surgery. Of particular interest in this report are the 534 participants who died over the course of the study.

The deceased participants were cross-classified by the method in which they lost weight and by the cause of their death. The possible outcomes for the method of losing weight were either "no surgery" or "surgery," and the possible outcomes for the cause of death were either "disease death" (such as heart disease, cancer, or diabetes) or a "nondisease death" (such as suicide or accident). Table 4.A shows the joint probability table.

**Sample Report—Linking Cause of Death with the Method of Losing Weight**

**TABLE 4.A** Joint Probability Table of Deaths Cross-Classified by Cause and Method of Losing Weight

Cause of Death	Method of Losing Weight		Total
	No Surgery	Surgery	
Death from Disease	0.53	0.28	0.81
Death from Nondisease	0.07	0.12	0.19
Total	0.60	0.40	1.00

The unconditional probabilities reveal that 0.60 of the deceased participants in the study did not have surgery, while 0.40 of those who died had opted for the stomach surgery. Of the 534 participants that died, the vast majority, 0.81, died from disease, whereas the cause of death for the remainder was from a nondisease cause.

Joint probabilities reveal that the probability that a deceased participant had no surgery and died from disease was 0.53; yet the probability that a deceased participant had surgery and died from disease was only 0.28. Using the unconditional probabilities and the joint probabilities, it is possible to calculate conditional probabilities. For example, given that a participant's cause of death was from disease, the probability that the participant did not have surgery was  $0.65 (= 0.53/0.81)$ . Similarly, of those participants who opted for no surgery, the likelihood that their death was from disease was  $0.88 (= 0.53/0.60)$ .

A comparison of the conditional probabilities with the unconditional probabilities can reveal whether or not the events "Death from Disease" and "No Surgery" are independent. For instance, there is an 81% chance that a randomly selected obese person dies from disease. However, given that an obese person chooses to lose weight without surgery, the likelihood that he/she dies from disease jumps to 88%. Thus, this initial research appears to suggest that a participant's cause of death is associated with his/her method of losing weight.

## CONCEPTUAL REVIEW

### LO 4.1 Describe fundamental probability concepts.

In order to assign the appropriate probability to an uncertain event, it is useful to establish some terminology. An **experiment** is a process that leads to one of several possible outcomes. A **sample space**, denoted  $S$ , of an experiment contains all possible outcomes of the experiment. An **event** is any subset of outcomes of an experiment, and is called a simple event if it contains a single outcome. Events are **exhaustive** if all possible outcomes of an experiment belong to the events. Events are **mutually exclusive** if they do not share any common outcome of an experiment.

A **probability** is a numerical value that measures the likelihood that an event occurs. It assumes a value between zero and one where a value zero indicates an impossible event and a value one indicates a definite event. The **two defining properties of a probability** are (1) the probability of any event  $A$  is a value between 0 and 1,  $0 \leq P(A) \leq 1$ , and (2) the sum of the probabilities of any list of mutually exclusive and exhaustive events equals 1.

### LO 4.2 Formulate and explain subjective, empirical, and classical probabilities.

A **subjective** probability is calculated by drawing on personal and subjective judgment. An **empirical probability** is calculated as a relative frequency of occurrence. A **classical probability** is based on logical analysis rather than on observation or personal judgment.

#### LO 4.3 Calculate and interpret the probability of the complement of an event.

Rules of probability allow us to calculate the probabilities of more complex events. The **complement rule** states that the probability of the complement of an event can be found by subtracting the probability of the event from one:  $P(A^c) = 1 - P(A)$ .

#### LO 4.4 Calculate and interpret the probability that at least one of two events will occur.

We calculate the probability that at least one of two events occurs by using the **addition rule**:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ . Since  $P(A \cap B) = 0$  for mutually exclusive events, the addition rule then simplifies in these instances to  $P(A \cup B) = P(A) + P(B)$ .

#### LO 4.5 Calculate and interpret a conditional probability and apply the multiplication rule.

The probability of event  $A$ , denoted  $P(A)$ , is an **unconditional probability**. It is the probability that  $A$  occurs without any additional information. The probability that  $A$  occurs given that  $B$  has already occurred, denoted  $P(A|B)$ , is a **conditional probability**. A conditional probability is computed as  $P(A|B) = \frac{P(A \cap B)}{P(B)}$ . We rearrange the conditional probability formula to arrive at the **multiplication rule**. When using this rule, we find the probability that two events,  $A$  and  $B$ , both occur; that is,  $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$ .

#### LO 4.6 Distinguish between independent and dependent events.

Two events,  $A$  and  $B$ , are **independent** if  $P(A|B) = P(A)$ , or if  $P(B|A) = P(B)$ . Otherwise, the events are **dependent**. For independent events, the multiplication rule simplifies to  $P(A \cap B) = P(A)P(B)$ .

#### LO 4.7 Calculate and interpret probabilities from a contingency table.

A **contingency table** generally shows frequencies for two qualitative (categorical) variables,  $x$  and  $y$ , where each cell represents a mutually exclusive combination of  $x$ - $y$  values. Empirical probabilities are easily calculated as the relative frequency of the occurrence of the event.

#### LO 4.8 Apply the total probability rule.

The **total probability rule** expresses the probability of an event  $A$  in terms of probabilities of the intersection of  $A$  with two mutually exclusive and exhaustive events,  $B$  and  $B^c$ :

$$P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

We can extend the above rule where the sample space is partitioned into  $n$  mutually exclusive and exhaustive events,  $B_1, B_2, \dots, B_n$ . The total probability rule is:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n), \text{ or equivalently,}$$

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n).$$

#### LO 4.9 Apply Bayes' theorem.

**Bayes' theorem** provides a procedure for updating probabilities based on new information. Let  $P(B)$  be the prior probability and  $P(B|A)$  be the posterior probability based on new information provided by  $A$ . Then:

$$P(B|A) = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

For the extended case, Bayes' theorem, for any  $i = 1, 2, \dots, n$ , is:

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n)}, \text{ or}$$

$$\text{equivalently, } P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n)}.$$

#### LO 4.10 Use a counting rule to calculate the probability of an event.

Shortcut rules for counting include the **factorial**, the **combination**, and the **permutation** formulas. When we are interested in arranging a given set of  $n$  items, we calculate  $n$  factorial as:  $n! = n \times (n-1) \times \dots \times 1$ . The combination and permutation formulas apply to two groups of predetermined size. We apply the combination formula when the order of the arrangement does not matter:  ${}_nC_x = \binom{n}{x} = \frac{n!}{(n-x)!x!}$ . We use the permutation formula when the order of the arrangement matters:  ${}_nP_x = \frac{n!}{(n-x)!}$ . In certain circumstances, counting rules can aid in calculating the probability of an event.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

72. According to a global survey of 4,400 parents of children between the ages of 14 to 17, 44% of parents spy on their teen's Facebook account (<http://msnbc.com>, April 25, 2012). Assume that American parents account for 10% of all parents of teens with Facebook accounts, of which 60% spy on their teen's Facebook account. Suppose a parent is randomly selected, and the following events are defined:  $A$  = selecting an American parent and  $B$  = selecting a spying parent.
  - a. Based on the above information, what are the probabilities that can be established? Would you label them as subjective, empirical, or classical?
  - b. Are the events  $A$  and  $B$  mutually exclusive and/or exhaustive? Explain.
  - c. Are the events  $A$  and  $B$  independent? Explain.
  - d. What is the probability of selecting an American parent given that she/he is a spying parent?
73. AccuScore calculated an 84% chance that there would be a fight during the game between the Anaheim Ducks and the Chicago Blackhawks, two of the National Hockey League's most pugnacious teams (*The Wall Street Journal*, March 3, 2009). What are the odds for a fight occurring?
74. According to a recent study, cell phones, especially text messaging, is the main medium for teenagers to stay connected with friends and family (*CNN*, March 19, 2012). It is found that only 23% of teens do not own a cell phone. Of those who own a cell phone, only one in four uses a smartphone. What proportion of all teenagers use a smartphone?
75. Henry Chow is a stockbroker working for Merrill Lynch. He knows from past experience that there is a 70% chance that his new client will want to include

U.S. equity in her portfolio and a 50% chance that she will want to include foreign equity. There is also a 40% chance that she will want to include both U.S. equity and foreign equity in her portfolio.

- a. What is the probability that the client will want to include U.S. equity if she already has foreign equity in her portfolio?
  - b. What is the probability that the client decides to include neither U.S. equity nor foreign equity in her portfolio?
76. The Easy Credit Company reports the following table representing a breakdown of customers according to the amount they owe and whether a cash advance has been made. An auditor randomly selects one of the accounts.

Amounts owed by customers	Cash Advance?	
	Yes	No
\$0 – 199.99	245	2890
\$200 – 399.99	380	1700
\$400 – 599.99	500	1425
\$600 – 799.99	415	940
\$800 – 999.99	260	480
\$1,000 or more	290	475
<b>Total Customers</b>	2090	7910

- a. What is the probability that a customer received a cash advance?
- b. What is the probability that a customer owed less than \$200 and received a cash advance?
- c. What is the probability that a customer owed less than \$200 or received a cash advance?
- d. Given that a customer received a cash advance, what is the probability that the customer owed \$1,000 or more?



- e. Given that a customer owed \$1,000 or more, what is the probability that the customer received a cash advance?
- f. Are the events “receiving a cash advance” and “owing \$1,000 or more” mutually exclusive? Explain using probabilities.
- g. Are the events “receiving a cash advance” and “owing \$1,000 or more” independent? Explain using probabilities.
77. The following frequency distribution shows the ages of India’s 40 richest individuals. One of these individuals is selected at random.
- | Ages        | Frequency |
|-------------|-----------|
| 30 up to 40 | 3         |
| 40 up to 50 | 8         |
| 50 up to 60 | 15        |
| 60 up to 70 | 9         |
| 70 up to 80 | 5         |
- SOURCE: [www.forbes.com](http://www.forbes.com).
- a. What is the probability that the individual is between 50 and 60 years of age?
- b. What is the probability that the individual is younger than 50 years of age?
- c. What is the probability that the individual is at least 60 years of age?
78. How much you smile in your younger days can predict your later success in marriage ([msnbc.com](http://msnbc.com), April 16, 2009). The analysis is based on the success rate in marriage of people over age 65 and their smiles when they were only 10 years old. Researchers found that only 11% of the biggest smilers had been divorced, while 31% of the biggest frowners had experienced a broken marriage.
- a. Suppose it is known that 2% of the people are the biggest smilers at age 10 and divorced in later years. What percent of people are the biggest smilers?
- b. If 25% of people are considered to be the biggest frowners, calculate the probability that a person is the biggest frowner at age 10 and divorced later in life.
79. Anthony Papantonis, owner of Nauset Construction, is bidding on two projects, A and B. The probability that he wins project A is 0.40 and the probability that he wins project B is 0.25. Winning Project A and winning Project B are independent events.
- a. What is the probability that he wins project A or project B?
- b. What is the probability that he does not win either project?
80. Wooden boxes are commonly used for the packaging and transportation of mangoes. A convenience store in Morganville, New Jersey, regularly buys mangoes from a wholesale dealer. For every shipment, the manager randomly inspects two mangoes from a box containing 20 mangoes for damages due to transportation. Suppose the chosen box contains exactly 3 damaged mangoes.
- a. Find the probability that the first mango is not damaged.
- b. Find the probability that neither of the mangoes is damaged.
- c. Find the probability that both mangoes are damaged.
81. A recent study shows that unemployment does not impact males and females in the same way (*Newsweek*, April 20, 2009). According to a Bureau of Labor Statistics report, 8.5% of those who are eligible to work are unemployed. The unemployment rate is 8.8% for eligible men and only 7.0% for eligible women. Suppose 52% of the eligible workforce in the U.S. consists of men.
- a. You have just heard that another worker in a large firm has been laid off. What is the probability that this worker is a man?
- b. You have just heard that another worker in a large firm has been laid off. What is the probability that this worker is a woman?
82. According to the CGMA Economic Index, which measures executive sentiment across the world, 18% of all respondents expressed optimism about the global economy ([www.aicpa.org](http://www.aicpa.org), March 29, 2012). Moreover, 22% of the respondents from the United States and 9% from Asia felt optimistic about the global economy.
- a. What is the probability that an Asian respondent is not optimistic about the global economy?
- b. If 28% of all respondents are from the United States, what is the probability that a respondent is from the United States and is optimistic about the global economy?
- c. Suppose 22% of all respondents are from Asia. If a respondent feels optimistic about the global economy, what is the probability that the respondent is from Asia?
83. A professor of management has heard that eight students in his class of 40 have landed an internship for the summer. Suppose he runs into two of his students in the corridor.
- a. Find the probability that neither of these students has landed an internship.
- b. Find the probability that both of these students have landed an internship.

84. It has generally been believed that it is not feasible for men and women to be just friends (*The New York Times*, April 12, 2012). Others argue that this belief may not be true anymore since gone are the days when men worked and women stayed at home and the only way they could get together was for romance. In a recent survey, 186 heterosexual college students were asked if it was feasible for men and women to be just friends. Thirty-two percent of females and 57% of males reported that it was not feasible for men and women to be just friends. Suppose the study consisted of 100 female and 86 male students.

- Construct a contingency table that shows frequencies for the qualitative variables Gender (men or women) and Feasible (yes or no).
- Find the probability that a student believes that men and women can be friends.
- If a student believes that men and women can be friends, what is the probability that this student is a male? Find the corresponding probability that this student is a female.

85. At a local bar in a small Midwestern town, beer and wine are the only two alcoholic options. The manager noted that of all male customers who visited over the weekend, 150 ordered beer, 40 ordered wine, and 20 asked for soft drinks. Of female customers, 38 ordered beer, 20 ordered wine, and 12 asked for soft drinks.

- Construct a contingency table that shows frequencies for the qualitative variables Gender (male or female) and Drink Choice (beer, wine, or soft drink).
- Find the probability that a customer orders wine.
- What is the probability that a male customer orders wine?
- Are the events “Wine” and “Male” independent? Explain using probabilities.

86. A recent study in the *Journal of the American Medical Association* (February 20, 2008) found that patients who go into cardiac arrest while in the hospital are more likely to die if it happens after 11 pm. The study investigated 58,593 cardiac arrests that occurred during the day or evening. Of those, 11,604 survived to leave the hospital. There were 28,155 cardiac arrests during the shift that began at 11 pm, commonly referred to as the graveyard shift. Of those, 4,139 survived for discharge. The following contingency table summarizes the results of the study.

	Survived for Discharge	Did not Survive for Discharge	Total
Day or Evening Shift	11,604	46,989	58,593
Graveyard Shift	4,139	24,016	28,155
Total	15,743	71,005	86,748

- What is the probability that a randomly selected patient experienced cardiac arrest during the graveyard shift?
- What is the probability that a randomly selected patient survived for discharge?
- Given that a randomly selected patient experienced cardiac arrest during the graveyard shift, what is the probability the patient survived for discharge?
- Given that a randomly selected patient survived for discharge, what is the probability the patient experienced cardiac arrest during the graveyard shift?
- Are the events “Survived for Discharge” and “Graveyard Shift” independent? Explain using probabilities. Given your answer, what type of recommendations might you give to hospitals?

87. It has been reported that women end up unhappier than men later in life, even though they start out happier (*Yahoo News*, August 1, 2008). Early in life, women are more likely to fulfill their family life and financial aspirations, leading to greater overall happiness. However, men report a higher satisfaction with their financial situation and family life, and are thus happier than women, in later life. Suppose the results of the survey of 300 men and 300 women are presented in the following table.

Response to the question “Are you satisfied with your financial and family life?”

	Age		
Response by Women	20 to 35	35 to 50	Over 50
Yes	73	36	32
No	67	54	38

	Age		
Response by Men	20 to 35	35 to 50	Over 50
Yes	58	34	38
No	92	46	32

- What is the probability that a randomly selected woman is satisfied with her financial and family life?
- What is the probability that a randomly selected man is satisfied with his financial and family life?

- c. For women, are the events “Yes” and “20 to 35” independent? Explain using probabilities.
  - d. For men, are the events “Yes” and “20 to 35” independent? Explain using probabilities.
88. An analyst predicts that there is a 40% chance that the U.S. economy will perform well. If the U.S. economy performs well, then there is an 80% chance that Asian countries will also perform well. On the other hand, if the U.S. economy performs poorly, the probability of Asian countries performing well goes down to 0.30.
- a. What is the probability that both the U.S. economy and the Asian countries will perform well?
  - b. What is the probability that the Asian countries will perform well?
  - c. What is the probability that the U.S. economy will perform well, given that the Asian countries perform well?
89. Apparently, depression significantly increases the risk of developing dementia later in life (*BBC News*, July 6, 2010). In a recent study, it was reported that 22% of those who had depression went on to develop dementia, compared to only 17% of those who did not have depression. Suppose 10% of all people suffer from depression.
- a. What is the probability of a person developing dementia?
  - b. If a person has developed dementia, what is the probability that the person suffered from depression earlier in life?
90. According to data from the *National Health and Nutrition Examination Survey*, 36.5% of adult women and 26.6% of adult men are at a healthy weight. Suppose 50.52% of the adult population consists of women.
- a. What proportion of adults is at a healthy weight?
  - b. If an adult is at a healthy weight, what is the probability that the adult is a woman?
  - c. If an adult is at a healthy weight, what is the probability that the adult is a man?
91. Suppose that 60% of the students do homework regularly. It is also known that 80% of students who had been doing homework regularly, end up doing well in the course (get a grade of A or B). Only 20% of students who had not been doing homework regularly, end up doing well in the course.
- a. What is the probability that a student does well in the course?
  - b. Given that the student did well in the course, what is the probability that the student had been doing homework regularly?
92. According to the Census’s Population Survey, the percentage of children with two parents at home is the highest for Asians and lowest for blacks (*USA Today*, February 26, 2009). It is reported that 85% of Asian children have two parents at home versus 78% of white, 70% of Hispanic, and 38% of black. Suppose there are 500 students in a representative school of which 280 are white, 50 are Asian, 100 are Hispanic, and 70 are black.
- a. What is the probability that a child has both parents at home?
  - b. If both parents are at home, what is the probability the child is Asian?
  - c. If both parents are at home, what is the probability the child is black?
93. Prior to the start of the season, a sports analyst is attempting to predict the end-of-season rankings of the 10 teams in a conference.
- a. How many different ways can the teams be ranked if ties are not considered?
  - b. How many different combinations are possible if only the first three places (first, second, and third) are considered?
  - c. If each of the combinations in part b is equally likely, what is the probability that the sports analyst selects the correct end-of-season combination?
  - d. How many different *rankings* are possible if only the first three places are considered?
  - e. If each of the rankings in part d is equally likely, what is the probability that the sports analyst selects the correct end-of-season ranking?
94. Assume high school coach Emily Williams has seven possible swimmers for a four-person relay team.
- a. If the order for the freestyle relay is unimportant, how many different relay teams are possible?
  - b. Assume for the medley relay team that order is important; how many different teams are possible?
  - c. Swimmer Michael P. is one of the seven swimmers. If the assignment of swimmers is random, how many of the teams in part a will include Michael P.? How many of the teams in part b will include Michael P.?

## CASE STUDIES

**CASE STUDY 4.1** Ever since the introduction of New Coke failed miserably in the 1980s, most food and beverage companies have been cautious about changing the taste or formula of their signature offerings. In an attempt to attract more business, Starbucks recently introduced a new milder brew, Pike Place Roast, as its main drip coffee at the majority of its locations nationwide. The idea was to offer a more approachable cup of coffee with a smoother finish. However, the strategy also downplayed the company's more established robust roasts; initially, the milder brew was the only option for customers after noon. Suppose on a recent afternoon, 100 customers were asked whether or not they would return in the near future for another cup of Pike Place Roast. The following contingency table (cross-classified by type of customer and whether or not the customer will return) lists the results:

**Data for Case Study 4.1**

Return in Near Future?	Customer Type	
	First-time Customer	Established Customer
Yes	35	10
No	5	50

In a report, use the sample information to:

1. Calculate and interpret unconditional probabilities.
2. Calculate the probability that a customer will return given that the customer is an established customer.
3. Determine whether the events “Customer will Return” and “Established Customer” are independent. Shortly after the introduction of Pike Place Roast, Starbucks decided to offer its bolder brew again in the afternoon at many of its locations. Do your results support Starbucks’ decision? Explain.

**CASE STUDY 4.2** It is common to ignore the thyroid gland of women during pregnancy (*The New York Times*, April 13, 2009). This gland makes hormones that govern metabolism, helping to regulate body weight, heart rate, and a host of other factors. If the thyroid malfunctions, it can produce too little or too much of these hormones. Hypothyroidism, caused by an untreated underactive thyroid in pregnant women, carries the risk of impaired intelligence in the child. According to one research study, 62 out of 25,216 pregnant women were identified with hypothyroidism. Nineteen percent of the children born to women with an untreated underactive thyroid had an I.Q. of 85 or lower, compared with only 5% of those whose mothers had a healthy thyroid. It was also reported that if mothers have their hypothyroidism treated, their children's intelligence would not be impaired.

In a report, use the sample information to:

1. Find the likelihood that a woman suffers from hypothyroidism during pregnancy and later has a child with an I.Q. of 85 or lower.
2. Determine the number of children in a sample of 100,000 that are likely to have an I.Q. of 85 or lower if the thyroid gland of pregnant women is ignored.
3. Compare and comment on your answer to part b with the corresponding number if all pregnant women are tested and treated for hypothyroidism.

**CASE STUDY 4.3** Enacted in 1998, the Children's Online Privacy Protection Act requires firms to obtain parental consent before tracking the information and the online

movement of children; however, the act applies to those children ages 12 and under. Teenagers are often oblivious to the consequences of sharing their lives online. Data reapers create huge libraries of digital profiles and sell these profiles to advertisers, who use it to detect trends and micro-target their ads back to teens. For example, a teen searching online for ways to lose weight could become enticed by an ad for dietary supplements, fed into his/her network by tracking cookies. As a preliminary step in gauging the magnitude of teen usage of social networking sites, an economist surveys 200 teen girls and 200 teen boys. Of teen girls, 166 use social networking sites; of teen boys, 156 use social networking sites.

In a report, use the sample information to:

1. Construct a contingency table that shows frequencies for the qualitative variables Gender (male or female) and Use of Social Networking Sites (Yes or No).
2. Determine the probability that a teen uses social networking sites.
3. Determine the probability that a teen girl uses a social networking site.
4. A bill before Congress would like to extend the Children's Online Privacy Protection Act to apply to 15-year-olds. In addition, the bill would also ban Internet companies from sending targeted advertising to children under 16 and give these children and their parents the ability to delete their digital footprint and profile with an "eraser button" (*The Boston Globe*, May 20, 2012). Given the probabilities that you calculated with respect to teen usage of social networking sites, do you think that this legislation is necessary? Explain.

**CASE STUDY 4.4** In 2008, it appeared that rising gas prices had made Californians less resistant to offshore drilling. A Field Poll survey showed that a higher proportion of Californians supported the idea of drilling for oil or natural gas along the state's coast than in 2005 (*The Wall Street Journal*, July 17, 2008). Assume that random drilling for oil only succeeds 5% of the time.

An oil company has just announced that it has discovered new technology for detecting oil. The technology is 80% reliable. That is, if there is oil, the technology will signal "oil" 80% of the time. Let there also be a 1% chance that the technology erroneously detects oil, when in fact no oil exists.

In a report, use the above information to:

1. Prepare a table that shows the relevant probabilities.
2. Find the probability that, on a recent expedition, oil actually existed but the technology detected "no oil" in the area.



# 5

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 5.1 Distinguish between discrete and continuous random variables.
- LO 5.2 Describe the probability distribution for a discrete random variable.
- LO 5.3 Calculate and interpret summary measures for a discrete random variable.
- LO 5.4 Distinguish between risk-neutral, risk-averse, and risk-loving consumers.
- LO 5.5 Calculate and interpret summary measures to evaluate portfolio returns.
- LO 5.6 Describe the binomial distribution and compute relevant probabilities.
- LO 5.7 Describe the Poisson distribution and compute relevant probabilities.
- LO 5.8 Describe the hypergeometric distribution and compute relevant probabilities.

# Discrete Probability Distributions

In this chapter, we extend our discussion about probability by introducing the concept of a random variable. A random variable summarizes the results of an experiment in terms of numerical values. It can be classified as discrete or continuous depending on the range of values that it assumes. A discrete random variable assumes a countable number of distinct values, whereas a continuous random variable is characterized by uncountable values. In this chapter, we focus on discrete random variables. Examples include the number of credit cards carried by consumers, the number of foreclosures in a sample of 100 households, and the number of cars lined up at a toll booth. The range of possible values that a random variable assumes, and the associated probabilities, are discussed. We calculate summary measures for a random variable, including its mean, variance, and standard deviation. Using properties of random variables, we are able to apply these summary measures to describe portfolio returns. Finally, we discuss three widely used discrete probability distributions: the binomial, the Poisson, and the hypergeometric distributions.





## INTRODUCTORY CASE

### Available Staff for Probable Customers

In addition to its previous plan to shut 100 stores, Starbucks announced plans in 2008 to close 500 more U.S. locations (*The Wall Street Journal*, July 9, 2008). Executives claimed that a weak economy and higher gas and food prices led to a drop in domestic store traffic. Others speculate that Starbucks' rapid expansion produced a saturated market. The locations that will close are not profitable, are not expected to be profitable, or are located near an existing company-operated Starbucks.

Anne Jones, a manager at a local Starbucks, has been reassured by headquarters that her store will remain open. She is concerned about how other nearby closings might affect business at her store. Anne knows that a typical Starbucks customer visits the chain between 15 and 18 times a month, making it among the nation's most frequented retailers. She believes that her loyal Starbucks customers, along with displaced customers, will average 18 visits to the store over a 30-day month. To decide staffing needs, Anne knows that she needs a solid understanding about the probability distribution of customer arrivals. If too many employees are ready to serve customers, some employees will be idle, which is costly to the store. However, if not enough employees are available to meet demand, this could result in losing angry customers who choose not to wait for service.

Anne wants to use the above information to:

1. Calculate the expected number of visits from a typical Starbucks customer in a specified time period.
2. Calculate the probability that a typical Starbucks customer visits the chain a certain number of times in a specified time period.

A synopsis of this case is provided at the end of Section 5.5.

## 5.1 RANDOM VARIABLES AND DISCRETE PROBABILITY DISTRIBUTIONS

### LO 5.1

Distinguish between discrete and continuous random variables.

We often have to make important decisions in the face of uncertainty. For example, a car dealership has to determine the number of cars to hold on its lot when the actual demand for cars is unknown. Similarly, an investor has to select a portfolio when the actual outcomes of investment returns are not known. This uncertainty is captured by what we call a **random variable**. A random variable summarizes outcomes of an experiment with numerical values.

A **random variable** is a function that assigns numerical values to the outcomes of an experiment.

We generally use the letter  $X$  to denote a random variable. A **discrete random variable** assumes a countable number of distinct values such as  $x_1, x_2, x_3$ , and so on. It may assume either a finite or an infinite number of values. A **continuous random variable**, on the other hand, is characterized by uncountable values. In other words, a continuous random variable can take on any value within an interval or collection of intervals.

A **discrete random variable** assumes a countable number of distinct values, whereas a **continuous random variable** is characterized by uncountable values in an interval.

Recall from Chapter 4, the sample space  $S$  is a set of all outcomes of an experiment. Whenever some numerical values are assigned to these outcomes, a random variable  $X$  can be defined. Consider the following experiments, and some examples of discrete random variables (with their possible values shown) that are associated with the experiments:

Experiment 1. Rolling a six-sided die;  $S = \{1, 2, 3, 4, 5, 6\}$ .

Let  $X$  = Win \$10 if odd number, lose \$10 if even number; possible values =  $\{-10, 10\}$

Let  $X$  = Win \$10 if number less than 3, lose \$10 if number more than 4; possible values =  $\{-10, 0, 10\}$

Experiment 2. Two shirts are selected from the production line and each is either defective (D) or nondefective (N);  $S = \{(D, D), (D, N), (N, D), (N, N)\}$ .

Let  $X$  = the number of defective shirts; possible values =  $\{0, 1, 2\}$

Let  $X$  = the proportion of defective shirts; possible values =  $\{0, 1/2, 1\}$

Experiment 3. Reviewing a single mortgage application and deciding whether the client gets approved (A) or denied (D);  $S = \{A, D\}$ .

Let  $X$  = 1 for A and 0 for D; possible values =  $\{0, 1\}$

Let  $X$  = 1 for A and  $-1$  for D; possible values =  $\{-1, 1\}$

Experiment 4. Reviewing multiple mortgage applications and, for each client, deciding whether the client gets approved (A) or denied (D);  $S$  = the set of all possible infinite sequences whose elements are A or D.

Let  $X$  = the number of approvals; possible values =  $\{0, 1, 2, 3, \dots\}$

Let  $X$  = the squared number of approvals; possible values =  $\{0, 1, 4, 9, \dots\}$

The random variables defined for Experiments 1, 2, and 3 have a finite and countable number of values, while the two random variables defined for Experiment 4 have an infinite but countable number of values.

Sometimes, we can define a random variable *directly* by identifying its values with some numerical outcomes. For example, we may be interested in the number of students who get financial aid out of the 100 students who applied. Then the set of possible values of the random variable, equivalent to the sample space, is  $\{0, 1, \dots, 100\}$ . In a similar way, we can define a discrete random variable with an infinite number of values that it may take. For example, consider the number of cars that cross the Brooklyn Bridge between 9:00 am and 10:00 am on a Monday morning. Here the discrete random variable takes an infinite but countable number of values from  $\{0, 1, 2, \dots\}$ . Note that we cannot specify an upper bound on the observed number of cars.

Although, we explore discrete random variables in this chapter, random variables can also be continuous. For example, the time taken by a student to complete a 60-minute exam may assume any value between 0 and 60 minutes. Thus, the set of such values is uncountable; that is, it is impossible to put all real numbers from the interval  $[0, 60]$  in a sequence. Here, the random variable is continuous because the outcomes are uncountable. Some students may think that time in the earlier example is countable in seconds; however, this is not the case once we consider fractions of a second. We will discuss the details of continuous random variables in the next chapter.

## The Discrete Probability Distribution

Every random variable is associated with a **probability distribution** that describes it completely. It is common to define discrete random variables in terms of their **probability mass function** and continuous random variables in terms of their **probability density function**. Both variables can also be defined in terms of their **cumulative distribution function**.

### LO 5.2

Describe the probability distribution for a discrete random variable.

The **probability mass function** for a discrete random variable  $X$  is a list of the values of  $X$  with the associated probabilities; that is, the list of all possible pairs  $(x, P(X = x))$ . The **cumulative distribution function** of  $X$  is defined as  $P(X \leq x)$ .

For convenience, we will use terms like “probability distribution” and “distribution” for the probability mass function. We will do the same in the next chapter for the probability density function. In both chapters, we will use “cumulative probability distribution” for the cumulative distribution function.

We can view a discrete probability distribution in several ways, including tabular, algebraic, and graphical forms. Example 5.1 shows one of two tabular forms. In general, we can construct a table in two different ways. The first approach directly specifies the probability that the random variable assumes a specific value.

### EXAMPLE 5.1

Refer back to Experiment 1 of rolling a fair six-sided die, with the random variable defined as the number rolled. Present the probability distribution in a tabular form.

**SOLUTION:** A probability distribution for rolling a six-sided die is shown in Table 5.1.

**TABLE 5.1** Probability Distribution for Example 5.1

$x$	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

From Table 5.1, we can deduce, for instance, that  $P(X = 5)$  equals  $1/6$ . For that matter, the probability that  $X$  assumes any of the six possible values is  $1/6$ .

The probability distribution defined in Example 5.1 illustrates two components of all discrete probability distributions.

**TWO KEY PROPERTIES OF DISCRETE PROBABILITY DISTRIBUTIONS**

- The probability of each value  $x$  is a value between 0 and 1, or equivalently,  $0 \leq P(X = x) \leq 1$ .
- The sum of the probabilities equals 1. In other words,  $\sum P(X = x_i) = 1$  where the sum extends over all values  $x$  of  $X$ .

The second tabular view of a probability distribution is based on the cumulative probability distribution. The cumulative probability distribution is convenient when we are interested in finding the probability that the random variable assumes a range of values rather than a specific value. For the random variable defined in Example 5.1, the cumulative probability distribution is shown in Table 5.2.

**TABLE 5.2** Cumulative Probability Distribution for Example 5.1

$x$	1	2	3	4	5	6
$P(X \leq x)$	1/6	2/6	3/6	4/6	5/6	6/6

If we are interested in finding the probability of rolling a four or less,  $P(X \leq 4)$ , we see from the cumulative probability distribution that this probability is 4/6. With the earlier probability representation, we would add up the probabilities to compute  $P(X \leq 4)$  as

$$P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1/6 + 1/6 + 1/6 + 1/6 = 4/6.$$

At the same time, we can use the cumulative probability distribution to find the probability that the random variable assumes a specific value. For example,  $P(X = 3)$  can be found as  $P(X \leq 3) - P(X \leq 2) = 3/6 - 2/6 = 1/6$ .

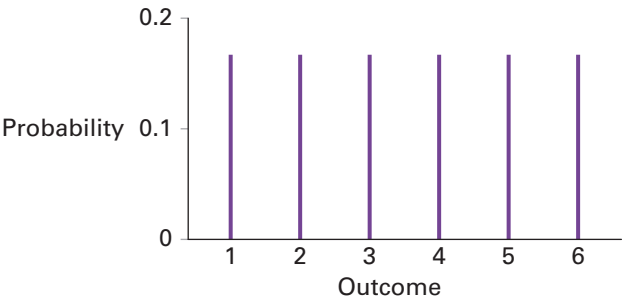
In many instances, we can express a probability distribution by applying an algebraic formula. A formula representation of the probability distribution for the random variable defined in Example 5.1 is:

$$P(X = x) = \begin{cases} 1/6 & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, from the formula we can ascertain that  $P(X = 5) = 1/6$  and  $P(X = 7) = 0$ .

In order to graphically depict a probability distribution, we place all values  $x$  of  $X$  on the horizontal axis and the associated probabilities  $P(X = x)$  on the vertical axis. We then draw a line segment that emerges from each  $x$  and ends where its height equals  $P(X = x)$ . Figure 5.1 graphically illustrates the probability distribution for the random variable defined in Example 5.1.

**FIGURE 5.1**  
Probability distribution  
when rolling a  
six-sided die



The probability distribution in Figure 5.1 is an example of a **discrete uniform distribution**, which has the following characteristics:

- The distribution has a finite number of specified values.
- Each value is equally likely.
- The distribution is symmetric.

### EXAMPLE 5.2

The number of homes that a realtor sells over a one-month period has the probability distribution shown in Table 5.3.

**TABLE 5.3** Probability Distribution for the Number of Houses Sold

Number of Houses Sold	Probability
0	0.30
1	0.50
2	0.15
3	0.05

- Is this a valid probability distribution?
- What is the probability that the realtor does not sell any houses in a one-month period?
- What is the probability that the realtor sells at most one house in a one-month period?
- What is the probability that the realtor sells at least two houses in a one-month period?
- Graphically depict the probability distribution and comment on its symmetry/skewness.

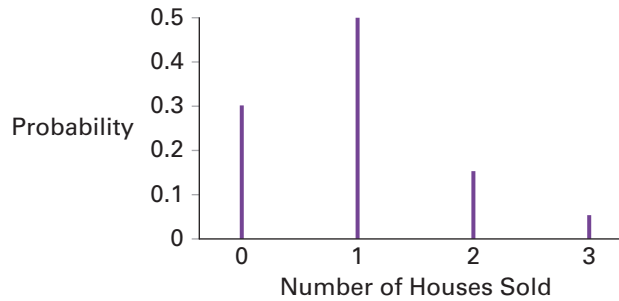
#### SOLUTION:

- We first note that the random variable  $X$  denotes the number of houses that the realtor sells over a one-month period. This variable assumes the values 0 through 3. The probability distribution is valid because it satisfies the following two conditions: (1) all probabilities fall between 0 and 1, and (2) the probabilities sum to 1 ( $0.30 + 0.50 + 0.15 + 0.05 = 1$ ).
- In order to find the probability that the realtor does not sell any houses in a one-month period, we find  $P(X = 0) = 0.30$ .
- We find the probability that a realtor sells at most one house as:  $P(X \leq 1) = P(X = 0) + P(X = 1) = 0.30 + 0.50 = 0.80$ .
- We find the probability that the realtor sells at least two houses as:  $P(X \geq 2) = P(X = 2) + P(X = 3) = 0.15 + 0.05 = 0.20$ .

Note that since the sum of the probabilities over all values of  $X$  equals 1, we can also find the above probability as  $P(X \geq 2) = 1 - P(X \leq 1) = 1 - 0.80 = 0.20$ .

- The graph in Figure 5.2 shows that the distribution is not symmetric; rather, it is positively skewed. There are small chances of selling two or three houses in a one-month period. The most likely outcome by far is selling one house over a one-month period, with a probability of 0.50.



**FIGURE 5.2** Probability distribution for the number of houses sold

## EXERCISES 5.1

### Mechanics

1. Consider the following discrete probability distribution.

$x$	15	22	34	40
$P(X = x)$	0.14	0.40	0.26	0.20

- Is this a valid probability distribution? Explain.
  - Graphically depict this probability distribution.
  - What is the probability that the random variable  $X$  is less than 40?
  - What is the probability that the random variable  $X$  is between 10 and 30?
  - What is the probability that the random variable  $X$  is greater than 20?
2. Consider the following discrete probability distribution.

$x$	-25	-15	10	20
$P(X = x)$	0.35	0.10		0.10

- Complete the probability distribution.
  - Graphically depict the probability distribution and comment on the symmetry of the distribution.
  - What is the probability that the random variable  $X$  is negative?
  - What is the probability that the random variable  $X$  is greater than -20?
  - What is the probability that the random variable  $X$  is less than 20?
3. Consider the following cumulative probability distribution.

$x$	0	1	2	3	4	5
$P(X \leq x)$	0.15	0.35	0.52	0.78	0.84	1

- Calculate  $P(X \leq 3)$ .
  - Calculate  $P(X = 3)$ .
  - Calculate  $P(2 \leq X \leq 4)$ .
4. Consider the following cumulative probability distribution.

$x$	-25	0	25	50
$P(X \leq x)$	0.25	0.50	0.75	1

- Calculate  $P(X \leq 0)$ .
- Calculate  $P(X = 50)$ .
- Is this a discrete uniform distribution? Explain.

### Applications

5. Identify the possible values of the following random variables. Which of the random variables are discrete?
- The numerical grade a student receives in a course.
  - The grade point average of a student.
  - The salary of an employee, defined in figures (4 figure, 5 figure, etc.).
  - The salary of an employee defined in dollars.
6. Identify the possible values of the following random variables. Which of the random variables are discrete?
- The advertised size of a round Domino's pizza.
  - The actual size of a round Domino's pizza.
  - The number of daily visitors to Yosemite National Park.
  - The age of a visitor to Yosemite National Park.
7. India is the second most populous country in the world, with a population of over 1 billion people. Although the government has offered various incentives for population control, some argue that the birth rate, especially in rural India, is still too high to be sustainable. A demographer assumes the following probability distribution for the household size in India.

Household Size	Probability
1	0.05
2	0.09
3	0.12
4	0.24
5	0.25
6	0.12
7	0.07
8	0.06



- What is the probability that there are less than 5 members in a household in India?
  - What is the probability that there are 5 or more members in a household in India?
  - What is the probability that the number of members in a household in India is strictly between 3 and 6?
  - Graphically depict this probability distribution and comment on its symmetry.
8. A financial analyst creates the following probability distribution for the performance of an equity income mutual fund.
- | Performance | Numerical Score | Probability |
|-------------|-----------------|-------------|
| Very poor   | 1               | 0.14        |
| Poor        | 2               | 0.43        |
| Neutral     | 3               | 0.22        |
| Good        | 4               | 0.16        |
| Very good   | 5               | 0.05        |
- Comment on the optimism or pessimism depicted in the analyst's estimates.
  - Convert the above probability distribution to a cumulative probability distribution.
  - What is the probability that this mutual fund will do at least "Good"?
9. A basketball player is fouled while attempting to make a basket and receives two free throws. The opposing coach believes there is a 55% chance that the player will miss both shots, a 25% chance that he will make one of the shots, and a 20% chance that he will make both shots.
- Construct the appropriate probability distribution.
  - What is the probability that he makes no more than one of the shots?
  - What is the probability that he makes at least one of the shots?
10. In early 2010, leading U.S. stock markets tumbled more than 2.5% as U.S. consumer confidence fell to its lowest level since August 2009 (*BBC News*, July 16, 2010). Given fresh economic data, an economist believes there is a 35% chance that consumer confidence will fall below 62 and only a 25% chance that it will rise above 65. The economist defines the confidence score as 1 if consumer confidence is below 62, 2 if it is between 62 and 65, and 3 if it is above 65.

- According to the economist, what is the probability that the confidence score is 2?
  - According to the economist, what is the probability that the confidence score is not 1?
11. Professor Sanchez has been teaching Principles of Economics for over 25 years. He uses the following scale for grading.

Grade	Numerical Score	Probability
A	4	0.10
B	3	0.30
C	2	0.40
D	1	0.10
F	0	0.10

- Depict the above probability distribution graphically. Comment on whether or not the probability distribution is symmetric.
  - Convert the above probability distribution to a cumulative probability distribution.
  - What is the probability of earning at least a B in Professor Sanchez's course?
  - What is the probability of passing Professor Sanchez's course?
12. Jane Wormley is a professor of management at a university. She expects to be able to use her grant money to fund up to two students for research assistance. While she realizes that there is a 5% chance that she may not be able to fund any student, there is an 80% chance that she will be able to fund two students.
- What is the probability that Jane will fund one student?
  - Construct a cumulative probability distribution of the random variable defined as the number of students that Jane will be able to fund.
13. Fifty percent of the customers who go to Sears Auto Center for tires buy four tires and 30% buy two tires. Moreover, 18% buy fewer than two tires, with 5% buying none.
- What is the probability that a customer buys three tires?
  - Construct a cumulative probability distribution for the number of tires bought.

## 5.2 EXPECTED VALUE, VARIANCE, AND STANDARD DEVIATION

### LO 5.3

The analysis of probability distributions is useful because it allows us to calculate various probabilities associated with the different values that the random variable assumes. In addition, it helps us calculate summary measures for a random variable. These summary measures include the mean, the variance, and the standard deviation.

Calculate and interpret summary measures for a discrete random variable.

## Expected Value

One of the most important probabilistic concepts in statistics is that of the **expected value**, also referred to as the **population mean**. The expected value of the discrete random variable  $X$ , denoted by  $E(X)$  or simply  $\mu$ , is a weighted average of all possible values of  $X$ . Before we present its formula, we would like to point out that the expected value of a random variable should not be confused with its most probable value. As we will see later, the expected value is, in general, not even one of the possible values of the random variable. We can think of the expected value as the long-run average value of the random variable over infinitely many independent repetitions of an experiment. Consider a simple experiment with a fair coin, where you win \$10 if it is heads and lose \$10 if it is tails. If you flip the coin many times, the expected gain is \$0, which is neither of the two possible values, namely \$10 or -\$10.

### EXPECTED VALUE OF A DISCRETE RANDOM VARIABLE

For a discrete random variable  $X$  with values  $x_1, x_2, x_3, \dots$ , which occur with probabilities  $P(X = x_i)$ , the **expected value** of  $X$  is calculated as

$$E(X) = \mu = \sum x_i P(X = x_i).$$

## Variance and Standard Deviation

The mean  $\mu$  of the random variable  $X$  provides us with a measure of the central location of the distribution of  $X$ , but it does not give us information on how the various values are dispersed from  $\mu$ . We need a measure that indicates whether the values of  $X$  are clustered about  $\mu$  or widely scattered from  $\mu$ .

### VARIANCE AND STANDARD DEVIATION OF A DISCRETE RANDOM VARIABLE

For a discrete random variable  $X$  with values  $x_1, x_2, x_3, \dots$ , which occur with probabilities  $P(X = x_i)$ , the **variance** of  $X$ , denoted as  $Var(X)$  or  $\sigma^2$ , is calculated as

$$Var(X) = \sigma^2 = \sum (x_i - \mu)^2 P(X = x_i).$$

The **standard deviation** of  $X$ , denoted as  $SD(X)$  or  $\sigma$ , is the positive square root of the variance of  $X$  or, equivalently,  $SD(X) = \sigma = \sqrt{\sigma^2}$ .

### EXAMPLE 5.3

Brad Williams is the owner of a large car dealership in Chicago. Brad decides to construct an incentive compensation program that equitably and consistently compensates employees on the basis of their performance. He offers an annual bonus of \$10,000 for superior performance, \$6,000 for good performance, \$3,000 for fair performance, and \$0 for poor performance. Based on prior records, he expects an employee to perform at superior, good, fair, and poor performance levels with probabilities 0.15, 0.25, 0.40, and 0.20, respectively. Table 5.4 lists the bonus amount, performance type, and the corresponding probabilities.

**TABLE 5.4** Probability Distribution for Compensation Program

Bonus (in \$1,000s)	Performance Type	Probability
\$10	Superior	0.15
6	Good	0.25
3	Fair	0.40
0	Poor	0.20

- Calculate the expected value of the annual bonus amount.
- Calculate the variance and the standard deviation of the annual bonus amount.
- What is the total annual amount that Brad can expect to pay in bonuses if he has 25 employees?

**SOLUTION:**

- Let the random variable  $X$  denote the bonus amount (in \$1,000s) for an employee. The first and second columns of Table 5.5 represent the probability distribution of  $X$ . The calculations for the mean are provided in the third column. We weigh each outcome by its respective probability,  $x_i P(X = x_i)$ , and then sum these weighted values. Thus, as shown at the bottom of the third column,  $E(X) = \mu = \sum x_i P(X = x_i) = 4.2$ , or \$4,200. Note that the expected value is not one of the possible values of  $X$ ; that is, none of the employees will earn a bonus of \$4,200. This outcome reinforces the interpretation of expected value as a long-run average.

**TABLE 5.5** Calculations for Example 5.3

Value, $x_i$	Probability, $P(X = x_i)$	Weighted Value, $x_i P(X = x_i)$	Weighted Squared Deviation, $(x_i - \mu)^2 P(X = x_i)$
10	0.15	$10 \times 0.15 = 1.5$	$(10 - 4.2)^2 \times 0.15 = 5.05$
6	0.25	$6 \times 0.25 = 1.5$	$(6 - 4.2)^2 \times 0.25 = 0.81$
3	0.40	$3 \times 0.40 = 1.2$	$(3 - 4.2)^2 \times 0.40 = 0.58$
0	0.20	$0 \times 0.20 = 0$	$(0 - 4.2)^2 \times 0.20 = 3.53$
		Total = 4.2	Total = 9.97

- The last column of Table 5.5 shows the calculation for the variance. We first calculate each  $x_i$ 's squared difference from the mean  $(x_i - \mu)^2$ , weigh each value by the appropriate probability,  $(x_i - \mu)^2 P(X = x_i)$ , and then sum these weighted squared differences. Thus, as shown at the bottom of the last column,  $\text{Var}(X) = \sigma^2 = \sum (x_i - \mu)^2 P(X = x_i) = 9.97$ , or 9.97 (in (\$1,000s)<sup>2</sup>). The standard deviation is the positive square root of the variance,  $SD(X) = \sigma = \sqrt{9.97} = 3.158$ , or \$3,158.
- Note that the expected bonus of an employee is \$4,200. Since Brad has 25 employees, he can expect to pay  $\$4,200 \times 25 = \$105,000$  in bonuses.

## Risk Neutrality and Risk Aversion

An important concept in economics, finance, and psychology relates to the behavior of consumers under uncertainty. Consumers are said to be **risk neutral** if they are indifferent to risk and care only about their expected gains. They are said to be **risk-averse** if they care about risk and, if confronted with two choices with the same expected gains, they prefer the one with lower risk. In other words, a risk-averse consumer will take a risk only if it entails a suitable compensation. Consider a seemingly fair gamble where you flip a coin and get \$10 if it is heads and lose \$10 if it is tails, resulting in an expected gain of zero ( $10 \times 0.5 - 10 \times 0.5 = 0$ ). A risk-neutral consumer is indifferent about participating in this gamble. For a risk-averse consumer, the pain associated with losing \$10 is more than the pleasure of winning \$10. Therefore, the consumer will not want to participate in this seemingly fair gamble because there is no reward to compensate for the risk. Example 5.4 expands on this type of consumer behavior.

### LO 5.4

Distinguish between risk-neutral, risk-averse, and risk-loving consumers.

A **risk-neutral consumer** completely ignores risk and makes his/her decisions solely on the basis of expected gains. A **risk-averse consumer** demands a positive expected gain as compensation for taking risk. This compensation increases with the level of risk taken and the degree of risk aversion. Finally, a **risk-loving consumer** may be willing to take a risk even if the expected gain is negative.

### EXAMPLE 5.4

You have a choice of receiving \$1,000 in cash or receiving a beautiful painting from your grandmother. The actual value of the painting is uncertain. You are told that the painting has a 20% chance of being worth \$2,000, a 50% chance of being worth \$1,000, and a 30% chance of being worth \$500. What should you do?

**SOLUTION:** Let the random variable  $X$  represent the worth of the painting. Given the above information, we define the probability distribution as shown in Table 5.6.

**TABLE 5.6** Probability Distribution for the Value of the Painting

$x$	$P(X = x)$
\$2,000	0.20
1,000	0.50
500	0.30

We calculate the expected value as

$$E(X) = \sum x_i P(X = x_i) = \$2,000 \times 0.20 + \$1,000 \times 0.50 + \$500 \times 0.30 = \$1,050.$$

Since the expected value of the painting is more than \$1,000, it may appear that the right choice is to pick the painting over \$1,000 in cash. This choice, however, is based entirely on the expected value of the painting, paying no attention to risk. While the expected value of \$1,050 is more than \$1,000, the painting entails some risk. For instance, there is a 30% chance that it may be worth only \$500. Therefore, a risk-neutral consumer will take the painting because its expected value exceeds the risk-free cash value of \$1,000. This consumer is not concerned with risk. A risk lover will be thrilled to take the painting. For a risk-averse consumer, however, the decision is not clear-cut. It depends on the risk involved in picking the painting and how much he/she wants to be compensated for this risk. One way to resolve this issue is to define the utility function of the consumer, which in essence conveys the degree of risk aversion. A risk-averse consumer will pick the risky prospect if the expected utility (not the expected money) of the risky prospect exceeds the utility of a risk-free alternative. Further details are beyond the scope of this text.

## EXERCISES 5.2

### Mechanics

14. Calculate the mean, the variance, and the standard deviation of the following discrete probability distribution.

$x$	5	10	15	20
$P(X = x)$	0.35	0.30	0.20	0.15

15. Calculate the mean, the variance, and the standard deviation of the following discrete probability distribution.

$x$	-23	-17	-9	-3
$P(X = x)$	0.50	0.25	0.15	0.10

## Applications

16. The number of homes that a realtor sells over a one-month period has the following probability distribution.

Number of Houses Sold	Probability
0	0.30
1	0.50
2	0.15
3	0.05

- On average, how many houses is the Realtor expected to sell over a one-month period?
  - What is the standard deviation of this probability distribution?
17. A marketing firm is considering making up to three new hires. Given its specific needs, the management feels that there is a 60% chance of hiring at least two candidates. There is only a 5% chance that it will not make any hires and a 10% chance that it will make all three hires.
- What is the probability that the firm will make at least one hire?
  - Find the expected value and the standard deviation of the number of hires.
18. An analyst has developed the following probability distribution for the rate of return for a common stock.

Scenario	Probability	Rate of Return
1	0.30	−5%
2	0.45	0%
3	0.25	10%

- Calculate the expected rate of return.
  - Calculate the variance and the standard deviation of this probability distribution.
19. Organizers of an outdoor summer concert in Toronto are concerned about the weather conditions on the day of the concert. They will make a profit of \$25,000 on a clear day and \$10,000 on a cloudy day. They will make a loss of \$5,000 if it rains. The weather channel has predicted a 60% chance of rain on the day of the concert. Calculate the expected profit from the concert if the likelihood is 10% that it will be sunny and 30% that it will be cloudy.
20. Mark Underwood is a professor of Economics at Indiana University. He has been teaching Principles of Economics for over 25 years. Professor Underwood uses the following scale for grading.

Grade	Probability
A	0.10
B	0.30
C	0.40
D	0.10
F	0.10

Calculate the expected numerical grade in Professor Underwood's class using 4.0 for A, 3.0 for B, etc.

- The manager of a publishing company plans to give a \$20,000 bonus to the top 15%, \$10,000 to the next 30%, and \$5,000 to the next 10% of sales representatives. If the publishing company has a total of 200 sales representatives, what is the expected bonus that the company will pay?
- An electronics store sells additional warranties on its Blu-ray players. Twenty percent of the buyers buy the limited warranty for \$25 and 5% buy the extended warranty for \$60. What is the expected revenue for the store if it sells 120 players?
- You are considering buying insurance for your new laptop computer, which you have recently bought for \$1,500. The insurance premium for three years is \$80. Over the three-year period there is an 8% chance that your laptop computer will require work worth \$400, a 3% chance that it will require work worth \$800, and a 2% chance that it will completely break down with a scrap value of \$100. Should you buy the insurance? (Assume risk neutrality.)
- Four years ago, Victor purchased a very reliable automobile (as rated by a reputable consumer advocacy publication). His warranty has just expired, but the manufacturer has just offered him a 5-year, bumper-to-bumper warranty extension. The warranty costs \$3,400. Victor constructs the following probability distribution with respect to anticipated costs if he chooses not to purchase the extended warranty.

Cost (in \$)	Probability
1,000	0.25
2,000	0.45
5,000	0.20
10,000	0.10

- Calculate Victor's expected cost.
  - Given your answer in part a, should Victor purchase the extended warranty? (Assume risk neutrality.) Explain.
25. An investor considers investing \$10,000 in the stock market. He believes that the probability is 0.30 that the economy will improve, 0.40 that it will stay the same, and 0.30 that it will deteriorate. Further, if the economy improves, he expects his investment to grow to \$15,000, but it can also go down to \$8,000 if the economy deteriorates. If the economy stays the same, his investment will stay at \$10,000.
- What is the expected value of his investment?
  - What should the investor do if he is risk neutral?
  - Is the decision clear-cut if he is risk averse? Explain.
26. You are considering two mutual funds as an investment. The possible returns for the funds are dependent on the state of the economy and are given in the accompanying table.

State of the Economy	Fund 1	Fund 2
Good	20%	40%
Fair	10%	20%
Poor	−10%	−40%

You believe that the likelihood is 20% that the economy will be good, 50% that it will be fair, and 30% that it will be poor.

- Find the expected value and the standard deviation of returns for Fund 1.
  - Find the expected value and the standard deviation of returns for Fund 2.
  - Which fund will you pick if you are risk averse? Explain.
27. Investment advisors recommend risk reduction through international diversification. International investing allows you to take advantage of the potential for growth in foreign economies, particularly in emerging markets. Janice Wong

is considering investment in either Europe or Asia. She has studied these markets and believes that both markets will be influenced by the U.S. economy, which has a 20% chance for being good, a 50% chance for being fair, and a 30% chance for being poor. Probability distributions of the returns for these markets are given in the accompanying table.

State of the U.S. Economy	Returns in Europe	Returns in Asia
Good	10%	18%
Fair	6%	10%
Poor	−6%	−12%

- Find the expected value and the standard deviation of returns in Europe and Asia.
- What will Janice pick as an investment if she is risk neutral?
- Discuss Janice's decision if she is risk averse.

## LO 5.5

## 5.3 PORTFOLIO RETURNS

Calculate and interpret summary measures to evaluate portfolio returns.

As discussed in Chapter 3, we often evaluate investment opportunities using expected return as a measure of reward, and variance or standard deviation of return as a measure of risk. Consider two assets where Asset A is expected to have a return of 12% and Asset B is expected to have a return of 8% for the year. While Asset A is attractive in terms of its reward, an investor may still choose Asset B over Asset A if the risk associated with Asset A is too high. In other words, both reward as well as risk are relevant for evaluating the investment.

So far we have considered assets separately. However, most investors hold a **portfolio** of assets, where a portfolio is defined as a collection of assets such as stocks and bonds. As in the case of an individual asset, an investor is concerned about the reward as well as the risk of a portfolio. The derivations of the expected return and the variance of a portfolio depend on some important results regarding the joint distribution of random variables.

Let  $X$  and  $Y$  represent two random variables of interest, denoting, say, the returns of two assets. Since an investor may have invested in both assets, we would like to evaluate the portfolio return formed by a linear combination of  $X$  and  $Y$ . The following properties for random variables are useful in evaluating portfolio returns.

### Properties of Random Variables

Given two random variables  $X$  and  $Y$ , the expected value of their sum,  $E(X + Y)$ , is equal to the sum of their individual expected values,  $E(X)$  and  $E(Y)$ , or

$$E(X + Y) = E(X) + E(Y).$$

Using algebra, it can be shown that the variance of the sum for two random variables,  $Var(X + Y)$ , yields

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y),$$

where  $Cov$  is the covariance between the random variables  $X$  and  $Y$ .

For given constants  $a$  and  $b$ , the above results are extended as:

$$E(aX + bY) = aE(X) + bE(Y), \text{ and} \\ Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y).$$



## Expected Return, Variance, and Standard Deviation for Portfolio Returns

We are now in a position to derive the expected return and the variance for a portfolio based on these properties. For the sake of simplicity, consider a portfolio consisting of only two assets, Asset A and Asset B. These assets, for instance, may represent stocks and bonds. Following popular notation in finance, let  $R_A$  and  $R_B$  be the random variables of interest, representing the returns for assets A and B, respectively. It is important to note that a portfolio is described not only by its assets but also by its **portfolio weights**. Consider a portfolio with a total value of \$5,000, with \$1,000 invested in Asset A and \$4,000 in Asset B. The portfolio weights are derived as

$$w_A = \frac{1,000}{5,000} = 0.20 \quad \text{and} \quad w_B = \frac{4,000}{5,000} = 0.80.$$

Note that the portfolio weights add up to one; that is,  $w_A + w_B = 0.20 + 0.80 = 1$ . We then define the portfolio return  $R_p$  as a linear combination of the individual returns,

$$R_p = w_A R_A + w_B R_B.$$

### PORTFOLIO EXPECTED RETURN

Given a portfolio with two assets, Asset A and Asset B, the **expected return for the portfolio**  $E(R_p)$  is computed as

$$E(R_p) = w_A E(R_A) + w_B E(R_B),$$

where  $w_A$  and  $w_B$  are the **portfolio weights** ( $w_A + w_B = 1$ ) and  $E(R_A)$  and  $E(R_B)$  are the expected returns on assets A and B, respectively.

### EXAMPLE 5.5

Consider an investment portfolio of \$40,000 in Stock A and \$60,000 in Stock B. Calculate the expected return for this portfolio based on the information in Table 5.7.

**TABLE 5.7** Data for Example 5.5

Stock A	Stock B
$E(R_A) = \mu_A = 9.5\%$	$E(R_B) = \mu_B = 7.6\%$
$SD(R_A) = \sigma_A = 12.93\%$	$SD(R_B) = \sigma_B = 8.20\%$
$Cov(R_A, R_B) = \sigma_{AB} = 18.60\%$	

**SOLUTION:** First we compute the portfolio weights. Since \$40,000 is invested in Stock A and \$60,000 in Stock B, we compute

$$w_A = \frac{40,000}{100,000} = 0.40 \quad \text{and} \quad w_B = \frac{60,000}{100,000} = 0.60.$$

Thus, using the formula for portfolio expected return, we solve:

$$E(R_p) = (0.40 \times 9.5\%) + (0.60 \times 7.6\%) = 3.80\% + 4.56\% = 8.36\%.$$

Note that the portfolio expected return of 8.36% is lower than the expected return of investing entirely in Stock A with an expected return of 9.5%, yet higher than the expected return of investing entirely in Stock B with an expected return of 7.6%.

The risk of the portfolio depends not only on the individual risks of the assets but also on the interplay between the asset returns. For example, if one asset does poorly, the second asset may serve as an offsetting factor to stabilize the risk of the overall portfolio. This result will work as long as the return of the second asset is not perfectly correlated with the return of the first asset. Similar to the covariance  $Cov(x, y) = \sigma_{xy}$  introduced in Chapter 3, the covariance  $Cov(R_A, R_B) = \sigma_{AB}$  helps determine whether the linear relationship between the asset returns is positive, negative, or zero. Recall that an easier measure to interpret is the correlation coefficient  $\rho$  which describes both the direction and the strength of the linear relationship between two random variables. The value of the correlation coefficient falls between  $-1$  and  $1$ . The closer the value is to  $1$ , the stronger is the positive relationship between the variables. Similarly, the closer the value is to  $-1$ , the stronger is the negative relationship between the variables. Let  $\rho_{AB} = \frac{\sigma_{AB}}{\sigma_A \sigma_B}$  denote the correlation coefficient between the returns  $R_A$  and  $R_B$ .

With information on either the covariance or the correlation coefficient for the two returns, we can now determine the portfolio variance of return.

#### PORTFOLIO VARIANCE

The **portfolio variance**,  $Var(R_p) = Var(w_A R_A + w_B R_B)$ , is calculated as

$$Var(R_p) = w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2w_A w_B \sigma_{AB}$$

or, equivalently,

$$Var(R_p) = w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2w_A w_B \rho_{AB} \sigma_A \sigma_B$$

where  $\sigma_A^2$  and  $\sigma_B^2$  are the variances of the returns for Asset A and Asset B, respectively,  $\sigma_{AB}$  is the covariance between the returns for Asset A and Asset B, and  $\rho_{AB}$  is the correlation coefficient between the returns for Asset A and Asset B.

The **standard deviation of return**  $SD(R_p)$  is then calculated as the positive square root of the portfolio variance.

#### EXAMPLE 5.6

Using the information in Example 5.5, answer the following questions.

- Calculate and interpret the correlation coefficient between the returns for Stock A and Stock B.
- Calculate the portfolio variance using both formulas.
- Calculate the portfolio standard deviation.
- Comment on the findings.

#### SOLUTION:

- We calculate the correlation coefficient as  $\rho_{AB} = \frac{\sigma_{AB}}{\sigma_A \sigma_B} = \frac{18.60}{12.93 \times 8.20} = 0.1754$ . This value implies that the returns have a positive linear relationship, though the magnitude of the relationship is weak ( $\rho_{AB}$  is well below 1).

- Using the first formula for portfolio variance, we calculate

$$\begin{aligned} Var(R_p) &= w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2w_A w_B \sigma_{AB} \\ &= (0.40)^2 (12.93)^2 + (0.60)^2 (8.20)^2 + 2(0.40)(0.60)(18.60) \\ &= 26.75 + 24.21 + 8.93 \\ &= 59.89. \end{aligned}$$

Using the alternative formula for portfolio variance, we calculate

$$\begin{aligned}
 \text{Var}(R_p) &= w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2w_A w_B \rho_{AB} \sigma_A \sigma_B \\
 &= (0.40)^2 (12.93)^2 + (0.60)^2 (8.20)^2 \\
 &\quad + 2(0.40)(0.60)(0.1754)(12.93)(8.20) \\
 &= 26.75 + 24.21 + 8.93 \\
 &= 59.89.
 \end{aligned}$$

Using either formula, the variance of portfolio return is 59.89 (%)<sup>2</sup>.

- c. The portfolio standard deviation is  $SD(R_p) = \sqrt{59.89} = 7.74$ , or 7.74%.
- d. We note how the portfolio standard deviation of 7.74%, a measure of risk, is lower than the risk of 12.93% of investing entirely in Stock A as well as the risk of 8.20% of investing entirely in Stock B. This occurs because the returns of Stock A and Stock B have a correlation of only 0.1754. This example highlights the benefits of properly diversifying your portfolio in order to reduce risk. In general, the benefits of diversification depend on the correlation between the assets: the lower the correlation, the larger the benefit.

## EXERCISES 5.3

28. What are the portfolio weights for a portfolio that has 100 shares of Stock X that sell for \$20 per share and 200 shares of Stock Y that sell for \$12 per share?
29. You own a portfolio that has \$4,400 invested in stocks and \$5,600 invested in bonds. What is the expected return of the portfolio if stocks and bonds are expected to yield a return of 9% and 5%, respectively?
30. A portfolio has \$200,000 invested in Asset X and \$300,000 in Asset Y. Consider the summary measures in the following table.

Measures	Asset X	Asset Y
Expected Return (%)	8	12
Standard deviation (%)	12	20
Correlation coefficient	0.40	

- a. Calculate the portfolio weights for assets X and Y.
  - b. Calculate the expected return for the portfolio.
  - c. Calculate the standard deviation for the portfolio.
31. An analyst has predicted the following returns for Stock A and Stock B in three possible states of the economy.

State	Probability	A	B
Boom	0.3	0.15	0.25
Normal	0.5	0.10	0.20
Recession	?	0.02	0.01

- a. What is the probability of a recession?
  - b. Calculate the expected return for Stock A and Stock B.
  - c. Calculate the expected return for a portfolio that is invested 55% in A and 45% in B.
32. A pension fund manager is considering three mutual funds for investment. The first one is a stock fund, the second is a bond fund and the third is a money market fund. The money market fund yields a risk-free return of 4%. The inputs for the risky funds are given in the following table.

Fund	Expected Return	Standard Deviation
Stock fund	14%	26%
Bond fund	8%	14%

The correlation coefficient between the stock and the bond funds is 0.20.

- a. What is the expected return and the variance for a portfolio that invests 60% in the stock fund and 40% in the bond fund?
  - b. What is the expected return and the variance for a portfolio that invests 60% in the stock fund and 40% in the money market fund? [Hint: Note that the correlation coefficient between the portfolio and the money market fund is zero.]
  - c. Compare the portfolios in parts a and b with a portfolio that is invested entirely in the bond fund.
33. You have \$400,000 invested in a well-diversified portfolio. You inherit a house that is presently worth \$200,000. Consider the summary measures in the following table:

Investment	Expected Return	Standard Deviation
Old portfolio	6%	16%
House	8%	20%

The correlation coefficient between your portfolio and the house is 0.38.

- a. What is the expected return and the standard deviation for your portfolio comprising your old portfolio and the house?
- b. Suppose you decide to sell the house and use the proceeds of \$200,000 to buy risk-free T-bills that promise a 3% rate of return. Calculate the expected return and the standard deviation for the resulting portfolio. [Hint: Note that the correlation coefficient between any asset and the risk-free T-bills is zero.]

Describe the binomial distribution and compute relevant probabilities.

Different types of experiments generate different probability distributions. In the next three sections, we discuss three special cases: the binomial, the Poisson, and the hypergeometric probability distributions. Here we focus on the binomial distribution. Before we can discuss the binomial distribution, we first must ensure that the experiment satisfies the conditions of a **Bernoulli process**, which is a particular type of experiment named after the person who first described it, the Swiss mathematician James Bernoulli (1654–1705).

A **Bernoulli process** consists of a series of  $n$  independent and identical trials of an experiment such that on each trial:

- There are only two possible outcomes, conventionally labeled success and failure; and
- The probabilities of success and failure remain the same from trial to trial.

We use  $p$  to denote the probability of success, and therefore,  $1 - p$  is the probability of failure.

A **binomial random variable** is defined as the number of successes achieved in the  $n$  trials of a Bernoulli process. The possible values of a binomial random variable include  $0, 1, \dots, n$ . Many experiments fit the conditions of a Bernoulli process. For instance:

- A bank grants or denies a loan to a mortgage applicant.
- A consumer either uses or does not use a credit card.
- An employee travels or does not travel by public transportation.
- A life insurance policy holder dies or does not die.
- A drug is either effective or ineffective.
- A college graduate applies or does not apply to graduate school.

Our goal is to attach probabilities to various outcomes of a Bernoulli process. The result is a **binomial probability distribution**, or simply, a **binomial distribution**.

A **binomial random variable**  $X$  is defined as the number of successes achieved in the  $n$  trials of a Bernoulli process. A **binomial distribution** shows the probabilities associated with the possible values of  $X$ .

We will eventually arrive at a general formula that helps us derive a binomial distribution. First, however, we will use a specific example and construct a **probability tree** in order to illustrate the possible outcomes and their associated probabilities.

### EXAMPLE 5.7

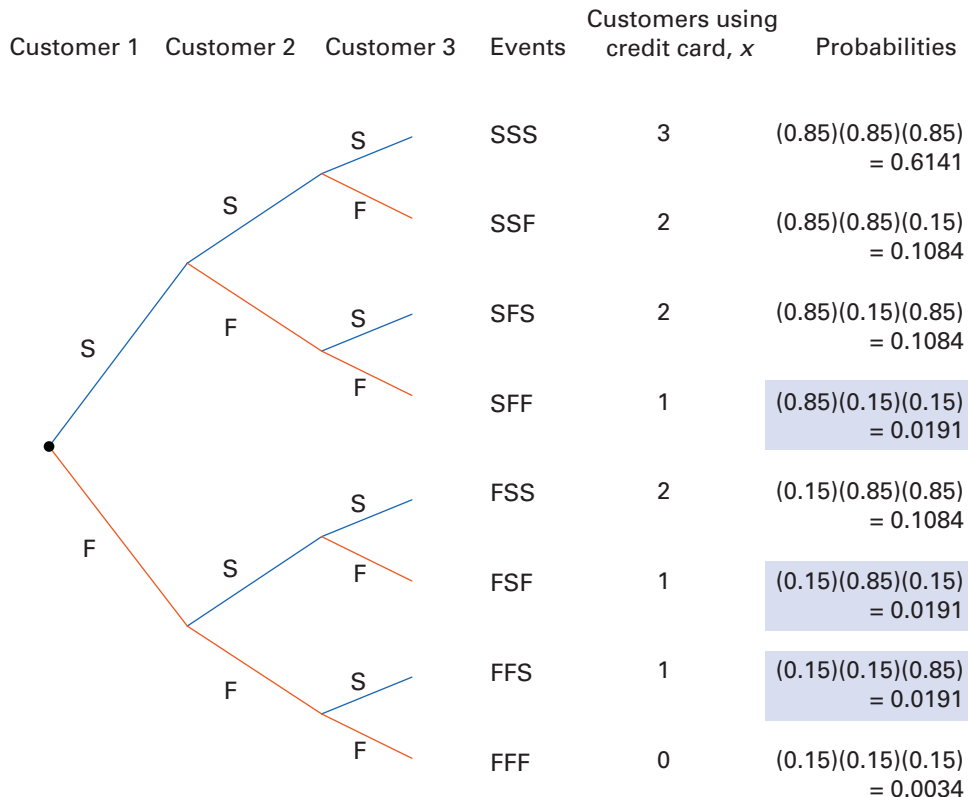
From past experience, a manager of an upscale shoe store knows that 85% of her customers will use a credit card when making purchases. Suppose three customers are in line to make a purchase.

- Does this example satisfy the conditions of a Bernoulli process?
- Construct a probability tree.
- Using the probability tree, derive the binomial probability distribution.

# SOLUTION:

- a. This example satisfies the conditions of a Bernoulli process because a customer either uses a credit card (labeled success), with an 85% likelihood, or does not use a credit card (labeled failure), with a 15% likelihood. Moreover, given a large number of customers, these probabilities of success and failure do not change from customer to customer.
- b. Recall from Chapter 4 that we can use a probability tree whenever an experiment can be broken down into stages. Here we can view each stage as a trial. The probability tree for Example 5.7 is shown in Figure 5.3. We let S denote the outcome that a customer uses a credit card and F denote the outcome that a customer does not use a credit card. Starting from the unlabeled node on the left, customer 1 has an 85% chance of using a credit card and a 15% chance of not using one. The branches emanating from customer 1 denote conditional probabilities of customer 2 using a credit card, given whether or not customer 1 used a credit card. However, since we assume that the trials of a Bernoulli process are independent, the conditional probability is the same as the unconditional probability. In other words, customer 2 has the same 85% chance of using a credit card and a 15% chance of not using one regardless of what customer 1 uses. The same holds for the probabilities for customer 3. The fourth column shows that there are eight possible events at the end of the probability tree. We are able to obtain relevant probabilities by using the multiplication rule for independent events. For instance, following the top branches throughout the probability tree, we calculate the probability that all three customers use a credit card as  $(0.85)(0.85)(0.85) = 0.6141$ . The probabilities for the remaining events are found in a similar manner.
- c. Since we are not interested in identifying the particular customer who uses a credit card, but rather the number of customers who use a credit card, we can

**FIGURE 5.3** Probability tree for Example 5.7



combine events with the same number of successes, using the addition rule for mutually exclusive events. For instance, in order to find the probability that one customer uses a credit card, we add the probabilities that correspond to the outcome  $x = 1$  (see shaded areas in Figure 5.3):  $0.0191 + 0.0191 + 0.0191 = 0.0573$ . Similarly, we calculate the remaining probabilities corresponding to the other values of  $X$  and construct the probability distribution shown in Table 5.8.

**TABLE 5.8** Binomial Probabilities for Example 5.7

$x$	$P(X = x)$
0	0.0034
1	0.0573
2	0.3252
3	0.6141
	Total = 1

Fortunately, we do not have to construct a probability tree each time we want to construct a binomial distribution. We can use the following formula for calculating probabilities associated with a binomial random variable.

#### THE BINOMIAL DISTRIBUTION

For a **binomial random variable**  $X$ , the probability of  $x$  successes in  $n$  Bernoulli trials is

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x}$$

for  $x = 0, 1, 2, \dots, n$ . By definition,  $0! = 1$ .

The formula consists of two parts:

- The first term,  $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ , tells us how many sequences with  $x$  successes and  $n - x$  failures are possible in  $n$  trials. We refer to the first term as the binomial coefficient, which is really the familiar combination formula used to find the number of ways to choose  $x$  objects from a total of  $n$  objects, where the order in which the  $x$  objects are listed *does not matter*. For instance, in order to calculate the number of sequences that contain exactly 1 credit card user in 3 trials, we substitute  $x = 1$  and  $n = 3$  into the formula and calculate  $\binom{3}{1} = \frac{3!}{1!(3-1)!} = \frac{3!}{1!(2)!} = \frac{3 \times 2 \times 1}{(1) \times (2 \times 1)} = 3$ . So there are three sequences having exactly 1 success—we can verify this result with Figure 5.3.
- The second part of the equation,  $p^x(1 - p)^{n-x}$ , represents the probability of any particular sequence with  $x$  successes and  $n - x$  failures. For example, we can obtain the probability of 1 success in 3 trials from rows 4, 6, or 7 in the last column of the probability tree in Figure 5.3 as (see shaded areas):

$$\left. \begin{array}{l} \text{row 4: } 0.85 \times 0.15 \times 0.15 \\ \text{row 6: } 0.15 \times 0.85 \times 0.15 \\ \text{row 7: } 0.15 \times 0.15 \times 0.85 \end{array} \right\} \text{ or } (0.85)^1 \times (0.15)^2 = 0.019$$

In other words, each sequence consisting of 1 success in 3 trials has a 1.91% chance of occurring.

In order to obtain the overall probability of getting 1 success in 3 trials, we then multiply the binomial coefficient by the probability of obtaining the particular



sequence, or here,  $3 \times 0.0191 = 0.0573$ . This is precisely the probability that we found for  $P(X = 1)$  using the probability tree.

Moreover, we could use the formulas shown in Section 5.2 to calculate the expected value, the variance, and the standard deviation for any binomial random variable. Fortunately, for the binomial distribution, these formulas simplify to  $E(X) = np$ ,  $Var(X) = np(1 - p)$ , and  $SD(X) = \sqrt{np(1 - p)}$ . The simplified formula for the expected value is rather intuitive in that if we know the probability of success  $p$  of an experiment and we repeat the experiment  $n$  times, then on average, we expect  $np$  successes.

#### EXPECTED VALUE, VARIANCE, AND STANDARD DEVIATION FOR A BINOMIAL RANDOM VARIABLE

If  $X$  is a binomial random variable, then

$$E(X) = \mu = np,$$

$$Var(X) = \sigma^2 = np(1 - p), \text{ and}$$

$$SD(X) = \sigma = \sqrt{np(1 - p)}.$$

For instance, for the binomial probability distribution assumed in Example 5.7, we can derive the expected value with the earlier general formula as

$$E(X) = \sum x_i P(X = x_i) = (0 \times 0.0034) + (1 \times 0.0573) + (2 \times 0.3252) + (3 \times 0.6141) = 2.55.$$

However, an easier way is to use  $E(X) = np$  and thus calculate the expected value as  $3 \times 0.85 = 2.55$ . Similarly, the variance and the standard deviation can be easily calculated as

$$Var(X) = \sigma^2 = np(1 - p) = 3 \times 0.85 \times 0.15 = 0.38.$$

$$SD(X) = \sigma = \sqrt{np(1 - p)} = \sqrt{0.38} = 0.62.$$

#### EXAMPLE 5.8

In the United States, about 30% of adults have four-year college degrees (*The Wall Street Journal*, April 26, 2012). Suppose five adults are randomly selected.

- What is the probability that none of the adults has a college degree?
- What is the probability that no more than two of the adults have a college degree?
- What is the probability that at least two of the adults have a college degree?
- Calculate the expected value, the variance, and the standard deviation of this binomial distribution.
- Graphically depict the probability distribution and comment on its symmetry/skewness.

**SOLUTION:** First, this problem satisfies the conditions for a Bernoulli process with a random selection of five adults,  $n = 5$ . Here, an adult either has a college degree, with probability  $p = 0.30$ , or does not have a college degree, with probability  $1 - p = 1 - 0.30 = 0.70$ . Given a large number of adults, it fulfills the requirement that the probability that an adult has a college degree stays the same from adult to adult.

- In order to find the probability that none of the adults has a college degree, we let  $x = 0$  and find

$$\begin{aligned} P(X = 0) &= \frac{5!}{0!(5 - 0)!} \times (0.30)^0 \times (0.70)^{5-0} \\ &= \frac{5 \times 4 \times \cdots \times 1}{(1) \times (5 \times 4 \times \cdots \times 1)} \times 1 \times (0.70)^5 = 1 \times 1 \times 0.1681 \\ &= 0.1681. \end{aligned}$$

In other words, there is a 16.81% chance that none of the adults has a college degree.

- b. We find the probability that no more than two adults have a college degree as:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2).$$

We have already found  $P(X = 0)$  from part a. So we now compute  $P(X = 1)$  and  $P(X = 2)$ :

$$P(X = 1) = \frac{5!}{1!(5-1)!} \times (0.30)^1 \times (0.70)^{5-1} = 0.3602$$

$$P(X = 2) = \frac{5!}{2!(5-2)!} \times (0.30)^2 \times (0.70)^{5-2} = 0.3087$$

Next we sum the three relevant probabilities and obtain  $P(X \leq 2) = 0.1681 + 0.3602 + 0.3087 = 0.8370$ . From a random sample of five adults, there is an 83.7% likelihood that no more than two of them will have a college degree.

- c. We find the probability that at least two adults have a college degree as:

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5).$$

We can solve this problem by calculating and then summing each of the four probabilities, from  $P(X = 2)$  to  $P(X = 5)$ . A simpler method uses one of the key properties of a probability distribution, which states that the sum of the probabilities over all values of  $X$  equals 1. Therefore,  $P(X \geq 2)$  can be written as  $1 - [P(X = 0) + P(X = 1)]$ . We have already calculated  $P(X = 0)$  and  $P(X = 1)$  from parts a and b, so

$$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - (0.1681 + 0.3602) = 0.4717.$$

- d. We use the simplified formulas to calculate the mean, the variance, and the standard deviation as

$$E(X) = np = 5 \times 0.30 = 1.5 \text{ adults,}$$

$$\text{Var}(X) = \sigma^2 = np(1-p) = 5 \times 0.30 \times 0.70 = 1.05 \text{ (adults)}^2, \text{ and}$$

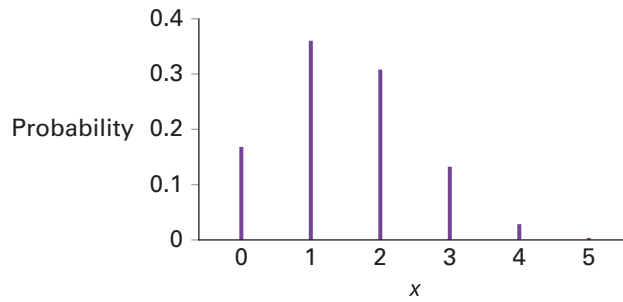
$$SD(X) = \sigma = \sqrt{np(1-p)} = \sqrt{1.05} = 1.02 = 1.02 \text{ adults.}$$

- e. Before we graph this distribution, we first show the complete binomial distribution for Example 5.8 in Table 5.9.

**TABLE 5.9** Binomial Distribution  
with  $n = 5$  and  $p = 0.30$

$x$	$P(X = x)$
0	0.1681
1	0.3602
2	0.3087
3	0.1323
4	0.0284
5	0.0024

This binomial distribution is graphically depicted in Figure 5.4. When randomly selecting five adults, the most likely outcome is that exactly one adult will have a college degree. The distribution is not symmetric; rather, it is positively skewed. In later chapters, we will learn that the binomial distribution is approximately symmetric when the sample size  $n$  is large.

**FIGURE 5.4**Binomial distribution with  $n = 5$  and  $p = 0.30$ 

## Using Excel to Obtain Binomial Probabilities

As you may have noticed, at times it is somewhat tedious and cumbersome to solve binomial distribution problems using the formulas. This issue becomes even more pronounced when we encounter large values for  $n$  and we wish to determine probabilities where  $X$  assumes a wide range of values. Some texts include probability tables to help with the calculations for important discrete probability distributions. We will rely on Excel to find cumbersome binomial probabilities. Consider the following problem.

### EXAMPLE 5.9

In 2007, approximately 4.7% of the households in the Detroit metropolitan area were in some stage of foreclosure, the highest foreclosure rate in the nation (*The Associated Press*, February 13, 2008). Suppose 100 mortgage-holding households in the Detroit area were sampled.

- What is the probability that exactly 5 of these households were in some stage of foreclosure?
- What is the probability that no more than 5 of these households were in some stage of foreclosure?
- What is the probability that more than 5 households were in some stage of foreclosure?

#### SOLUTION:

- It is possible to use the binomial formula and find the probability that exactly 5 households were in some stage of foreclosure as  $P(X = 5) = \frac{100!}{5!95!} \times (0.047)^5 \times (0.953)^{95}$ ; however, we would quickly find the arithmetic quite unwieldy. Fortunately, we can use Excel's BINOM.DIST function to calculate this probability. In general, we find an empty cell and insert '=BINOM.DIST( $x$ ,  $n$ ,  $p$ , 0 or 1)', where  $x$  is the number of successes,  $n$  is the number of trials, and  $p$  is the probability of success. For the last argument in the function, if we enter 0, then we are prompting Excel to return the probability of a specific value,  $P(X = x)$ ; if we enter 1, then we are prompting Excel to return a cumulative probability  $P(X \leq x)$ . In this example, we input '=BINOM.DIST(5, 100, 0.047, 0)'. After choosing <Enter>, Excel returns a value of 0.1783; thus,  $P(X = 5) = 0.1783$ .
- We write the probability that no more than 5 of these households are in some stage of foreclosure as  $P(X \leq 5)$ . Here, we input '=BINOM.DIST(5, 100, 0.047, 1)'. Excel returns the probability 0.6697; thus,  $P(X \leq 5) = 0.6697$ .
- The probability that more than five households are in some stage of foreclosure is written as  $P(X > 5)$ . Using the information in part b, we find this as  $P(X > 5) = 1 - P(X \leq 5) = 1 - 0.6697 = 0.3303$ .

## EXERCISES 5.4

### Mechanics

34. Assume that  $X$  is a binomial random variable with  $n = 5$  and  $p = 0.35$ . Calculate the following probabilities.
- $P(X = 0)$
  - $P(X = 1)$
  - $P(X \leq 1)$
35. Assume that  $X$  is a binomial random variable with  $n = 6$  and  $p = 0.68$ . Calculate the following probabilities.
- $P(X = 5)$
  - $P(X = 4)$
  - $P(X \geq 4)$
36. Assume that  $X$  is a binomial random variable with  $n = 8$  and  $p = 0.32$ . Calculate the following probabilities.
- $P(3 < X < 5)$
  - $P(3 < X \leq 5)$
  - $P(3 \leq X \leq 5)$
37. Let the probability of success on a Bernoulli trial be 0.30. In five Bernoulli trials, what is the probability that there will be (a) 4 failures, (b) more than the expected number of failures?
38. (Use computer) Let  $X$  represent a binomial random variable with  $n = 150$  and  $p = 0.36$ . Find the following probabilities.
- $P(X \leq 50)$
  - $P(X = 40)$
  - $P(X > 60)$
  - $P(X \geq 55)$
39. (Use computer) Let  $X$  represent a binomial random variable with  $n = 200$  and  $p = 0.77$ . Find the following probabilities.
- $P(X \leq 150)$
  - $P(X > 160)$
  - $P(155 \leq X \leq 165)$
  - $P(X = 160)$
- e. Calculate the variance and the standard deviation for this probability distribution.
41. At a local community college, 40% of students who enter the college as freshmen go on to graduate. Ten freshmen are randomly selected.
- What is the probability that none of them graduates from the local community college?
  - What is the probability that at most nine will graduate from the local community college?
  - What is the expected number that will graduate?
42. The percentage of Americans who have confidence in U.S. banks dropped to 23% in June 2010, which is far below the pre-recession level of 41% reported in June 2007 (gallup.com).
- What is the probability that fewer than half of 10 Americans in 2010 have confidence in U.S. banks?
  - What would have been the corresponding probability in 2007?
43. In recent analyses of Census figures, one in four American counties has passed or is approaching the tipping point where black, Hispanic, and Asian children constitute a majority of the under-20 population (*The New York Times*, August 6, 2008). Racial and ethnic minorities now account for 43% of Americans under 20.
- What is the expected number of whites in a random sample of 5,000 under-20 Americans? What is the corresponding standard deviation?
  - What is the expected number of racial and ethnic minorities in a random sample of 5,000 under-20 Americans? What is the corresponding standard deviation?
  - If you randomly sample six American counties, what is the probability that for the under-20 population, whites have a majority in all of the counties?
44. Approximately 76% of baby boomers aged 43 to 61 are still in the workforce (*The Boston Globe*, July 10, 2008). Six baby boomers are selected at random.
- What is the probability that exactly one of the baby boomers is still in the workforce?
  - What is the probability that at least five of the baby boomers are still in the workforce?
  - What is the probability that less than two of the baby boomers are still in the workforce?
  - What is the probability that more than the expected number of the baby boomers are still in the workforce?
45. Sikhism, a religion founded in the 15th century in India, is going through turmoil due to a rapid decline in the number of Sikh youths who wear turbans (*The Washington Post*, March 29, 2009). The tedious task of combing and

### Applications

40. According to a report from the Center for Studying Health System Change, 20% of Americans delay or go without medical care because of concerns about cost (*The Wall Street Journal*, June 26, 2008). Suppose eight individuals are randomly selected.
- What is the probability that none will delay or go without medical care?
  - What is the probability that no more than two will delay or go without medical care?
  - What is the probability that at least seven will delay or go without medical care?
  - What is the expected number of individuals who will delay or go without medical care?

tying up long hair and a desire to assimilate has led to approximately 25% of Sikh youths giving up the turban.

- a. What is the probability that exactly two in a random sample of five Sikh youths wear a turban?
  - b. What is the probability that two or more in a random sample of five Sikh youths wear a turban?
  - c. What is the probability that more than the expected number of Sikh youths wear a turban in a random sample of five Sikh youths?
  - d. What is the probability that more than the expected number of Sikh youths wear a turban in a random sample of 10 Sikh youths?
46. According to the U.S. Census, roughly half of all marriages in the United States end in divorce. Researchers from leading universities have shown that the emotions aroused by one person's divorce can transfer like a virus, making divorce contagious (*CNN*, June 10, 2010). A splitup between immediate friends increases a person's own chances of getting divorced from 36% to 63%, an increase of 75%. Use these findings to answer the following questions.
- a. Compute the probability that more than half of four randomly selected marriages will end in divorce.
  - b. Redo part a if it is known that the couple's immediate friends have split up.
  - c. Redo part a if it is known that none of the couple's immediate friends has split up.
47. Sixty percent of a firm's employees are men. Suppose four of the firm's employees are randomly selected.
- a. What is more likely, finding three men and one woman or two men and two women?
  - b. Do you obtain the same answer as in part a if 70% of the firm's employees had been men?
48. The principal of an architecture firm tells her client that there is at least a 50% chance of having an acceptable design by the end of the week. She knows that there is only a 25% chance that any one designer would be able to do so by the end of the week.
- a. Would she be correct in her statement to the client if she asks two of her designers to work on the design, independently?
  - b. If not, what if she asks three of her designers to work on the design, independently?
49. (Use computer) Suppose 40% of recent college graduates plan on pursuing a graduate degree. Fifteen recent college graduates are randomly selected.
- a. What is the probability that no more than four of the college graduates plan to pursue a graduate degree?
  - b. What is the probability that exactly seven of the college graduates plan to pursue a graduate degree?
  - c. What is the probability that at least six but no more than nine of the college graduates plan to pursue a graduate degree?
50. (Use computer) At the University of Notre Dame Mendoza College of Business, 40% of the students seeking a master's degree specialize in finance (*Kiplinger's Personal Finance*, March 2009). Twenty master's degree students are randomly selected.
- a. What is the probability that exactly 10 of the students specialize in finance?
  - b. What is the probability that no more than 10 of the students specialize in finance?
  - c. What is the probability that at least 15 of the students specialize in finance?
51. (Use computer) The Washington, DC, region has one of the fastest-growing foreclosure rates in the nation, as 15,613 homes went into foreclosure during the one-year period ending in February 2008 (*The Washington Post*, June 19, 2008). Over the past year, the number of foreclosures per 10,000 homes is 131 for the Washington area, while it is 87 nationally. In other words, the foreclosure rate is 1.31% for the Washington, DC area and 0.87% for the nation. Assume that the foreclosure rates remain stable.
- a. What is the probability that in a given year, fewer than 2 out of 100 houses in the Washington, DC area will go up for foreclosure?
  - b. What is the probability that in a given year, fewer than 2 out of 100 houses in the nation will go up for foreclosure?
  - c. Comment on the above findings.

## 5.5 THE POISSON DISTRIBUTION

### LO 5.7

Another important discrete probability distribution is the **Poisson distribution**, named after the French mathematician Simeon Poisson (1781–1849). It is particularly useful in problems that deal with finding the number of occurrences of a certain event over time or space, where space refers to area or region.

Describe the Poisson distribution and compute relevant probabilities.

A **Poisson random variable** counts the number of occurrences of a certain event over a given interval of time or space.

For simplicity, we call these occurrences “successes.” We first must ensure that our experiment satisfies the conditions of a **Poisson process**.

An experiment satisfies a **Poisson process** if:

- The number of successes within a specified time or space interval equals any integer between zero and infinity.
- The number of successes counted in nonoverlapping intervals are independent.
- The probability that success occurs in any interval is the same for all intervals of equal size and is proportional to the size of the interval.

For a Poisson process, we define the number of successes achieved in a specified time or space interval as a Poisson random variable. Like the Bernoulli process, many experiments fit the conditions of a Poisson process. Consider the following examples of Poisson random variables categorized by those relating to time and those relating to space.

#### Examples of Poisson Random Variables with Respect to Time

- The number of cars that cross the Brooklyn Bridge between 9:00 am and 10:00 am on a Monday morning.
- The number of customers that use a McDonald’s drive-thru in a day.
- The number of bankruptcies that are filed in a month.
- The number of homicides that occur in a year.

#### Examples of Poisson Random Variables with Respect to Space

- The number of defects in a 50-yard roll of fabric.
- The number of schools of fish in 100 square miles.
- The number of leaks in a specified stretch of a pipeline.
- The number of bacteria in a specified culture.

We use the following formula for calculating probabilities associated with a Poisson random variable.

#### THE POISSON DISTRIBUTION

For a **Poisson random variable**  $X$ , the probability of  $x$  successes over a given interval of time or space is

$$P(X = x) = \frac{e^{-\mu} \mu^x}{x!},$$

for  $x = 0, 1, 2, \dots$ , where  $\mu$  is the mean number of successes and  $e \approx 2.718$  is the base of the natural logarithm.

As with the binomial random variable, we have simplified formulas to calculate the variance and the standard deviation of a Poisson random variable. An interesting fact is that the mean of the Poisson random variable is equal to the variance.

#### EXPECTED VALUE, VARIANCE, AND STANDARD DEVIATION OF A POISSON RANDOM VARIABLE

If  $X$  is a Poisson random variable, then

$$E(X) = \mu,$$

$$Var(X) = \sigma^2 = \mu, \quad \text{and}$$

$$SD(X) = \sigma = \sqrt{\mu}.$$



### EXAMPLE 5.10

We can now address questions first posed by Anne Jones in the introductory case of this chapter. Recall that Anne is concerned about staffing needs at the Starbucks that she manages. She has specific questions about the probability distribution of customer arrivals at her store. Anne believes that the typical Starbucks customer averages 18 visits to the store over a 30-day month. She has the following questions:

- How many visits should Anne expect in a 5-day period from a typical Starbucks customer?
- What is the probability that a customer visits the chain five times in a 5-day period?
- What is the probability that a customer visits the chain no more than two times in a 5-day period?
- What is the probability that a customer visits the chain at least three times in a 5-day period?

**SOLUTION:** In applications of the Poisson distribution, we first determine the mean number of successes in the relevant time or space interval. We use the Poisson process condition that the probability that success occurs in any interval is the same for all intervals of equal size and is proportional to the size of the interval. Here, the relevant mean will be based on the rate of 18 visits over a 30-day month.

- Given the rate of 18 visits over a 30-day month, we can write the mean for the 30-day period as  $\mu_{30} = 18$ . For this problem, we compute the proportional mean for a 5-day period as  $\mu_5 = 3$  because  $\frac{18 \text{ visits}}{30 \text{ days}} = \frac{3 \text{ visits}}{5 \text{ days}}$ .

In other words, on average, a typical Starbucks customer visits the store three times over a 5-day period.

- In order to find the probability that a customer visits the chain five times in a 5-day period, we calculate

$$P(X = 5) = \frac{e^{-3}3^5}{5!} = \frac{(0.0498)(243)}{120} = 0.1008.$$

- For the probability that a customer visits the chain no more than two times in a 5-day period, we find  $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$ . We calculate the individual probabilities, and then find the sum:

$$P(X = 0) = \frac{e^{-3}3^0}{0!} = \frac{(0.0498)(1)}{1} = 0.0498,$$

$$P(X = 1) = \frac{e^{-3}3^1}{1!} = \frac{(0.0498)(3)}{1} = 0.1494, \quad \text{and}$$

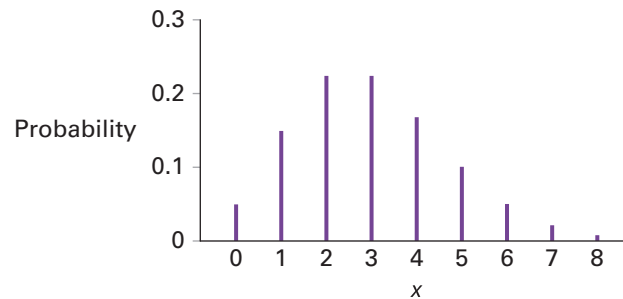
$$P(X = 2) = \frac{e^{-3}3^2}{2!} = \frac{(0.0498)(9)}{2} = 0.2241.$$

Thus,  $P(X \leq 2) = 0.0498 + 0.1494 + 0.2241 = 0.4233$ . There is approximately a 42% chance that a customer visits the chain no more than two times in a 5-day period.

- We write the probability that a customer visits the chain at least three times in a 5-day period as  $P(X \geq 3)$ . Initially, we might attempt to solve this problem by evaluating  $P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) + \dots$ . However, given the infinite number of possible values, we cannot solve a Poisson problem this way. Here, we find  $P(X \geq 3)$  as  $1 - [P(X = 0) + P(X = 1) + P(X = 2)]$ . Based on the probabilities in part c, we have  $P(X \geq 3) = 1 - [0.0498 + 0.1494 + 0.2241] = 1 - 0.4233 = 0.5767$ . Thus, there is about a 58% chance that a customer will frequent the chain at least three times in a 5-day period.

Figure 5.5 graphs the Poisson distribution  $P(X = x)$  with  $\mu = 3$ , for  $x$  ranging from 0 to 8. The most likely outcomes are when  $x$  equals 2 and  $x$  equals 3, and the distribution is positively skewed. Remember that, theoretically, the values that the Poisson random variable assumes are infinitely countable, but the probabilities approach zero beyond those shown here.

**FIGURE 5.5**  
Poisson distribution  
with  $\mu = 3$



## Using Excel to Obtain Poisson Probabilities

Like the binomial formula, the manual use of the Poisson formula can become quite cumbersome, especially when the values of  $x$  and  $\mu$  become large. Excel again proves useful when calculating Poisson probabilities, as the next example shows.

### EXAMPLE 5.11

Even as a recession gripped the country, 114 microbreweries and brewpubs opened in the United States in 2008 (*The Wall Street Journal*, March 18, 2009). Assume this number represents an average and remains constant over time. Find the following probabilities with Excel.

- What is the probability that no more than 100 microbreweries or brewpubs open in a given year?
- What is the probability that exactly 115 microbreweries or brewpubs open in a given year?

#### SOLUTION:

- We wish to determine the probability that no more than 100 microbreweries or brewpubs open in a given year; that is,  $P(X \leq 100)$ . We use Excel's POISSON.DIST function to calculate this probability. In general, when finding a Poisson probability, we find an empty cell and insert “=POISSON.DIST( $x, \mu, 0$  or  $1$ )”, where  $x$  is the number of successes over some interval and  $\mu$  is the mean over that interval. As similarly defined for Excel's binomial function, the term 0 will prompt Excel to return the probability of a specific value,  $P(X = x)$ , whereas 1 prompts Excel to return the cumulative probability  $P(X \leq x)$ . In this example, we input “=POISSON.DIST(100, 114, 1)”. Excel returns a value of 0.1012; thus,  $P(X \leq 100) = 0.1012$ . There is about a 10% chance that no more than 100 microbreweries or brewpubs will open in any given year.
- Here we wish to find  $P(X = 115)$ ; that is, the probability that exactly 115 microbreweries or brewpubs open in any given year. We input “=POISSON.DIST(115, 114, 0)” and Excel returns 0.0370. Thus, there is a 3.7% chance that exactly 115 microbreweries or brewpubs will open in any given year.

## SYNOPSIS OF INTRODUCTORY CASE

Anne Jones, the manager of a Starbucks store, is concerned about how other nearby store closings might affect foot traffic at her store. A solid understanding of the likelihood of customer arrivals is necessary before she can make further statistical inference. Historical data allow her to assume that a typical Starbucks customer averages 18 visits to a Starbucks store over a 30-day month. With this information and the knowledge that she can model customer arrivals using the Poisson distribution, she deduces that a typical customer averages three visits in a 5-day period. The likelihood that a typical customer frequents her store five times in a 5-day period is approximately 10%. Moreover, there is approximately a 42% chance that a typical customer goes to Starbucks no more than two times in a 5-day period, while the chances that this customer visits the chain at least three times is approximately 58%. These preliminary probabilities will prove vital as Anne plans her future staffing needs.



## EXERCISES 5.5

### Mechanics

52. Assume that  $X$  is a Poisson random variable with  $\mu = 1.5$ . Calculate the following probabilities.
  - a.  $P(X = 1)$
  - b.  $P(X = 2)$
  - c.  $P(X \geq 2)$
53. Assume that  $X$  is a Poisson random variable with  $\mu = 4$ . Calculate the following probabilities.
  - a.  $P(X = 4)$
  - b.  $P(X = 2)$
  - c.  $P(X \leq 1)$
54. Let the mean success rate of a Poisson process be 8 successes per hour.
  - a. Find the expected number of successes in a half-hour period.
  - b. Find the probability of at least two successes in a given half-hour period.
  - c. Find the expected number of successes in a two-hour period.
  - d. Find the probability of 10 successes in a given two-hour period.
55. (Use computer) Assume that  $X$  is a Poisson random variable with  $\mu = 15$ . Calculate the following probabilities.
  - a.  $P(X \leq 10)$
  - b.  $P(X = 13)$
  - c.  $P(X > 15)$
  - d.  $P(12 \leq X \leq 18)$
56. (Use computer) Assume that  $X$  is a Poisson random variable with  $\mu = 20$ . Calculate the following probabilities.
  - a.  $P(X < 14)$
  - b.  $P(X \geq 20)$

- c.  $P(X = 25)$
- d.  $P(18 \leq X \leq 23)$

### Applications

57. Which of the following probabilities are likely to be found using a Poisson distribution?
  - a. The probability that there will be six leaks in a specified stretch of a pipeline.
  - b. The probability that at least 10 students in a class of 40 will land a job right after graduation.
  - c. The probability that at least 50 families will visit the Acadia National Park over the weekend.
  - d. The probability that no customer will show up in the next five minutes.
58. Which of the following scenarios are likely to represent Poisson random variables?
  - a. The number of violent crimes in New York over a six-week period.
  - b. The number of customers of a bank manager who will default.
  - c. The number of scratches on a 2-by-1-foot portion of a large wooden table.
  - d. The number of patients of a doctor for whom the drug will be effective.
59. On average, there are 12 potholes per mile on a particular stretch of the state highway. Suppose the potholes are distributed evenly on the highway.
  - a. Find the probability of finding fewer than two potholes in a quarter-mile stretch of the highway.
  - b. Find the probability of finding more than one pothole in a quarter-mile stretch of the highway.

60. A tollbooth operator has observed that cars arrive randomly at an average rate of 360 cars per hour.
  - a. Find the probability that two cars arrive during a specified one-minute period.
  - b. Find the probability that at least two cars arrive during a specified one-minute period.
  - c. Find the probability that 40 cars arrive between 10:00 am and 10:10 am.
61. A textile manufacturing process finds that on average, two flaws occur per every 50 yards of material produced.
  - a. What is the probability of exactly two flaws in a 50-yard piece of material?
  - b. What is the probability of no more than two flaws in a 50-yard piece of material?
  - c. What is the probability of no flaws in a 25-yard piece of material?
62. Motorists arrive at a Gulf gas station at the rate of two per minute during morning hours.
  - a. What is the probability that more than two motorists will arrive at the Gulf gas station during a one-minute interval in the morning?
  - b. What is the probability that exactly six motorists will arrive at the Gulf gas station during a five-minute interval in the morning?
  - c. How many motorists can an employee expect in her three-hour morning shift?
63. Airline travelers should be ready to be more flexible as airlines once again cancel thousands of flights this summer. The Coalition for Airline Passengers Rights, Health, and Safety averages 400 calls a day to help stranded travelers deal with airlines (<http://seattlepi.com>, July 10, 2008). Suppose the hotline is staffed for 16 hours a day.
  - a. Calculate the average number of calls in a one-hour interval; 30-minute interval; 15-minute interval.
  - b. What is the probability of exactly six calls in a 15-minute interval?
  - c. What is the probability of no calls in a 15-minute interval?
  - d. What is the probability of at least two calls in a 15-minute interval?
64. (Use computer) On average, 400 people a year are struck by lightning in the United States (*The Boston Globe*, July 21, 2008).
  - a. What is the probability that at most 425 people are struck by lightning in a year?
  - b. What is the probability that at least 375 people are struck by lightning in a year?
65. According to a recent government report, the aging of the U.S. population is translating into many more visits to doctors' offices and hospitals (*USA Today*, August 7, 2008). It is estimated that an average person makes four visits a year to doctors' offices and hospitals.
  - a. What are the mean and the standard deviation of an average person's number of monthly visits to doctors' offices and hospitals?
  - b. What is the probability that an average person does not make any monthly visits to doctors' offices and hospitals?
  - c. What is the probability that an average person makes at least one monthly visit to doctors' offices and hospitals?
66. (Use computer) According to Nielsen, the average teenager sends 3,339 texts per month (*CNN*, October 15, 2010).
  - a. Find the probability that an average teenager sends more than 1,000 texts per week.
  - b. Find the probability that an average teenager sends fewer than 500 texts per week.
67. (Use computer) In the fiscal year that ended September 30, 2008, there were 24,584 age-discrimination claims filed with the Equal Employment Opportunity Commission, an increase of 29% from the previous year (*The Wall Street Journal*, March 7–8, 2009). Assume there were 260 working days in the fiscal year for which a worker could file a claim.
  - a. Calculate the average number of claims filed on a working day.
  - b. What is the probability that exactly 100 claims were filed on a working day?
  - c. What is the probability that no more than 100 claims were filed on a working day?

## LO 5.8

## 5.6 THE HYPERGEOMETRIC DISTRIBUTION

Describe the hypergeometric distribution and compute relevant probabilities.

In Section 5.4, we defined a binomial random variable  $X$  as the number of successes in the  $n$  trials of a Bernoulli process. The trials, according to a Bernoulli process, are independent and the probability of success does not change from trial to trial. The **hypergeometric distribution** is appropriate in applications where we cannot assume the trials are independent.

Consider a box full of production items, of which 10% are known to be defective. Let success be labeled as the draw of a defective item. The probability of success may not be the same from trial to trial; it will depend on the size of the population and whether the sampling

was done with or without replacement. Suppose the box consists of 20 items of which 10%, or 2, are defective. The probability of success in the first draw is  $0.10 (= 2/20)$ . However, the probability of success in subsequent draws will depend on the outcome of the first draw. For example, if the first item was defective, the probability of success in the second draw will be  $0.0526 (= 1/19)$ , while if the first item was not defective, the probability of success in the second draw will be  $0.1053 (= 2/19)$ . Therefore, the binomial distribution is not appropriate because the trials are not independent and the probability of success changes from trial to trial.

We use the **hypergeometric distribution** in place of the binomial distribution when we are sampling **without replacement** from a population whose size  $N$  is not significantly larger than the sample size  $n$ .

In the preceding example, we assumed sampling without replacement; in other words, after an item is drawn, it is not put back in the box for subsequent draws. The binomial distribution would be appropriate if we sample with replacement since, in that case, for each draw there will be 20 items of which 2 are defective, resulting in an unchanging probability of success. Moreover, the dependence of the trials can be ignored if the population size is very large relative to the sample size. For instance, if the box consists of 10,000 items of which 10%, or 1,000, are defective, then the probability of success in the second draw will be either  $999/9,999$  or  $1,000/9,999$ , which are both approximately equal to 0.10.

#### THE HYPERGEOMETRIC DISTRIBUTION

For a **hypergeometric random variable**  $X$ , the probability of  $x$  successes in a random selection of  $n$  items is

$$P(X = x) = \frac{\binom{S}{x} \binom{N-S}{n-x}}{\binom{N}{n}},$$

for  $x = 0, 1, 2, \dots, n$  if  $n \leq S$  or  $x = 0, 1, 2, \dots, S$  if  $n > S$ , where  $N$  denotes the number of items in the population of which  $S$  are successes.

The formula consists of three parts:

- The first term in the numerator,  $\binom{S}{x} = \frac{S!}{x!(S-x)!}$ , represents the number of ways  $x$  successes can be selected from  $S$  successes in the population.
- The second term in the numerator,  $\binom{N-S}{n-x} = \frac{(N-S)!}{(n-x)!(N-S-n+x)!}$ , represents the number of ways  $(n-x)$  failures can be selected from  $(N-S)$  failures in the population.
- The denominator,  $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ , represents the number of ways a sample of size  $n$  can be selected from the population of size  $N$ .

As with the binomial and Poisson distributions, simplified formulas can be used to calculate the mean, the variance, and the standard deviation of a hypergeometric random variable.

#### EXPECTED VALUE, VARIANCE, AND STANDARD DEVIATION OF A HYPERGEOMETRIC RANDOM VARIABLE

If  $X$  is a hypergeometric random variable, then

$$\begin{aligned} E(X) &= \mu = n \left( \frac{S}{N} \right), \\ Var(X) &= \sigma^2 = n \left( \frac{S}{N} \right) \left( 1 - \frac{S}{N} \right) \left( \frac{N-n}{N-1} \right), \quad \text{and} \\ SD(X) &= \sigma = \sqrt{n \left( \frac{S}{N} \right) \left( 1 - \frac{S}{N} \right) \left( \frac{N-n}{N-1} \right)}. \end{aligned}$$

### EXAMPLE 5.12

Wooden boxes are commonly used for the packaging and transportation of mangoes. A convenience store in Morganville, New Jersey, regularly buys mangoes from a wholesale dealer. For every shipment, the manager randomly inspects five mangoes from a box containing 20 mangoes for damages due to transportation. Suppose the chosen box contains exactly two damaged mangoes.

- What is the probability that one out of five mangoes used in the inspection is damaged?
- If the manager decides to reject the shipment if one or more mangoes are damaged, what is the probability that the shipment will be rejected?
- Calculate the expected value, the variance, and the standard deviation of the number of damaged mangoes used in the inspection.

**SOLUTION:** The hypergeometric distribution is appropriate because the probability of finding a damaged mango changes from draw to draw (sampling is without replacement and the population size  $N$  is not significantly more than the sample size  $n$ ). We use the following values to answer the questions:  $N = 20$ ,  $n = 5$ ,  $S = 2$ .

- The probability that one out of five mangoes is damaged is  $P(X = 1)$ . We calculate

$$P(X = 1) = \frac{\binom{2}{1} \binom{20-2}{5-1}}{\binom{20}{5}} = \frac{\left(\frac{2!}{1!1!}\right) \left(\frac{18!}{4!14!}\right)}{\left(\frac{20!}{5!15!}\right)} = \frac{(2)(3060)}{15,504} = 0.3947.$$

Therefore, the likelihood that exactly one out of five mangoes is damaged is 39.47%.

- In order to find the probability that one or more mangoes are damaged, we need to calculate  $P(X \geq 1)$ . We note that  $P(X \geq 1) = 1 - P(X = 0)$  where

$$P(X = 0) = \frac{\binom{2}{0} \binom{20-2}{5-0}}{\binom{20}{5}} = \frac{\left(\frac{2!}{0!2!}\right) \left(\frac{18!}{5!13!}\right)}{\left(\frac{20!}{5!15!}\right)} = \frac{(1)(8568)}{15504} = 0.5526.$$

Therefore, the probability that the shipment will be rejected equals  $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.5526 = 0.4474$ .

- We use the simplified formulas to obtain the mean, the variance, and the standard deviation as

$$E(X) = n \left( \frac{S}{N} \right) = 5 \left( \frac{2}{20} \right) = 0.50,$$

$$Var(X) = n \left( \frac{S}{N} \right) \left( 1 - \frac{S}{N} \right) \left( \frac{N-n}{N-1} \right) = 5 \left( \frac{2}{20} \right) \left( 1 - \frac{2}{20} \right) \left( \frac{20-5}{20-1} \right) = 0.3553, \text{ and}$$

$$SD(X) = \sqrt{0.3553} = 0.5960.$$

### Using Excel to Obtain Hypergeometric Probabilities

Since it is tedious to solve hypergeometric distribution problems using the formula, we typically use Excel's HYPGEOM.DIST function to aid in the calculations. In general, when finding a hypergeometric probability, we find an empty cell and insert “=HYPGEOM.DIST( $x$ ,  $n$ ,  $S$ ,  $N$ , 0 or 1)”, where  $x$  is the number of successes in the sample,  $n$  is the sample size,  $S$  is the number of successes in the population, and  $N$  is the population size. As noted in Sections 5.4 and 5.5, if we enter the value 0 as the last term in the command, then Excel returns  $P(X = x)$ ; if we enter the value 1 as the last term, then Excel returns  $P(X \leq x)$ . In Example 5.12a, we input “=HYPGEOM.DIST(1, 5, 2, 20, 0)”. Excel returns a value of 0.3947; thus,  $P(X = 1) = 0.3947$ , which is the value that we calculated manually.



## EXERCISES 5.6

### Mechanics

68. Assume that  $X$  is a hypergeometric random variable with  $N = 25$ ,  $S = 3$ , and  $n = 4$ . Calculate the following probabilities.
- $P(X = 0)$
  - $P(X = 1)$
  - $P(X \leq 1)$
69. Assume that  $X$  is a hypergeometric random variable with  $N = 15$ ,  $S = 4$ , and  $n = 3$ . Calculate the following probabilities.
- $P(X = 1)$
  - $P(X = 2)$
  - $P(X \geq 2)$
70. Compute the probability of no successes in a random sample of three items obtained from a population of 12 items that contains two successes. What are the expected number and the standard deviation of the number of successes from the sample?
71. (Use computer) Assume that  $X$  is a hypergeometric random variable with  $N = 50$ ,  $S = 20$ , and  $n = 5$ . Calculate the following probabilities.
- $P(X = 2)$
  - $P(X \geq 2)$
  - $P(X \leq 3)$
72. (Use computer) Compute the probability of at least eight successes in a random sample of 20 items obtained from a population of 100 items that contains 25 successes. What are the expected number and the standard deviation of the number of successes?

### Applications

73. Suppose you have an urn of ten marbles, of which five are red and five are green. If you draw two marbles from this urn, what is the probability that both marbles are red? What is the probability that at least one of the marbles is red?
74. A professor of management has heard that eight students in his class of 40 have landed an internship for the summer. Suppose he runs into three of his students in the corridor.
- Find the probability that none of these students has landed an internship.
  - Find the probability that at least one of these students has landed an internship.
75. Despite the repeated effort by the government to reform how Wall Street pays its executives, some of the nation's biggest banks are continuing to pay out bonuses nearly as large as those in the best years before the crisis (*The Washington Post*, January 15, 2010).
- It is known that 10 out of 15 members of the board of directors of a company were in favor of a bonus. Suppose three members were randomly selected by the media.
- What is the probability that all of them were in favor of a bonus?
  - What is the probability that at least two members were in favor of a bonus?
76. Many programming teams work independently at a large software company. The management has been putting pressure on these teams to finish a project on time. The company currently has 18 large programming projects, of which only 12 are likely to finish on time. Suppose the manager decides to randomly supervise three such projects.
- What is the probability that all three projects finish on time?
  - What is the probability that at least two projects finish on time?
77. David Barnes and his fiancée Valerie Shah are visiting Hawaii. There are 20 guests registered for orientation. It is announced that 12 randomly selected registered guests will receive a free lesson of Tahitian dance.
- What is the probability that both David and Valerie get picked for the Tahitian dance lesson?
  - What is the probability that neither of them gets picked for the Tahitian dance lesson?
78. The National Science Foundation is fielding applications for grants to study climate change. Twenty universities apply for a grant, and only four of them will be awarded. If Syracuse University and Auburn University are among the 20 applicants, what is the probability that these two universities will receive a grant? Assume that the selection is made randomly.
79. (Use computer) A committee of 40 members consists of 24 men and 16 women. A subcommittee consisting of 10 randomly selected members will be formed.
- What are the expected number of men and women on the subcommittee?
  - What is the probability that at least half of the members on the subcommittee will be women?
80. (Use computer) Powerball is a jackpot game with a grand prize starting at \$20 million and often rolling over into the hundreds of millions. In 2006, the jackpot was \$365 million. The winner may choose to receive the jackpot prize paid over 29 years or as a lump-sum payment. For \$1 the player selects six numbers for the base game of Powerball. There are two independent stages of the game. Five balls are randomly drawn from 59 consecutively numbered white balls. Moreover, one ball, called the Powerball, is

randomly drawn from 39 consecutively numbered red balls. To be a winner, the numbers selected by the player must match the numbers on the randomly drawn white balls as well as the Powerball.

- a. What is the probability that the player is able to match the numbers of two out of five randomly drawn white balls?
- b. What is the probability that the player is able to match the numbers of all five randomly drawn white balls?
- c. What is the probability that the player is able to match the Powerball for a randomly drawn red ball?
- d. What is the probability of winning the jackpot? [*Hint: Remember that the two stages of drawing white and red balls are independent.*]

## WRITING WITH STATISTICS



Senior executives at Skyhigh Construction, Inc., participate in a pick-your-salary plan. They choose salaries in a range between \$125,000 and \$150,000. By choosing a lower salary, an executive has an opportunity to make a larger bonus. If Skyhigh does not generate an operating profit during the year, then no bonuses are paid. Skyhigh has just hired two new senior executives, Allen Grossman and Felicia Arroyo. Each must decide whether to choose *Option 1*: a base pay of \$125,000 with a possibility of a large bonus or *Option 2*: a base pay of \$150,000 with a possibility of a bonus, but the bonus would be one-half of the bonus under Option 1.

Grossman, 44 years old, is married with two young children. He bought his home at the height of the market and has a rather large monthly mortgage payment. Arroyo, 32 years old, just completed her MBA at a prestigious Ivy League university. She is single and has no student loans due to a timely inheritance upon entering graduate school. Arroyo just moved to the area so she has decided to rent an apartment for at least one year. Given their personal profiles, inherent perceptions of risk, and subjective views of the economy, Grossman and Arroyo construct their individual probability distributions with respect to bonus outcomes shown in Table 5.10.

**TABLE 5.10** Grossman's and Arroyo's Probability Distributions

Bonus (in \$)	Probability	
	Grossman	Arroyo
0	0.35	0.20
50,000	0.45	0.25
100,000	0.10	0.35
150,000	0.10	0.20

Jordan Lake, an independent human resources specialist, is asked to summarize the payment plans with respect to each executive's probability distribution.

Jordan would like to use the above probability distributions to

1. Compute expected values to evaluate payment plans for Grossman and Arroyo.
2. Help Grossman and Arroyo decide whether to choose Option 1 or Option 2 for his/her compensation package.

Skyhigh Construction, Inc., has just hired two new senior executives, Allen Grossman and Felicia Arroyo, to oversee planned expansion of operations. As senior executives, they participate in a pick-your-salary plan. Each executive is given two options for compensation:

*Option 1:* A base pay of \$125,000 with a possibility of a large bonus.

*Option 2:* A base pay of \$150,000 with a possibility of a bonus, but the bonus would be one-half of the bonus under Option 1.

Grossman and Arroyo understand that if the firm does not generate an operating profit in the fiscal year, then no bonuses are paid. Each executive has constructed a probability distribution given his/her personal background, underlying risk preferences, and subjective view of the economy.

Given the probability distributions and with the aid of expected values, the following analysis will attempt to choose the best option for each executive. Grossman, a married father with two young children, believes that Table 5.A best reflects his bonus payment expectations.

**TABLE 5.A** Calculating Grossman's Expected Bonus

Bonus (in \$), $x_i$	Probability, $P(x_i)$	Weighted Value, $x_i P(x_i)$
0	0.35	$0 \times 0.35 = 0$
50,000	0.45	$50,000 \times 0.45 = 22,500$
100,000	0.10	$100,000 \times 0.10 = 10,000$
150,000	0.10	$150,000 \times 0.10 = 15,000$
		Total = \$47,500

Expected bonus,  $E(X)$ , is calculated as a weighted average of all possible bonus values and is shown at the bottom of the third column of Table 5.A. Grossman's expected bonus is \$47,500. Using this value for his bonus, his salary options are

*Option 1:*  $\$125,000 + \$47,500 = \$172,500$

*Option 2:*  $\$150,000 + (1/2 \times \$47,500) = \$173,750$

Grossman should choose *Option 2* as his salary plan.

Arroyo is single with few financial constraints. Table 5.B shows the expected value of her bonus given her probability distribution.

**TABLE 5.B** Calculating Arroyo's Expected Bonus

Bonus (in \$), $x_i$	Probability, $P(x_i)$	Weighted Value, $x_i P(x_i)$
0	0.20	$0 \times 0.20 = 0$
50,000	0.25	$50,000 \times 0.25 = 12,500$
100,000	0.35	$100,000 \times 0.35 = 35,000$
150,000	0.20	$150,000 \times 0.20 = 30,000$
		Total = \$77,500

Arroyo's expected bonus amounts to \$77,500. Thus, her salary options are

*Option 1:*  $\$125,000 + \$77,500 = \$202,500$

*Option 2:*  $\$150,000 + (1/2 \times \$77,500) = \$188,750$

Arroyo should choose *Option 1* as her salary plan.

## CONCEPTUAL REVIEW

### LO 5.1 Distinguish between discrete and continuous random variables.

A **random variable** summarizes outcomes of an experiment with numerical values. A random variable is either discrete or continuous. A **discrete random variable** assumes a countable number of distinct values, whereas a **continuous random variable** is characterized by uncountable values in an interval.

### LO 5.2 Describe the probability distribution for a discrete random variable.

The **probability distribution function** for a discrete random variable  $X$  is a list of the values of  $X$  with the associated probabilities; that is, the list of all possible pairs  $(x, P(X = x))$ . The **cumulative distribution function** of  $X$  is defined as  $P(X \leq x)$ .

### LO 5.3 Calculate and interpret summary measures for a discrete random variable.

For a discrete random variable  $X$  with values  $x_1, x_2, x_3, \dots$ , which occur with probabilities  $P(X = x_i)$ , the **expected value** of  $X$  is calculated as  $E(X) = \mu = \sum x_i P(X = x_i)$ . We interpret the expected value as the long-run average value of the random variable over infinitely many independent repetitions of an experiment. Measures of dispersion indicate whether the values of  $X$  are clustered about  $\mu$  or widely scattered from  $\mu$ . The **variance** of  $X$  is calculated as  $Var(X) = \sigma^2 = \sum (x_i - \mu)^2 P(X = x_i)$ . The **standard deviation** of  $X$  is  $SD(X) = \sigma = \sqrt{\sigma^2}$ .

### LO 5.4 Distinguish between risk-neutral, risk-averse, and risk-loving consumers.

In general, a **risk-averse consumer** expects a reward for taking risk. A risk-averse consumer may decline a risky prospect even if it offers a positive expected gain. A **risk-neutral consumer** completely ignores risk and always accepts a prospect that offers a positive expected gain. Finally, a **risk-loving consumer** may accept a risky prospect even if the expected gain is negative.

### LO 5.5 Calculate and interpret summary measures to evaluate portfolio returns.

**Portfolio return**  $R_p$  is represented as a linear combination of the individual returns. With two assets,  $R_p = w_A R_A + w_B R_B$ , where  $R_A$  and  $R_B$  represent asset returns and  $w_A$  and  $w_B$  are the corresponding portfolio weights. The **expected return** and the **variance** for the portfolio are  $E(R_p) = w_A E(R_A) + w_B E(R_B)$  and  $Var(R_p) = w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2w_A w_B \sigma_{AB}$ , or equivalently,  $Var(R_p) = w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2w_A w_B \rho_{AB} \sigma_A \sigma_B$ .

### LO 5.6 Describe the binomial distribution and compute relevant probabilities.

A **Bernoulli process** is a series of  $n$  independent and identical trials of an experiment such that on each trial there are only two possible outcomes, conventionally labeled “success” and “failure.” The probabilities of success and failure, denoted  $p$  and  $1 - p$ , remain the same from trial to trial.

For a **binomial random variable**  $X$ , the probability of  $x$  successes in  $n$  Bernoulli trials is  $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x}$  for  $x = 0, 1, 2, \dots, n$ .

The **expected value**, the **variance**, and the **standard deviation** of a binomial random variable are  $E(X) = np$ ,  $Var(X) = \sigma^2 = np(1 - p)$ , and  $SD(X) = \sigma = \sqrt{np(1 - p)}$ , respectively.

### LO 5.7 Describe the Poisson distribution and compute relevant probabilities.

A **Poisson random variable** counts the number of occurrences of a certain event over a given interval of time or space. For simplicity, we call these occurrences “successes.”

For a Poisson random variable  $X$ , the probability of  $x$  successes over a given interval of time or space is  $P(X = x) = \frac{e^{-\mu} \mu^x}{x!}$  for  $x = 0, 1, 2, \dots$ , where  $\mu$  is the mean number of successes and  $e \approx 2.718$  is the base of the natural logarithm. The **expected value**, the **variance**, and the **standard deviation** of a Poisson distribution are  $E(X) = \mu$ ,  $Var(X) = \sigma^2 = \mu$ , and  $SD(X) = \sigma = \sqrt{\mu}$ , respectively.

**LO 5.8 Describe the hypergeometric distribution and compute relevant probabilities.**

The hypergeometric distribution is appropriate in applications where the trials are not independent and the probability of success changes from trial to trial. We use it in place of the binomial distribution when we are **sampling without replacement** from a population whose size  $N$  is not significantly larger than the sample size  $n$ . For a **hypergeometric random variable**  $X$ , the probability of  $x$  successes in a random selection of  $n$  items is

$$P(X = x) = \frac{\binom{S}{x} \binom{N-S}{n-x}}{\binom{N}{n}} \text{ for } x = 0, 1, 2, \dots, n \text{ if } n \leq S \text{ or } x = 0, 1, 2, \dots, S \text{ if } n > S, \text{ where } N$$

denotes the number of items in the population of which  $S$  are successes. The **expected value**, the **variance**, and the **standard deviation** of a hypergeometric distribution are  $E(X) = n\left(\frac{S}{N}\right)$ ,  $Var(X) = \sigma^2 = n\left(\frac{S}{N}\right)\left(1 - \frac{S}{N}\right)\left(\frac{N-n}{N-1}\right)$ , and  $SD(X) = \sigma = \sqrt{n\left(\frac{S}{N}\right)\left(1 - \frac{S}{N}\right)\left(\frac{N-n}{N-1}\right)}$ , respectively.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

81. Facing the worst economic climate since the dot-com bust in the early 2000s, high-tech companies in the United States search for investment opportunities with cautious optimism (*USA TODAY*, February 17, 2009). Suppose the investment team at Microsoft is considering an innovative start-up project. According to its estimates, Microsoft can make a profit of \$5 million if the project is very successful and \$2 million if it is somewhat successful. It also stands to lose \$4 million if the project fails. Calculate the expected profit or loss for Microsoft if the probabilities that the project is very successful and somewhat successful are 0.10 and 0.40, respectively, with the remaining amount being the failure probability.
82. An analyst developed the following probability distribution for the rate of return for a common stock.

Scenario	Probability	Rate of Return
1	0.25	−15%
2	0.35	5%
3	0.40	10%

- Calculate the expected rate of return.
  - Calculate the variance and the standard deviation of this probability distribution.
83. Consider the following information on the expected return for companies X and Y.

Economy	Probability	X	Y
Boom	0.20	30%	10%
Neutral	0.50	10%	20%
Poor	0.30	−30%	5%

- Calculate the expected value and the standard deviation of returns for companies X and Y.
  - Calculate the correlation coefficient if the covariance between X and Y is 88.
84. An investor owns a portfolio consisting of two mutual funds, A and B, with 35% invested in A. The following table lists the inputs for these funds.

Measures	Fund A	Fund B
Expected Value	10	5
Variance	98	26
Covariance	22	

- Calculate the expected value for the portfolio return.
  - Calculate the standard deviation for the portfolio return.
85. A professor uses a relative scale for grading. She announces that 60% of the students will get at least a B, with 15% getting A's. Also, 5% will get a D and another 5% will get an F. Assume that no incompletes are given in the course. Let Score be defined by 4 for A, 3 for B, 2 for C, 1 for D, and 0 for F.



- a. Find the probability that a student gets a B.
  - b. Find the probability that a student gets at least a C.
  - c. Compute the expected value and the standard deviation of Score.
86. Fifty percent of the customers who go to Sears Auto Center for tires buy four tires and 30% buy two tires. Moreover, 18% buy fewer than two tires, with 5% buying none.
- a. Find the expected value and the standard deviation of the number of tires a customer buys.
  - b. If Sears Auto Center makes a \$15 profit on every tire it sells, what is its expected profit if it services 120 customers?
87. Rent-to-own (RTO) stores allow consumers immediate access to merchandise in exchange for a series of weekly or monthly payments. The agreement is for a fixed time period. At the same time, the customer has the flexibility to terminate the contract by returning the merchandise. Suppose the RTO store makes a \$200 profit on appliances when the customer ends up owning the merchandise by making all payments. It makes a \$20 profit when the customer returns the product and a loss of \$600 when the customer defaults. Let the return and default probabilities be 0.60 and 0.05, respectively.
- a. Construct a probability distribution for the profit per appliance.
  - b. What is the expected profit for a store that sells 200 rent-to-own contracts?
88. Forty-four percent of consumers with credit cards carry balances from month to month (bankrate.com, February 20, 2007). Four consumers with credit cards are randomly selected.
- a. What is the probability that all consumers carry a credit card balance?
  - b. What is the probability that fewer than two consumers carry a credit card balance?
  - c. Calculate the expected value, the variance, and the standard deviation for this binomial distribution.
89. According to the Department of Transportation, 27% of domestic flights were delayed in 2007 (*Money*, May 2008). At New York's John F. Kennedy Airport, five flights are randomly selected.
- a. What is the probability that all five flights are delayed?
  - b. What is the probability that all five are on time?
90. Apple products have become a household name in America with 51% of all households owning at least one Apple product (*CNN*, March 19, 2012).
- a. What is the probability that two in a random sample of four households own an Apple product?
  - b. What is the probability that all four in a random sample of four households own an Apple product?
  - c. In a random sample of 100 households, find the expected value and the standard deviation for the number of households that own an Apple product.
91. (Use computer) Twenty percent of U.S. mortgages are "underwater" (*The Boston Globe*, March 5, 2009). A mortgage is considered underwater if the value of the home is less than what is owed on the mortgage. Suppose 100 mortgage holders are randomly selected.
- a. What is the probability that exactly 15 of the mortgages are underwater?
  - b. What is the probability that more than 20 of the mortgages are underwater?
  - c. What is the probability that at least 25 of the mortgages are underwater?
92. (Use computer) According to a survey by consulting firm Watson Wyatt, approximately 19% of employers have eliminated perks or plan to do so in the next year (*Kiplinger's Personal Finance*, February 2009). Suppose 30 employers are randomly selected.
- a. What is the probability that exactly ten of the employers have eliminated or plan to eliminate perks?
  - b. What is the probability that at least ten employers, but no more than 20 employers, have eliminated or plan to eliminate perks?
  - c. What is the probability that at most eight employers have eliminated or plan to eliminate perks?
93. Studies have shown that bats can consume an average of ten mosquitoes per minute (berkshiremuseum.org).
- a. Calculate the average number of mosquitoes that a bat consumes in a 30-second interval.
  - b. What is the probability that a bat consumes four mosquitoes in a 30-second interval?
  - c. What is the probability that a bat does not consume any mosquitoes in a 30-second interval?
  - d. What is the probability that a bat consumes at least one mosquito in a 30-second interval?
94. (Use computer) Despite the fact that home prices seem affordable and mortgage rates are at historic lows, real estate agents say they are showing more homes, but not selling more (*The Boston Globe*, March 7, 2009). A real estate company estimates that an average of five people show up at an open house to view a property. There is going to be an open house on Sunday.
- a. What is the probability that at least five people will show up to view the property?
  - b. What is the probability that fewer than five people will show up to view the property?
95. (Use computer) The police have estimated that there are 12 major accidents per day on a particular 10-mile stretch of a national highway. Suppose the



- incidence of accidents is evenly distributed on this 10-mile stretch of the highway.
- Find the probability that there will be fewer than eight major accidents on this 10-mile stretch of the highway.
  - Find the probability that there will be more than two accidents on a 1-mile stretch of this highway.
96. Suppose you draw three cards, without replacement, from a deck of well shuffled cards. Remember that each deck consists of 52 cards, with 13 each of spades, hearts, clubs, and diamonds.
- What is the probability that you draw all spades?
  - What is the probability that you draw two or fewer spades?
  - What is the probability that you draw all spades or hearts?
97. A professor has learned that three students in her class of 20 will cheat on the exam. She decides to focus her attention on four randomly chosen students during the exam.
- What is the probability that she finds at least one of the students cheating?
  - What is the probability that she finds at least one of the students cheating if she focuses on six randomly chosen students?
98. (Use computer) Many U.S. households still do not have Internet access. Suppose 20 out of 80 households in a small southern town do not have Internet access. A company that provides high-speed Internet has recently entered the market. As part of the marketing campaign, the company decides to randomly select ten households and offer them free laptops along with a brochure that describes their services. The aim is to build goodwill and, with a free laptop, tempt nonusers into getting Internet access.
- What is the probability that six laptop recipients do not have Internet access?
  - What is the probability that at least five laptop recipients do not have Internet access?
  - What is the probability that two or fewer laptop recipients do not have Internet access?
  - What is the expected number of laptop recipients who do not have Internet access?

## CASE STUDIES

**CASE STUDY 5.1** An extended warranty is a prolonged warranty offered to consumers by the warranty administrator, the retailer, or the manufacturer. A recent report in *The New York Times* (November 23, 2009) suggests that 20.4% of laptops fail over three years. Roberto D'Angelo is interested in an extended warranty for his laptop. A good extended warranty is being offered at Compuvest.com for \$74. It will cover any repair job that his laptop may need in the next three years. Based on his research, he determines that the likelihood of a repair job in the next three years is 13% for a minor repair, 8% for a major repair, and 3% for a catastrophic repair. The extended warranty will save him \$80 for a minor repair, \$320 for a major repair, and \$500 for a catastrophic repair. These results are summarized in the following probability distribution.

**Data for Case Study 5.1** Probability Distribution for Repair Cost

Type of Repair	Probability	Repair Cost
None	0.76	\$0
Minor	0.13	\$80
Major	0.08	\$320
Catastrophic	0.03	\$500

In a report, use the above information to

- Calculate and interpret the expected value of the repair cost.
- Analyze the expected gain or loss for a consumer who buys the extended warranty.
- Determine what kind of a consumer (risk neutral, risk averse, or both) will buy this extended warranty.

**CASE STUDY 5.2** According to figures released by the New York City government, smoking among New York City teenagers is on a decline, continuing a trend that began more than a decade ago (*The New York Times*, January 2, 2008). According to the

New York City Youth Risk Behavior Survey, the teenage smoking rate dropped to 8.5% in 2007 from about 17.6% in 2001 and 23% in 1997. City officials attribute the lower smoking rate to factors including a cigarette tax increase, a ban on workplace smoking, and television and subway ads that graphically depict tobacco-related illnesses. In a report, use the above information to

1. Calculate the probability that at least one in a group of 10 New York City teenagers smoked in 2007.
2. Calculate the probability that at least one in a group of 10 New York City teenagers smoked in 2001.
3. Calculate the probability that at least one in a group of 10 New York City teenagers smoked in 1997.
4. Comment on the smoking trend between 1997 and 2007.

**CASE STUDY 5.3** Disturbing news regarding Scottish police concerns the number of crashes involving vehicles on operational duties (*BBC News*, March 10, 2008). Statistics showed that Scottish forces' vehicles had been involved in traffic accidents at the rate of 1,000 per year. The statistics included vehicles involved in 999 calls (the equivalent of 911 in the United States) and pursuits. Fire service and ambulance vehicles were not included in the figures.

In a report, use the above information to

1. Calculate and interpret the expected number of traffic accidents per day involving vehicles on operational duties.
2. Use this expected value to construct the probability distribution table that lists the probability of 0, 1, 2, . . . , 10 traffic accidents per day. Graph this distribution and summarize your findings.

## APPENDIX 5.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor.

### Minitab

#### The Binomial Distribution

- A. (Replicating Example 5.9a) From the menu choose **Calc > Probability Distributions > Binomial**.
- B. Select **Probability** since we are finding  $P(X = 5)$ . (For cumulative probabilities, select **Cumulative probability**.) Enter 100 as the **Number of trials** and 0.047 as the **Event probability**. Select **Input constant** and enter the value 5.

#### The Poisson Distribution

- A. (Replicating Example 5.11a) From the menu choose **Calc > Probability Distributions > Poisson**.
- B. Select **Cumulative probability** since we are finding  $P(X \leq 100)$ . (For calculating  $P(X = x)$ , select **Probability**.) Enter 114 for the **Mean**. Select **Input constant** and enter the value 100.

#### The Hypergeometric Distribution

- A. (Replicating Example 5.12a) From the menu choose **Calc > Probability Distributions > Hypergeometric**.
- B. Select **Probability** since we are finding  $P(X = 1)$ . (For cumulative probabilities, select **Cumulative probability**.) Enter 20 for the **Population size (N)**, 2 for **Event count in population (M)**, and 5 for the **Sample size (n)**. Select **Input constant** and enter 1.

## SPSS

Note: In order for the calculated probability to be seen on the spreadsheet, SPSS must first “view” data on the spreadsheet. For this purpose, enter a value of zero in the first cell of the first column.

### The Binomial Distribution

- A. (Replicating Example 5.9a) From the menu choose **Transform > Compute Variable**.
- B. Under **Target Variable**, type pdfbinomial. Under **Function group**, select **PDF & Noncentral PDF**, and under **Functions and Special Variables**, double-click on **Pdf.Binom.** (For cumulative probabilities, under **Function group** select **CDF & Noncentral CDF**, and under **Functions and Special Variables** double-click on **Cdf.Binom.**) In the **Numeric Expression** box, enter 5 for **quant**, 100 for **n**, and 0.047 for **prob**.

### The Poisson Distribution

- A. (Replicating Example 5.11a) From the menu choose **Transform > Compute Variable**.
- B. Under **Target Variable**, type cdfpoisson. Under **Function group**, select **CDF & Noncentral CDF**, and under **Functions and Special Variables**, double-click on **Pdf.Poisson.** (For calculating  $P(X = x)$ , under **Function group** select **PDF & Noncentral PDF**, and under **Functions and Special Variables**, double-click on **Pdf.Poisson.**) In the **Numeric Expression** box, enter 100 for **quant** and 114 for **Mean**.

### The Hypergeometric Distribution

- A. (Replicating Example 5.12a) From the menu choose **Transform > Compute Variable**.
- B. Under **Target Variable**, type pdfhyper. Under **Function group**, select **PDF & Noncentral PDF**, and under **Functions and Special Variables**, double-click on **Pdf.Hyper.** (For cumulative probabilities, under **Function group** select **CDF & Noncentral CDF**, and under **Functions and Special Variables** double-click on **Cdf.Hyper.**) In the **Numeric Expression** box, enter 1 for **quant**, 20 for **total**, 5 for **sample**, and 2 for **hits**.

## JMP

Note: In order for the calculated probability to be seen on the spreadsheet, JMP must first “view” data on the spreadsheet. For this purpose, enter a value of zero in the first cell of the first column.

### The Binomial Distribution

- A. (Replicating Example 5.9a) Right-click at the top of the column in the spreadsheet view and select **Formula**. Under **Functions (grouped)**, choose **Discrete Probability > Binomial Probability**. (For cumulative probabilities, select **Binomial Distribution**.)
- B. Enter 0.047 for **p**, 100 for **n**, and 5 for **k**.

### The Poisson Distribution

- A. (Replicating Example 5.11a) Right-click at the top of the column in the spreadsheet view and select **Formula**. Under **Functions (grouped)**, choose **Discrete Probability > Poisson Distribution**. (For calculating  $P(X = x)$ , select **Poisson Probability**.)
- B. Enter 114 for **lambda** and 100 for **k**.

### The Hypergeometric Distribution

- A. (Replicating Example 5.12a) Right-click at the top of the column in the spreadsheet view and select **Formula**. Under **Functions (grouped)**, choose **Discrete Probability > Hypergeometric Probability**. (For cumulative probabilities, select **Hypergeometric Distribution**.)
- B. Enter 20 for **N**, 2 for **K**, 5 for **n**, and 1 for **x**.

# 6

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 6.1** Describe a continuous random variable.
- LO 6.2** Calculate and interpret probabilities for a random variable that follows the continuous uniform distribution.
- LO 6.3** Explain the characteristics of the normal distribution.
- LO 6.4** Use the standard normal table (*z* table).
- LO 6.5** Calculate and interpret probabilities for a random variable that follows the normal distribution.
- LO 6.6** Calculate and interpret probabilities for a random variable that follows the exponential distribution.
- LO 6.7** Calculate and interpret probabilities for a random variable that follows the lognormal distribution.

# Continuous Probability Distributions

In Chapter 5, we defined a random variable and discussed its numerical values. We classified the random variable as a discrete or a continuous random variable, depending on the range of numerical values that it assumes. A discrete random variable assumes a countable number of distinct values, such as the number of houses that a realtor sells in a month, the number of foreclosures in a sample of 100 households, and the number of cars lined up at a toll booth. A continuous random variable, on the other hand, is characterized by uncountable values because it can take on any value within an interval. Examples of a continuous random variable include the investment return on a mutual fund, the waiting time at a toll booth, and the amount of soda in a cup. In all of these examples, it is impossible to list all possible values of the random variable. In this chapter, we focus on continuous random variables. Most of this chapter is devoted to the discussion of the normal distribution, which is the most extensively used continuous probability distribution and is the cornerstone of statistical inference. Other important continuous distributions discussed are the continuous uniform, the exponential, and the lognormal distributions.





## INTRODUCTORY CASE

### Demand for Salmon

Akiko Hamaguchi is the manager of a small sushi restaurant called Little Ginza in Phoenix, Arizona. As part of her job, Akiko has to purchase salmon every day for the restaurant. For the sake of freshness, it is important that she buys the right amount of salmon daily. Buying too much may result in wastage, and buying too little may disappoint some customers on high-demand days.

Akiko has estimated that the daily consumption of salmon is normally distributed with a mean of 12 pounds and a standard deviation of 3.2 pounds. She has always bought 20 pounds of salmon every day. Lately, she has been criticized by the owners because this amount of salmon was too often resulting in wastage. As part of cost cutting, Akiko is considering a new strategy. She will buy salmon that is sufficient to meet the daily demand of customers on 90% of the days.

Akiko wants to use the above information to:

1. Calculate the probability that the demand for salmon at Little Ginza is above 20 pounds.
2. Calculate the probability that the demand for salmon at Little Ginza is below 15 pounds.
3. Determine the amount of salmon that should be bought daily so that the restaurant meets demand on 90% of the days.

A synopsis of this case is provided at the end of Section 6.3.

## 6.1 CONTINUOUS RANDOM VARIABLES AND THE UNIFORM DISTRIBUTION

### LO 6.1

Describe a continuous random variable.

As discussed in Chapter 5, a discrete random variable  $X$  assumes a countable number of distinct values such as  $x_1, x_2, x_3$ , and so on. A **continuous random variable**, on the other hand, is characterized by uncountable values because it can take on any value within an interval or collection of intervals. Unlike the case of a discrete random variable, we cannot describe the possible values of a continuous random variable  $X$  with a list  $x_1, x_2, \dots$  because the value  $(x_1 + x_2)/2$ , not in the list, might also be possible. Consider, for example, a continuous random variable defined by the amount of time a student takes to finish the exam. Here, it is impossible to put in a sequence all possible values of the random variable. Some students may think that time is countable in seconds; however, this may not be the case once we consider fractions of a second. Similarly, other continuous random variables, such as the investment return on a mutual fund and the amount of soda in a cup, are characterized by uncountable values.

For a discrete random variable, we can compute the probability that it assumes a particular value  $x$ , or written as a probability statement,  $P(X = x)$ . For instance, for a binomial random variable, we can calculate the probability of exactly one success in  $n$  trials; that is,  $P(X = 1)$ . We cannot make this calculation with a continuous random variable. The probability that a continuous random variable assumes a particular value  $x$  is zero; that is,  $P(X = x) = 0$ . This occurs because we cannot assign a nonzero probability to each of the uncountable values and still have the probabilities sum to one. Thus, for a continuous random variable it is only meaningful to calculate the probability that the value of the random variable falls within some specified interval. Therefore, for a continuous random variable,  $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$  since  $P(X = a)$  and  $P(X = b)$  are both zero.

In Chapter 5, we learned that a probability mass function for a discrete random variable  $X$  is a list of the values of  $X$  with the associated probabilities. For a continuous random variable, the counterpart to the probability mass function is called the **probability density function**, denoted by  $f(x)$ . In this text this book we often use the term “probability distribution” to refer to both functions. The graph of  $f(x)$  approximates the relative frequency polygon for the population. Unlike the probability mass function,  $f(x)$  does not provide probabilities directly. The probability that the variable assumes a value within an interval, say  $P(a \leq X \leq b)$ , is defined as the area under  $f(x)$  between points  $a$  and  $b$ . Moreover, the entire area under  $f(x)$  over all values of  $x$  must equal one; this is equivalent to the fact that, for discrete random variables, the probabilities add up to one.

### THE PROBABILITY DENSITY FUNCTION

The probability density function  $f(x)$  for a continuous random variable  $X$  has the following properties:

- $f(x) \geq 0$  for all possible values  $x$  of  $X$ , and
- the area under  $f(x)$  over all values of  $x$  equals one.

As in the case for a discrete random variable, we can use the **cumulative distribution function**, denoted by  $F(x)$ , to compute probabilities for a continuous random variable. For a value  $x$  of the random variable  $X$ ,  $F(x) = P(X \leq x)$  is simply the area under the probability density function up to the value  $x$ .

### THE CUMULATIVE DISTRIBUTION FUNCTION

For any value  $x$  of the random variable  $X$ , the cumulative distribution function  $F(x)$  is defined as

$$F(x) = P(X \leq x).$$

If you are familiar with calculus, then you will recognize that this cumulative probability is the integral of  $f(u)$  for values less than or equal to  $x$ . Similarly,  $P(a \leq X \leq b) = F(b) - F(a)$  is the integral of  $f(u)$  between points  $a$  and  $b$ . Fortunately, we do not necessarily need



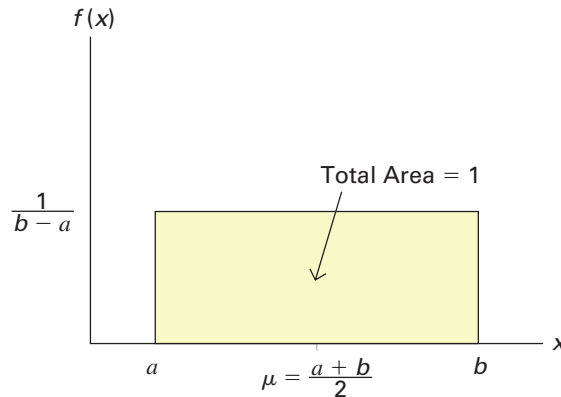
the knowledge of integral calculus to compute probabilities for the continuous random variables discussed in this text.

## The Continuous Uniform Distribution

One of the simplest continuous probability distributions is called the **continuous uniform distribution**. This distribution is appropriate when the underlying random variable has an equally likely chance of assuming a value within a specified range. Examples of uniformly distributed random variables include the delivery time of an appliance, the scheduled flight time between cities, and the waiting time for a campus bus. Any specified range for each of the above random variables can be assumed to be equally probable.

Suppose you are informed that your new refrigerator will be delivered between 2:00 pm and 3:00 pm. Let the random variable  $X$  denote the delivery time of your refrigerator. This variable is bounded below by 2:00 pm and above by 3:00 pm for a total range of 60 minutes. It is reasonable to infer that the probability of delivery between 2:00 pm and 2:30 pm equals 0.50 ( $=30/60$ ), as does the probability of delivery between 2:30 pm and 3:00 pm. Similarly, the probability of delivery in any 15-minute interval equals 0.25 ( $=15/60$ ), and so on.

Figure 6.1 depicts the probability density function for a continuous uniform random variable. The values  $a$  and  $b$  on the horizontal axis represent its lower and upper limits, respectively. The continuous uniform distribution is symmetric around its mean  $\mu$ , computed as  $\frac{a+b}{2}$ . In the refrigerator delivery example, the mean is  $\mu = \frac{2+3}{2} = 2.5$ , implying that you expect the delivery at 2:30 pm. The standard deviation  $\sigma$  of a continuous uniform variable equals  $\sqrt{(b-a)^2/12}$ .



**FIGURE 6.1** Continuous uniform probability density function

It is important to emphasize that the height of the probability density function does not directly represent a probability. As mentioned earlier, for all continuous random variables, it is the area under  $f(x)$  that corresponds to probability. For the continuous uniform distribution, the probability is essentially the area of a rectangle, which is the base times the height. Therefore, the probability is easily computed by multiplying the length of a specified interval (base) with  $f(x) = \frac{1}{b-a}$  (height).

### THE CONTINUOUS UNIFORM DISTRIBUTION

A random variable  $X$  follows the **continuous uniform distribution** if its probability density function is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \text{ and} \\ 0 & \text{for } x < a \text{ or } x > b, \end{cases}$$

where  $a$  and  $b$  represent the lower and upper limits of values, respectively, that the random variable assumes.

The expected value and the standard deviation of  $X$  are computed as

$$E(X) = \mu = \frac{a+b}{2} \quad \text{and} \quad SD(X) = \sigma = \sqrt{(b-a)^2/12}.$$

### LO 6.2

Calculate and interpret probabilities for a random variable that follows the continuous uniform distribution.

### EXAMPLE 6.1

A manager of a local drugstore is projecting next month's sales for a particular cosmetic line. She knows from historical data that sales follow a continuous uniform distribution with a lower limit of \$2,500 and an upper limit of \$5,000.

- What are the mean and the standard deviation for this continuous uniform distribution?
- What is the probability that sales exceed \$4,000?
- What is the probability that sales are between \$3,200 and \$3,800?

#### SOLUTION:

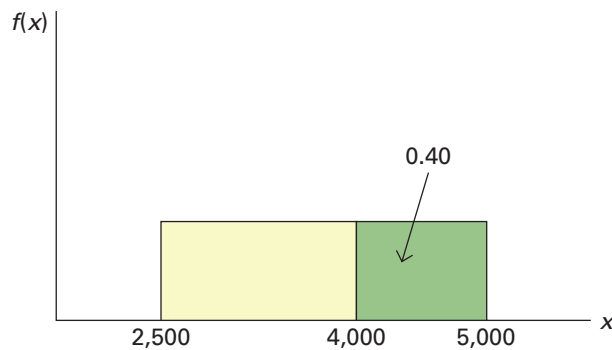
- With a value for the lower limit of  $a = \$2,500$  and a value for the upper limit of  $b = \$5,000$ , we calculate the mean and the standard deviation for this continuous uniform distribution as

$$\mu = \frac{a + b}{2} = \frac{\$2,500 + \$5,000}{2} = \$3,750, \text{ and}$$

$$\sigma = \sqrt{(b - a)^2 / 12} = \sqrt{(5,000 - 2,500)^2 / 12} = \$721.69.$$

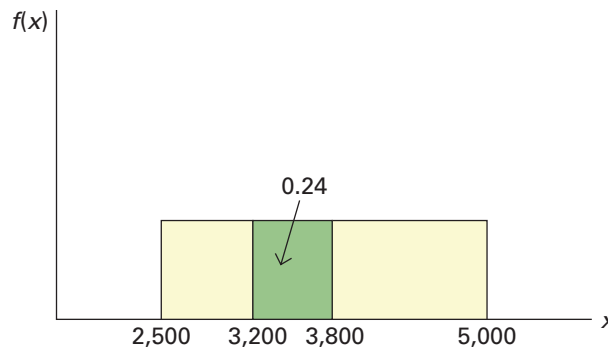
- When solving for the probability that sales exceed \$4,000, we find  $P(X > 4,000)$ , which is the area between \$4,000 and \$5,000, as shown in Figure 6.2. The base of the rectangle equals  $5,000 - 4,000 = 1,000$  and the height equals  $\frac{1}{5,000 - 2,500} = 0.0004$ . Thus,  $P(X > 4,000) = 1,000 \times 0.0004 = 0.40$ .

**FIGURE 6.2** Area to the right of 4,000 (Example 6.1b)



- When solving for the probability that sales are between \$3,200 and \$3,800, we find  $P(3,200 \leq X \leq 3,800)$ . Using the same methodology as in part b, we multiply the base times the height of the rectangle, as shown in Figure 6.3. Therefore, we obtain the probability as  $(3,800 - 3,200) \times 0.0004 = 0.24$ .

**FIGURE 6.3** Area between 3,200 and 3,800 (Example 6.1c)



## EXERCISES 6.1

### Mechanics

- The cumulative probabilities for a continuous random variable  $X$  are  $P(X \leq 10) = 0.42$  and  $P(X \leq 20) = 0.66$ . Calculate the following probabilities.
  - $P(X > 10)$
  - $P(X > 20)$
  - $P(10 < X < 20)$
- For a continuous random variable  $X$  with an upper bound of 4,  $P(0 \leq X \leq 2.5) = 0.54$  and  $P(2.5 \leq X \leq 4) = 0.16$ . Calculate the following probabilities.
  - $P(X < 0)$
  - $P(X > 2.5)$
  - $P(0 \leq X \leq 4)$
- For a continuous random variable  $X$ ,  $P(20 \leq X \leq 40) = 0.15$  and  $P(X > 40) = 0.16$ . Calculate the following probabilities.
  - $P(X < 40)$
  - $P(X < 20)$
  - $P(X = 40)$
- A random variable  $X$  follows the continuous uniform distribution with a lower bound of 5 and an upper bound of 35.
  - What is the height of the density function  $f(x)$ ?
  - What are the mean and the standard deviation for the distribution?
  - Calculate  $P(X > 10)$ .
- A random variable  $X$  follows the continuous uniform distribution with a lower bound of  $-2$  and an upper bound of 4.
  - What is the height of the density function  $f(x)$ ?
  - What are the mean and the standard deviation for the distribution?
  - Calculate  $P(X \leq -1)$ .
- A random variable  $X$  follows the continuous uniform distribution with a lower limit of 10 and an upper limit of 30.
  - Calculate the mean and the standard deviation for the distribution.
  - What is the probability that  $X$  is greater than 22?
  - What is the probability that  $X$  is between 15 and 23?
- A random variable  $X$  follows the continuous uniform distribution with a lower limit of 750 and an upper limit of 800.
  - Calculate the mean and the standard deviation for the distribution.
  - What is the probability that  $X$  is less than 770?
- Calculate the average price of electricity for a New England customer.
- What is the probability that a New England customer pays less than 15.5 cents per kilowatt-hour?
- A local carnival is not able to operate its rides if the average price of electricity is more than 14 cents per kilowatt-hour. What is the probability that the carnival will need to close?
- The arrival time of an elevator in a 12-story dormitory is equally likely at any time range during the next 4 minutes.
  - Calculate the expected arrival time.
  - What is the probability that an elevator arrives in less than  $1\frac{1}{2}$  minutes?
  - What is the probability that the wait for an elevator is more than  $1\frac{1}{2}$  minutes?
- The Netherlands is one of the world leaders in the production and sale of tulips. Suppose the heights of the tulips in the greenhouse of Rotterdam's Fantastic Flora follow a continuous uniform distribution with a lower bound of 7 inches and an upper bound of 16 inches. You have come to the greenhouse to select a bouquet of tulips, but only tulips with a height greater than 10 inches may be selected. What is the probability that a randomly selected tulip is tall enough to pick?
- The scheduled arrival time for a daily flight from Boston to New York is 9:25 am. Historical data show that the arrival time follows the continuous uniform distribution with an early arrival time of 9:15 am and a late arrival time of 9:55 am.
  - Calculate the mean and the standard deviation of the distribution.
  - What is the probability that a flight arrives late (later than 9:25 am)?
- You were informed at the nursery that your peach tree will definitely bloom sometime between March 18 and March 30. Assume that the bloom times follow a continuous uniform distribution between these specified dates.
  - What is the probability that the tree does not bloom until March 25?
  - What is the probability that the tree will bloom by March 20?
- You have been informed that the assessor will visit your home sometime between 10:00 am and 12:00 pm. It is reasonable to assume that his visitation time is uniformly distributed over the specified two-hour interval. Suppose you have to run a quick errand at 10:00 am.
  - If it takes 15 minutes to run the errand, what is the probability that you will be back before the assessor visits?
  - If it takes 30 minutes to run the errand, what is the probability that you will be back before the assessor visits?

### Applications

- Suppose the average price of electricity for a New England customer follows the continuous uniform distribution with a lower bound of 12 cents per kilowatt-hour and an upper bound of 20 cents per kilowatt-hour.

## 6.2 THE NORMAL DISTRIBUTION

The **normal probability distribution**, or simply the **normal distribution**, is the familiar **bell-shaped distribution**. It is also referred to as the Gaussian distribution.<sup>1</sup> The normal distribution is the most extensively used probability distribution in statistical work. One reason for this common use is that the normal distribution closely approximates the probability distribution for a wide range of random variables of interest. Examples of random variables that closely follow a normal distribution include:

- Heights and weights of newborn babies
- Scores on the SAT
- Cumulative debt of college graduates
- Advertising expenditure of firms
- Rate of return on an investment

Whenever possible, it is instructive to analyze the underlying data to determine if the normal distribution is appropriate for a given application. There are various ways to do this including inspecting histograms (Chapter 2) and boxplots (Chapter 3) for symmetry and bell shape. In this chapter, we simply assume that the random variable in question is normally distributed and focus on finding probabilities associated with this type of random variable. The computation of these probabilities is easy and direct. Another important function of the normal distribution is that it serves as the cornerstone of statistical inference. Recall from Chapter 1 that the study of statistics is divided into two branches: Descriptive Statistics and Inferential Statistics. Statistical inference is generally based on the assumption of the normal distribution and serves as the major topic in the remainder of this text.

### LO 6.3

Explain the characteristics of the normal distribution.

### Characteristics of the Normal Distribution

- The normal distribution is **bell-shaped** and **symmetric** around its mean; that is, one side of the mean is just the mirror image of the other side. In other words, the mean, the median, and the mode are all equal for a normally distributed random variable.
- The normal distribution is **completely described by two parameters**—the population mean  $\mu$  and the population variance  $\sigma^2$ . The population mean describes the central location and the population variance describes the dispersion of the distribution.
- The normal distribution is **asymptotic** in the sense that the tails get closer and closer to the horizontal axis but never touch it. Thus, theoretically, a normal random variable can assume any value between minus infinity and plus infinity.

The following definition mathematically expresses the probability density function of the normal distribution.

#### THE NORMAL DISTRIBUTION

A random variable  $X$  with mean  $\mu$  and variance  $\sigma^2$  follows the normal distribution if its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where  $\pi$  equals approximately 3.14159 and  $\exp(w) = e^w$  is the exponential function where  $e \approx 2.718$  is the base of the natural logarithm.

<sup>1</sup>The discovery of the normal (Gaussian) distribution is often credited to Carl Friedrich Gauss (1777–1855), even though some attribute the credit to De Moivre (1667–1754), who had earlier discovered it in the context of simplifying the binomial distribution calculations.

A graph depicting the normal probability density function is often referred to as the **normal curve** or the **bell curve**. The following example relates the normal curve to the location and the dispersion of the normally distributed random variable.

### EXAMPLE 6.2

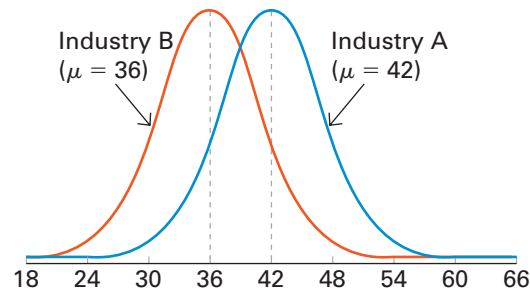
Suppose we know that the ages of employees in Industries A, B, and C are normally distributed. We are given the following information on the relevant parameters:

Industry A	Industry B	Industry C
$\mu = 42$ years	$\mu = 36$ years	$\mu = 42$ years
$\sigma = 5$ years	$\sigma = 5$ years	$\sigma = 8$ years

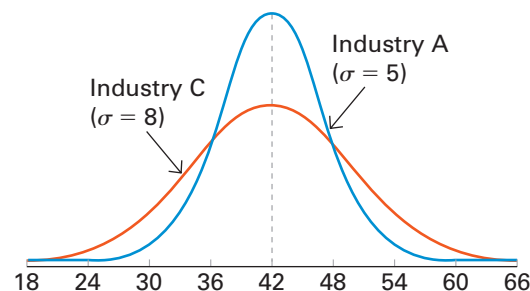
Graphically compare the ages of employees in Industry A with Industry B. Repeat the comparison for Industry A with Industry C.

**SOLUTION:** Figure 6.4 illustrates the difference in location given that the mean age of employees of Industry A is greater than that of Industry B. Both distributions show the same dispersion since the standard deviation is the same. Figure 6.5 compares the dispersion given that the standard deviation of age in Industry A is less than that of Industry C. Here, the peak of Industry A is higher than the peak of Industry C, reflecting the fact that an employee's age is likelier to be closer to the mean age in Industry A. These graphs also serve to point out that we can capture the entire distribution of any normally distributed random variable based on its mean and variance (or standard deviation).

**FIGURE 6.4** Normal probability density function for two values of  $\mu$  along with  $\sigma = 5$



**FIGURE 6.5** Normal probability density function for two values of  $\sigma$  along with  $\mu = 42$



We generally use the cumulative distribution function  $F(x)$  to compute probabilities for a normally distributed random variable, where  $F(x) = P(X \leq x)$  is simply the area under  $f(x)$  up to the value  $x$ . As mentioned earlier, we do not necessarily need the knowledge of integral calculus to compute probabilities for the normal distribution. Instead, we rely on a table to find probabilities. We can also compute probabilities with certain calculators, Excel, and other statistical packages. The specifics of how to use the table are delineated next.

Use the standard normal table ( $z$  table).

## The Standard Normal Variable

The **standard normal distribution** is a special case of the normal distribution with a mean equal to zero and a standard deviation (or variance) equal to one. Using the letter  $Z$  to denote a random variable with the standard normal distribution, we have  $E(Z) = 0$  and  $SD(Z) = 1$ . As usual, we use the lowercase letter  $z$  to denote the value that the random variable  $Z$  may assume.

The value  $z$  is actually the  $z$ -score that we discussed in Chapter 3. It measures the number of standard deviations a given value is away from the mean. For example, a  $z$ -score of 2 implies that the given value is 2 standard deviations above the mean. Similarly, a  $z$ -score of  $-1.5$  implies that the given value is 1.5 standard deviations below the mean. As mentioned in Chapter 3, converting values into  $z$ -scores is called standardizing the data.

In this section, we focus on solving problems related to the standard normal distribution. In the next section, we will show that any normal distribution is equivalent to the standard normal distribution when the unit of measurement is changed to measure standard deviations from the mean. Therefore, while most real-world normally distributed variables are not standard normal, we can always transform (standardize) them into standard normal to compute the relevant probabilities.

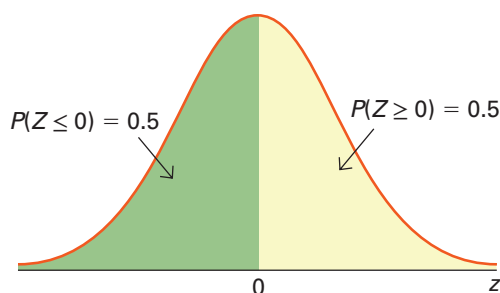
All introductory statistics texts include a **standard normal table**, also referred to as the  **$z$  table**, that provide areas (probabilities) under the  $z$  curve. However, the format of these tables is sometimes different. In this book, the  $z$  table provides cumulative probabilities  $P(Z \leq z)$ ; this table appears on two pages in Appendix A and is labeled Table 1. The left-hand page provides cumulative probabilities for  $z$  values less than or equal to zero. The right-hand page shows cumulative probabilities for  $z$  values greater than or equal to zero. Given the symmetry of the normal distribution and the fact that the area under the entire curve is one, other probabilities can be easily computed.

### STANDARD NORMAL DISTRIBUTION

The standard normal random variable  $Z$  is a normal random variable with  $E(Z) = 0$  and  $SD(Z) = 1$ . The  $z$  table provides cumulative probabilities  $P(Z \leq z)$  for positive and for negative values of  $z$ .

Figure 6.6 represents the standard normal distribution ( $z$  distribution). Since the random variable  $Z$  is symmetric around its mean of zero,  $P(Z < 0) = P(Z > 0) = 0.5$ . As is the case with all continuous random variables, we can also write the probabilities as  $P(Z \leq 0) = P(Z \geq 0) = 0.5$ .

**FIGURE 6.6**  
Standard normal probability density function



## Finding a Probability for a Given $z$ Value

As mentioned earlier, the  $z$  table provides cumulative probabilities  $P(Z \leq z)$  for a given  $z$ . Consider, for example, a cumulative probability  $P(Z \leq 1.52)$ . Since  $z = 1.52$  is positive, we can look up this probability from the right-hand page of the  $z$  table; Table 6.1 shows a portion of the table.

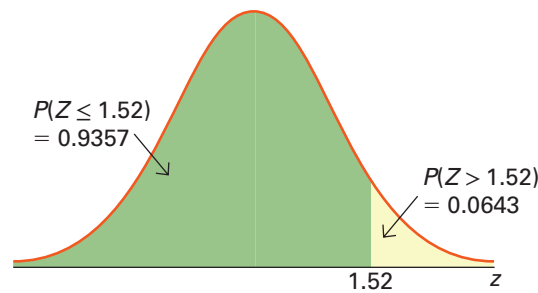


**TABLE 6.1** Portion of the Right-Hand Page of the  $z$  Table

$z$	0.00	0.01	0.02
0.0	0.5000	0.5040	↓
0.1	0.5398	0.5438	↓
⋮	⋮	⋮	⋮
1.5	→	→	0.9357

The first column of the table, denoted as the  $z$  column, shows values of  $z$  up to the tenth decimal point, while the first row of the table, denoted as the  $z$  row, shows hundredths values. Thus, for  $z = 1.52$ , we match 1.5 on the  $z$  column with 0.02 on the  $z$  row to find a corresponding probability of 0.9357. The arrows in Table 6.1 indicate that  $P(Z \leq 1.52) = 0.9357$ .

In Figure 6.7, the cumulative probability corresponding to  $z = 1.52$  is highlighted. Note that  $P(Z \leq 1.52) = 0.9357$  represents the area under the  $z$  curve to the left of 1.52. Therefore, the area to the right of 1.52 can be computed as  $P(Z > 1.52) = 1 - P(Z \leq 1.52) = 1 - 0.9357 = 0.0643$ . Note that since  $P(Z > 1.52) = P(Z < -1.52)$ , an alternative way to find the probability of 0.0643 is to use the left-hand page of the  $z$  table as discussed next.

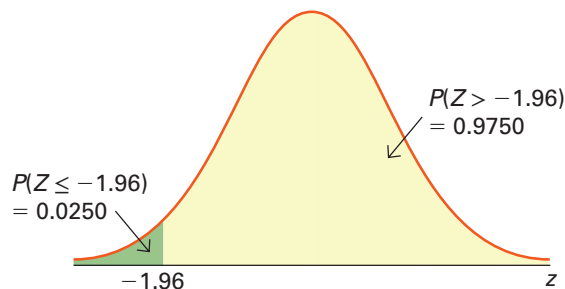
**FIGURE 6.7**

Cumulative probability with respect to  $z = 1.52$

Suppose we want to find  $P(Z \leq -1.96)$ . Since  $z$  is a negative value, we can look up this probability from the left-hand page of the  $z$  table; Table 6.2 shows a portion of the table with arrows indicating that  $P(Z \leq -1.96) = 0.0250$ . Figure 6.8 highlights the corresponding probability. As before, the area to the right of  $-1.96$  can be computed as  $P(Z > -1.96) = 1 - P(Z \leq -1.96) = 1 - 0.0250 = 0.9750$ . Note that since  $P(Z > -1.96) = P(Z < 1.96)$ , an alternative way is to use the right-hand page of the  $z$  table to find the probability as 0.9750.

**TABLE 6.2** Portion of the Left-Hand Page of the  $z$  Table

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06
-3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	↓
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	↓
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-1.9	→	→	→	→	→	→	0.0250

**FIGURE 6.8**

Cumulative probability with respect to  $z = -1.96$

### EXAMPLE 6.3

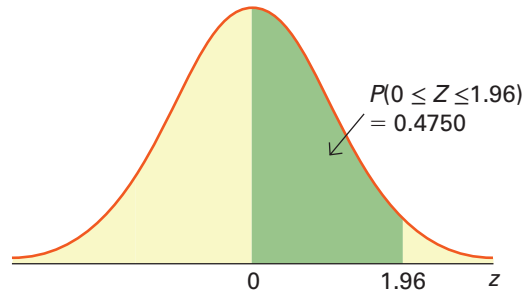
Find the following probabilities for the standard normal random variable  $Z$ .

- a.  $P(0 \leq Z \leq 1.96)$
- b.  $P(1.52 \leq Z \leq 1.96)$
- c.  $P(-1.52 \leq Z \leq 1.96)$
- d.  $P(Z > 4)$

**SOLUTION:** It always helps to start by highlighting the relevant probability in the  $z$  graph.

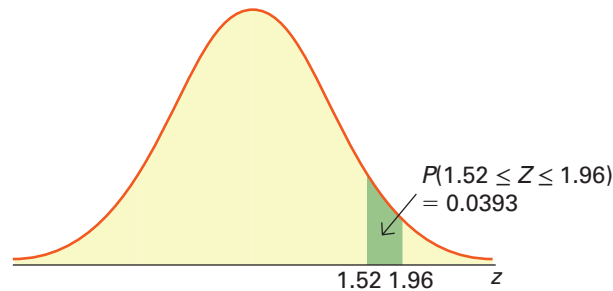
- a. As shown in Figure 6.9, the area between 0 and 1.96 is equivalent to the area to the left of 1.96 minus the area to the left of 0. Therefore,  $P(0 \leq Z \leq 1.96) = P(Z \leq 1.96) - P(Z < 0) = 0.9750 - 0.50 = 0.4750$ .

**FIGURE 6.9** Finding the probability between 0 and 1.96



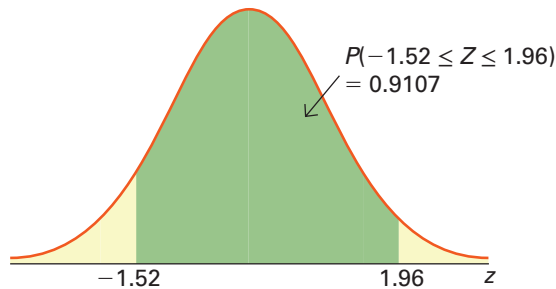
- b. As in part a and shown in Figure 6.10,  $P(1.52 \leq Z \leq 1.96) = P(Z \leq 1.96) - P(Z < 1.52) = 0.9750 - 0.9357 = 0.0393$ .

**FIGURE 6.10** Finding the probability between 1.52 and 1.96



- c. From Figure 6.11,  $P(-1.52 \leq Z \leq 1.96) = P(Z \leq 1.96) - P(Z < -1.52) = 0.9750 - 0.0643 = 0.9107$ .

**FIGURE 6.11** Finding the probability between -1.52 and 1.96



- d.  $P(Z > 4) = 1 - P(Z \leq 4)$ . However, the  $z$  table only goes up to 3.99 with  $P(Z \leq 3.99) = 1.0$  (approximately). In fact, for any  $z$  value greater than 3.99, it is acceptable to treat  $P(Z \leq z) = 1.0$ . Therefore,  $P(Z > 4) = 1 - P(Z \leq 4) = 1 - 1 = 0$ .

## Finding a $z$ Value for a Given Probability

So far we have computed probabilities for given  $z$  values. Now we will evaluate  $z$  values for given probabilities.

### EXAMPLE 6.4

For the standard normal variable  $Z$ , find the  $z$  values that satisfy the following probability statements.

- a.  $P(Z \leq z) = 0.6808$
- b.  $P(Z \leq z) = 0.90$
- c.  $P(Z \leq z) = 0.0643$
- d.  $P(Z > z) = 0.0212$
- e.  $P(-z \leq Z \leq z) = 0.95$

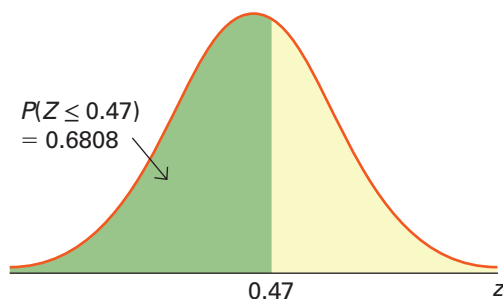
**SOLUTION:** As mentioned earlier, it is useful to draw a graph to set up a problem. Recall too that the  $z$  table lists  $z$  values along with the corresponding cumulative probabilities. Noncumulative probabilities can be evaluated using symmetry.

- a. Since the probability is already in a cumulative format—that is,  $P(Z \leq z) = 0.6808$ —we simply look up 0.6808 from the body of the table (right-hand side) to find the corresponding  $z$  value from the row/column of  $z$ . Table 6.3 shows the relevant portion of the  $z$  table and Figure 6.12 depicts the corresponding area. Therefore,  $z = 0.47$ .

**TABLE 6.3** Portion of the  $z$  Table for Example 6.4a

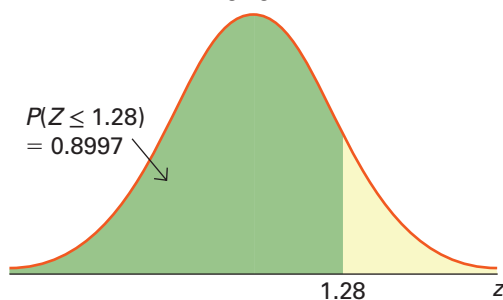
$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	↑
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	↑
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.4	←	←	←	←	←	←	←	0.6808

**FIGURE 6.12** Finding  $z$  given  $P(Z \leq z) = 0.6808$



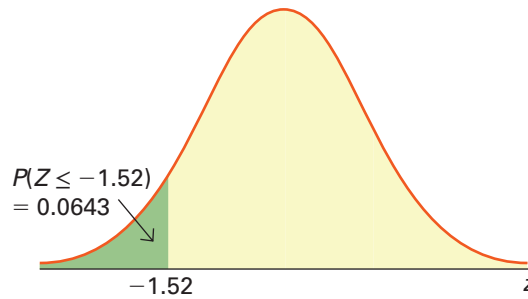
- b. When deriving  $z$  for  $P(Z \leq z) = 0.90$ , we find that the  $z$  table (right-hand side) does not contain the cumulative probability 0.90. In such cases, we use the closest cumulative probability to solve the problem. Therefore,  $z$  is approximately equal to 1.28, which corresponds to a cumulative probability of 0.8997. Figure 6.13 shows this result graphically.

**FIGURE 6.13** Finding  $z$  given  $P(Z \leq z) = 0.90$



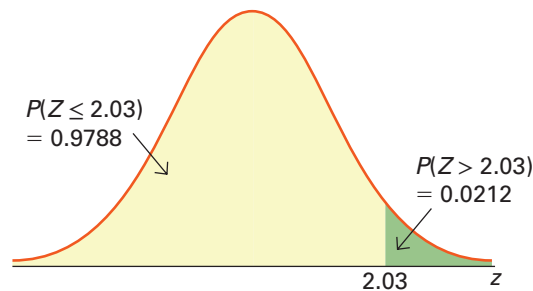
- c. As shown in Figure 6.14, the  $z$  value that solves  $P(Z \leq z) = 0.0643$  must be negative because the probability to its left is less than 0.50. We look up the cumulative probability 0.0643 in the table (left-hand side) to get  $z = -1.52$ .

**FIGURE 6.14** Finding  $z$  given  $P(Z \leq z) = 0.0643$



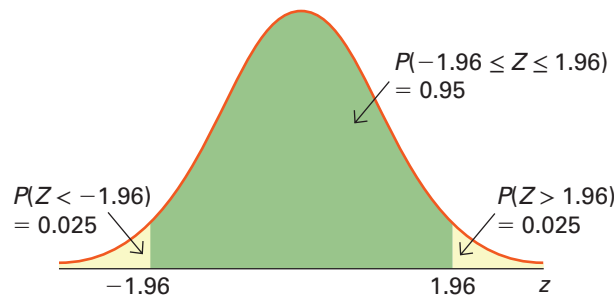
- d. When deriving  $z$  for  $P(Z > z) = 0.0212$ , we have to find a  $z$  value such that the probability to the right of this value is 0.0212. Since the table states cumulative probabilities, we look up  $P(Z \leq z) = 1 - 0.0212 = 0.9788$  in the table (right-hand side) to get  $z = 2.03$ . Figure 6.15 shows the results.

**FIGURE 6.15** Finding  $z$  given  $P(Z > z) = 0.0212$



- e. Since we know that the total area under the curve equals one, and we want to find  $-z$  and  $z$  such that the area between the two values equals 0.95, we can conclude that the area in either tail is 0.025; that is,  $P(Z < -z) = 0.025$  and  $P(Z > z) = 0.025$ . Figure 6.16 shows these results. We then use the cumulative probability,  $P(Z \leq z) = 0.95 + 0.025 = 0.975$ , to find  $z = 1.96$ .

**FIGURE 6.16** Finding  $z$  given  $P(-z \leq Z \leq z) = 0.95$



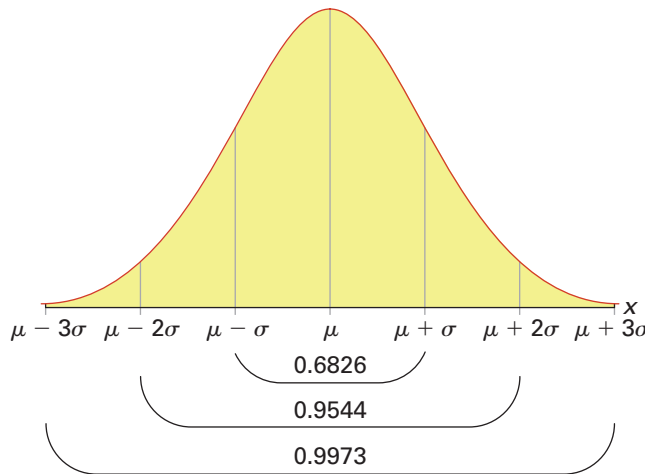
## Revisiting the Empirical Rule

In Chapter 3, we used the empirical rule to approximate the percentage of values that fall within 1, 2, or 3 standard deviations of the mean. Approximate percentages are

appropriate for many real-world applications where the normal distribution is used only as an approximation. For normally distributed random variables, we can find the exact percentages.

The empirical rule for normal distributions is shown in Figure 6.17. Given a normal random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$ :

- 68.26% of the values fall within 1 standard deviation of the mean; that is,  $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6826$ ,
- 95.44% of the values fall within 2 standard deviations of the mean; that is,  $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9544$ , and
- 99.73% of the values fall within 3 standard deviations of the mean; that is,  $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973$ .



**FIGURE 6.17**  
Graphical description  
of the empirical rule

For the standard normal variable  $Z$ ,  $P(-1 \leq Z \leq 1)$  refers to the probability within 1 standard deviation of the mean since  $\mu = 0$  and  $\sigma = 1$ . From the  $z$  table, we can show that  $P(-1 \leq Z \leq 1)$  equals  $P(Z \leq 1) - P(Z < -1) = 0.8413 - 0.1587 = 0.6826$ . Therefore, the exact probability that  $Z$  falls within 1 standard deviation of the mean is 0.6826. Similarly, the exact probabilities that  $Z$  falls within 2 and 3 standard deviations of the mean are  $P(-2 \leq Z \leq 2) = 0.9544$  and  $P(-3 \leq Z \leq 3) = 0.9973$ , respectively.

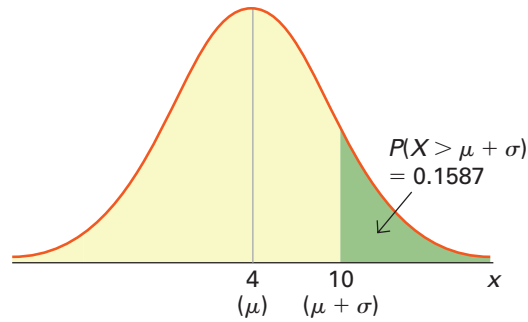
### EXAMPLE 6.5

An investment strategy has an expected return of 4% and a standard deviation of 6%. Assume that investment returns are normally distributed.

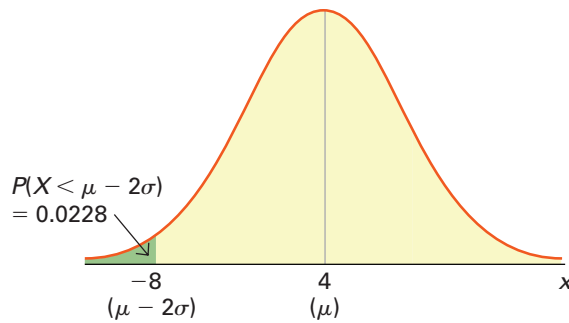
- What is the probability of earning a return greater than 10%?
- What is the probability of earning a return less than -8%?

**SOLUTION:** We use the empirical rule for normal distributions with  $\mu = 4$  and  $\sigma = 6$  to solve these questions.

- A return of 10% is one standard deviation above the mean because  $10 = 4 + 6$ . Since 68.26% of observations fall within one standard deviation of the mean, 31.74% ( $100\% - 68.26\%$ ) of the observations are outside the range. Using symmetry, we conclude that 15.87% (half of 31.74%) of the observations are greater than 10% (see Figure 6.18).

**FIGURE 6.18** Finding  $P(X > 10)$ 

- b. A return of  $-8\%$  is two standard deviations below the mean, or  $-8 = 4 - (2 \times 6)$ . Since 95.44% of the observations fall within two standard deviations of the mean, only 2.28% (half of 4.56%) are below  $-8\%$  (see Figure 6.19).

**FIGURE 6.19** Finding  $P(X < -8)$ 

## EXERCISES 6.2

### Mechanics

14. Find the following probabilities based on the standard normal variable  $Z$ .
  - a.  $P(Z > 1.32)$
  - b.  $P(Z \leq -1.32)$
  - c.  $P(1.32 \leq Z \leq 2.37)$
  - d.  $P(-1.32 \leq Z \leq 2.37)$
15. Find the following probabilities based on the standard normal variable  $Z$ .
  - a.  $P(Z > 0.74)$
  - b.  $P(Z \leq -1.92)$
  - c.  $P(0 \leq Z \leq 1.62)$
  - d.  $P(-0.90 \leq Z \leq 2.94)$
16. Find the following probabilities based on the standard normal variable  $Z$ .
  - a.  $P(-0.67 \leq Z \leq -0.23)$
  - b.  $P(0 \leq Z \leq 1.96)$
  - c.  $P(-1.28 \leq Z \leq 0)$
  - d.  $P(Z > 4.2)$
17. Find the following  $z$  values for the standard normal variable  $Z$ .
  - a.  $P(Z \leq z) = 0.9744$
  - b.  $P(Z > z) = 0.8389$
  - c.  $P(-z \leq Z \leq z) = 0.95$
  - d.  $P(0 \leq Z \leq z) = 0.3315$
18. Find the following  $z$  values for the standard normal variable  $Z$ .
  - a.  $P(Z \leq z) = 0.1020$
  - b.  $P(z \leq Z \leq 0) = 0.1772$
  - c.  $P(Z > z) = 0.9929$
  - d.  $P(0.40 \leq Z \leq z) = 0.3368$

### Applications

19. The historical returns on a balanced portfolio have had an average return of  $8\%$  and a standard deviation of  $12\%$ . Assume that returns on this portfolio follow a normal distribution. Use the empirical rule for normal distributions to answer the following questions.
  - a. What percentage of returns were greater than  $20\%$ ?
  - b. What percentage of returns were below  $-16\%$ ?



20. Assume that IQ scores follow a normal distribution with a mean of 100 and a standard deviation of 16. Use the empirical rule for normal distributions to answer the following questions.
  - a. What percentage of people score between 84 and 116?
  - b. What percentage of people score less than 68?
21. The average rent in a city is \$1,500 per month with a standard deviation of \$250. Assume rent follows the normal distribution. Use the empirical rule for normal distributions to answer the following questions.
  - a. What percentage of rents are between \$1,250 and \$1,750?
  - b. What percentage of rents are less than \$1,250?
  - c. What percentage of rents are greater than \$2,000?
22. A professional basketball team averages 80 points per game with a standard deviation of 10 points. Assume points per game follow the normal distribution. Use the empirical rule for normal distributions to answer the following questions.
  - a. What percentage of scores are between 60 and 100 points?
  - b. What percentage of scores are more than 100 points? If there are 82 games in a regular season, in how many games will the team score more than 100 points?

## 6.3 SOLVING PROBLEMS WITH NORMAL DISTRIBUTIONS

LO 6.5

In the preceding section, we found probabilities for the standard normal distribution, which is a normal distribution with mean zero and standard deviation one. For other normal distributions, we found probabilities using the empirical rule. However, in many applications, the underlying distribution is not standard normal and the interval for computing a probability cannot be expressed within one, two, or three standard deviations of the mean. In this section, we examine problems in these situations.

Calculate and interpret probabilities for a random variable that follows the normal distribution.

### The Transformation of Normal Random Variables

The importance of the standard normal distribution arises from the fact that any normal random variable can be transformed into the standard normal random variable to derive the relevant probabilities. In other words, any normally distributed random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$  can be transformed (standardized) into the standard normal variable  $Z$  with mean zero and standard deviation one. We transform  $X$  into  $Z$  by subtracting from  $X$  its mean and dividing by its standard deviation.

#### THE STANDARD TRANSFORMATION: CONVERTING $X$ INTO $Z$

Any normally distributed random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$  can be transformed into the standard normal random variable  $Z$  as

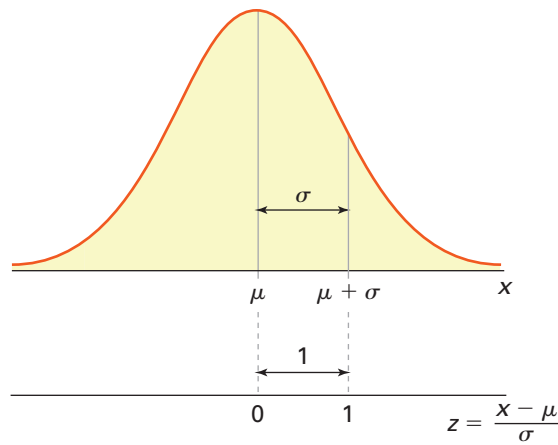
$$Z = \frac{X - \mu}{\sigma}.$$

Therefore, any value  $x$  has a corresponding value  $z$  given by

$$z = \frac{x - \mu}{\sigma}.$$

As illustrated in Figure 6.20, if the  $x$  value is at the mean—that is,  $x = \mu$ —then the corresponding  $z$  value is  $z = \frac{\mu - \mu}{\sigma} = 0$ . Similarly, if the  $x$  value is at one standard deviation above the mean—that is,  $x = \mu + \sigma$ —then the corresponding  $z$  value is  $z = \frac{\mu + \sigma - \mu}{\sigma} = 1$ . Therefore, by construction,  $E(Z) = 0$  and  $SD(Z) = 1$ .

**FIGURE 6.20**  
Transforming  $X$  with mean  $\mu$  and standard deviation  $\sigma$  to  $Z$  deviation  $\sigma$



### EXAMPLE 6.6

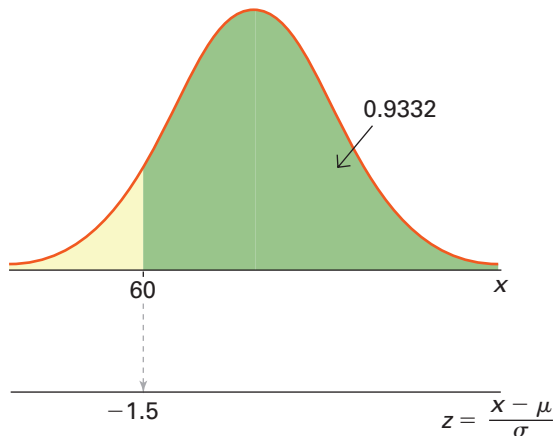
Scores on a management aptitude exam are normally distributed with a mean of 72 and a standard deviation of 8.

- What is the probability that a randomly selected manager will score above 60?
- What is the probability that a randomly selected manager will score between 68 and 84?

**SOLUTION:** Let  $X$  represent scores with  $\mu = 72$  and  $\sigma = 8$ . We will use the standard transformation  $z = \frac{x - \mu}{\sigma}$  to solve these problems.

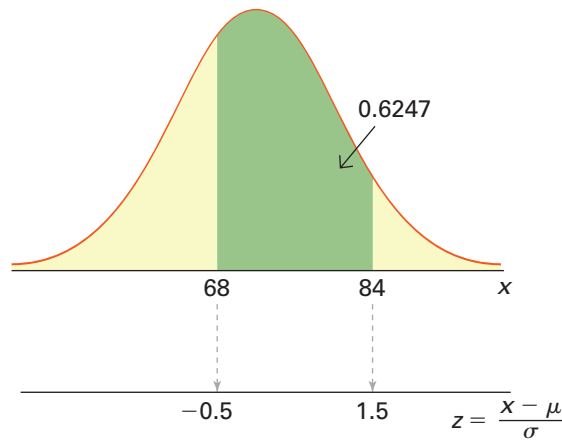
- The probability that a manager scores above 60 is  $P(X > 60)$ . Figure 6.21 shows the probability as the shaded area to the right of 60. We derive  $P(X > 60) = P\left(Z > \frac{60 - 72}{8}\right) = P(Z > -1.5)$ . Since  $P(Z > -1.5) = 1 - P(Z \leq -1.5)$ , we look up  $-1.50$  in the  $z$  table (left-hand side) to get this probability as  $1 - 0.0668 = 0.9332$ .

**FIGURE 6.21** Finding  $P(X > 60)$



- When solving for the probability that a manager scores between 68 and 84, we find  $P(68 \leq X \leq 84)$ . The shaded area in Figure 6.22 shows this probability. We derive  $P(68 \leq X \leq 84) = P\left(\frac{68 - 72}{8} \leq Z \leq \frac{84 - 72}{8}\right) = P(-0.5 \leq Z \leq 1.5)$ . We compute this probability using the  $z$  table as  $P(Z \leq 1.5) - P(Z < -0.5) = 0.9332 - 0.3085 = 0.6247$ .

**FIGURE 6.22** Finding  $P(68 \leq X \leq 84)$



## The Inverse Transformation

So far we have used the standard transformation to compute probabilities for given  $x$  values. We can use the **inverse transformation**,  $x = \mu + z\sigma$ , to compute  $x$  values for given probabilities.

### THE INVERSE TRANSFORMATION: CONVERTING $Z$ INTO $X$

The standard normal variable  $Z$  can be transformed to the normally distributed random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$  as  $X = \mu + Z\sigma$ .

Therefore, any value  $z$  has a corresponding value  $x$  given by  $x = \mu + z\sigma$ .

### EXAMPLE 6.7

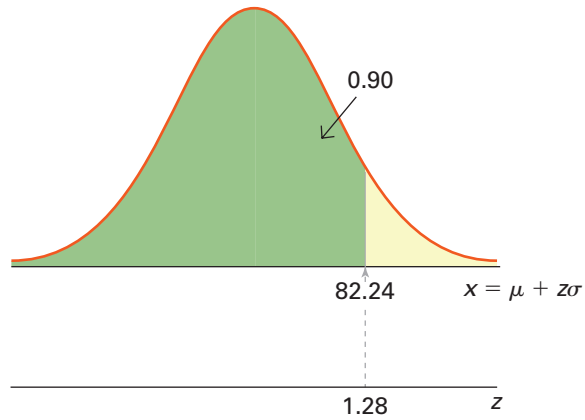
Scores on a management aptitude examination are normally distributed with a mean of 72 and a standard deviation of 8.

- What is the lowest score that will place a manager in the top 10% (90th percentile) of the distribution?
- What is the highest score that will place a manager in the bottom 25% (25th percentile) of the distribution?

**SOLUTION:** Let  $X$  represent scores on a management aptitude examination with  $\mu = 72$  and  $\sigma = 8$ . We will use the inverse transformation  $x = \mu + z\sigma$  to solve these problems.

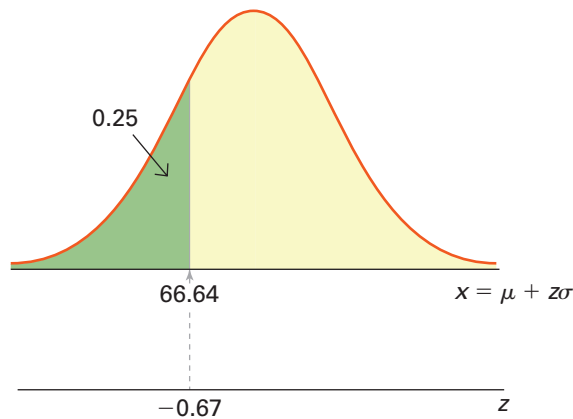
- The 90th percentile is a numerical value  $x$  such that  $P(X < x) = 0.90$ . We look up 0.90 (or the closest value to 0.90) in the  $z$  table (right-hand side) to get  $z = 1.28$  and use the inverse transformation to find  $x = 72 + 1.28(8) = 82.24$ . Therefore, a score of 82.24 or higher will place a manager in the top 10% of the distribution (see Figure 6.23).

**FIGURE 6.23** Finding  $x$  given  $P(X < x) = 0.90$



- b. The 25th percentile is a numerical value  $x$  such that  $P(X < x) = 0.25$ . Using the  $z$  table (left-hand side), we find the corresponding  $z$  value that satisfies  $P(Z < z) = 0.25$  as  $-0.67$ . We then solve  $x = 72 - 0.67(8) = 66.64$ . Therefore, a score of 66.64 or lower will place a manager in the bottom 25% of the distribution (see Figure 6.24).

**FIGURE 6.24** Finding  $x$  given  $P(X < x) = 0.25$



### EXAMPLE 6.8

We can now answer the questions first posed by Akiko Hamaguchi in the introductory case of this chapter. Recall that Akiko would like to buy the right amount of salmon for daily consumption at Little Ginza. Akiko has estimated that the daily consumption of salmon is normally distributed with a mean of 12 pounds and a standard deviation of 3.2 pounds. She wants to answer the following questions:

- What is the probability that the demand for salmon at Little Ginza is above 20 pounds?
- What is the probability that the demand for salmon at Little Ginza is below 15 pounds?
- How much salmon should be bought so that it meets customer demand on 90% of the days?

**SOLUTION:** Let  $X$  denote customer demand for salmon at the restaurant. We know that  $X$  is normally distributed with  $\mu = 12$  and  $\sigma = 3.2$ .

- a. When solving for the probability that the demand for salmon is more than 20 pounds, we find  $P(X > 20) = P\left(Z > \frac{20 - 12}{3.2}\right) = P(Z > 2.50) = 1 - 0.9938 = 0.0062$ .

- b. When solving for the probability that the demand for salmon is less than 15 pounds, we find  $P(X < 15) = P\left(Z < \frac{15 - 12}{3.2}\right) = P(Z < 0.94) = 0.8264$ .
- c. In order to compute the required amount of salmon that should be purchased to meet demand on 90% of the days, we solve for  $x$  in  $P(X \leq x) = 0.90$ . Since  $P(X \leq x) = 0.90$  is equivalent to  $P(Z \leq z) = 0.90$ , we first derive  $z = 1.28$ . Given  $x = \mu + z\sigma$ , we find  $x = 12 + 1.28(3.2) = 16.10$ . Therefore, Akiko should buy 16.10 pounds of salmon daily to ensure that customer demand is met on 90% of the days.

## Using Excel for the Normal Distribution

### The Standard Transformation

We can easily find normal probabilities using Excel's NORM.DIST function. In general, in order to find  $P(X \leq x)$ , we input “=NORM.DIST( $x, \mu, \sigma, 1$ )”, where  $x$  is the value for which we want to evaluate the cumulative normal probability,  $\mu$  is the mean of the distribution,  $\sigma$  is the standard deviation of the distribution, and 1 is prompting Excel to return a cumulative probability. If we enter 0 as the fourth argument, Excel returns the height of the normal probability distribution at the point  $x$ . This feature is particularly useful if we want to plot the normal curve. Let's revisit Example 6.8a, where we want to find  $P(X > 20)$ . We know that the data are normally distributed with a mean of 12 and a standard deviation of 3.2. We input “=NORM.DIST(20, 12, 3.2, 1)”. Excel returns a cumulative probability of 0.9938, which means that  $P(X \leq 20) = 0.9938$ . Since we want to find  $P(X > 20)$ , we compute  $1 - 0.9938 = 0.0062$ .

### The Inverse Transformation

We can use Excel's NORM.INV function if we want to find a particular  $x$  value for a given cumulative probability. In general, we input “NORM.INV(*probability*,  $\mu, \sigma$ )”, where *probability* is the given cumulative probability,  $\mu$  is the mean of the distribution, and  $\sigma$  is the standard deviation of the distribution. Let's revisit Example 6.8c, where we want to find  $x$  such that  $P(X \leq x) = 0.90$ . We input “=NORM.INV(0.90, 12, 3.2)”. Excel returns the value 16.10, which is the same as we found manually.

We also note that Excel has related formulas if we assume  $\mu = 0$  and  $\sigma = 1$ ; that is, the variable has already been standardized. The formula NORM.S.DIST finds  $P(Z \leq z)$  and NORM.S.INV finds a particular  $z$  value for a given cumulative probability. For example, to find  $P(Z \leq 1.52)$ , we input “=NORM.S.DIST(1.52, 1)” and Excel returns 0.9357—this is the same probability that we found in Section 6.2 when we used the  $z$  table (Figure 6.7). Similarly, to find the  $z$  value such that  $P(Z \leq z) = 0.6808$  (Figure 6.12), we input “=NORM.S.INV(0.6808)” and Excel returns 0.47, or equivalently,  $z = 0.47$ . We will find these functions quite useful when solving problems in later chapters.

## A Note on the Normal Approximation of the Binomial Distribution

Recall from Chapter 5 that it is tedious to compute binomial probabilities with the formula when we encounter large values for  $n$ . As it turns out, with large values for  $n$ , the binomial distribution can be approximated by the normal distribution. Based on this normal distribution approximation, with mean  $\mu = np$  and standard deviation  $\sigma = \sqrt{npq}$ , we can use the  $z$  table to compute relevant binomial probabilities. Some researchers believe that the discovery of the normal distribution in the 18th century was due to the need to simplify the binomial probability calculations. The popularity of this method, however, has been greatly reduced by the advent of computers. As we learned in Chapter 5, it is easy to compute exact binomial probabilities with Excel; thus, there is no reason to approximate. The normal distribution approximation, however, is extremely important when making an inference for the population proportion  $p$ , which is a key parameter of the binomial distribution. In later chapters, we will study the details of this approximation and how it is used for making inferences.

## SYNOPSIS OF INTRODUCTORY CASE



Akiko Hamaguchi is a manager at a small sushi restaurant called Little Ginza in Phoenix, Arizona. She is aware of the importance of purchasing the right amount of salmon daily. While purchasing too much salmon results in wastage, purchasing too little can disappoint customers who may choose not to frequent the restaurant in the future. In the past, she has always bought 20 pounds of salmon daily. A careful analysis of her purchasing habits and customer demand reveals that Akiko is buying too much salmon. The probability that the demand for salmon would exceed 20 pounds is very small at 0.0062. Even a purchase of 15 pounds satisfies customer demand on 82.64% of the days. In order to execute her new strategy of meeting daily demand of customers on 90% of the days, Akiko should purchase approximately 16 pounds of salmon daily.

## EXERCISES 6.3

### Mechanics

23. Let  $X$  be normally distributed with mean  $\mu = 10$  and standard deviation  $\sigma = 6$ .
  - a. Find  $P(X \leq 0)$ .
  - b. Find  $P(X > 2)$ .
  - c. Find  $P(4 \leq X \leq 10)$ .
  - d. Find  $P(6 \leq X \leq 14)$ .
24. Let  $X$  be normally distributed with mean  $\mu = 10$  and standard deviation  $\sigma = 4$ .
  - a. Find  $P(X \leq 0)$ .
  - b. Find  $P(X > 2)$ .
  - c. Find  $P(4 \leq X \leq 10)$ .
  - d. Find  $P(6 \leq X \leq 14)$ .
25. Let  $X$  be normally distributed with mean  $\mu = 120$  and standard deviation  $\sigma = 20$ .
  - a. Find  $P(X \leq 86)$ .
  - b. Find  $P(80 \leq X \leq 100)$ .
  - c. Find  $x$  such that  $P(X \leq x) = 0.40$ .
  - d. Find  $x$  such that  $P(X > x) = 0.90$ .
26. Let  $X$  be normally distributed with mean  $\mu = 2.5$  and standard deviation  $\sigma = 2$ .
  - a. Find  $P(X > 7.6)$ .
  - b. Find  $P(7.4 \leq X \leq 10.6)$ .
  - c. Find  $x$  such that  $P(X > x) = 0.025$ .
  - d. Find  $x$  such that  $P(x \leq X \leq 2.5) = 0.4943$ .
27. Let  $X$  be normally distributed with mean  $\mu = 2500$  and standard deviation  $\sigma = 800$ .
  - a. Find  $x$  such that  $P(X \leq x) = 0.9382$ .
  - b. Find  $x$  such that  $P(X > x) = 0.025$ .
  - c. Find  $x$  such that  $P(2500 \leq X \leq x) = 0.1217$ .
  - d. Find  $x$  such that  $P(X \leq x) = 0.4840$ .
28. The random variable  $X$  is normally distributed. Also, it is known that  $P(X > 150) = 0.10$ .
  - a. Find the population mean  $\mu$  if the population standard deviation  $\sigma = 15$ .
  - b. Find the population mean  $\mu$  if the population standard deviation  $\sigma = 25$ .
  - c. Find the population standard deviation  $\sigma$  if the population mean  $\mu = 136$ .
  - d. Find the population standard deviation  $\sigma$  if the population mean  $\mu = 128$ .
29. (Use Excel) Let  $X$  be normally distributed with  $\mu = 254$  and  $\sigma = 11$ . In addition to providing the answer, state the relevant Excel commands.
  - a. Find  $P(X \leq 266)$ .
  - b. Find  $P(250 < X < 270)$ .
  - c. Find  $x$  such that  $P(X \leq x) = 0.33$ .
  - d. Find  $x$  such that  $P(X > x) = 0.33$ .
30. (Use Excel) Let  $X$  be normally distributed with  $\mu = -15$  and  $\sigma = 9$ . In addition to providing the answer, state the relevant Excel commands.
  - a. Find  $P(X > -12)$ .
  - b. Find  $P(0 \leq X \leq 5)$ .
  - c. Find  $x$  such that  $P(X \leq x) = 0.25$ .
  - d. Find  $x$  such that  $P(X > x) = 0.25$ .
31. The average high school teacher annual salary is \$43,000 (Payscale.com, August 20, 2010). Let teacher salary be normally distributed with a standard deviation of \$18,000.
  - a. What percentage of high school teachers make between \$40,000 and \$50,000?
  - b. What percentage of high school teachers make more than \$80,000?

### Applications

31. The average high school teacher annual salary is \$43,000 (Payscale.com, August 20, 2010). Let teacher salary be normally distributed with a standard deviation of \$18,000.
  - a. What percentage of high school teachers make between \$40,000 and \$50,000?
  - b. What percentage of high school teachers make more than \$80,000?



32. Americans are increasingly skimping on their sleep (*National Geographic News*, February 24, 2005). A health expert believes that American adults sleep an average of 6.2 hours on weekdays with a standard deviation of 1.2 hours. To answer the following questions, assume that sleep time on weekdays is normally distributed.
- What percentage of American adults sleep more than 8 hours on weekdays?
  - What percentage of American adults sleep less than 6 hours on weekdays?
  - What percentage of American adults sleep between 6 and 8 hours on weekdays?
33. The weight of turkeys is normally distributed with a mean of 22 pounds and a standard deviation of 5 pounds.
- Find the probability that a randomly selected turkey weighs between 20 and 26 pounds.
  - Find the probability that a randomly selected turkey weighs less than 12 pounds.
34. Suppose that the miles-per-gallon (mpg) rating of passenger cars is a normally distributed random variable with a mean and a standard deviation of 33.8 mpg and 3.5 mpg, respectively.
- What is the probability that a randomly selected passenger car gets at least 40 mpg?
  - What is the probability that a randomly selected passenger car gets between 30 and 35 mpg?
  - An automobile manufacturer wants to build a new passenger car with an mpg rating that improves upon 99% of existing cars. What is the minimum mpg that would achieve this goal?
35. According to a company's website, the top 25% of the candidates who take the entrance test will be called for an interview. You have just been called for an interview. The reported mean and standard deviation of the test scores are 68 and 8, respectively. What is the possible range for your test score if you assume that the scores are normally distributed?
36. A financial advisor informs a client that the expected return on a portfolio is 8% with a standard deviation of 12%. There is a 25% chance that the return would be negative and a 15% chance that the return would be above 16%. If the advisor is right about her assessment, is it reasonable to assume that the underlying return distribution is normal?
37. A packaging system fills boxes to an average weight of 18 ounces with a standard deviation of 0.2 ounce. It is reasonable to assume that the weights are normally distributed. Calculate the 1st, 2nd, and 3rd quartiles of the box weight.
38. According to the Bureau of Labor Statistics, it takes an average of 22 weeks for someone over 55 to find a new job, compared with 16 weeks for younger workers (*The Wall Street Journal*, September 2, 2008). Assume that the probability distributions are normal and that the standard deviation is 2 weeks for both distributions.
- What is the probability that it takes a worker over the age of 55 more than 19 weeks to find a job?
  - What is the probability that it takes a younger worker more than 19 weeks to find a job?
  - What is the probability that it takes a worker over the age of 55 between 23 and 25 weeks to find a job?
  - What is the probability that it takes a younger worker between 23 and 25 weeks to find a job?
39. Loans that are 60 days or more past due are considered seriously delinquent. The Mortgage Bankers Association reported that the rate of seriously delinquent loans has an average of 9.1% (*The Wall Street Journal*, August 26, 2010). Let the rate of seriously delinquent loans follow a normal distribution with a standard deviation of 0.80%.
- What is the probability that the proportion of seriously delinquent loans has a rate above 8%?
  - What is the probability that the proportion of seriously delinquent loans has a rate between 9.5% and 10.5%?
40. The time required to assemble an electronic component is normally distributed with a mean and a standard deviation of 16 minutes and 4 minutes, respectively.
- Find the probability that a randomly picked assembly takes between 10 and 20 minutes.
  - It is unusual for the assembly time to be above 24 minutes or below 6 minutes. What proportion of assembly times fall in these unusual categories?
41. Recent research suggests that Americans make an average of 10 phone calls per day (*CNN*, August 26, 2010). Let the number of calls be normally distributed with a standard deviation of 3 calls.
- What is the probability that an average American makes between 4 and 12 calls per day?
  - What is the probability that an average American makes more than 6 calls per day?
  - What is the probability that an average American makes more than 16 calls per day?
42. The manager of a night club in Boston stated that 95% of the customers are between the ages of 22 and 28 years. If the age of customers is normally distributed with a mean of 25 years, calculate its standard deviation.
43. An estimated 1.8 million students take on student loans to pay ever-rising tuition and room and board (*The New York Times*, April 17, 2009). It is also known that the average cumulative debt of recent college graduates is about \$22,500. Let the cumulative debt among recent college graduates be normally distributed with a standard deviation of \$7,000. Approximately how many recent college graduates have accumulated a student loan of more than \$30,000?

44. Scores on a marketing exam are known to be normally distributed with mean and standard deviation of 60 and 20, respectively.
- Find the probability that a randomly selected student scores between 50 and 80.
  - Find the probability that a randomly selected student scores between 20 and 40.
  - The syllabus suggests that the top 15% of the students will get an A in the course. What is the minimum score required to get an A?
  - What is the passing score if 10% of the students will fail the course?
45. Average talk time between charges of a cell phone is advertised as 4 hours. Assume that talk time is normally distributed with a standard deviation of 0.8 hour.
- Find the probability that talk time between charges for a randomly selected cell phone is below 3.5 hours.
  - Find the probability that talk time between charges for a randomly selected cell phone is either more than 4.5 hours or below 3.5 hours.
  - Twenty-five percent of the time, talk time between charges is below the 1st quartile value. What is this value?
46. A young investment manager tells his client that the probability of making a positive return with his suggested portfolio is 90%. If it is known that returns are normally distributed with a mean of 5.6%, what is the risk, measured by standard deviation, that this investment manager assumes in his calculation?
47. A construction company in Naples, Florida, is struggling to sell condominiums. In order to attract buyers, the company has made numerous price reductions and better financing offers. Although condominiums were once listed for \$300,000, the company believes that it will be able to get an average sale price of \$210,000. Let the price of these condominiums in the next quarter be normally distributed with a standard deviation of \$15,000.
- What is the probability that the condominium will sell at a price (i) below \$200,000?, (ii) above \$240,000?
  - The company is also trying to sell an artist's condo. Potential buyers will find the unusual features of this condo either pleasing or objectionable. The manager expects the average sale price of this condo to be the same as others at \$210,000, but with a higher standard deviation of \$20,000. What is the probability that this condo will sell at a price (i) below \$200,000?, (ii) above \$240,000?
48. You are considering the risk-return profile of two mutual funds for investment. The relatively risky fund promises an expected return of 8% with a standard deviation of 14%.
- The relatively less risky fund promises an expected return and standard deviation of 4% and 5%, respectively. Assume that the returns are approximately normally distributed.
- Which mutual fund will you pick if your objective is to minimize the probability of earning a negative return?
  - Which mutual fund will you pick if your objective is to maximize the probability of earning a return above 8%?
49. First introduced in Los Angeles, the concept of Korean-style tacos sold from a catering truck has been gaining popularity nationally (*The New York Times*, July 27, 2010). This taco is an interesting mix of corn tortillas with Korean-style beef, garnished with onion, cilantro, and a hash of chili-soy-dressed lettuce. Suppose one such taco truck operates in the Detroit area. The owners have estimated that the daily consumption of beef is normally distributed with a mean of 24 pounds and a standard deviation of 6 pounds. While purchasing too much beef results in wastage, purchasing too little can disappoint customers.
- Determine the amount of beef the owners should buy so that it meets demand on 80% of the days.
  - How much should the owners buy if they want to meet demand on 95% of the days?
50. A new car battery is sold with a two-year warranty whereby the owner gets the battery replaced free of cost if it breaks down during the warranty period. Suppose an auto store makes a net profit of \$20 on batteries that stay trouble-free during the warranty period; it makes a net loss of \$10 on batteries that break down. The life of batteries is known to be normally distributed with a mean and a standard deviation of 40 and 16 months, respectively.
- What is the probability that a battery will break down during the warranty period?
  - What is the expected profit of the auto store on a battery?
  - What is the expected monthly profit on batteries if the auto store sells an average of 500 batteries a month?
51. (Use Excel) While Massachusetts is no California when it comes to sun, the solar energy industry is flourishing in this state (*The Boston Globe*, May 27, 2012). The state's capital, Boston, averages 211.7 sunny days per year. Assume that the number of sunny days follows a normal distribution with a standard deviation of 20 days. In addition to providing the answer, state the relevant Excel commands.
- What is the probability that Boston has less than 200 sunny days in a given year?
  - Los Angeles averages 266.5 sunny days per year. What is the probability that Boston has at least as many sunny days as Los Angeles?
  - Suppose a dismal year in Boston is one where the number of sunny days is in the bottom 10% for that

- year. At most, how many sunny days must occur annually for it to be a dismal year in Boston?
- d. In 2012, Boston experienced unusually warm, dry, and sunny weather. Suppose this occurs only 1% of the time. What is the minimum number of sunny days that would satisfy the criteria for being an unusually warm, dry, and sunny year in Boston?
52. (Use Excel) A certain brand of refrigerators has a length of life that is normally distributed with a mean and a standard deviation of 15 years and 2 years, respectively. In addition to providing the answer, state the relevant Excel commands.
- What is the probability a refrigerator will last less than 6.5 years?
  - What is the probability that a refrigerator will last more than 23 years?
  - What length of life should the retailer advertise for these refrigerators so that only 3% of the refrigerators fail before the advertised length of life?

## 6.4 OTHER CONTINUOUS PROBABILITY DISTRIBUTIONS

As discussed earlier, the normal distribution is the most extensively used probability distribution in statistical work. One reason that this occurs is because the normal distribution accurately describes numerous random variables of interest. However, there are applications where other continuous distributions are more appropriate.

### The Exponential Distribution

A useful nonsymmetric continuous probability distribution is the **exponential distribution**. The exponential distribution is related to the Poisson distribution, even though the Poisson distribution deals with discrete random variables. Recall from Chapter 5 that the Poisson random variable counts the number of occurrences of an event over a given interval of time or space. For instance, the Poisson distribution is used to calculate the likelihood of a specified number of cars arriving at a McDonald's drive-thru over a particular time period or the likelihood of a specified number of defects in a 50-yard roll of fabric. Sometimes we are less interested in the *number* of occurrences over a given interval of time or space, but rather in the time that has elapsed or space encountered *between* such occurrences. For instance, we might be interested in the length of time that elapses between car arrivals at the McDonald's drive-thru or the distance between defects in a 50-yard roll of fabric. We use the exponential distribution for describing these times or distances. The exponential random variable is nonnegative; that is, the underlying variable  $X$  is defined for  $x \geq 0$ .

In order to better understand the connection between the Poisson and the exponential distributions, consider the introductory case of Chapter 5 where Anne was concerned about staffing needs at the Starbucks that she managed. Recall that Anne believed that the typical Starbucks customer averaged 18 visits to the store over a 30-day period. The Poisson random variable appropriately captures the number of visits, with the expected value (mean), over a 30-day period, as

$$\mu_{\text{Poisson}} = 18.$$

Since the number of visits follows the Poisson distribution, the time between visits has an exponential distribution. In addition, given the expected number of 18 visits over a 30-day month, the expected time between visits is derived as

$$\mu_{\text{Exponential}} = \frac{30}{18} = 1.67.$$

It is common to define the exponential probability distribution in terms of its *rate parameter*  $\lambda$  (the Greek letter lambda), which is the inverse of its mean. In the above example,

$$\lambda = \frac{1}{\mu} = \frac{1}{1.67} = 0.60.$$

We can think of the mean of the exponential distribution as the average time between arrivals, whereas the rate parameter measures the average number of arrivals per unit of time.

#### LO 6.6

Calculate and interpret probabilities for a random variable that follows the exponential distribution.

Note that the rate parameter is the same as the mean of the Poisson distribution, when defined per unit of time. For a Poisson process, the mean of 18 visits over a 30-day period is equivalent to a mean of  $18/30 = 0.60$  per day, which is the same as the rate parameter  $\lambda$ .

### THE EXPONENTIAL DISTRIBUTION

A random variable  $X$  follows the **exponential distribution** if its probability density function is

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0,$$

where  $\lambda$  is a rate parameter and  $e \approx 2.718$  is the base of the natural logarithm.

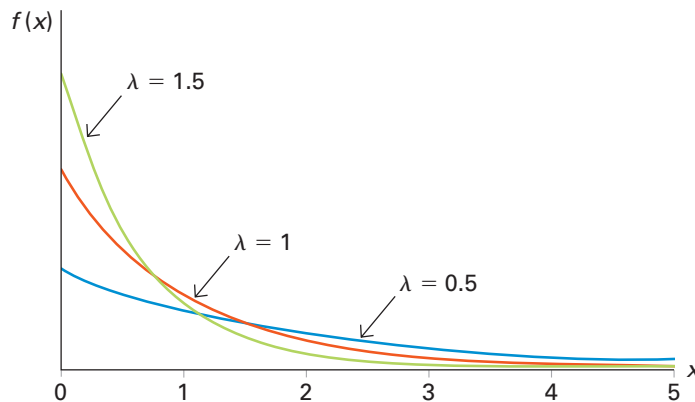
The mean and the standard deviation of  $X$  are equal:  $E(X) = SD(X) = \frac{1}{\lambda}$ . For  $x \geq 0$ , the **cumulative distribution function** of  $X$  is

$$P(X \leq x) = 1 - e^{-\lambda x}.$$

Therefore,  $P(X > x) = 1 - P(X \leq x) = e^{-\lambda x}$ .

The curves in Figure 6.25 show the shapes of the exponential probability density function based on various values of the rate parameter  $\lambda$ .

**FIGURE 6.25**  
Exponential probability  
density function for various  
values of  $\lambda$



### EXAMPLE 6.9

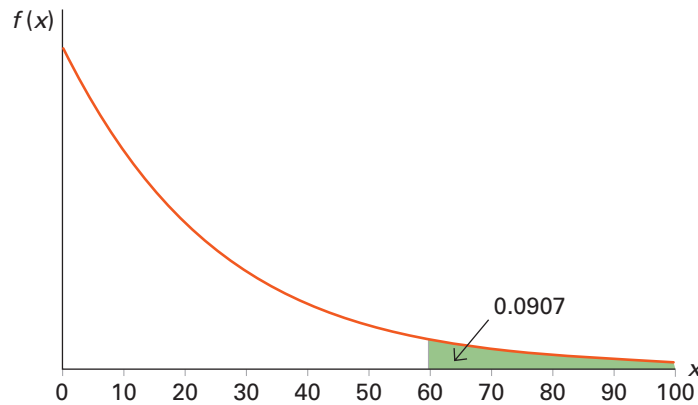
Let the time between e-mail messages during work hours be exponentially distributed with a mean of 25 minutes.

- Calculate the rate parameter  $\lambda$ .
- What is the probability that you do not get an e-mail for more than one hour?
- What is the probability that you get an e-mail within 10 minutes?

#### SOLUTION:

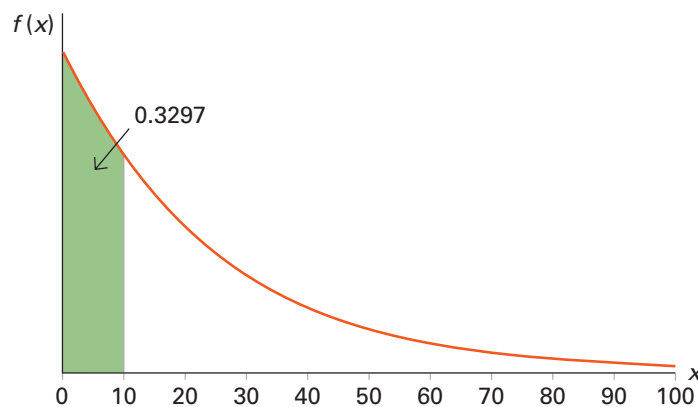
- Since the mean  $E(X)$  equals  $\frac{1}{\lambda}$ , we compute  $\lambda = \frac{1}{E(X)} = \frac{1}{25} = 0.04$ .
- The probability that you do not get an e-mail for more than an hour is  $P(X > 60)$ . We use  $P(X > x) = 1 - P(X \leq x) = e^{-\lambda x}$  to compute  $P(X > 60) = e^{-0.04(60)} = e^{-2.40} = 0.0907$ . The probability of not getting an e-mail for more than one hour is 0.0907. Figure 6.26 highlights this probability.

**FIGURE 6.26** Finding  $P(X > 60)$  (Example 6.9b)



- c. The probability that you get an e-mail within 10 minutes is  $P(X \leq 10) = 1 - e^{-0.04(10)} = 1 - 0.6703 = 0.3297$ . Figure 6.27 highlights this probability.

**FIGURE 6.27** Finding  $P(X \leq 10)$  (Example 6.9c)



The exponential distribution is also used in modeling lifetimes or failure times. For example, an electric bulb with a rated life of 1,000 hours is expected to fail after about 1,000 hours of use. However, the bulb may burn out either before or after 1,000 hours. Thus, the lifetime of an electric bulb is a random variable with an expected value of 1,000. A noted feature of the exponential distribution is that it is “memoryless,” thus implying a constant failure rate. In the electric bulb example, it implies that the probability that the bulb will burn out on a given day is independent of whether the bulb has already been used for 10, 100, or 1,000 hours.

## Using Excel for the Exponential Distribution

We can find exponential probabilities using Excel’s EXPON.DIST function. In general, in order to find  $P(X \leq x)$ , we input “=EXPON.DIST( $x$ ,  $\lambda$ , 1)”, where  $x$  is the value for which we want to evaluate the cumulative probability,  $\lambda$  is the rate parameter, and 1 is prompting Excel to return a cumulative probability. If we enter 0 as the third argument, Excel returns the height of the exponential distribution at the point  $x$ . This option is useful if we want to plot the exponential distribution. Let’s revisit Example 6.9b where we want to find  $P(X > 60)$ . We input “=EXPON.DIST(60, 0.04, 1)”. Excel returns a cumulative probability of 0.9093. Since we want to find  $P(X > 60)$ , we compute  $1 - 0.9093 = 0.0907$ .

**LO 6.7**

Calculate and interpret probabilities for a random variable that follows the lognormal distribution.

## The Lognormal Distribution

The **lognormal distribution** is defined with reference to the normal distribution. However, unlike the normal distribution, the lognormal distribution is relevant for a positive random variable and it is also positively skewed. Thus, it is useful for describing variables such as income, real estate values, and asset prices. Unlike the exponential distribution whose failure rate is constant, the failure rate of the lognormal distribution may increase or decrease over time. This flexibility has led to broad applications of the lognormal distribution ranging from modeling the failure time of new equipment to the lifetime of cancer patients. For instance, in the break-in period of new equipment, the failure rate is high. However, if it survives this initial period, the subsequent failure rate is greatly reduced. The same is true for cancer survivors.

A random variable  $Y$  is lognormal if its natural logarithm  $X = \ln(Y)$  is normally distributed. Alternatively, if  $X$  is a normal random variable, the lognormal variable is defined as  $Y = e^X$ .

### THE LOGNORMAL DISTRIBUTION

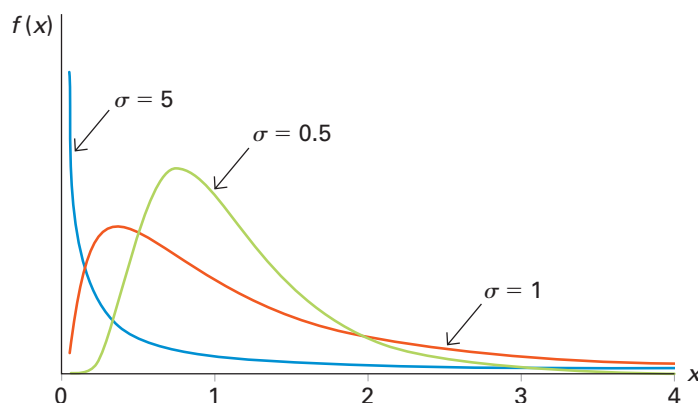
Let  $X$  be a normally distributed random variable with mean  $\mu$  and standard deviation  $\sigma$ . The random variable  $Y = e^X$  follows the **lognormal distribution** with a probability density function defined as

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(y) - \mu)^2}{2\sigma^2}\right) \quad \text{for } y > 0,$$

where  $\pi$  equals approximately 3.14159,  $\exp(x) = e^x$  is the exponential function, and  $e \approx 2.718$  is the base of the natural logarithm.

The graphs in Figure 6.28 show the shapes of the lognormal density function based on various values of  $\sigma$ . The lognormal distribution is clearly positively skewed for  $\sigma > 1$ . For  $\sigma < 1$ , the lognormal distribution somewhat resembles the normal distribution.

**FIGURE 6.28**  
Lognormal probability density function for various values of  $\sigma$  along with  $\mu = 0$



The mean and the variance of the lognormal random variable  $Y$  are related to the mean and the standard deviation of the corresponding normal random variable  $X$ .



### EXPECTED VALUES AND STANDARD DEVIATIONS OF THE LOGNORMAL AND NORMAL DISTRIBUTIONS

Let  $X$  be a normal random variable with mean  $\mu$  and standard deviation  $\sigma$  and let  $Y = e^X$  be the corresponding lognormal variable. The mean  $\mu_Y$  and the standard deviation  $\sigma_Y$  of  $Y$  are derived as

$$\mu_Y = \exp\left(\frac{2\mu + \sigma^2}{2}\right) \quad \text{and} \quad \sigma_Y = \sqrt{(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)}.$$

Equivalently, the mean and the standard deviation of the normal variable  $X = \ln(Y)$  are derived as

$$\mu = \ln\left(\frac{\mu_Y^2}{\sqrt{\mu_Y^2 + \sigma_Y^2}}\right) \quad \text{and} \quad \sigma = \sqrt{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}.$$

#### EXAMPLE 6.10

Compute the mean and the standard deviation of a lognormal random variable if the mean and the standard deviation of the underlying normal random variable are as follows:

- a.  $\mu = 0, \sigma = 1$
- b.  $\mu = 2, \sigma = 1$
- c.  $\mu = 2, \sigma = 1.5$

**SOLUTION:** Since  $X$  is normal,  $Y = e^X$  is lognormal with mean  $\mu_Y = \exp\left(\frac{2\mu + \sigma^2}{2}\right)$  and standard deviation  $\sigma_Y = \sqrt{(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)}$ .

- a. Given  $\mu = 0$  and  $\sigma = 1$ , we compute  $\mu_Y = \exp\left(\frac{0 + 1^2}{2}\right) = 1.65$  and  $\sigma_Y = \sqrt{(\exp(1^2) - 1)\exp(0 + 1^2)} = 2.16$ .
- b. Given  $\mu = 2$  and  $\sigma = 1$ , we compute  $\mu_Y = \exp\left(\frac{4 + 1^2}{2}\right) = 12.18$  and  $\sigma_Y = \sqrt{(\exp(1^2) - 1)\exp(4 + 1^2)} = 15.97$ .
- c. Given  $\mu = 2$  and  $\sigma = 1.5$ , we compute  $\mu_Y = \exp\left(\frac{4 + 1.5^2}{2}\right) = 22.76$  and  $\sigma_Y = \sqrt{(\exp(1.5^2) - 1)\exp(4 + 1.5^2)} = 66.31$ .

The popularity of the lognormal distribution is also due to the fact that the probabilities of a lognormal random variable are easily evaluated by reference to the normal distribution. This is illustrated in the following example.

#### EXAMPLE 6.11

Let  $Y = e^X$  where  $X$  is normally distributed with mean  $\mu = 5$  and standard deviation  $\sigma = 1.2$ .

- a. Find  $P(Y \leq 200)$ .
- b. Find the 90th percentile of  $Y$ .

**SOLUTION:** We solve these problems by first converting them into the corresponding normal distribution problems.

- a. Note that  $P(Y \leq 200) = P(\ln(Y) \leq \ln(200)) = P(X \leq 5.30)$ . We transform  $x = 5.30$  in the usual way to get  $z = \frac{5.30 - 5}{1.2} = 0.25$ . From the  $z$  table, we get

$P(Z \leq 0.25) = 0.5987$ . Therefore,  $P(Y \leq 200) = P(X \leq 5.30) = P(Z \leq 0.25) = 0.5987$ .

- b. The 90th percentile is a value  $y$  such that  $P(Y < y) = 0.90$ . We first note that  $P(Y < y) = 0.90$  is equivalent to  $P(\ln(Y) < \ln(y)) = P(X < x) = 0.90$  where  $x = \ln(y)$ . We look up the cumulative probability of 0.90 in the  $z$  table to get  $z = 1.28$ . We use the inverse transformation to derive  $x = \mu + z\sigma = 5 + 1.28(1.2) = 6.54$ . Finally, we compute  $y = e^x = e^{6.54} = 692.29$ . Therefore, the 90th percentile of the distribution is 692.29.

## Using Excel for the Lognormal Distribution

### The Standard Transformation

We can find lognormal probabilities using Excel's LOGNORM.DIST function. Suppose the random variable  $Y = e^X$  follows the lognormal distribution. In general, to find  $P(Y \leq y)$ , we input “=LOGNORM.DIST( $y, \mu, \sigma, 1$ )”, where  $y$  is the nonnegative value for which we want to find the cumulative probability,  $\mu$  is the mean of the underlying normal distribution,  $\sigma$  is the standard deviation of the underlying normal distribution, and 1 is prompting Excel to return a cumulative probability. If we enter the value 0 for the fourth argument, Excel returns the height of the lognormal distribution at the point  $y$ . This option is useful if we want to plot the lognormal distribution. Let's revisit 6.11a where we want to find  $P(Y \leq 200)$ , given  $\mu = 5$  and  $\sigma = 1.2$ . We input “=LOGNORM.DIST(200, 5, 1.2, 1)” and Excel returns 0.5982. Note that the Excel output differs slightly from the manual calculation due to rounding.

### The Inverse Transformation

We can use Excel's LOGNORM.INV function if we want to find a particular  $y$  value for a given cumulative probability. In general, we input “=LOGNORM.INV(*probability*,  $\mu, \sigma$ )”, where *probability* is the given cumulative probability,  $\mu$  is the mean of the underlying normal distribution, and  $\sigma$  is the standard deviation of the underlying normal distribution. Let's revisit 6.11b where we want to find the 90th percentile of  $Y$  given  $\mu = 5$  and  $\sigma = 1.2$ . We input “=LOGNORM.INV(0.90, 5, 1.2)” and Excel returns 690.81. Again, the Excel calculation differs slightly from the manual calculation due to rounding.

## EXERCISES 6.4

### Mechanics

53. Assume a Poisson random variable has a mean of 6 successes over a 120-minute period.
  - a. Find the mean of the random variable, defined by the time between successes.
  - b. What is the rate parameter of the appropriate exponential distribution?
  - c. Find the probability that the time to success will be more than 60 minutes.
54. Assume a Poisson random variable has a mean of four arrivals over a 10-minute interval.
  - a. What is the mean of the random variable, defined by the time between arrivals?
  - b. Find the probability that the next arrival would be within the mean time.
  - c. Find the probability that the next arrival would be between one and two minutes.
55. A random variable  $X$  is exponentially distributed with a mean of 0.1.
  - a. What is the rate parameter  $\lambda$ ? What is the standard deviation of  $X$ ?
  - b. Compute  $P(X > 0.20)$ .
  - c. Compute  $P(0.10 \leq X \leq 0.20)$ .
56. A random variable  $X$  is exponentially distributed with an expected value of 25.
  - a. What is the rate parameter  $\lambda$ ? What is the standard deviation of  $X$ ?
  - b. Compute  $P(20 \leq X \leq 30)$ .
  - c. Compute  $P(15 \leq X \leq 35)$ .
57. A random variable  $X$  is exponentially distributed with a probability density function of  $f(x) = 5e^{-5x}$ . Calculate the mean and the standard deviation of  $X$ .
58. (Use Excel) Let  $X$  be exponentially distributed with  $\lambda = 0.5$ . In addition to providing the answer, state the relevant Excel commands.
  - a.  $P(X \leq 1)$
  - b.  $P(2 < X < 4)$
  - c.  $P(X > 10)$

59. (Use Excel) Let  $X$  be exponentially distributed with  $\mu = 1.25$ . In addition to providing the answer, state the relevant Excel commands.
- $P(X < 2.3)$
  - $P(1.5 \leq X \leq 5.5)$
  - $P(X > 7)$
60. Compute the mean and the variance of a lognormal variable  $Y = e^X$  where  $X$  is normally distributed with the following mean and variance:
- $\mu = 3, \sigma^2 = 2$
  - $\mu = 5, \sigma^2 = 2$
  - $\mu = 5, \sigma^2 = 3$
61. Let  $Y = e^X$ , where  $X$  is normally distributed. Compute the mean and the variance of  $X$  given the following information.
- $\mu_Y = 14, \sigma_Y^2 = 22$
  - $\mu_Y = 20, \sigma_Y^2 = 22$
  - $\mu_Y = 20, \sigma_Y^2 = 120$
62. Let  $Y = e^X$  where  $X$  is normally distributed with  $\mu = 1.8$  and  $\sigma = 0.80$ . Compute the following values.
- $P(Y \leq 7.5)$
  - $P(8 < Y < 9)$
  - The 90th percentile of  $Y$
63. Let  $Y$  have the lognormal distribution with mean 82.8 and variance 156.25. Compute the following probabilities.
- $P(Y > 100)$
  - $P(80 < Y < 100)$

## Applications

64. Studies have shown that bats can consume an average of 10 mosquitoes per minute (<http://berkshitemuseum.org>). Assume that the number of mosquitoes consumed per minute follows a Poisson distribution.
- What is the mean time between eating mosquitoes?
  - Find the probability that the time between eating mosquitoes is more than 15 seconds.
  - Find the probability that the time between eating mosquitoes is between 15 and 20 seconds.
65. According to *Daily Mail* (February 28, 2012), there was an average of one complaint every 12 seconds against Britain's biggest banks in 2011. It is reasonable to assume that the time between complaints is exponentially distributed.
- What is the mean time between complaints?
  - What is the probability that the next complaint will take less than the mean time?
  - What is the probability that the next complaint will take between 5 and 10 seconds?
66. A tollbooth operator has observed that cars arrive randomly at an average rate of 360 cars per hour.
- What is the mean time between car arrivals at this tollbooth?
  - What is the probability that the next car will arrive within ten seconds?
67. Customers make purchases at a convenience store, on average, every six minutes. It is fair to assume that the time between customer purchases is exponentially distributed. Jack operates the cash register at this store.
- What is the rate parameter  $\lambda$ ? What is the standard deviation of this distribution?
  - Jack wants to take a five-minute break. He believes that if he goes right after he has serviced a customer, he will lower the probability of someone showing up during his five-minute break. Is he right in this belief?
  - What is the probability that a customer will show up in less than five minutes?
  - What is the probability that nobody shows up for over half an hour?
68. When crossing the Golden Gate Bridge, traveling into San Francisco, all drivers must pay a toll. Suppose the amount of time (in minutes) drivers wait in line to pay the toll follows an exponential distribution with a probability density function of  $f(x) = 0.2e^{-0.2x}$ .
- What is the mean waiting time that drivers face when entering San Francisco via the Golden Gate Bridge?
  - What is the probability that a driver spends more than the average time to pay the toll?
  - What is the probability that a driver spends more than 10 minutes to pay the toll?
  - What is the probability that a driver spends between 4 and 6 minutes to pay the toll?
69. A hospital administrator worries about the possible loss of electric power as a result of a power blackout. The hospital, of course, has a standby generator, but it too is subject to failure, having a mean time between failures of 500 hours. It is reasonable to assume that the time between failures is exponentially distributed.
- What is the probability that the standby generator fails during the next 24-hour blackout?
  - Suppose the hospital owns two standby generators that work independently of one another. What is the probability that both generators fail during the next 24-hour blackout?
70. (Use Excel) On average, the state police catch eight speeders per hour at a certain location on Interstate I-90. Assume that the number of speeders per hour follows the Poisson distribution. In addition to providing the answer, state the relevant Excel commands.
- What is the probability that the state police wait less than 10 minutes for the next speeder?
  - What is the probability that the state police wait between 15 and 20 minutes for the next speeder?
  - What is the probability that the state police wait more than 25 minutes for the next speeder?

71. (Use Excel) Motorists arrive at a Gulf station at the rate of two per minute during morning hours. Assume that the arrival of motorists at the station follows a Poisson distribution. In addition to providing the answer, state the relevant Excel commands.
- What is the probability that the next car's arrival is in less than one minute?
  - What is the probability that the next car's arrival is in more than five minutes?
72. The Bahamas is a tropical paradise made up of 700 islands sprinkled over 100,000 square miles of the Atlantic Ocean. According to the figures released by the government of the Bahamas, the mean household income in the Bahamas is \$39,626 and the median income is \$33,600. A demographer decides to use the lognormal random variable to model this nonsymmetric income distribution. Let  $Y$  represent household income, where for a normally distributed  $X$ ,  $Y = e^X$ . In addition, suppose the standard deviation of household income is \$10,000. Use this information to answer the following questions.
- Compute the mean and the standard deviation of  $X$ .
  - What proportion of the people in the Bahamas have household income above the mean?
  - What proportion of the people in the Bahamas have household income below \$20,000?
  - Compute the 75th percentile of the income distribution in the Bahamas.
73. It is well documented that a typical washing machine can last anywhere between 5 to 12 years. Let the life of a washing machine be represented by a lognormal variable,  $Y = e^X$  where  $X$  is normally distributed. In addition, let the mean and standard deviation of the life of a washing machine be 8 years and 4 years, respectively.
- Compute the mean and the standard deviation of  $X$ .
  - What proportion of the washing machines will last for more than 10 years?
  - What proportion of the washing machines will last for less than 6 years?
  - Compute the 90th percentile of the life of the washing machines.

## WRITING WITH STATISTICS



Professor Lang is a professor of Economics at Salem State University. She has been teaching a course in Principles of Economics for over 25 years. Professor Lang has never graded on a curve since she believes that relative grading may unduly penalize (benefit) a good (poor) student in an unusually strong (weak) class. She always uses an absolute scale for making grades, as shown in the two left columns of Table 6.4.

**TABLE 6.4** Grading Scales with Absolute Grading versus Relative Grading

Absolute Grading		Relative Grading	
Grade	Score	Grade	Probability
A	92 and above	A	0.10
B	78 up to 92	B	0.35
C	64 up to 78	C	0.40
D	58 up to 64	D	0.10
F	Below 58	F	0.05

A colleague of Professor Lang's has convinced her to move to relative grading, since it corrects for unanticipated problems. Professor Lang decides to experiment with grading based on the relative scale as shown in the two right columns of Table 6.4. Using this relative grading scheme, the top 10% of students will get A's, the next 35% B's, and so on. Based on her years of teaching experience, Professor Lang believes that the scores in her course follow a normal distribution with a mean of 78.6 and a standard deviation of 12.4.

Professor Lang wants to use the above information to:

- Calculate probabilities based on the absolute scale. Compare these probabilities to the relative scale.
- Calculate the range of scores for various grades based on the relative scale. Compare these ranges to the absolute scale.
- Determine which grading scale makes it harder to get higher grades.

Many teachers would confess that grading is one of the most difficult tasks of their profession. Two common grading systems used in higher education are relative and absolute. Relative grading systems are norm referenced or curve-based, in which a grade is based on the student's relative position in class. Absolute grading systems, on the other hand, are criterion referenced, in which a grade is related to the student's absolute performance in class. In short, with absolute grading, the student's score is compared to a predetermined scale whereas with relative grading, the score is compared to the scores of other students in the class.

Let  $X$  represent the grade in Professor Lang's class, which is normally distributed with a mean of 78.6 and a standard deviation of 12.4. This information is used to derive the grade probabilities based on the absolute scale. For instance, the probability of receiving an A is derived as  $P(X \geq 92) = P(Z \geq 1.08) = 0.14$ . Other probabilities, derived similarly, are presented in Table 6.A.

**TABLE 6.A** Probabilities Based on Absolute Scale and Relative Scale

Grade	Probability Based on Absolute Scale	Probability Based on Relative Scale
A	0.14	0.10
B	0.38	0.35
C	0.36	0.40
D	0.07	0.10
F	0.05	0.05

The second column of Table 6.A shows that 14% of students are expected to receive A's, 38% B's, and so on. Although these numbers are generally consistent with the relative scale restated in the third column of Table 6.A, it appears that the relative scale makes it harder for students to get higher grades. For instance, 14% get A's with the absolute scale compared to only 10% with the relative scale.

Alternatively, we can compare the two grading methods on the basis of the range of scores for various grades. The second column of Table 6.B restates the range of scores based on absolute grading. In order to obtain the range of scores based on relative grading, it is once again necessary to apply concepts from the normal distribution. For instance, the minimum score required to earn an A with relative grading is derived by solving for  $x$  in  $P(X \geq x) = 0.10$ . Since  $P(X \geq x) = 0.10$  is equivalent to  $P(Z \geq z) = 0.10$ , it follows that  $z = 1.28$ . Inserting the proper values of the mean, the standard deviation, and  $z$  into  $x = \mu + z\sigma$  yields a value of  $x$  equal to 94.47. Ranges for other grades, derived similarly, are presented in the third column of Table 6.B.

**TABLE 6.B** Range of Scores with Absolute Grading versus Relative Grading

Grade	Range of Scores Based on Absolute Grading	Range of Scores Based on Relative Grading
A	92 and above	94.47 and above
B	78 up to 92	80.21 up to 94.47
C	64 up to 78	65.70 up to 80.21
D	58 up to 64	58.20 up to 65.70
F	Below 58	Below 58.20

Once again comparing the results in Table 6.B, the use of the relative scale makes it harder for students to get higher grades in Professor Lang's courses. For instance, in order to receive an A with relative grading, a student must have a score of at least 94.47 versus a score of at least 92 with absolute grading. Both absolute and relative grading methods have their merits and teachers often make the decision on the basis of their teaching philosophy. However, if Professor Lang wants to keep the grades consistent with her earlier absolute scale, she should base her relative scale on the probabilities computed in the second column of Table 6.A.

## CONCEPTUAL REVIEW

### LO 6.1 Describe a continuous random variable.

A **continuous random variable** is characterized by uncountable values because it can take on any value within an interval. The probability that a continuous random variable  $X$  assumes a particular value  $x$  is zero; that is,  $P(X = x) = 0$ . Thus, for a continuous random variable, we calculate the probability within a specified interval. Moreover, the following equalities hold:  $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$ .

The **probability density function**  $f(x)$  of a continuous random variable  $X$  is nonnegative and the entire area under this function equals one. The probability  $P(a \leq X \leq b)$  is the area under  $f(x)$  between points  $a$  and  $b$ .

For any value  $x$  of the random variable  $X$ , the **cumulative distribution function**  $F(x)$  is defined as  $F(x) = P(X \leq x)$ .

### LO 6.2 Calculate and interpret probabilities for a random variable that follows the continuous uniform distribution.

The **continuous uniform distribution** describes a random variable that has an equally likely chance of assuming a value within a specified range. The probability is essentially the area of a rectangle, which is the base times the height; that is, the length of a specified interval times the probability density function  $f(x) = \frac{1}{b-a}$ , where  $a$  and  $b$  are the lower and upper bounds of the interval, respectively.

### LO 6.3 Explain the characteristics of the normal distribution.

The **normal distribution** is the most extensively used continuous probability distribution and is the cornerstone of statistical inference. It is the familiar bell-shaped distribution, which is symmetric around the mean; that is, one side of the mean is just the mirror image of the other side. The normal distribution is completely described by two parameters: the population mean  $\mu$  and the population variance  $\sigma^2$ .

The **standard normal distribution**, also referred to as the **z distribution**, is a special case of the normal distribution, with mean equal to zero and standard deviation (or variance) equal to one.

### LO 6.4 Use the standard normal table (z table).

The **standard normal table**, also called the **z table**, provides **cumulative probabilities**  $P(Z \leq z)$ ; this table appears on two pages in Table 1 of Appendix A. The left-hand page provides cumulative probabilities for  $z$  values less than or equal to zero. The right-hand page shows cumulative probabilities for  $z$  values greater than or equal to zero. We also use the table to compute  $z$  values for given cumulative probabilities.

### LO 6.5 Calculate and interpret probabilities for a random variable that follows the normal distribution.

Any normally distributed random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$  can be transformed into the standard normal random variable  $Z$  as  $Z = \frac{X - \mu}{\sigma}$ . This standard transformation implies that any value  $x$  has a corresponding value  $z$  given by  $z = \frac{x - \mu}{\sigma}$ .

The standard normal variable  $Z$  can be transformed to the normally distributed random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$  as  $X = \mu + Z\sigma$ . This inverse transformation implies that any value  $z$  has a corresponding value  $x$  given by  $x = \mu + z\sigma$ .



**LO 6.6 Calculate and interpret probabilities for a random variable that follows the exponential distribution.**

A useful nonsymmetric continuous probability distribution is the **exponential distribution**. A random variable  $X$  follows the exponential distribution if its probability density function is  $f(x) = \lambda e^{-\lambda x}$  for  $x \geq 0$ , where  $\lambda$  is a rate parameter and  $e \approx 2.718$  is the base of the natural logarithm. The mean and the standard deviation of the distribution are both equal to  $1/\lambda$ . For  $x \geq 0$ , the **cumulative probability** is computed as  $P(X \leq x) = 1 - e^{-\lambda x}$ .

**LO 6.7 Calculate and interpret probabilities for a random variable that follows the lognormal distribution.**

The **lognormal distribution** is another useful positively skewed distribution. Let  $X$  be a normal random variable with mean  $\mu$  and variance  $\sigma^2$  and let  $Y = e^X$  be the corresponding lognormal variable. The mean  $\mu_Y$  and standard deviation  $\sigma_Y$  of  $Y$  are derived as  $\mu_Y = \exp\left(\frac{2\mu + \sigma^2}{2}\right)$  and  $\sigma_Y = \sqrt{(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)}$ , respectively. Equivalently, the mean and standard deviation of the normal variable  $X = \ln(Y)$  are derived as  $\mu = \ln\left(\frac{\mu_Y^2}{\sqrt{\mu_Y^2 + \sigma_Y^2}}\right)$  and  $\sigma = \sqrt{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}$ , respectively. Probabilities for a lognormal random variable are easily evaluated by reference to the normal distribution.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

74. A florist makes deliveries between 1:00 pm and 5:00 pm daily. Assume delivery times follow the continuous uniform distribution.
  - a. Calculate the mean and the variance of this distribution.
  - b. Determine the percentage of deliveries that are made after 4:00 pm.
  - c. Determine the percentage of deliveries that are made prior to 2:30 pm.
75. A worker at a landscape design center uses a machine to fill bags with potting soil. Assume that the quantity put in each bag follows the continuous uniform distribution with low and high filling weights of 10 pounds and 12 pounds, respectively.
  - a. Calculate the expected value and the standard deviation of this distribution.
  - b. Find the probability that the weight of a randomly selected bag is no more than 11 pounds.
  - c. Find the probability that the weight of a randomly selected bag is at least 10.5 pounds.
76. The revised guidelines from the National High Blood Pressure Education Program define normal blood pressure as readings below 120/80 millimeters of mercury (*The New York Times*, May 14, 2003). Prehypertension is suspected when the top number (systolic) is between 120 and 139 or when the bottom number (diastolic) is between 80 and 90. A recent survey reported that the mean systolic reading of Canadians is 125 with a standard deviation of 17 and the mean diastolic reading is 79 with a standard deviation of 10. Assume that diastolic as well as systolic readings are normally distributed.
  - a. What proportion of Canadians are suffering from prehypertension caused by high diastolic readings?
  - b. What proportion of Canadians are suffering from prehypertension caused by high systolic readings?
77. U.S. consumers are increasingly viewing debit cards as a convenient substitute for cash and checks. The average amount spent annually on a debit card is \$7,790 (*Kiplinger's*, August 2007). Assume that the average amount spent on a debit card is normally distributed with a standard deviation of \$500.
  - a. A consumer advocate comments that the majority of consumers spend over \$8,000 on a debit card. Find a flaw in this statement.
  - b. Compute the 25th percentile of the amount spent on a debit card.
  - c. Compute the 75th percentile of the amount spent on a debit card.
  - d. What is the interquartile range of this distribution?

78. On St. Patrick's Day, men spend an average of \$43.87 while women spend an average of \$29.54 (*USA TODAY*, March 17, 2009). Assume the standard deviations of spending for men and women are \$3 and \$11, respectively, and that both distributions are normally distributed.
- What is the probability that men spend over \$50 on St. Patrick's Day?
  - What is the probability that women spend over \$50 on St. Patrick's Day?
  - Are men or women more likely to spend over \$50 on St. Patrick's Day?
79. Lisa Mendes and Brad Lee work in the sales department of an AT&T Wireless Store. Lisa has been signing in an average of 48 new cell phone customers every month with a standard deviation of 22, while Brad signs in an average of 56 new customers with a standard deviation of 17. The store manager offers both Lisa and Brad a \$100 incentive bonus if they can sign in more than 100 new customers in a month. Assume a normal distribution to answer the following questions.
- What is the probability that Lisa will earn the \$100 incentive bonus?
  - What is the probability that Brad will earn the \$100 incentive bonus?
  - Are you surprised by the results? Explain.
80. The car speeds on a certain stretch of the interstate highway I-95 are known to be normally distributed with a mean of 72 and a standard deviation of 15. You have just heard a policeman comment that about 3% of the drivers drive at extremely dangerous speeds. What is the minimum speed that the policeman considers extremely dangerous?
81. The average household income in a community is known to be \$80,000. Also, 20% of the households have an income below \$60,000 and another 20% have an income above \$90,000. Is it reasonable to use the normal distribution to model the household income in this community?
82. The length of components produced by a company is normally distributed with a mean of 6 cm and a standard deviation of 0.02 cm. Calculate the 1st, 2nd, and 3rd quartiles of the component length.
83. Entrance to a prestigious MBA program in India is determined by a national test where only the top 10% of the examinees are admitted to the program. Suppose it is known that the scores on this test are normally distributed with a mean of 420 and a standard deviation of 80. Parul Monga is trying desperately to get into this program. What is the minimum score that she must earn to get admitted?
84. A new water filtration system is sold with a 10-year warranty that includes all parts and repairs. Suppose the life of this water filtration system is normally distributed with mean and standard deviation of 16 and 5 years, respectively.
- What is the probability that the water filtration system will require a repair during the warranty period?
  - Suppose the water filtration firm makes a \$300 profit for every new system it installs. This profit, however, is reduced to \$50 if the system requires repair during the warranty period. Find the expected profit of the firm if it installs 1,000 new water filtration systems.
85. (Use Excel) Suppose that the average IQ score is normally distributed with a mean of 100 and a standard deviation of 16. In addition to providing the answer, state the relevant Excel commands.
- What is the probability a randomly selected person will have an IQ score of less than 80?
  - What is the probability that a randomly selected person will have an IQ score greater than 125?
  - What minimum IQ score does a person have to achieve to be in the top 2.5% of IQ scores?
86. (Use Excel) Suppose that the annual household income in a small Midwestern community is normally distributed with a mean of \$55,000 and a standard deviation of \$4,500. In addition to providing the answer, state the relevant Excel commands.
- What is the probability that a randomly selected household will have an income between \$50,000 and \$65,000?
  - What is the probability that a randomly selected household will have an income of more than \$70,000?
  - What minimum income does a household need to earn to be in the top 5% of incomes?
  - What maximum income does a household need to earn to be in the bottom 40% of incomes?
87. On a particularly busy section of the Garden State Parkway in New Jersey, police use radar guns to detect speeders. Assume the time that elapses between successive speeders is exponentially distributed with a mean of 15 minutes.
- Calculate the rate parameter  $\lambda$ .
  - What is the probability of a waiting time less than 10 minutes between successive speeders?
  - What is the probability of a waiting time in excess of 25 minutes between successive speeders?

88. According to the Federal Bureau of Investigation, there is a violent crime in the United States every 22 seconds (*ABC News*, September 25, 2007). Assume that the time between successive violent crimes is exponentially distributed.
- What is the probability that there is a violent crime in the United States in the next one minute?
  - If there has not been a violent crime in the previous minute, what is the probability that there will be a violent crime in the subsequent minute?
89. In a local law office, jobs to a printer are sent at a rate of 8 jobs per hour. Suppose that the number of jobs sent to a printer follows the Poisson distribution.
- What is the expected time between successive jobs?
  - What is the probability that the next job will be sent within five minutes?
90. Disturbing news regarding Scottish police concerns the number of crashes involving vehicles on operational duties (*BBC News*, March 10, 2008). Statistics showed that Scottish forces' vehicles had been involved in traffic accidents at the rate of 1,000 per year. Suppose the number of crashes involving vehicles on operational duties follows a Poisson distribution.
- What is the average number of days between successive crashes?
  - What is the rate parameter of the appropriate exponential distribution?
  - What is the probability that the next vehicle will crash within a day?
91. A large technology firm receives an average of 12 new job applications every 10 days for positions that are not even advertised. Suppose the number of job applications received follows a Poisson distribution.
- What is the average number of days between successive job applications?
  - What is the probability that the next job application is received within a day?
  - What is the probability that the next job application is received between 1 and 2 days?
92. (Use Excel) The mileage (in 1,000s of miles) that car owners get with a certain kind of radial tire is a random variable having an exponential distribution with a mean of 50. In addition to providing the answer, state the relevant Excel commands.
- What is the probability that a tire will last at most 40,000 miles?
  - What is the probability that a tire will last at least 65,000 miles?
  - What is the probability that a tire will last between 70,000 and 80,000 miles?
93. (Use Excel) On average, a certain kind of kitchen appliance requires repairs once every four years. Assume that the times between repairs are exponentially distributed. In addition to providing the answer, state the relevant Excel commands.
- What is the probability that the appliance will work no more than three years without requiring repairs?
  - What is the probability that the appliance will work at least six years without requiring repairs?
94. The relief time provided by a standard dose of a popular children's allergy medicine averages six hours with a standard deviation of two hours.
- Determine the percentage of children who experience relief for less than four hours if the relief time follows a normal distribution.
  - Determine the percentage of children who experience relief for less than four hours if the relief time follows a lognormal distribution.
  - Compare the results based on these two distributions.
95. The mileage (in 1,000s of miles) that car owners get with a certain kind of radial tire is a random variable  $Y$  having a lognormal distribution such that  $Y = e^X$  where  $X$  is normally distributed. Let the mean and the standard deviation of the life of a radial tire be 40,000 miles and 5,000 miles, respectively.
- Compute the mean and standard deviation of  $X$ .
  - What proportion of the tires will last for more than 50,000 miles?
  - What proportion of the tires will last for no more than 35,000 miles?
  - Compute the 95th percentile of the life expectancy of the tire.

## CASE STUDIES

**CASE STUDY 6.1** Body mass index (BMI) is a reliable indicator of body fat for most children and teens. BMI is calculated from a child's weight and height and is used as an easy-to-perform method of screening for weight categories that may lead to health problems. For children and teens, BMI is age- and sex-specific and is often referred to as BMI-for-age.

The Centers for Disease Control and Prevention (CDC) reports BMI-for-age growth charts for girls as well as boys to obtain a percentile ranking. Percentiles are the most commonly used indicator to assess the size and growth patterns of individual children in the United States.

The following table provides weight status categories and the corresponding percentiles and BMI ranges for 10-year-old boys in the United States.

Weight Status Category	Percentile Range	BMI Range
Underweight	Less than 5th	Less than 14.2
Healthy Weight	Between 5th and 85th	Between 14.2 and 19.4
Overweight	Between 85th and 95th	Between 19.4 and 22.2
Obese	More than 95th	More than 22.2

Health officials of a Midwestern town are concerned about the weight of children in their town. They believe that the BMI of their 10-year-old boys is normally distributed with mean 19.2 and standard deviation 2.6.

In a report, use the sample information to:

1. Compute the proportion of 10-year-old boys in this town that are in the various weight status categories given the BMI ranges.
2. Discuss whether the concern of health officials is justified.

**CASE STUDY 6.2** Vanguard's Precious Metals and Mining fund (Metals) and Fidelity's Strategic Income fund (Income) were two top-performing mutual funds for the years 2000 through 2009. An analysis of annual return data for these two funds provided important information for any type of investor. Over the past 10 years, the Metals fund posted a mean return of 24.65% with a standard deviation of 37.13%. On the other hand, the mean and the standard deviation of return for the Income fund were 8.51% and 11.07%, respectively. It is reasonable to assume that the returns of the Metals and the Income funds are both normally distributed, where the means and the standard deviations are derived from the 10-year sample period.

In a report, use the sample information to compare and contrast the Metals and Income funds from the perspective of an investor whose objective is to:

1. Minimize the probability of earning a negative return.
2. Maximize the probability of earning a return between 0% and 10%.
3. Maximize the probability of earning a return greater than 10%.

**CASE STUDY 6.3** A variety of packaging solutions exist for products that must be kept within a specific temperature range. A cold chain distribution is a temperature-controlled supply chain. An unbroken cold chain is an uninterrupted series of storage and distribution activities that maintain a given temperature range. Cold chains are particularly useful in the food and pharmaceutical industries. A common suggested temperature range for a cold chain distribution in pharmaceutical industries is between 2 and 8 degrees Celsius.

Gopal Vasudeva works in the packaging branch of Merck & Co. He is in charge of analyzing a new package that the company has developed. With repeated trials, Gopal has determined that the mean temperature that this package is able to maintain during its use is 5.6°C with a standard deviation of 1.2°C. He is not sure if the distribution of temperature is symmetric or skewed to the right.

In a report, use the sample information to:

1. Calculate and interpret the probability that temperature goes (a) below 2°C and (b) above 8°C using a normal distribution approximation.

2. Calculate the probability that temperature goes (a) below 2°C and (b) above 8°C using a lognormal distribution approximation.
3. Compare the results from the two distributions used in the analysis.

## APPENDIX 6.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor.

### Minitab

#### The Uniform Distribution

- A. (Replicating Example 6.1b) From the menu choose **Calc > Probability Distributions > Uniform**.
- B. Select **Cumulative probability**. Enter 2500 as the **Lower endpoint** and 5000 as the **Upper endpoint**. Select **Input constant** and enter 4000. Because Minitab calculates  $P(X \leq 4000) = 0.60$ , and we want to find  $P(X > 4000)$ , we calculate  $1 - 0.60 = 0.40$ .

#### The Normal Distribution

##### *The Standard Transformation*

- A. (Replicating Example 6.8a) From the menu choose **Calc > Probability Distributions > Normal**.
- B. Select **Cumulative probability**. Enter 12 for the **Mean** and 3.2 for the **Standard deviation**. Select **Input constant** and enter 20. Because Minitab returns  $P(X \leq 20) = 0.9938$ , and we want to find  $P(X > 20)$ , we calculate  $1 - 0.9938 = 0.0062$ .

##### *The Inverse Transformation*

- A. (Replicating Example 6.8c) From the menu choose **Calc > Probability Distributions > Normal**.
- B. Select **Inverse cumulative probability**. Enter 12 for the **Mean** and 3.2 for the **Standard deviation**. Select **Input constant** and enter 0.90.

#### The Exponential Distribution

- A. (Replicating Example 6.9b) Choose **Calc > Probability Distributions > Exponential**.
- B. Select **Cumulative probability**. Enter 25 for **Scale** (since  $\text{Scale} = E(X) = 25$ ) and 0.0 for **Threshold**. Select **Input constant** and enter 60. Because Minitab returns  $P(X \leq 60) = 0.9093$ , and we want to find  $P(X > 60)$ , we calculate  $1 - 0.9093 = 0.0907$ .

#### The Lognormal Distribution

##### *The Lognormal Transformation*

- A. (Replicating 6.11a) From the menu choose **Calc > Probability Distributions > Lognormal**.
- B. Select **Cumulative probability**. Enter 5 for the **Location** and 1.2 for the **Scale**. Select **Input constant** and enter 200.



### *The Inverse Transformation*

- A. (Replicating 6.11b) From the menu choose **Calc > Probability Distributions > Lognormal**.
- B. Select **Inverse cumulative probability**. Enter 5 for the **Location** and 1.2 for the **Scale**. Select **Input constant** and enter 0.90.

## SPSS

Note: In order for the calculated probability to be seen on the spreadsheet, SPSS must first “view” data on the spreadsheet. For this purpose, enter a value of zero in the first cell of the first column.

### The Uniform Distribution

- A. (Replicating Example 6.1b) From the menu choose **Transform > Compute Variable**.
- B. Under **Target Variable**, type **cdfuniform**. Under **Function group**, select **CDF & Noncentral CDF** and under **Functions and Special Variables**, double-click on **Cdf.Uniform**. In the **Numeric Expression** box, enter 4000 for **quant**, 2500 for **min**, and 5000 for **max**. Because SPSS returns  $P(X \leq 4000) = 0.60$ , and we want to find  $P(X > 4000)$ , we calculate  $1 - 0.60 = 0.40$ .

### The Normal Distribution

#### *The Standard Transformation*

- A. (Replicating Example 6.8a) From the menu choose **Transform > Compute Variable**.
- B. Under **Target Variable**, type **cdfnorm**. Under **Function group**, select **CDF & Noncentral CDF** and under **Functions and Special Variables**, double-click on **Cdf.Normal**. In the **Numeric Expression** box, enter 20 for **quant**, 12 for **mean**, and 3.2 for **stddev**. Since SPSS returns  $P(X \leq 20) = 0.9938$ , and we want to find  $P(X > 20)$ , we calculate  $1 - 0.9938 = 0.0062$ .

#### *The Inverse Transformation*

- A. (Replicating Example 6.8c) From the menu choose **Transform > Compute Variable**.
- B. Under **Target Variable**, type **invnorm**. Under **Function group**, select **Inverse DF** and under **Functions and Special Variables**, double-click on **Idf.Normal**. In the **Numeric Expression** box, enter 0.9 for **prob**, 12 for **mean**, and 3.2 for **stddev**.

### The Exponential Distribution

- A. (Replicating Example 6.9b) From the menu choose **Transform > Compute Variable**.
- B. Under **Target Variable**, type **cdfexp**. Under **Function group**, select **CDF & Noncentral CDF** and under **Functions and Special Variables**, double-click on **Cdf.Exp**. In the **Numeric Expression** box, enter 60 for **quant** and 0.04 for **scale**. Since SPSS returns  $P(X \leq 60) = 0.9093$ , and we want to find  $P(X > 60)$ , we calculate  $1 - 0.9093 = 0.0907$ .

### The Lognormal Distribution

#### *The Lognormal Transformation*

- A. (Replicating 6.11a) The easiest way to solve lognormal distribution problems in SPSS is to modify the normal distribution instructions. So from the menu choose **Transform > Compute Variable**.
- B. Under **Target Variable**, type **cdflognorm**. Under **Function group**, select **CDF & Noncentral CDF** and under **Function and Special Variables**, double-click on **Cdf.Normal**. In the **Numeric Expression** box, enter  $\ln(200)$  for **quant**, 5 for **mean**, and 1.2 for **stddev**.



### *The Inverse Transformation*

- A. (Replicating 6.11b) From the menu choose **Transform > Compute Variable**.
- B. Under **Target Variable**, type `invlognorm`. In the **Numeric Expression** box, type `exp(IDF.NORMAL(0.90, 5, 1.2))`.

## **JMP**

Note: In order for the calculated probability to be seen on the spreadsheet, JMP must first “view” data on the spreadsheet. For this purpose, enter a value of zero in the first cell of the first column.

## **The Normal Distribution**

### *The Standard Transformation*

- A. (Replicating Example 6.8a) Right-click on the header at the top of the column in the spreadsheet view and select **Formula**. Under **Functions (grouped)**, choose **Probability > Normal Distribution**.
- B. Put the insertion marker on the box for **x** and click the insert button (shown as a caret ^ with the mathematical operations) until you see **mean** and **std dev** next to **x**. Enter 20 for **x**, 12 for **mean**, and 3.2 for **std dev**. Because JMP returns  $P(X \leq 20) = 0.9938$ , and we want to find  $P(X > 20)$ , we calculate  $1 - 0.9938 = 0.0062$ .

### *The Inverse Transformation*

- A. (Replicating Example 6.8c) Right-click on the header at the top of the column in the spreadsheet view and select **Formula**. Under **Functions (grouped)**, choose **Probability > Normal Quantile**.
- B. Put the insertion marker on the box for **p** and click the insert button (shown as a caret ^ with the mathematical operations) until you see **mean** and **std dev** next to **p**. Enter 0.90 for **p**, 12 for **mean**, and 3.2 for **std dev**.

## **The Exponential Distribution**

- A. (Replicating Example 6.9b) Right-click on the header at the top of the column in the spreadsheet view and select **Formula**. Under **Functions (grouped)**, choose **Probability > Weibull Distribution**. (The exponential distribution is a special case of the Weibull distribution when the shape parameter, see next step, equals 1.)
- B. Put the insertion marker on the box for **x** and click the insert button (shown as a caret ^ with the mathematical operations) until you see **shape** and **scale** next to **x**. Enter 60 for **x**, 1 for **shape**, and 25 for **scale**. Because JMP returns  $P(X \leq 60) = 0.9093$ , and we want to find  $P(X > 60)$ , we calculate  $1 - 0.9093 = 0.0907$ .

## **The Lognormal Distribution**

### *The Lognormal Transformation*

- A. (Replicating 6.11a) Right-click on the header at the top of the column in the spreadsheet view and select **Formula**. Under **Function (grouped)**, choose **Probability > GLog Distribution**.
- B. Put the insertion marker on the box for **x** and click the insert button (shown as a caret ^ with the mathematical operations) until you see **mu**, **sigma**, and **lambda** next to **x**. Enter 200 for **x**, 5 for **mu**, 1.2 for **sigma**, and 0 for **lambda**.

### *The Inverse Transformation*

- A. (Replicating 6.11b) Right-click on the header at the top of the column in the spreadsheet view and select **Formula**. Under **Function (grouped)**, choose **Probability > GLog Quantile**.
- B. Put the insertion marker on the box for **p** and click the insert button (shown as a caret ^ with the mathematical operations) until you see **mu**, **sigma**, and **lambda** next to **p**. Enter 0.90 for **p**, 5 for **mu**, 1.2 for **sigma**, and 0 for **lambda**.

# 7

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 7.1 Explain common sample biases.
- LO 7.2 Describe various sampling methods.
- LO 7.3 Describe the sampling distribution of the sample mean.
- LO 7.4 Explain the importance of the central limit theorem.
- LO 7.5 Describe the sampling distribution of the sample proportion.
- LO 7.6 Use a finite population correction factor.
- LO 7.7 Construct and interpret control charts for quantitative and qualitative data.

# Sampling and Sampling Distributions

In the last few chapters, we had information on the population parameters, such as the population proportion and the population mean, for the analysis of discrete and continuous random variables. In many instances, we do not have information on the parameters, so we make statistical inferences on the basis of sample statistics. The credibility of any statistical inference depends on the quality of the sample on which it is based. In this chapter, we discuss various ways to draw a good sample and also highlight cases in which the sample misrepresents the population. It is important to note that any given statistical problem involves only one population, but many possible samples from which a statistic can be derived. Therefore, while the population parameter is a constant, the sample statistic is a random variable whose value depends on the choice of the random sample. We will discuss how to evaluate the properties of sample statistics. In particular, we will study the probability distributions of the sample mean and the sample proportion based on simple random sampling. Finally, we will use these distributions to construct control charts, which are popular statistical tools for monitoring and improving quality.



## INTRODUCTORY CASE

### Marketing Iced Coffee

Although hot coffee is still Americans' drink of choice, the market share of iced coffee is growing steadily. Thirty percent of coffee drinkers had at least one iced, frozen, or blended coffee drink in 2009, up from 28% in 2008 (*The Boston Globe*, April 6, 2010). In response to this growing change in taste, the coffee chains have ramped up their offerings: Starbucks recently introduced an upgraded Frappuccino; Dunkin' Donuts launched a new iced dark roast; and McDonald's unveiled new blended coffee iced drinks and smoothies.

In order to capitalize on this trend, Starbucks advertised a Happy Hour from May 7 through May 16 whereby customers enjoyed a half-price Frappuccino beverage between 3 pm and 5 pm (<http://starbucks.com>). Anne Jones, a manager at a local Starbucks (see the Chapter 5 introductory case), wonders how this marketing campaign has affected her business. She knows that women and teenage girls comprise the majority of the iced-coffee market, since they are willing to spend more on indulgences. In fact, Anne reviews her records prior to the promotion and finds that 43% of iced-coffee customers were women and 21% were teenage girls. She also finds that customers spent an average of \$4.18 on iced coffee with a standard deviation of \$0.84.

One month after the marketing period ends, Anne surveys 50 of her iced-coffee customers and finds that they had spent an average of \$4.26. In addition, 23 (46%) of the customers were women and 17 (34%) were teenage girls. Anne wants to determine if the marketing campaign has had a lingering effect on the amount of money customers spend on iced coffee and on the proportion of customers who are women and teenage girls. Anne wonders if Starbucks would have gotten such business if it had chosen not to pursue the marketing campaign.

Anne wants to use the above survey information to:

1. Calculate the probability that customers spend an average of \$4.26 or more on iced coffee.
2. Calculate the probability that 46% or more of iced-coffee customers are women.
3. Calculate the probability that 34% or more of iced-coffee customers are teenage girls.

A synopsis of this case is provided at the end of Section 7.3.

Explain common sample biases.

A major portion of statistics is concerned with statistical inference, where we examine the problem of estimating population parameters or testing hypotheses about such parameters. Recall that a population consists of all items of interest in the statistical problem. If we had access to data that encompass the entire population, then the values of the parameters would be known and no statistical inference would be needed. Since it is generally not feasible to gather data on an entire population, we use a subset of the population, or a sample, and use this information to make statistical inference. We can think of a census and survey data as representative of population and sample data, respectively. While a census captures almost everyone in the country, a survey captures a small number of people who fit a particular category. We regularly use survey data to analyze government and business activities.

#### POPULATION VERSUS SAMPLE

A **population** consists of all items of interest in a statistical problem, whereas a **sample** is a subset of the population. We use a **sample statistic**, or simply **statistic**, to make inferences about the unknown population **parameter**.

In later chapters, we explore estimation and hypothesis testing, which are based on sample information. It is important to note that no matter how sophisticated the statistical methods are, the credibility of statistical inference depends on the quality of the sample on which it is based. A primary requisite for a “good” sample is that it be **representative** of the population we are trying to describe. When the information from a sample is not typical of information in the population in a systematic way, we say that **bias** has occurred.

**Bias** refers to the tendency of a sample statistic to systematically over- or underestimate a population parameter. It is often caused by samples that are not representative of the population.

### Classic Case of a “Bad” Sample: The *Literary Digest* Debacle of 1936

In theory, drawing conclusions about a population based on a good sample sounds logical; however, in practice, what constitutes a “good” sample? Unfortunately, there are many ways to collect a “bad” sample. One way is to inadvertently pick a sample that represents only a portion of the population. The *Literary Digest*’s attempt to predict the 1936 presidential election is a classic example of an embarrassingly inaccurate poll.

In 1932 and amid the Great Depression, Herbert Hoover was voted out of the White House and Franklin Delano Roosevelt (FDR) was elected the 32nd president of the United States. Although FDR’s attempts to end the Great Depression within four years were largely unsuccessful, he retained the general public’s faith. In 1936, FDR ran for reelection against Alf Landon, the governor of Kansas and the Republican nominee. The *Literary Digest*, an influential, general-interest weekly magazine, wanted to predict the next U.S. president, as it had done successfully five times before.

After conducting the largest poll in history, the *Literary Digest* predicted a landslide victory for Alf Landon: 57% of the vote to FDR’s 43%. Moreover, the *Literary Digest* claimed that its prediction would be within a fraction of 1% of the actual vote. Instead, FDR won in a landslide: 62% to 38%. So what went wrong?

The *Literary Digest* sent postcards to 10 million people (one-quarter of the voting population at the time) and received responses from 2.4 million people. The response rate of 24% (2.4 million/10 million) might seem low to some, but in reality it is a reasonable response rate given this type of polling. What was atypical of the poll is the manner in which the *Literary Digest* obtained the respondents' names. The *Literary Digest* randomly sampled its own subscriber list, club membership rosters, telephone directories, and automobile registration rolls. This sample reflected predominantly middle- and upper-class people; that is, the vast majority of those polled were wealthier people, who were more inclined to vote for the Republican candidate. Back in the 1930s, owning a phone, for instance, was far from universal. Only 11 million residential phones were in service in 1936, and these homes were disproportionately well-to-do and in favor of Landon. The sampling methodology employed by the *Literary Digest* suffered from **selection bias**. Selection bias occurs when portions of the population are underrepresented in the sample. FDR's support came from lower-income classes whose opinion was not reflected in the poll. The sample, unfortunately, misrepresented the general electorate.

**Selection bias** refers to a systematic underrepresentation of certain groups from consideration for the sample.

What should the *Literary Digest* have done differently? At a minimum, most would agree that names should have been obtained from voter registration lists rather than telephone directory lists and car registrations.

In addition to selection bias, the *Literary Digest* survey also had a great deal of **nonresponse bias**. This occurs when those responding to a survey or poll differ systematically from the nonrespondents. In the survey, a larger percentage of educated people mailed back the questionnaires. During that time period, the more educated tended to come from affluent families that again favored the Republican candidate. Problems with nonresponse bias persist today. Most people do not want to spend time carefully reading and responding to polls conducted by mail. Only those who care a great deal about an election or a particular issue take the time to read the instructions, fill out the questionnaire, and mail it back. Those who do respond may be atypical of the population as a whole.

**Nonresponse bias** refers to a systematic difference in preferences between respondents and nonrespondents to a survey or a poll.

The most effective way to deal with nonresponse bias is to reduce nonresponse rates. Paying attention to survey design, wording, and ordering of the questions can increase the response rate. Sometimes, rather than sending out a very large number of surveys, it may be preferable to use a smaller representative sample for which the response rate is likely to be high.

It turns out that someone did accurately predict the 1936 presidential election. From a sample of 50,000 with a response rate of 10% (5,000 respondents), a young pollster named George Gallup predicted that FDR would win 56% of the vote to Landon's 44%. Despite using a far smaller sample with a lower response rate, it was far more *representative* of the true voting population. Gallup later founded the Gallup Organization, one of the leading polling companies of all time.

## Sampling Methods

As mentioned earlier, a primary requisite for a “good” sample is that it be representative of the population you are trying to describe. The basic type of sample that can be used to draw statistically sound conclusions about a population is a **simple random sample**.

### LO 7.2

Describe various sampling methods.



### SIMPLE RANDOM SAMPLE

A **simple random sample** is a sample of  $n$  observations that has the same probability of being selected from the population as any other sample of  $n$  observations. Most statistical methods presume simple random samples.

### EXAMPLE 7.1

A recent analysis shows a dramatic decline in studying time among today's college students (*The Boston Globe*, July 4, 2010). In 1961, students invested 24 hours per week in their academic pursuits, whereas today's students study an average of 14 hours per week. A dean at a large university in California wonders if this trend applies to the students at her university. The university has 20,000 students and the dean would like a sample of 100. Use Excel to generate a simple random sample of 100 students.

**SOLUTION:** We can use Excel's RANDBETWEEN function to generate random integers within some interval. In general, we input “=RANDBETWEEN(Bottom, Top)”, where Bottom and Top refer to the smallest and largest integers, respectively, that Excel might return. In order to randomly select a student from a list of 20,000 students, we input “=RANDBETWEEN(1, 20000)”. Suppose Excel returns the value 6,319. The dean can then choose the 6,319th student from the list. In order to generate the remaining 99 random numbers, we can select the cell with the value 6,319, drag it down 99 cells, and then from the menu choose **Home > Fill > Down**.

While a simple random sample is the most commonly used sampling method, in some situations other sampling methods have an advantage over simple random samples. Two alternative methods for forming a sample are stratified random sampling and cluster sampling.

Political pollsters often employ **stratified random sampling** in an attempt to ensure that each area of the country, each ethnic group, each religious group, and so forth, is appropriately represented in the sample. With stratified random sampling, the population is divided into groups (strata) based on one or more classification criteria. Simple random samples are then drawn from each stratum in sizes proportional to the relative size of each stratum in the population. These samples are then pooled.

### STRATIFIED RANDOM SAMPLING

In **stratified random sampling**, the population is first divided up into mutually exclusive and collectively exhaustive groups, called *strata*. A stratified sample includes randomly selected observations from each stratum. The number of observations per stratum is proportional to the stratum's size in the population. The data for each stratum are eventually pooled.

Stratified random sampling has two advantages. First, it guarantees that the population subdivisions of interest are represented in the sample. Second, the estimates of parameters produced from stratified random sampling have greater precision than estimates obtained from simple random sampling.

Even stratified random sampling, however, can fall short with its predictive ability. One of the nagging mysteries of the 2008 Democratic presidential primaries was: Why were the polls so wrong in New Hampshire? All nine major polling groups predicted that Barack Obama would beat Hillary Clinton in the New Hampshire primary by an average



of 8.3 percentage points. When the votes were counted, Clinton won by 2.6%. Several factors contributed to the wrong prediction by the polling industry. First, pollsters overestimated the turnout of young voters, who overwhelmingly favored Obama in exit polls but did not surge to vote as they had in the Iowa caucus. Second, Clinton's campaign made a decision to target female Democrats, especially single women. This focus did not pay off in Iowa, but it did in New Hampshire. Finally, on the eve of the primary, a woman in Portsmouth asked Clinton: "How do you do it?" Clinton's teary response was powerful and warm. Voters, who rarely saw Clinton in such an emotional moment, found her response humanizing and appealing. Most polls had stopped phoning voters over the weekend, too soon to catch the likely voter shift.

**Cluster sampling** is another method for forming a representative sample. A cluster sample is formed by dividing the population into groups (clusters), such as geographic areas, and then selecting a sample of the groups for the analysis. The technique works best when most of the variation in the population is within the groups and not between the groups. In such instances, a cluster is a miniversion of the population.

#### CLUSTER SAMPLING

In **cluster sampling**, the population is first divided up into mutually exclusive and collectively exhaustive groups, called *clusters*. A cluster sample includes observations from randomly selected clusters.

In general, cluster sampling is cheaper as compared to other sampling methods. However, for a given sample size, it provides less precision than either simple random sampling or stratified sampling. Cluster sampling is useful in applications where the population is concentrated in natural clusters such as city blocks, schools, and other geographic areas. It is especially attractive when constructing a complete list of the population members is difficult and/or costly. For example, since it may not be possible to create a full list of customers who go to Walmart, we can form a sample that includes customers only from selected stores.

#### STRATIFIED VERSUS CLUSTER SAMPLING

In stratified sampling, the sample consists of observations from each group, whereas in cluster sampling, the sample consists of observations from the selected groups. Stratified sampling is preferred when the objective is to increase precision, and cluster sampling is preferred when the objective is to reduce costs.

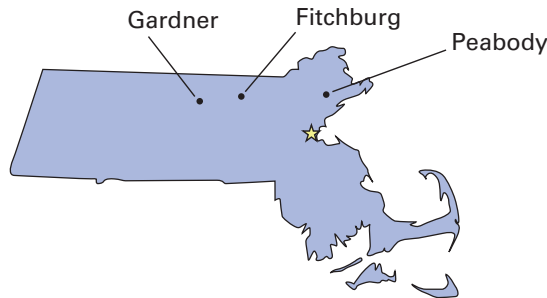
## The Special Election to Fill Ted Kennedy's Senate Seat

On January 19, 2010, Scott Brown, the Republican candidate, beat Martha Coakley, the Democratic candidate, in a special election to fill the U.S. Senate seat for Massachusetts that had been vacated with the death of Senator Ted Kennedy. Given that Kennedy, the "Liberal Lion," had held the seat for over 40 years, the election was one of the biggest upsets in Massachusetts' political history. Nine days prior to the election, a *Boston Globe* poll gave Coakley, the state's attorney general, a 15-point lead over Brown. Critics accused the *Globe*, which had endorsed Coakley, of purposely running a bad poll to discourage voters from coming out for Brown. In reality, by the time the *Globe* released the poll, it contained old information from January 2–6. In addition, the *Globe* partnered with the University of New Hampshire for the poll, and unfortunately included people in the poll who said that they were unlikely to vote! Eighty years after the *Literary Digest* fiasco, pollsters are still making predictions based on samples with a great deal of selection bias.

The first poll that foretold Brown's stunning victory over Coakley was released by Suffolk University on January 14. The poll had Brown ahead by 50% to Coakley's 46%,

approximately one percentage point off the Election Day results (52% to 47%). How did Suffolk University arrive at its findings? It conducted a statewide poll and, in addition, implemented a form of cluster sampling. As mentioned earlier, the technique works best when most of the variation in the population is within the groups and not between the groups. The pollsters from Suffolk University selected three bellwethers, or towns that would indicate the way that the state would vote. In choosing the bellwethers, the pollsters spent enormous amounts of time examining the results of similar elections over many years. Figure 7.1 shows a map of Massachusetts and the three bellwethers: Gardner, Fitchburg, and Peabody. The statewide poll and the results from the bellwethers were reported separately but yielded the same results.

**FIGURE 7.1**  
Map of Massachusetts  
with three bellwethers  
(towns)



In practice, it is extremely difficult to obtain a truly random sample that is representative of the underlying population. As researchers, we need to be aware of the population from which the sample was selected and then limit our conclusions to that population. For the remainder of the text, we assume that the sample data are void of “human error”; that is, we have sampled from the correct population (no selection bias); we have no response bias; and we have collected, analyzed, and reported the data properly.

## EXERCISES 7.1

1. In 2010, Apple introduced the iPad, a tablet-style computer that its former CEO Steve Jobs called a “a truly magical and revolutionary product” (*CNN*, January 28, 2010). Suppose you are put in charge of determining the age profile of people who purchased the iPad in the United States. Explain in detail the following sampling strategies that you could use to select a representative sample.
  - a. Simple random sampling
  - b. Stratified random sampling
  - c. Cluster sampling
2. A marketing firm opens a small booth at a local mall over the weekend, where shoppers are asked how much money they spent at the food court. The objective is to determine the average monthly expenditure of shoppers at the food court. Has the marketing firm committed any sampling bias? Discuss.
3. Natalie Min is an undergraduate in the Haas School of Business at Berkeley. She wishes to pursue an MBA from Berkeley and wants to know the profile of other students who are likely to apply to the Berkeley MBA program. In particular, she wants to know the GPA of students with whom she might be competing. She randomly surveys 40 students from her accounting class for the analysis. Discuss in detail whether or not Natalie’s analysis is based on a representative sample.
4. Vons, a large supermarket in Grover Beach, California, is considering extending its store hours from 7:00 am to midnight, seven days a week, to 6:00 am to midnight. Discuss the sampling bias in the following sampling strategies:
  - a. Mail a prepaid envelope to randomly selected residents in the Grover Beach area, asking for their preference for the store hours.
  - b. Ask the customers who frequent the store in the morning if they would prefer an earlier opening time.
  - c. Place an ad in the local newspaper, requesting people to submit their preference for store hours on the store’s website.
5. In the previous question regarding Vons’ store hours, explain how you can obtain a representative sample based on the following sampling strategies:
  - a. Simple random sampling.
  - b. Stratified random sampling.
  - c. Cluster sampling.

## 7.2 THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

As mentioned earlier, we are generally interested in the characteristics of a population. For instance, a student is interested in the average starting salary (population mean) of business graduates. Similarly, a banker is interested in the default probability (population proportion) of mortgage holders. Recall that the population mean and the population proportion are parameters that describe quantitative and qualitative data, respectively. Since it is cumbersome, if not impossible, to analyze the entire population, we generally make inferences about the characteristics of the population on the basis of a random sample drawn from the population.

It is important to note that there is only one population, but many possible samples of a given size can be drawn from the population. Therefore, a population parameter is a constant, even though its value may be unknown. On the other hand, a statistic, such as the sample mean or the sample proportion, is a random variable whose value depends on the particular sample that is randomly drawn from the population.

A **parameter** is a **constant**, although its value may be unknown. A **statistic** is a **random variable** whose value depends on the chosen random sample.

Consider the starting salary of business graduates as the variable of interest. If you decide to make inferences about the population mean salary on the basis of a random draw of 38 recent business graduates, then the sample mean  $\bar{X}$  is the relevant statistic. Note that the value of  $\bar{X}$  will change if you choose a different random sample of 38 business graduates. In other words,  $\bar{X}$  is a random variable whose value depends on the chosen random sample. The sample mean is commonly referred to as the **estimator**, or the **point estimator**, of the population mean.

### ESTIMATOR AND ESTIMATE

When a statistic is used to estimate a parameter, it is referred to as an **estimator**. A particular value of the estimator is called an **estimate**.

In the above example, the sample mean  $\bar{X}$  is the estimator of the mean starting salary of business graduates. If the average derived from a specific sample is \$54,000, then  $\bar{x} = 54,000$  is the estimate of the population mean. Similarly, if the variable of interest is the default probability of mortgage holders, then the sample proportion of defaults, denoted by  $\bar{P}$ , from a random sample of 80 mortgage holders is the estimator of the population proportion. If 10 out of 80 mortgage holders in a given sample default, then  $\bar{p} = 10/80 = 0.125$  is the estimate of the population proportion.

In this section, we will focus on the probability distribution of the sample mean  $\bar{X}$ , which is also referred to as the **sampling distribution** of  $\bar{X}$ . Since  $\bar{X}$  is a random variable, its sampling distribution is simply the probability distribution derived from all possible samples of a given size from the population. Consider, for example, a mean derived from a sample of  $n$  observations. Another mean can similarly be derived from a different sample of  $n$  observations. If we repeat this process a very large number of times, then the frequency distribution of the sample means can be thought of as its sampling distribution. In particular, we will discuss the expected value and the standard deviation of the sample mean. We will also study the conditions under which the sampling distribution of the sample mean is normally distributed.

**LO 7.3**

Describe the sampling distribution of the sample mean.

## The Expected Value and the Standard Error of the Sample Mean

Let the random variable  $X$  represent a certain characteristic of a population under study, with an expected value,  $E(X) = \mu$ , and a variance,  $\text{Var}(X) = \sigma^2$ . For example,  $X$  could represent the salary of business graduates or the return on investment. We can think of  $\mu$  and  $\sigma^2$  as the mean and the variance of an individual observation drawn randomly from the population of interest, or simply as the population mean and the population variance. Let the sample mean  $\bar{X}$  be based on a random sample of  $n$  observations from this population. It is easy to derive the expected value and the variance of  $\bar{X}$  (see Appendix 7.1 for the derivations).

The **expected value** of  $\bar{X}$  is the same as the expected value of the individual observation—that is,  $E(\bar{X}) = E(X) = \mu$ . In other words, if we were to sample repeatedly from a given population, the average value of the sample means will equal the population mean from the underlying population. This is an important property of an estimator, called unbiasedness, that holds irrespective of whether the sample mean is based on a small or a large sample. An estimator is **unbiased** if its expected value equals the population parameter. Other desirable properties of an estimator are described in Appendix 7.2.

### THE EXPECTED VALUE OF THE SAMPLE MEAN

The **expected value** of the sample mean  $\bar{X}$  equals the population mean, or  $E(\bar{X}) = \mu$ . In other words, the sample mean is an **unbiased** estimator of the population mean.

It is important to note that we estimate the population mean on the basis of just one sample. The above result shows that we are not systematically under- or overestimating the population parameter.

The **variance** of  $\bar{X}$  is equal to  $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ . In other words, if we were to sample repeatedly from a given population, the variance of the sample mean will equal the variance of the individual observation, drawn from the underlying population, divided by the sample size. Note that  $\text{Var}(\bar{X})$  is smaller than the variance of  $X$ , which is equal to  $\text{Var}(X) = \sigma^2$ . This is an intuitive result, suggesting that the variability between sample means is less than the variability between observations. Since each sample is likely to contain both high and low observations, the highs and lows cancel one another, making the variance of  $\bar{X}$  smaller than the variance of  $X$ . As usual, the **standard deviation** of  $\bar{X}$  is calculated as the positive square root of the variance. However, in order to distinguish the variability between samples from the variability between individual observations, we refer to the standard deviation of  $\bar{X}$  as the **standard error of the sample mean** computed as  $se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ .

### THE STANDARD ERROR OF THE SAMPLE MEAN

The standard deviation of the sample mean  $\bar{X}$  is referred to as the **standard error of the sample mean**. It equals the population standard deviation divided by the square root of the sample size—that is,  $se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ .

In Chapter 8, we will discuss that the exact standard error of an estimator is often not known and, therefore, must be estimated from the given sample data. For convenience, we use “ $se$ ” to denote both the exact and the estimated standard error of an estimator.

### EXAMPLE 7.2

The chefs at a local pizza chain in Cambria, California, strive to maintain the suggested size of their 16-inch pizzas. Despite their best efforts, they are unable to make every pizza exactly 16 inches in diameter. The manager has determined that the size of the pizzas is normally distributed with a mean of 16 inches and a standard deviation of 0.8 inch.

- What are the expected value and the standard error of the sample mean derived from a random sample of 2 pizzas?
- What are the expected value and the standard error of the sample mean derived from a random sample of 4 pizzas?
- Compare the expected value and the standard error of the sample mean with those of an individual pizza.

**SOLUTION:** We know that the population mean  $\mu = 16$  and the population standard deviation  $\sigma = 0.8$ . We use  $E(\bar{X}) = \mu$  and  $se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$  to calculate the following results.

- With the sample size  $n = 2$ ,  $E(\bar{X}) = 16$  and  $se(\bar{X}) = \frac{0.8}{\sqrt{2}} = 0.57$ .
- With the sample size  $n = 4$ ,  $E(\bar{X}) = 16$  and  $se(\bar{X}) = \frac{0.8}{\sqrt{4}} = 0.40$ .
- The expected value of the sample mean for both sample sizes is identical to the expected value of the individual pizza. However, the standard error of the sample mean with  $n = 4$  is lower than the one with  $n = 2$ . For both sample sizes, the standard error of the sample mean is lower than the standard deviation of the individual pizza. This result confirms that averaging reduces variability.

## Sampling from a Normal Population

An important feature of the sampling distribution of the sample mean  $\bar{X}$  is that, irrespective of the sample size  $n$ ,  $\bar{X}$  is normally distributed if the population  $X$  from which the sample is drawn is normal. In other words, if  $X$  is normal with expected value  $\mu$  and standard deviation  $\sigma$ , then  $\bar{X}$  is also normal with expected value  $\mu$  and standard error  $\sigma/\sqrt{n}$ .

### SAMPLING FROM A NORMAL POPULATION

For any sample size  $n$ , the sampling distribution of  $\bar{X}$  is **normal** if the population  $X$  from which the sample is drawn is normally distributed.

If  $\bar{X}$  is normal, we can transform it into a **standard normal random variable** as:

$$Z = \frac{\bar{X} - E(\bar{X})}{se(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Therefore, any value  $\bar{x}$  has a corresponding value  $z$  given by  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ .

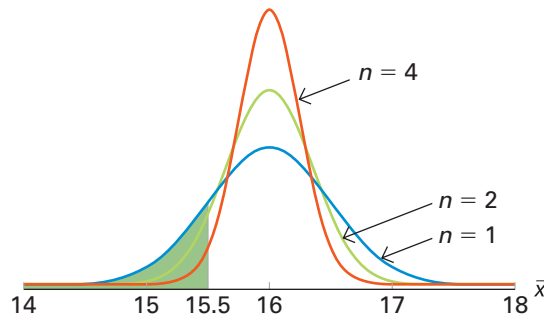
### EXAMPLE 7.3

Use the information in Example 7.2 to answer the following questions:

- What is the probability that a randomly selected pizza is less than 15.5 inches?
- What is the probability that 2 randomly selected pizzas average less than 15.5 inches?
- What is the probability that 4 randomly selected pizzas average less than 15.5 inches?
- Comment on the computed probabilities.

**SOLUTION:** Since the population is normally distributed, the sampling distribution of the sample mean is also normal. Figure 7.2 depicts the shapes of the three distributions based on the population mean  $\mu = 16$  and the population standard deviation  $\sigma = 0.8$ .

**FIGURE 7.2** Normal distribution of the sample mean



Note that when the sample size  $n = 1$ , the sample mean  $\bar{x}$  is the same as the individual observation  $x$ .

- a. We use the standard transformation to derive  $P(X < 15.5) = P\left(Z < \frac{15.5 - 16}{0.8}\right) = P(Z < -0.63) = 0.2643$ . There is a 26.43% chance that an individual pizza is less than 15.5 inches.
- b. Here we use the standard transformation to derive  $P(\bar{X} < 15.5) = P\left(Z < \frac{15.5 - 16}{0.8/\sqrt{2}}\right) = P(Z < -0.88) = 0.1894$ . In a random sample of 2 pizzas, there is an 18.94% chance that the average size is less than 15.5 inches.
- c. Again we find  $P(\bar{X} < 15.5)$ , but now  $n = 4$ . Therefore,  $P(\bar{X} < 15.5) = P\left(Z < \frac{15.5 - 16}{0.8/\sqrt{4}}\right) = P(Z < -1.25) = 0.1056$ . In a random sample of 4 pizzas, there is a 10.56% chance that the average size is less than 15.5 inches.
- d. The probability that the average size is under 15.5 inches, for 4 randomly selected pizzas, is less than half of that for an individual pizza. This is due to the fact that while  $X$  and  $\bar{X}$  have the same expected value of 16, the variance of  $\bar{X}$  is less than that of  $X$ .

#### LO 7.4

Explain the importance of the central limit theorem.

## The Central Limit Theorem

For making statistical inferences, it is essential that the sampling distribution of  $\bar{X}$  is normally distributed. So far we have only considered the case where  $\bar{X}$  is normally distributed because the population  $X$  from which the sample is drawn is normal. What if the underlying population is not normal? Here we present the **central limit theorem (CLT)**, which perhaps is the most remarkable result of probability theory. The CLT states that the sum or the average of a large number of independent observations from the same underlying distribution has an approximate normal distribution. The approximation steadily improves as the number of observations increases. In other words, irrespective of whether or not the population  $X$  is normal, the sample mean  $\bar{X}$  computed from a random sample of size  $n$  will be approximately normally distributed as long as  $n$  is sufficiently large.

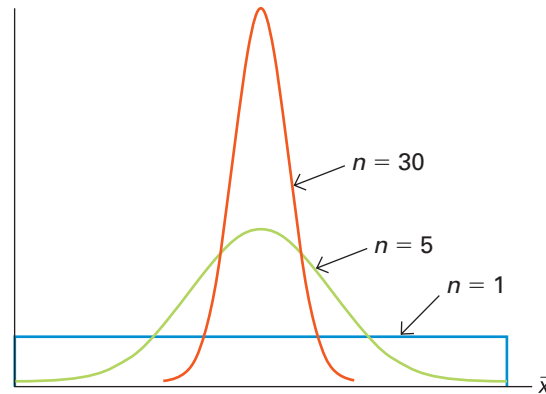
### THE CENTRAL LIMIT THEOREM FOR THE SAMPLE MEAN

For any population  $X$  with expected value  $\mu$  and standard deviation  $\sigma$ , the sampling distribution of  $\bar{X}$  will be **approximately normal if the sample size  $n$  is sufficiently large**. As a general guideline, the normal distribution approximation is justified when  $n \geq 30$ .

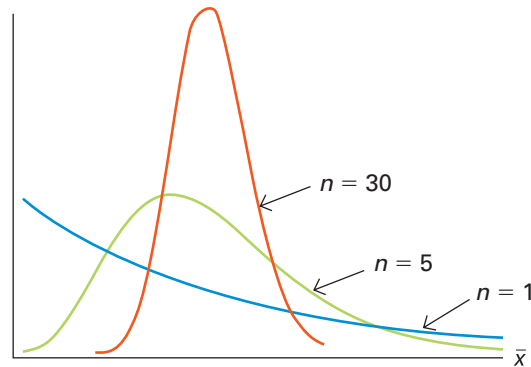
As before, if  $\bar{X}$  is approximately normal, then we can transform any value  $\bar{x}$  to its corresponding value  $z$  given by  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ .



Figure 7.2, discussed in Example 7.3, is not representative of the CLT principle because, for a normal population, the sampling distribution of  $\bar{X}$  is normal irrespective of the sample size. Figures 7.3 and 7.4, however, illustrate the CLT by using random samples of various sizes drawn from nonnormal populations. The relative frequency polygon of  $\bar{X}$ , which essentially represents its distribution, is generated from repeated draws (computer simulations) from the continuous uniform distribution (Figure 7.3) and the exponential distribution (Figure 7.4). Both of these nonnormal distributions were discussed in Chapter 6.



**FIGURE 7.3**  
Sampling distribution of  $\bar{X}$   
when the population has a  
uniform distribution



**FIGURE 7.4**  
Sampling distribution of  $\bar{X}$   
when the population has an  
exponential distribution

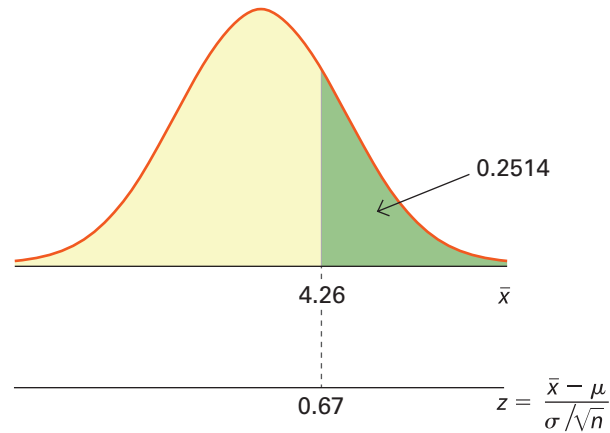
Note that when the sample size  $n = 1$ , the sample mean is the same as the individual observation (population) with the familiar uniform and exponential shapes. With  $n = 5$ , the sampling distribution of  $\bar{X}$  begins to resemble the shape of the normal distribution. With  $n = 30$ , the shapes of the sampling distribution of  $\bar{X}$  are approximately normal with the uniform distribution as well as the exponential distribution. The CLT can similarly be illustrated with other distributions of the population. How large a sample is necessary for normal convergence depends on the magnitude of the departure of the population from normality. As mentioned earlier, practitioners often use the normal distribution approximation when  $n \geq 30$ .

### EXAMPLE 7.4

Consider the information presented in the introductory case of this chapter. Recall that Anne wants to determine if the marketing campaign has had a lingering effect on the amount of money customers spend on iced coffee. Before the campaign, customers spent an average of \$4.18 on iced coffee with a standard deviation of \$0.84. Anne reports that the average amount, based on 50 customers sampled after the campaign, is \$4.26. If Starbucks chose not to pursue the marketing campaign, how likely is it that customers will spend an average of \$4.26 or more on iced coffee?

**SOLUTION:** If Starbucks did not pursue the marketing campaign, spending on iced coffee would still have mean  $\mu = 4.18$  and standard deviation  $\sigma = 0.84$ . Anne needs to calculate the probability that the sample mean is at least 4.26, or,  $P(\bar{X} \geq 4.26)$ . The population from which the sample is drawn is not known to be normal. However, since  $n \geq 30$ , from the central limit theorem, we know that  $\bar{X}$  is approximately normal. Therefore, as shown in Figure 7.5,  $P(\bar{X} \geq 4.26) = P\left(Z \geq \frac{4.26 - 4.18}{0.84/\sqrt{50}}\right) = P(Z \geq 0.67) = 1 - 0.7486 = 0.2514$ . It is quite plausible (probability = 0.2514) that in a sample of 50 customers, the sample mean is \$4.26 or more even if Starbucks did not pursue the marketing campaign.

**FIGURE 7.5** Finding  $P(\bar{X} \geq 4.26)$



## EXERCISES 7.2

### Mechanics

6. A random sample is drawn from a normally distributed population with mean  $\mu = 12$  and standard deviation  $\sigma = 1.5$ .
  - a. Comment on the sampling distribution of the sample mean with  $n = 20$  and  $n = 40$ .
  - b. Can you use the standard normal distribution to calculate the probability that the sample mean is less than 12.5 for both sample sizes?
  - c. Report the probability if you answered yes to the previous question for either sample size.
7. A random sample is drawn from a population with mean  $\mu = 66$  and standard deviation  $\sigma = 5.5$ .
  - a. Comment on the sampling distribution of the sample mean with  $n = 16$  and  $n = 36$ .
  - b. Can you use the standard normal distribution to calculate the probability that the sample mean falls between 66 and 68 for both sample sizes?
  - c. Report the probability if you answered yes to the previous question for either sample size.
8. A random sample of size  $n = 100$  is taken from a population with mean  $\mu = 80$  and standard deviation  $\sigma = 14$ .
  - a. Calculate the expected value and the standard error for the sampling distribution of the sample mean.
  - b. What is the probability that the sample mean falls between 77 and 85?
  - c. What is the probability that the sample mean is greater than 84?
9. A random sample of size  $n = 50$  is taken from a population with mean  $\mu = -9.5$  and standard deviation  $\sigma = 2$ .
  - a. Calculate the expected value and the standard error for the sampling distribution of the sample mean.
  - b. What is the probability that the sample mean is less than  $-10$ ?
  - c. What is the probability that the sample mean falls between  $-10$  and  $-9$ ?

### Applications

10. According to a recent survey, high school girls average 100 text messages daily (*The Boston Globe*, April 21, 2010). Assume the population standard deviation is 20 text messages. Suppose a random sample of 50 high school girls is taken.
  - a. What is the probability that the sample mean is more than 105?

- b. What is the probability that the sample mean is less than 95?
  - c. What is the probability that the sample mean is between 95 and 105?
11. Beer bottles are filled so that they contain an average of 330 ml of beer in each bottle. Suppose that the amount of beer in a bottle is normally distributed with a standard deviation of 4 ml.
  - a. What is the probability that a randomly selected bottle will have less than 325 ml of beer?
  - b. What is the probability that a randomly selected 6-pack of beer will have a mean amount less than 325 ml?
  - c. What is the probability that a randomly selected 12-pack of beer will have a mean amount less than 325 ml?
  - d. Comment on the sample size and the corresponding probabilities.
12. Despite its nutritional value, seafood is only a tiny part of the American diet, with the average American eating just 16 pounds of seafood per year. Janice and Nina both work in the seafood industry and they decide to create their own random samples and document the average seafood diet in their sample. Let the standard deviation of the American seafood diet be 7 pounds.
  - a. Janice samples 42 Americans and finds an average seafood consumption of 18 pounds. How likely is it to get an average of 18 pounds or more if she had a representative sample?
  - b. Nina samples 90 Americans and finds an average seafood consumption of 17.5 pounds. How likely is it to get an average of 17.5 pounds or more if she had a representative sample?
  - c. Which of the two women is likely to have used a more representative sample? Explain.
13. The weight of people in a small town in Missouri is known to be normally distributed with a mean of 180 pounds and a standard deviation of 28 pounds. On a raft that takes people across the river, a sign states, "Maximum capacity 3,200 pounds or 16 persons." What is the probability that a random sample of 16 persons will exceed the weight limit of 3,200 pounds?
14. The weight of turkeys is known to be normally distributed with a mean of 22 pounds and a standard deviation of 5 pounds.
  - a. Discuss the sampling distribution of the sample mean based on a random draw of 16 turkeys.
  - b. Find the probability that the mean weight of 16 randomly selected turkeys is more than 25 pounds.
  - c. Find the probability that the mean weight of 16 randomly selected turkeys is between 18 and 24 pounds.
15. A small hair salon in Denver, Colorado, averages about 30 customers on weekdays with a standard deviation of 6. It is safe to assume that the underlying distribution is normal. In an attempt to increase the number of weekday customers, the manager offers a \$2 discount on 5 consecutive weekdays. She reports that her strategy has worked since the sample mean of customers during this 5 weekday period jumps to 35.
  - a. How unusual would it be to get a sample average of 35 or more customers if the manager had not offered the discount?
  - b. Do you feel confident that the manager's discount strategy has worked? Explain.
16. Last year, the typical college student graduated with \$27,200 in debt (*The Boston Globe*, May 27, 2012). Let debt among recent college graduates be normally distributed with a standard deviation of \$7,000.
  - a. What is the probability that the average debt of four recent college graduates is more than \$25,000?
  - b. What is the probability that the average debt of four recent college graduates is more than \$30,000?
17. Forty families gathered for a fund-raising event. Suppose the individual contribution for each family is normally distributed with a mean and a standard deviation of \$115 and \$35, respectively. The organizers would call this event a success if the total contributions exceed \$5,000. What is the probability that this fund-raising event is a success?
18. A doctor is getting sued for malpractice by four of her former patients. It is believed that the amount that each patient will sue her for is normally distributed with a mean of \$800,000 and a standard deviation of \$250,000.
  - a. What is the probability that a given patient sues the doctor for more than \$1,000,000?
  - b. If the four patients sue the doctor independently, what is the probability that the total amount they sue for is over \$4,000,000?
19. Suppose that the miles-per-gallon (mpg) rating of passenger cars is a normally distributed random variable with a mean and a standard deviation of 33.8 and 3.5 mpg, respectively.
  - a. What is the probability that a randomly selected passenger car gets more than 35 mpg?
  - b. What is the probability that the average mpg of four randomly selected passenger cars is more than 35 mpg?
  - c. If four passenger cars are randomly selected, what is the probability that all of the passenger cars get more than 35 mpg?
20. Suppose that IQ scores are normally distributed with a mean of 100 and a standard deviation of 16.
  - a. What is the probability that a randomly selected person will have an IQ score of less than 90?
  - b. What is the probability that the average IQ score of four randomly selected people is less than 90?
  - c. If four people are randomly selected, what is the probability that all of them have an IQ score of less than 90?

## 7.3 THE SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

### LO 7.5

Describe the sampling distribution of the sample proportion.

Our discussion thus far has focused on the population mean, but many business, socioeconomic, and political matters are concerned with the population proportion. For instance, a banker is interested in the default probability of mortgage holders; a superintendent may note the proportion of students suffering from the flu when determining whether to keep school open; an incumbent up for reelection cares about the proportion of constituents that will ultimately cast a vote for him/her. In all of these examples, the parameter of interest is the population proportion  $p$ . However, analogous to our discussion concerning the mean, we almost always make inferences about the population proportion on the basis of sample data. Here, the relevant statistic (estimator) is the sample proportion,  $\bar{P}$ ; a particular value (estimate) is denoted by  $\bar{p}$ . Since  $\bar{P}$  is a random variable, we need to discuss its sampling distribution.

### The Expected Value and the Standard Error of the Sample Proportion

We first introduced the population proportion  $p$  in Chapter 5, when we discussed the binomial distribution. It turns out that the sampling distribution of  $\bar{P}$  is closely related to the binomial distribution. Recall that the binomial distribution describes the number of successes  $X$  in  $n$  trials of a Bernoulli process where  $p$  is the probability of success; thus,  $\bar{P} = \frac{X}{n}$  is the number of successes  $X$  divided by the sample size  $n$ . We can derive the **expected value** and the **variance** of the sampling distribution of  $\bar{P}$  as  $E(\bar{P}) = p$  and  $Var(\bar{P}) = \frac{p(1-p)}{n}$ , respectively. (See Appendix 7.1 for the derivations.) Note that since  $E(\bar{P}) = p$ , it implies that  $\bar{P}$  is an unbiased estimator of  $p$ . Analogous to our discussion in the last section, we refer to the standard deviation of the sample proportion as the **standard error of the sample proportion**—that is,  $se(\bar{P}) = \sqrt{\frac{p(1-p)}{n}}$ .

#### EXPECTED VALUE AND STANDARD ERROR OF THE SAMPLE PROPORTION

The **expected value** of the sample proportion  $\bar{P}$  is equal to the population proportion, or,  $E(\bar{P}) = p$ .

The standard deviation of the sample proportion  $\bar{P}$  is referred to as the **standard error of the sample proportion**. It equals  $se(\bar{P}) = \sqrt{\frac{p(1-p)}{n}}$ .

#### EXAMPLE 7.5

Many people apply for jobs to serve as paramedics or firefighters, yet they cannot complete basic physical fitness standards. A recent study found that 77% of all candidates for paramedic and firefighter positions were overweight or obese (*Obesity*, March 19, 2009).

- What are the expected value and the standard error of the sample proportion derived from a random sample of 100 candidates for paramedic or firefighter positions?
- What are the expected value and the standard error of the sample proportion derived from a random sample of 200 candidates for paramedic or firefighter positions?
- Comment on the value of the standard error as the sample size gets larger.

**SOLUTION:** Given that  $p = 0.77$ , we can derive the expected value and the standard error of  $\bar{P}$  as follows.

- a. With  $n = 100$ ,  $E(\bar{P}) = 0.77$  and  $se(\bar{P}) = \sqrt{\frac{0.77(1-0.77)}{100}} = 0.042$ .
- b. With  $n = 200$ ,  $E(\bar{P}) = 0.77$  and  $se(\bar{P}) = \sqrt{\frac{0.77(1-0.77)}{200}} = 0.030$ .
- c. As in the case of the sample mean, while the expected value of the sample proportion is unaffected by the sample size, the standard error of the sample proportion is reduced as the sample size increases.

In this text, we make statistical inferences about the population proportion only when the sampling distribution of  $\bar{P}$  is approximately normal. From the CLT stated in Section 7.2, we can conclude that  $\bar{P}$  is approximately normally distributed when the sample size is sufficiently large.

#### THE CENTRAL LIMIT THEOREM FOR THE SAMPLE PROPORTION

For any population proportion  $p$ , the sampling distribution of  $\bar{P}$  is **approximately normal if the sample size  $n$  is sufficiently large**. As a general guideline, the normal distribution approximation is justified when  $np \geq 5$  and  $n(1-p) \geq 5$ .

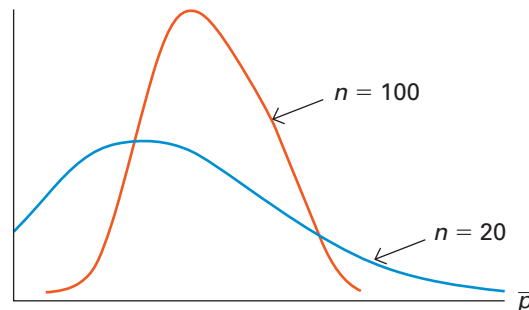
If  $\bar{P}$  is normal, we can transform it into the **standard normal random variable** as

$$Z = \frac{\bar{P} - E(\bar{P})}{se(\bar{P})} = \frac{\bar{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Therefore, any value  $\bar{p}$  has a corresponding value  $z$  given by

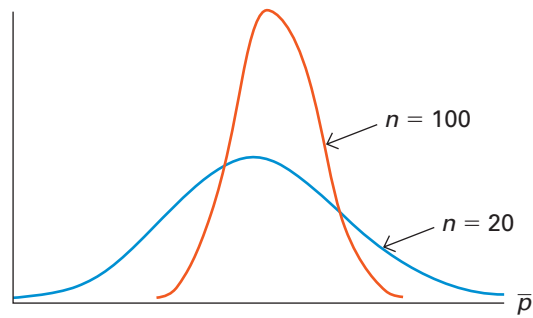
$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

According to the CLT, the sampling distribution of  $\bar{P}$  approaches the normal distribution as the sample size increases. However, as the population proportion deviates from  $p = 0.50$ , we need a larger sample size for the approximation. We illustrate these results by generating the sampling distribution of  $\bar{P}$  from repeated draws from a population with various values of the population proportion and sample sizes. As in the case of  $\bar{X}$ , we use the relative frequency polygon to represent the distribution of  $\bar{P}$ . The simulated sampling distribution of  $\bar{P}$  is based on the population proportion  $p = 0.10$  (Figure 7.6) and  $p = 0.30$  (Figure 7.7).



**FIGURE 7.6** Sampling distribution of  $\bar{P}$  when the population proportion is  $p = 0.10$

**FIGURE 7.7**  
Sampling distribution of  $\bar{P}$  when the population proportion is  $p = 0.30$



When  $p = 0.10$ , the sampling distribution of  $\bar{P}$  does not resemble the shape of the normal distribution with  $n = 20$  since the approximation condition  $np \geq 5$  and  $n(1 - p) \geq 5$  is not satisfied. However, the curve becomes close to normal with  $n = 100$ . When  $p = 0.30$ , the shape of the sampling distribution of  $\bar{P}$  is approximately normal since the approximation condition is satisfied with both sample sizes. In empirical work, it is common to work with large survey data, and as a result, the normal distribution approximation is justified.

### EXAMPLE 7.6

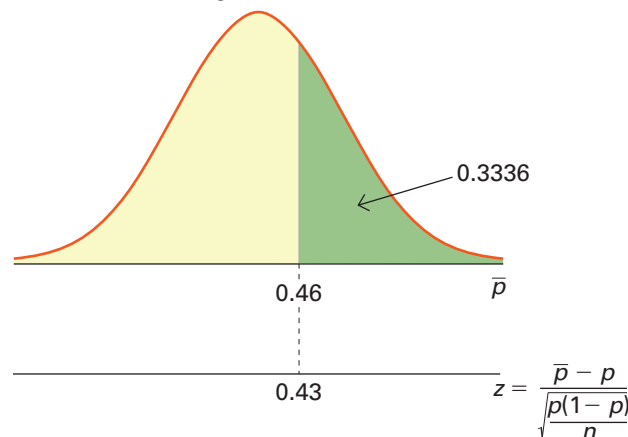
Consider the information presented in the introductory case of this chapter. Recall that Anne Jones wants to determine if the marketing campaign has had a lingering effect on the proportion of customers who are women and teenage girls. Prior to the campaign, 43% of the customers were women and 21% were teenage girls. Based on a random sample of 50 customers after the campaign, these proportions increase to 46% for women and 34% for teenage girls. Anne has the following questions.

- If Starbucks chose not to pursue the marketing campaign, how likely is it that 46% or more of iced-coffee customers are women?
- If Starbucks chose not to pursue the marketing campaign, how likely is it that 34% or more of iced-coffee customers are teenage girls?

**SOLUTION:** If Starbucks had not pursued the marketing campaign, the proportion of customers would still be  $p = 0.43$  for women and  $p = 0.21$  for teenage girls. With  $n = 50$ , the normal approximation for the sample proportion is justified for both population proportions.

- As shown in Figure 7.8, we find that  $P(\bar{P} \geq 0.46) = P\left(Z \geq \frac{0.46 - 0.43}{\sqrt{\frac{0.43(1 - 0.43)}{50}}}\right) = P(Z \geq 0.43) = 1 - 0.6664 = 0.3336$ .

**FIGURE 7.8** Finding  $P(\bar{P} \geq 0.46)$

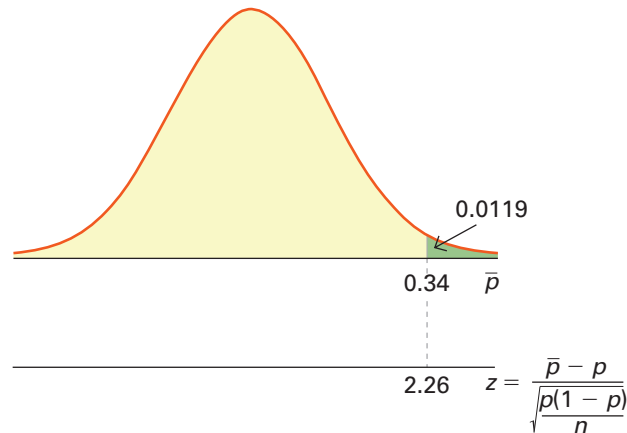




With a chance of 33.36%, it is quite plausible that the proportion of iced coffee purchased by women is at least 0.46 even if Starbucks did not pursue the marketing campaign.

- b. As shown in Figure 7.9, we find  $P(\bar{P} \geq 0.34) = P\left(Z \geq \frac{0.34 - 0.21}{\sqrt{\frac{0.21(1 - 0.21)}{50}}}\right) = P(Z \geq 2.26) = 1 - 0.9881 = 0.0119$ .

**FIGURE 7.9** Finding  $P(\bar{P} \geq 0.34)$



With only a 1.19% chance, it is unlikely that the proportion of iced coffee purchased by teenage girls is at least 0.34 if Starbucks did not pursue the marketing campaign.

Therefore, Anne can use this sample information to infer that the increase in the proportion of iced-coffee sales to women may not necessarily be due to the marketing campaign. However, the marketing campaign may have been successful in increasing the proportion of iced-coffee sales to teenage girls.

## SYNOPSIS OF INTRODUCTORY CASE

Iced coffee, traditionally a warm-weather and warm-region drink, has broadened its appeal over the years. According to a May 13, 2010 report in *Bloomberg Businessweek*, the number of servings of iced coffee surged from 300 million in 2001 to 1.2 billion in 2009. Large corporations have taken notice and have engaged in various strategies to capitalize on the growing trend. Starbucks, for instance, recently promoted a happy hour where customers paid half-price for a Frappuccino beverage between 3:00 pm and 5:00 pm for a 10-day period in May. One month after the marketing period ended, Anne Jones, the manager at a local Starbucks, surveys 50 of her customers. She reports an increase in spending in the sample, as well as an increase in the proportion of customers who are women and teenage girls. Anne wants to determine if the increase is due to chance or due to the marketing campaign. Based on an analysis with probabilities, Anne finds that higher spending in a sample of 50 customers is plausible even if Starbucks had not pursued the marketing campaign. Using a similar analysis with proportions, she infers that while the marketing campaign may not have necessarily increased the proportion of women customers, it seems to have attracted more teenage girls. The findings are consistent with current market research, which has shown that teenage girls have substantial income of their own to spend and often purchase items that are perceived as indulgences.



## EXERCISES 7.3

### Mechanics

21. Consider a population proportion  $p = 0.68$ .
  - a. Calculate the expected value and the standard error of  $\bar{P}$  with  $n = 20$ . Is it appropriate to use the normal distribution approximation for  $\bar{P}$ ? Explain.
  - b. Calculate the expected value and the standard error of  $\bar{P}$  with  $n = 50$ . Is it appropriate to use the normal distribution approximation for  $\bar{P}$ ? Explain.
22. Consider a population proportion  $p = 0.12$ .
  - a. Discuss the sampling distribution of the sample proportion with  $n = 20$  and  $n = 50$ .
  - b. Can you use the normal approximation to calculate the probability that the sample proportion is between 0.10 and 0.12 for both sample sizes?
  - c. Report the probabilities if you answered yes to the previous question.
23. A random sample of size  $n = 200$  is taken from a population with population proportion  $p = 0.75$ .
  - a. Calculate the expected value and the standard error for the sampling distribution of the sample proportion.
  - b. What is the probability that the sample proportion is between 0.70 and 0.80?
  - c. What is the probability that the sample proportion is less than 0.70?
- b. What is the probability that the sample proportion is less than 0.80?
- c. What is the probability that the sample proportion is within  $\pm 0.02$  of the population proportion?
26. According to a recent FCC survey, one in six cell phone users has experienced “bill shock” from unexpectedly high cell phone bills (*Tech Daily Dose*, May 26, 2010).
  - a. Discuss the sampling distribution of the sample proportion based on a sample of 200 cell phone users. Is it appropriate to use the normal distribution approximation for the sample proportion?
  - b. What is the probability that more than 20% of cell phone users in the sample have experienced “bill shock”?
27. A car manufacturer is concerned about poor customer satisfaction at one of its dealerships. The management decides to evaluate the satisfaction surveys of its next 40 customers. The dealer will be fined if the number of customers who report favorably is between 22 and 26. The dealership will be dissolved if fewer than 22 customers report favorably. It is known that 70% of the dealer’s customers report favorably on satisfaction surveys.
  - a. What is the probability that the dealer will be fined?
  - b. What is the probability that the dealership will be dissolved?

### Applications

24. Europeans are increasingly upset at their leaders for making deep budget cuts to many social programs that are becoming too expensive to sustain. For instance, the popularity of then President Nicolas Sarkozy of France plummeted in 2010, giving him an approval rating of just 26% (*The Wall Street Journal*, July 2, 2010).
  - a. What is the probability that fewer than 60 of 200 French people gave President Sarkozy a favorable rating?
  - b. What is the probability that more than 150 of 200 French people gave President Sarkozy an unfavorable rating?
25. A recent study by Allstate Insurance Co. finds that 82% of teenagers have used cell phones while driving (*The Wall Street Journal*, May 5, 2010). Suppose a random sample of 100 teen drivers is taken.
  - a. Discuss the sampling distribution of the sample proportion.
28. At a new exhibit in the Museum of Science, people are asked to choose between 50 or 100 random draws from a machine. The machine is known to have 60 green balls and 40 red balls. After each draw, the color of the ball is noted and the ball is put back for the next draw. You win a prize if more than 70% of the draws result in a green ball. Would you choose 50 or 100 draws for the game? Explain.
29. After years of rapid growth, illegal immigration into the United States has declined, perhaps owing to the recession and increased border enforcement by the United States (*Los Angeles Times*, September 1, 2010). While its share has declined, California still accounts for 23% of the nation’s estimated 11.1 million undocumented immigrants.
  - a. In a sample of 50 illegal immigrants, what is the probability that more than 20% live in California?
  - b. In a sample of 200 illegal immigrants, what is the probability that more than 20% live in California?
  - c. Comment on the reason for the difference between the computed probabilities in parts a and b.

## 7.4 THE FINITE POPULATION CORRECTION FACTOR

### LO 7.6

Use a finite population correction factor.

One of the implicit assumptions we have made thus far is that the sample size  $n$  is much smaller than the population size  $N$ . In many applications, the size of the population is not even known. For instance, we do not have information on the total number of pizzas made

at a local pizza chain in Cambria (Examples 7.2 and 7.3) or the total number of customers at the local Starbucks store (Examples 7.4 and 7.6). If the population size is known and is relatively small (finite), then it is preferable to use a correction factor in the standard error of the estimators, which accounts for the added precision gained by sampling a larger percentage of the population. As a general guideline, we use the finite population correction factor  $\sqrt{\frac{N-n}{N-1}}$  when the sample constitutes at least 5% of the population—that is,  $n \geq 0.05N$ .

#### THE FINITE POPULATION CORRECTION FACTOR FOR THE SAMPLE MEAN

We use the **finite population correction factor** to reduce the sampling variation of the sample mean  $\bar{X}$ . The resulting standard error of  $\bar{X}$  is  $se(\bar{X}) = \frac{\sigma}{\sqrt{n}} \left( \sqrt{\frac{N-n}{N-1}} \right)$ . The transformation for any value  $\bar{x}$  to its corresponding  $z$  value is made accordingly.

Note that the correction factor is always less than one; when  $N$  is large relative to  $n$ , the correction factor is close to one and the difference between the formulas with and without the correction is negligible.

#### EXAMPLE 7.7

A large introductory marketing class has 340 students. The class is divided up into groups for the final course project. Connie is in a group of 34 students. These students had averaged 72 on the midterm, when the class as a whole had an average score of 73 with a standard deviation of 10.

- Calculate the expected value and the standard error of the sample mean based on a random sample of 34 students.
- How likely is it that a random sample of 34 students will average 72 or lower?

**SOLUTION:** The population mean is  $\mu = 73$  and the population standard deviation is  $\sigma = 10$ .

- The expected value of the sample mean is  $E(\bar{X}) = \mu = 73$ . We use the finite population correction factor because the sample size  $n = 34$  is more than 5% of the population size  $N = 340$ . Therefore, the standard error of the sample mean is  $se(\bar{X}) = \frac{\sigma}{\sqrt{n}} \left( \sqrt{\frac{N-n}{N-1}} \right) = \frac{10}{\sqrt{34}} \left( \sqrt{\frac{340-34}{340-1}} \right) = 1.63$ .  
Note that without the correction factor, the standard error would be higher at  $se(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{34}} = 1.71$ .

- We find  $P(\bar{X} \leq 72) = P\left(Z \leq \frac{72-73}{1.63}\right) = P(Z \leq -0.61) = 0.2709$ .  
That is, the likelihood of 34 students averaging 72 or lower is 27.09%.

We can use a similar finite population correction factor for a sample proportion when the sample size is at least 5% of the population size.

#### THE FINITE POPULATION CORRECTION FACTOR FOR THE SAMPLE PROPORTION

We use the **finite population correction factor** to reduce the sampling variation of the sample proportion  $\bar{P}$ . The resulting standard error of  $\bar{P}$  is  $se(\bar{P}) = \sqrt{\frac{p(1-p)}{n}} \left( \sqrt{\frac{N-n}{N-1}} \right)$ . The transformation for any value  $\bar{p}$  to its corresponding  $z$  value is made accordingly.

### EXAMPLE 7.8

The home ownership rate during 2009 declined to approximately 67% becoming comparable to the rate in early 2000 (*U.S. Census Bureau News*, February 2, 2010). A random sample of 80 households is taken from a small island community with 1,000 households. The home ownership rate on the island is equivalent to the national home ownership rate of 67%.

- a. Calculate the expected value and the standard error for the sampling distribution of the sample proportion. Is it necessary to apply the finite population correction factor? Explain.
- b. What is the probability that the sample proportion is within 0.02 of the population proportion?

#### SOLUTION:

- a. We must apply the finite population correction factor because the sample size  $n = 80$  is at least 5% of the population size  $N = 1,000$ . Therefore,  $E(\bar{P}) = p = 0.67$  and

$$se(\bar{P}) = \sqrt{\frac{p(1-p)}{n}} \left( \sqrt{\frac{N-n}{N-1}} \right) = \sqrt{\frac{0.67(1-0.67)}{80}} \left( \sqrt{\frac{1,000-80}{1,000-1}} \right) = 0.0505.$$

- b. The probability that the sample proportion is within 0.02 of the population proportion is  $P(0.65 \leq \bar{P} \leq 0.69)$ . We find that  $P(0.65 \leq \bar{P} \leq 0.69) = P\left(\frac{0.65-0.67}{0.0505} \leq Z \leq \frac{0.69-0.67}{0.0505}\right) = P(-0.40 \leq Z \leq 0.40) = 0.6554 - 0.3446 = 0.3108$ . The likelihood that the home ownership rate is within 0.02 of the population proportion is 31.08%.

## EXERCISES 7.4

### Mechanics

30. A random sample of size  $n = 100$  is taken from a population of size  $N = 2,500$  with mean  $\mu = -45$  and variance  $\sigma^2 = 81$ .
  - a. Is it necessary to apply the finite population correction factor? Explain. Calculate the expected value and the standard error of the sample mean.
  - b. What is the probability that the sample mean is between  $-47$  and  $-43$ ?
  - c. What is the probability that the sample mean is greater than  $-44$ ?
31. A random sample of size  $n = 70$  is taken from a finite population of size  $N = 500$  with mean  $\mu = 220$  and variance  $\sigma^2 = 324$ .
  - a. Is it necessary to apply the finite population correction factor? Explain. Calculate the expected value and the standard error of the sample mean.
  - b. What is the probability that the sample mean is less than 210?
  - c. What is the probability that the sample mean lies between 215 and 230?
32. A random sample of size  $n = 100$  is taken from a population of size  $N = 3,000$  with a population proportion of  $p = 0.34$ .
  - a. Is it necessary to apply the finite population correction factor? Explain. Calculate the expected value and the standard error of the sample proportion.
  - b. What is the probability that the sample proportion is greater than 0.37?
33. A random sample of size  $n = 80$  is taken from a population of size  $N = 600$  with a population proportion  $p = 0.46$ .
  - a. Is it necessary to apply the finite population correction factor? Explain. Calculate the expected value and the standard error of the sample proportion.
  - b. What is the probability that the sample mean is less than 0.40?

### Applications

34. A recent study finds that companies are setting aside a large chunk of their IT spending for green technology projects (*BusinessWeek*, March 5, 2009). Two out of three of the large companies surveyed by Deloitte said they have at least 5% of their IT budget earmarked for green IT projects. Suppose that the survey was based on 1,000 large companies. What is the probability that more

than 75 of 120 large companies will have at least 5% of their IT expenditure earmarked for green IT projects?

35. The issues surrounding the levels and structure of executive compensation have gained added prominence in the wake of the financial crisis that erupted in the fall of 2008. Based on the 2006 compensation data obtained from the Securities and Exchange Commission (SEC) website, it was determined that the mean and the standard deviation of compensation for the 500 highest paid CEOs in publicly traded U.S. companies are \$10.32 million and \$9.78 million, respectively. An analyst randomly chooses 32 CEO compensations for 2006.
- Is it necessary to apply the finite population correction factor? Explain.
  - Is the sampling distribution of the sample mean approximately normally distributed? Explain.
  - Calculate the expected value and the standard error of the sample mean.
  - What is the probability that the sample mean is more than \$12 million?
36. Suppose in the previous question that the analyst had randomly chosen 12 CEO compensations for 2006.
- Is it necessary to apply the finite population correction factor? Explain.
  - Is the sampling distribution of the sample mean approximately normally distributed? Explain.
  - Calculate the expected value and the standard error of the sample mean.
  - Can you use the normal approximation to calculate the probability that the sample mean is more than \$12 million? Explain.
37. Given the recent economic downturn, only 60% in a graduating class of 250 will find employment in the first round of a job search. You have 20 friends who have recently graduated.
- Discuss the sampling distribution of the sample proportion of your friends who will find employment in the first round of a job search.
  - What is the probability that less than 50% of your friends will find employment in the first round of a job search?

## 7.5 STATISTICAL QUALITY CONTROL

LO 7.7

Now more than ever, a successful firm must focus on the quality of the products and services it offers. Global competition, technological advances, and consumer expectations are all factors contributing to the quest for quality. In order to ensure the production of high-quality goods and services, a successful firm implements some form of quality control. In this section, we give a brief overview of the field of **statistical quality control**.

Construct and interpret control charts for quantitative and qualitative data.

**Statistical quality control** involves statistical techniques used to develop and maintain a firm's ability to produce high-quality goods and services.

In general, two approaches are used for statistical quality control. A firm uses **acceptance sampling** if it produces a product (or offers a service) and at the completion of the production process, the firm then inspects a portion of the products. If a particular product does not conform to certain specifications, then it is either discarded or repaired. There are several problems with this approach to quality control. First, it is costly to discard or repair a product. Second, the detection of all defective products is not guaranteed. Defective products may be delivered to customers, thus damaging the firm's reputation.

A preferred approach to quality control is the **detection approach**. A firm using the detection approach inspects the production process and determines at which point the production process does not conform to specifications. The goal is to determine whether the production process should be continued or adjusted before a large number of defects are produced. In this section, we focus on the detection approach to quality control.

In general, no two products or services are identical. In any production process, variation in the quality of the end product is inevitable. Two types of variation occur. **Chance variation** is caused by a number of randomly occurring events that are part of the production process. This type of variation is not generally considered under the control

of the individual worker or machine. For example, suppose a machine fills one-gallon jugs of milk. It is unlikely that the filling weight of each jug is exactly 128 ounces. Very slight differences in the production process lead to minor differences in the weights of one jug to the next. Chance variation is expected and is not a source of alarm in the production process so long as its magnitude is tolerable and the end product meets acceptable specifications.

The other source of variation is referred to as **assignable variation**. This type of variation in the production process is caused by specific events or factors that can usually be identified and eliminated. Suppose in the milk example that the machine is “drifting” out of alignment. This causes the machine to overfill each jug—a costly expense for the firm. Similarly, it is bad for the firm in terms of its reputation, if the machine begins to underfill each jug. The firm wants to identify and correct these types of variations in the production process.

## Control Charts

Walter A. Shewhart, a researcher at Bell Telephone Laboratories during the 1920s, is often credited as being the first to apply statistics to improve the quality of output. He developed the **control chart**—a tool used to monitor the behavior of a production process.

### THE CONTROL CHART

The most commonly used statistical tool in quality control is the **control chart**, a plot of calculated statistics of the production process over time. If the calculated statistics fall in an expected range, then the production process is in control. If the calculated statistics reveal an undesirable trend, then adjustment of the production process is likely necessary.

We can construct a number of different control charts where each differs by either the variable of interest and/or the type of data that are available. For quantitative data, examples of control charts include

- The  $\bar{x}$  **chart**, which monitors the *central tendency* of a production process, and
- The **R chart** and the **s chart**, which monitor the *variability* of a production process.

For qualitative data, examples of control charts include

- The  $\bar{p}$  **chart**, which monitors the *proportion* of defectives (or some other characteristic) in a production process,
- The **c chart**, which monitors the *count* of defects per item, such as the number of blemishes on a sampled piece of furniture.

In general, all of these control charts (and others that we have not mentioned) have the following characteristics:

1. A control chart plots the sample estimates, such as  $\bar{x}$  or  $\bar{p}$ . So as more and more samples are taken, the resulting control chart provides one type of safeguard when assessing if the production process is operating within predetermined guidelines.
2. All sample estimates are plotted with reference to a **centerline**. The centerline represents the variable’s expected value when the production process is in control.
3. In addition to the centerline, all control charts include an **upper control limit** and a **lower control limit**. These limits indicate excessive deviation above (upper control limit) or below (lower control limit) the expected value of the variable of interest. A control chart is valid only if the sampling distribution of the relevant estimator is (approximately) normal. Under this assumption, the control limits are generally set at three standard deviations from the centerline. As we observed in Chapter 6, the



area under the normal curve that corresponds to  $\pm 3$  standard deviations from the expected value is 0.9973. Thus, there is only a  $1 - 0.9973 = 0.0027$  chance that the sample estimates will fall outside the limit boundaries. In general, we define the upper and lower control limits as follows:

Upper Control Limit (UCL): Expected Value +  $(3 \times \text{Standard Error})$

Lower Control Limit (LCL): Expected Value –  $(3 \times \text{Standard Error})$

If the sample estimates fall randomly within the upper and lower control limits, then the production process is deemed in control. Any sample estimate that falls above the upper control limit or below the lower control limit is considered evidence that the production process is out of control and should be adjusted. In addition, any type of patterns within the control limits may suggest possible problems with the process. One indication of a process that is potentially heading out of control is unusually long runs above or below the centerline. Another possible problem is any evidence of a trend within the control limits.

In the next example, we focus on quantitative data and illustrate the  $\bar{x}$  chart. We then turn to qualitative data and construct the  $\bar{p}$  chart.

### EXAMPLE 7.9

A firm that produces one-gallon jugs of milk wants to ensure that the machine is operating properly. Every two hours, the company samples 25 jugs and calculates the following sample mean filling weights (in ounces):

128.7	128.4	128.0	127.8	127.5	126.9
-------	-------	-------	-------	-------	-------

Assume that when the machine is operating properly,  $\mu = 128$  and  $\sigma = 2$ , and that filling weights follow the normal distribution. Can the firm conclude that the machine is operating properly? Should the firm have any concerns with respect to this machine?

**SOLUTION:** Here the firm is interested in monitoring the population mean. To answer these questions, we construct an  $\bar{x}$  chart. As mentioned earlier, this chart relies on the normal distribution for the sampling distribution of the estimator  $\bar{X}$ . Recall that if we are sampling from a normal population, then  $\bar{X}$  is normally distributed even for small sample sizes. In this example, we are told that filling weights follow the normal distribution, a common assumption in the literature on quality control.

For the  $\bar{x}$  chart, the centerline is the mean when the process is in control. Here, we are given that  $\mu = 128$ . We then calculate the upper and lower control limits as plus and minus three standard deviations from the mean:

$$\text{Upper Control Limit, UCL: } \mu + 3 \frac{\sigma}{\sqrt{n}} = 128 + 3 \frac{2}{\sqrt{25}} = 129.2$$

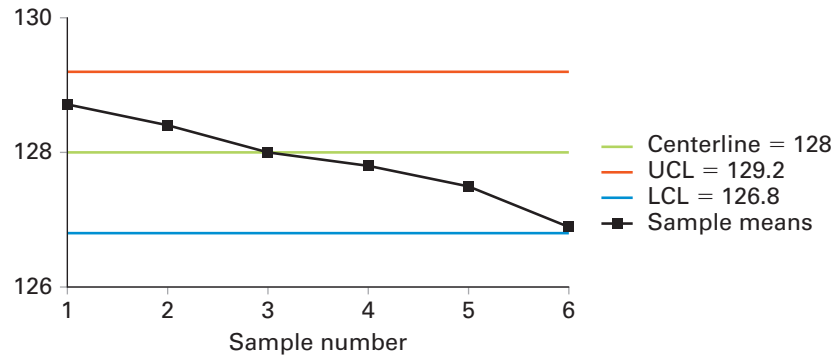
$$\text{Lower Control Limit, LCL: } \mu - 3 \frac{\sigma}{\sqrt{n}} = 128 - 3 \frac{2}{\sqrt{25}} = 126.8$$

Figure 7.10 shows the centerline and the control limits as well as the sample means for Example 7.9.

All of the sample means fall within the upper control and the lower control limits, which indicates, at least initially, that the production process is in control. However, the sample means should be randomly spread between these limits; there should be no pattern. In this example, there is clearly a downward trend in the sample means. It appears as though the machine is beginning to underfill the one-gallon jugs. So even though none of the sample means lies beyond the control limits, the production process is likely veering out of control and the firm would be wise to inspect the machine sooner rather than later.

**FIGURE 7.10**

Mean chart for  
milk production  
process



A firm may be interested in the stability of the proportion of its goods or services possessing a certain attribute or characteristic. For example, most firms strive to produce high-quality goods (or services) and thus hope to keep the proportion of defects at a minimum. When a production process is to be assessed based on sample proportions—here, the proportion of defects—then a  $\bar{p}$  chart proves quite useful. Since the primary purpose of the  $\bar{p}$  chart is to track the proportion of defects in a production process, it is also referred to as a fraction defective chart or a percent defective chart. Consider the next example.

### EXAMPLE 7.10

A production process has a 5% defective rate. A quality inspector takes 6 samples of  $n = 500$ . The following sample proportions are obtained:

0.065	0.075	0.082	0.086	0.090	0.092
-------	-------	-------	-------	-------	-------

- Construct a  $\bar{p}$  chart. Plot the sample proportions on the  $\bar{p}$  chart.
- Is the production process in control? Explain.

#### SOLUTION:

- The  $\bar{p}$  chart relies on the central limit theorem for the normal approximation for the sampling distribution of the estimator  $\bar{P}$ . Recall that so long as  $np$  and  $n(1 - p)$  are greater than or equal to five, then the sampling distribution of  $\bar{P}$  is approximately normally distributed. This condition is satisfied in Example 7.10. Since the expected proportion of defects is equal to 0.05, we set the centerline at  $p = 0.05$ . We then calculate the upper control limit and lower control limit as follows:

$$\text{UCL: } p + 3\sqrt{\frac{p(1-p)}{n}} = 0.05 + 3\sqrt{\frac{0.05(1-0.05)}{500}} = 0.079$$

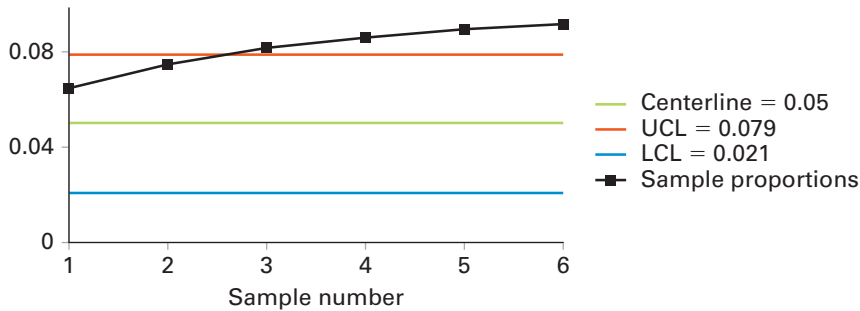
$$\text{LCL: } p - 3\sqrt{\frac{p(1-p)}{n}} = 0.05 - 3\sqrt{\frac{0.05(1-0.05)}{500}} = 0.021$$

We note that if UCL is a value greater than one, then we reset UCL to one in the control chart. Similarly, if the LCL is a negative value, we reset LCL to zero in the control chart.

Plotting the values for the centerline, UCL, and LCL, as well as the sample proportions, yields Figure 7.11.

- Four of the most recent sample proportions fall above the upper control limit. This provides evidence that the process is out of control and needs adjustment.

**FIGURE 7.11** Proportion of defects



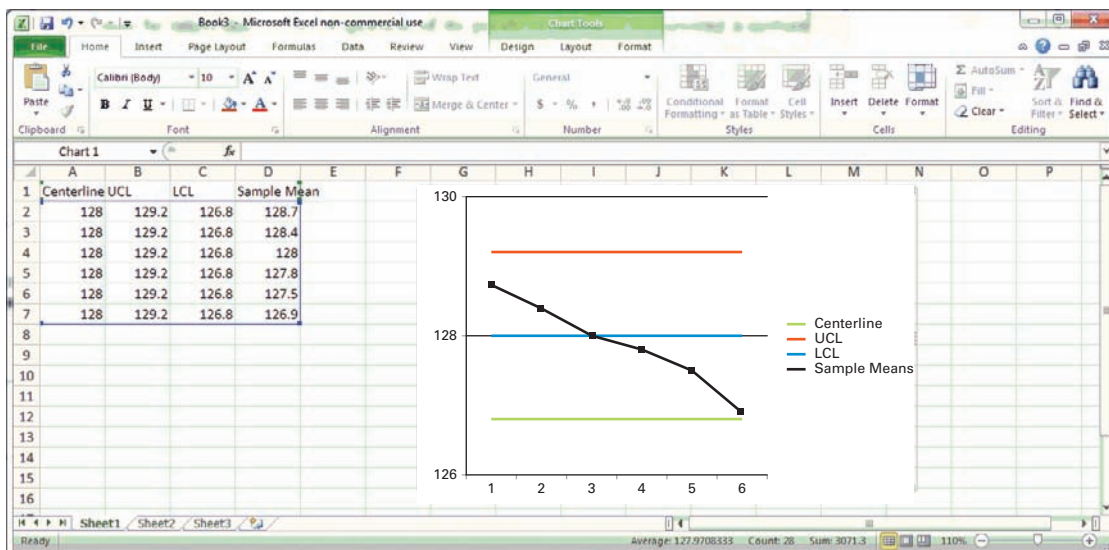
## Using Excel to Create a Control Chart

Even though Excel does not have a built-in function to create a control chart, it is still relatively easy to construct one. The added step when using Excel is that if we are not given values for the centerline, UCL, LCL, and the sample means, then we have to calculate these values first—other software packages do these calculations for us. We will replicate Figure 7.10 using the values that we have calculated (or were given) in Example 7.9.

- Enter Headings for the Centerline, UCL, LCL, and Sample Mean as shown in the first row of the Excel spreadsheet in Figure 7.12.
- Enter the relevant values under each of the headings. For columns with many repeated values (Centerline, UCL, and LCL), it is useful to select the respective value, drag it down a certain number of cells, and then from the menu choose **Home > Fill > Down**. For instance, for the Centerline value of 128, select 128, drag the cursor down five more cells (since we want it repeated six times), and choose **Home > Fill > Down**.
- After all the data have been entered into the spreadsheet, select all the data with the headings and choose **Insert > Line > 2-D Line** (choose the option on the top left). Figure 7.12 shows the embedded control chart.
- Formatting regarding colors, axes, grids, etc. can be done by selecting **Layout** from the menu.

In order to construct a  $\bar{p}$  chart using Excel, you would follow the same steps as those outlined above for the  $\bar{x}$  chart.

**FIGURE 7.12** Using Excel to create a control chart



## EXERCISES 7.5

### Mechanics

38. Consider a normally distributed population with mean  $\mu = 80$  and standard deviation  $\sigma = 14$ .
  - a. Construct the centerline and the upper and lower control limits for the  $\bar{x}$  chart if samples of size 5 are used.
  - b. Repeat the analysis with samples of size 10.
  - c. Discuss the effect of the sample size on the control limits.
39. Random samples of size  $n = 250$  are taken from a population with  $p = 0.04$ .
  - a. Construct the centerline and the upper and lower control limits for the  $\bar{p}$  chart.
  - b. Repeat the analysis with  $n = 150$ .
  - c. Discuss the effect of the sample size on the control limits.
40. Random samples of size  $n = 25$  are taken from a normally distributed population with mean  $\mu = 20$  and standard deviation  $\sigma = 10$ .
  - a. Construct the centerline and the upper and lower control limits for the  $\bar{x}$  chart.
  - b. Suppose six samples of size 25 produced the following sample means: 18, 16, 19, 24, 28, and 30. Plot these values on the  $\bar{x}$  chart.
  - c. Are any points outside the control limits? Does it appear that the process is under control? Explain.
41. Random samples of size  $n = 36$  are taken from a population with mean  $\mu = 150$  and standard deviation  $\sigma = 42$ .
  - a. Construct the centerline and the upper and lower control limits for the  $\bar{x}$  chart.
  - b. Suppose five samples of size 36 produced the following sample means: 133, 142, 150, 165, and 169. Plot these values on the  $\bar{x}$  chart.
  - c. Are any points outside the control limits? Does it appear that the process is under control? Explain.
42. Random samples of size  $n = 500$  are taken from a population with  $p = 0.34$ .
  - a. Construct the centerline and the upper and lower control limits for the  $\bar{p}$  chart.
  - b. Suppose six samples of size 500 produced the following sample proportions: 0.28, 0.30, 0.33, 0.34, 0.37, and 0.39. Plot these values on the  $\bar{p}$  chart.
  - c. Are any points outside the control limits? Does it appear that the process is under control? Explain.
43. Random samples of size  $n = 400$  are taken from a population with  $p = 0.10$ .
  - a. Construct the centerline and the upper and lower control limits for the  $\bar{p}$  chart.

- b. Suppose six samples of size 400 produced the following sample proportions: 0.06, 0.11, 0.09, 0.08, 0.14, and 0.16. Plot these values on the  $\bar{p}$  chart.
- c. Is the production process under control? Explain.

### Applications

44. A production process is designed to fill boxes with an average of 14 ounces of cereal. The population of filling weights is normally distributed with a standard deviation of 2 ounces. Inspectors take periodic samples of 10 boxes. The following sample means are obtained.

13.7	14.2	13.9	14.1	14.3	13.9
------	------	------	------	------	------

- a. Construct an  $\bar{x}$  chart. Plot the sample means on the  $\bar{x}$  chart.
  - b. Can the firm conclude that the production process is operating properly? Explain.
45. Major League Baseball Rule 1.09 states that "the baseball shall weigh not less than 5 or more than 5¼ ounces" (www.mlb.com). Use these values as the lower and the upper control limits, respectively. Assume the centerline equals 5.125 ounces. Periodic samples of 50 baseballs produce the following sample means:

5.05	5.10	5.15	5.20	5.22	5.24
------	------	------	------	------	------

- a. Construct an  $\bar{x}$  chart. Plot the sample means on the  $\bar{x}$  chart.
  - b. Are any points outside the control limits? Does it appear that the process is under control? Explain.
46. **FILE Cricket.** Fast bowling, also known as pace bowling, is an important component of the bowling attack in the sport of cricket. The objective is to bowl at a high speed and make the ball turn in the air and off the ground so that it becomes difficult for the batsman to hit it cleanly. Kalwant Singh is a budding Indian cricketer in a special bowling camp. While his coach is happy with Kalwant's average bowling speed, he feels that Kalwant lacks consistency. He records his bowling speed on the next four overs, where each over consists of six balls.

Over 1	Over 2	Over 3	Over 4
96.8	99.2	88.4	98.4
99.5	100.2	97.8	91.4
88.8	90.1	82.8	85.5
81.9	98.7	91.2	87.6
100.1	96.4	94.2	90.3
96.8	98.8	89.8	85.9

It is fair to assume that Kalwant's bowling speed is normally distributed with a mean and a standard deviation of 94 miles and 2.8 miles per hour, respectively.

- a. Construct the centerline and the upper and lower control limits for the  $\bar{x}$  chart. Plot the average speed of Kalwant's four overs on the  $\bar{x}$  chart.
  - b. Is there any pattern in Kalwant's bowling that justifies his coach's concerns that he is not consistent in bowling? Explain.
47. A manufacturing process produces steel rods in batches of 1,000. The firm believes that the percent of defective items generated by this process is 5%.
- a. Construct the centerline and the upper and lower control limits for the  $\bar{p}$  chart.
  - b. An engineer inspects the next batch of 1,000 steel rods and finds that 6.2% are defective. Is the manufacturing process under control? Explain.
48. A firm produces computer chips for personal computers. From past experience, the firm knows that 4% of the chips are defective. The firm collects a sample of the first 500 chips manufactured at 1:00 pm for the past two weeks. The following sample proportions are obtained:

0.044	0.052	0.060	0.036	0.028	0.042	0.034	0.054	0.048	0.025
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

- a. Construct a  $\bar{p}$  chart. Plot the sample proportions on the  $\bar{p}$  chart.
- b. Can the firm conclude that the process is operating properly?

49. The college admissions office at a local university usually admits 750 students and knows from previous experience that 25% of these students choose not to enroll at the university.
- a. Construct the centerline and the upper and lower control limits for the  $\bar{p}$  chart.
  - b. Assume that this year the university admits 750 students and 240 choose not to enroll at the university. Should the university be concerned? Explain.
50. Following customer complaints about the quality of service, Dell stopped routing corporate customers to a technical support call center in Bangalore, India (*USA TODAY*, November 24, 2003). Suppose Dell's decision to direct customers to call centers outside of India was based on consumer complaints in the last six months. Let the number of complaints per month for 80 randomly selected customers be given below.

Month	Number of Complaints
1	20
2	12
3	24
4	14
5	25
6	22

- a. Construct the centerline and the upper and lower control limits for the  $\bar{p}$  chart if management allows a 15% complaint rate.
- b. Can you justify Dell's decision to direct customers to call centers outside of India?

## WRITING WITH STATISTICS

Barbara Dwyer, the manager at Lux Hotel, makes every effort to ensure that customers attempting to make phone reservations wait an average of only 60 seconds to speak with a reservations specialist. She knows that this is likely to be the customer's first impression of the hotel and she wants the initial interaction to be a positive one. Since the hotel accepts phone reservations 24 hours a day, Barbara wonders if this quality service is consistently maintained throughout the day. She takes six samples of  $n = 4$  calls during each of four shifts over one 24-hour period and records the wait time of each call. A portion of the data, in seconds, is presented in Table 7.1.

Barbara assumes that wait times are normally distributed with a mean and standard deviation of 60 seconds and 30 seconds, respectively. She wants to use the sample information to:

1. Prepare a control chart for wait times.
2. Use the control chart to determine whether quality service is consistently maintained throughout the day.



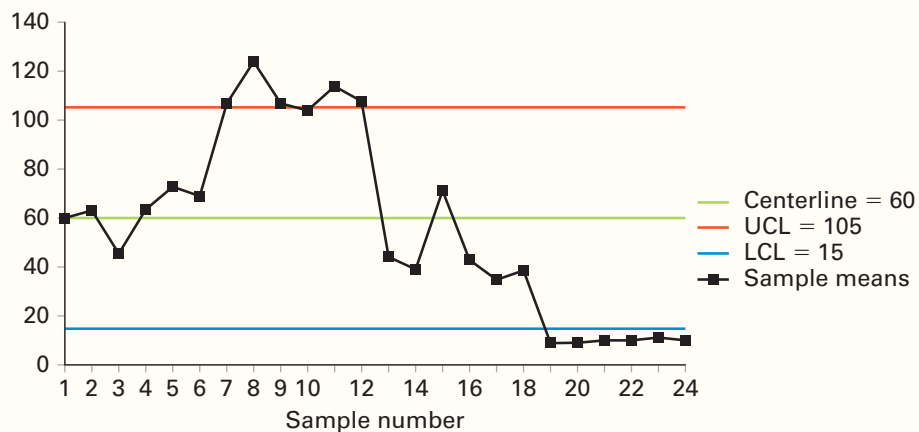
**TABLE 7.1** Wait times for phone reservations

Shift	Sample	Wait Time (in seconds)				Sample Mean, $\bar{x}$
Shift 1: 12:00 am–6:00 am	1	67	48	52	71	60
	2	57	68	60	66	63
	3	37	41	60	41	45
	4	83	59	49	66	64
	5	82	63	64	83	73
	6	87	53	66	69	69
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Shift 4: 6:00 pm–12:00 am	19	6	11	8	9	9
	20	10	8	10	9	9
	21	11	7	14	7	10
	22	8	9	9	12	10
	23	9	12	9	14	11
	24	5	8	15	11	10

## Sample Report— Customer Wait Time

When a potential customer phones Lux Hotel, it is imperative for the reservations specialist to set a tone that relays the high standard of service that the customer will receive if he/she chooses to stay at the Lux. For this reason, management at the Lux strives to minimize the time that elapses before a potential customer speaks with a reservations specialist; however, management also recognizes the need to use its resources wisely. If too many reservations specialists are on duty, then resources are wasted due to idle time; yet if too few reservations specialists are on duty, the result might mean angry first-time customers or, worse, lost customers. In order to ensure customer satisfaction as well as an efficient use of resources, a study is conducted to determine whether a typical customer waits an average of 60 seconds to speak with a reservations specialist. Before data are collected, a control chart is constructed. The upper control limit (UCL) and the lower control limit (LCL) are set three standard deviations from the desired average of 60 seconds. In Figure 7.A, the desired average of 60 seconds is denoted as the centerline and the upper and lower control limits amount to 105 seconds and 15 seconds ( $\mu \pm 3 \frac{\sigma}{\sqrt{n}} = 60 \pm 3 \frac{30}{\sqrt{4}} = 60 \pm 45$ ), respectively. The reservation process is deemed under control if the sample means fall randomly within the upper and lower control limits; otherwise the process is out of control and adjustments should be made.

**FIGURE 7.A** Sample mean wait times





During each of four shifts, six samples of  $n = 4$  calls are randomly selected over one 24-hour period and the average wait time of each sample is recorded. All six sample means from the first shift (1st shift: 12:00 am–6:00 am, sample numbers 1 through 6) fall within the control limits, indicating that the reservation process is in control. However, five sample means from the second shift (2nd shift: 6:00 am–12:00 pm, sample numbers 7 through 12) lie above the upper control limit. Customers calling during the second shift are waiting too long before they speak with a specialist. In terms of quality standards, this is unacceptable from the hotel's perspective. All six sample means from the third shift fall within the control limits (3rd shift: 12:00 pm–6:00 pm, sample numbers 13 through 18), yet all sample means for the fourth shift fall below the lower control limit (4th shift: 6:00 pm–12:00 am, sample numbers 19 through 24). Customers are waiting for very short periods of time to speak with a reservations specialist, but reservations specialists may have too much idle time. Perhaps one solution is to shift some reservations specialists from shift four to shift two.

## CONCEPTUAL REVIEW

### LO 7.1 Explain common sample biases.

A sampling **bias** occurs when the information from a sample is not typical of that in the population in a systematic way. It is often caused by samples that are not representative of the population. **Selection bias** refers to a systematic underrepresentation of certain groups from consideration for the sample. **Nonresponse bias** refers to a systematic difference in preferences between respondents and nonrespondents to a survey or a poll.

### LO 7.2 Describe various sampling methods.

A **simple random sample** is a sample of  $n$  observations that has the same probability of being selected from the population as any other sample of  $n$  observations. Most statistical methods presume simple random samples.

A **stratified random sample** is formed when the population is divided into groups (strata) based on one or more classification criteria. A stratified random sample includes randomly selected observations from each stratum. The number of observations per stratum is proportional to the stratum's size in the population. The data for each stratum are eventually pooled. A **cluster sample** is formed when the population is divided into groups (clusters) based on geographic areas. Whereas a stratified random sample consists of elements from each group, a cluster sample includes observations from randomly selected clusters. Stratified random sampling is preferred when the objective is to **increase precision** and cluster sampling is preferred when the objective is to **reduce costs**.

### LO 7.3 Describe the sampling distribution of the sample mean.

A particular characteristic of a population, such as the mean or the proportion, is called a **parameter**, which is a constant even though its value may be unknown. A **statistic**, such as the sample mean or the sample proportion, is a **random variable** whose value depends on the chosen random sample. When a statistic is used to estimate a parameter, it is referred to as an **estimator**. A particular value of the estimator is called an **estimate**.

Since the statistic  $\bar{X}$  is a random variable, its sampling distribution is the probability distribution of sample means derived from all possible samples of a given size from the population. The **expected value** of the sample mean  $\bar{X}$  equals  $E(\bar{X}) = \mu$  and the standard deviation, commonly referred to as the **standard error**, equals  $se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ . For any sample size, the sampling distribution of  $\bar{X}$  is normal if the **population is normally distributed**.

If  $\bar{X}$  is normally distributed, then any value  $\bar{x}$  can be transformed to its corresponding  $z$  value as:  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ .

**LO 7.4 Explain the importance of the central limit theorem.**

The **central limit theorem (CLT)** is used when the random sample is drawn from a nonnormal population. It states that for any population  $X$  with expected value  $\mu$  and standard deviation  $\sigma$ , the sampling distribution of  $\bar{X}$  is approximately normal if the sample size  $n$  is **sufficiently large**. As a general guideline, the normal distribution approximation is justified when  $n \geq 30$ .

**LO 7.5 Describe the sampling distribution of the sample proportion.**

The **expected value** and the **standard error** of the sample proportion  $\bar{P}$  are  $E(\bar{P}) = p$  and  $se(\bar{P}) = \sqrt{\frac{p(1-p)}{n}}$ . From the CLT, we can conclude that for any population proportion  $p$ , the sampling distribution of  $\bar{P}$  is approximately normal if the sample size  $n$  is **sufficiently large**. As a general guideline, the normal distribution approximation is justified when  $np \geq 5$  and  $n(1-p) \geq 5$ . If  $\bar{P}$  is normally distributed, then any value  $\bar{p}$  can be transformed to its corresponding  $z$  value as:  $z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ .

**LO 7.6 Use a finite population correction factor.**

If the population size is relatively small (finite) and its value is known, then it is preferable to use the correction factor in the standard error of the estimators. As a general guideline, we use the finite correction factor when the sample constitutes at least 5% of the population—that is,  $n \geq 0.05N$ . With the correction factor,  $se(\bar{X}) = \frac{\sigma}{\sqrt{n}} \left( \sqrt{\frac{N-n}{N-1}} \right)$  and  $se(\bar{P}) = \sqrt{\frac{p(1-p)}{n}} \left( \sqrt{\frac{N-n}{N-1}} \right)$ . The transformation to the corresponding  $z$  value is made accordingly.

**LO 7.7 Construct and interpret control charts for quantitative and qualitative data.**

**Statistical quality control** involves statistical techniques used to develop and maintain a firm's ability to produce high-quality goods and services. The most commonly used statistical tool in quality control is the **control chart**. A control chart specifies a centerline as well as an upper control limit (UCL) and a lower control limit (LCL). In general, the UCL and the LCL are set within three standard deviations of the centerline.

The upper and lower control limits for the  $\bar{x}$  **chart** are defined as  $\mu + 3 \frac{\sigma}{\sqrt{n}}$  and  $\mu - 3 \frac{\sigma}{\sqrt{n}}$ , respectively. For the  $\bar{p}$  **chart**, these limits are defined as  $p + 3 \sqrt{\frac{p(1-p)}{n}}$  and  $p - 3 \sqrt{\frac{p(1-p)}{n}}$ , respectively. In general, if the sample means or the sample proportions fall within the control limits, then the process is under control; otherwise it is out of control and adjustment is necessary. However, even if these sample estimates fall within the control limits, they must be randomly spread between the limits. If there is a trend or unusually long runs above or below the centerline, then the process may be veering out of control.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

51. A seminal study conducted by scientists at the University of Illinois found evidence of improved memory and reasoning for those who took three

vigorous 40-minute walks a week over six months (*Newsweek*, June 28–July 5, 2010). As an assistant manager working for a public health institute based in Florida, you would like to estimate the

- proportion of adults in Miami, Florida, who follow such a walking regimen. Discuss the sampling bias in the following strategies where people are asked if they walk regularly:
- Randomly selected adult beachgoers in Miami.
  - Randomly selected Miami residents who are requested to disclose the information in prepaid envelopes.
  - Randomly selected Miami residents who are requested to disclose the information on the firm's website.
  - Randomly selected adult patients at all hospitals in Miami.
52. In the previous question regarding walking regimens of the residents of Miami, explain how you can obtain a representative sample based on the following sampling strategies:
- Simple random sampling.
  - Stratified random sampling.
  - Cluster sampling.
53. According to the Bureau of Labor Statistics, it takes an average of 22 weeks for someone over 55 to find a new job, compared with 16 weeks for younger workers (*The Wall Street Journal*, September 2, 2008). Assume that the probability distributions are normal and that the standard deviation is 2 weeks for both distributions.
- What is the probability that 8 workers over the age of 55 take an average of more than 20 weeks to find a job?
  - What is the probability that 20 younger workers average less than 15 weeks to find a job?
54. Presidential job approval is the most-watched statistic in American politics. According to the June 2010 NBC/*Wall Street Journal* public opinion poll, president Barack Obama had reached his lowest approval rating since taking office in January of 2009. The poll showed that 48% of people disapproved of the job Obama was doing as president of the United States, while only 45% approved. Experts attributed the drop in approval ratings to a poor economy and the government's reaction to the massive oil spill in the Gulf of Mexico. Use the June 2010 approval and disapproval ratings to answer the following questions.
- What is the probability that President Obama gets a majority support in a random sample of 50 Americans?
  - What is the probability that President Obama gets a majority disapproval in a random sample of 50 Americans?
55. While starting salaries have fallen for college graduates in many of the top hiring fields, there is some good news for business undergraduates with concentrations in accounting and finance (*Bloomberg Businessweek*, July 1, 2010). According to the National Association of Colleges and Employers' Summer 2010 Salary Survey, accounting graduates commanded the second highest salary at \$50,402, followed by finance graduates at \$49,703. Let the standard deviation for accounting and finance graduates be \$6,000 and \$10,000, respectively.
- What is the probability that 100 randomly selected accounting graduates will average more than \$52,000 in salary?
  - What is the probability that 100 randomly selected finance graduates will average more than \$52,000 in salary?
  - Comment on the above probabilities.
56. An automatic machine in a manufacturing process is operating properly if the length of an important subcomponent is normally distributed with a mean  $\mu = 80$  cm and a standard deviation  $\sigma = 2$  cm.
- Find the probability that the length of one randomly selected unit is less than 79 cm.
  - Find the probability that the average length of 10 randomly selected units is less than 79 cm.
  - Find the probability that the average length of 30 randomly selected units is less than 79 cm.
57. Trader Joe's is a privately held chain of specialty grocery stores in the United States. Starting out as a small chain of convenience stores, it has expanded to over 340 stores as of June 2010 (<http://Traderjoe.com>). It has developed a reputation as a unique grocery store selling products such as gourmet foods, beer and wine, bread, nuts, cereal, and coffee. One of their best-selling nuts is Raw California Almonds, which are priced at \$4.49 for 16 ounces. Since it is impossible to pack exactly 16 ounces in each packet, a researcher has determined that the weight of almonds in each packet is normally distributed with a mean and a standard deviation equal to 16.01 ounces and 0.08 ounces, respectively.
- Discuss the sampling distribution of the sample mean based on any given sample size.
  - Find the probability that a random sample of 20 bags of almonds will average less than 16 ounces.
  - Suppose your cereal recipe calls for no less than 48 ounces of almonds. What is the probability that three packets of almonds will meet your requirement?
58. Georgia residents spent an average of \$470.73 on the lottery in 2010, or 1% of their personal income ([www.msn.com](http://www.msn.com), May 23, 2012). Suppose the amount spent on the lottery follows a normal distribution with a standard deviation of \$50.
- What is the probability that a randomly selected Georgian spent more than \$500 on the lottery?

- b. If four Georgians are randomly selected, what is the probability that the average amount spent on the lottery was more than \$500?
- c. If four Georgians are randomly selected, what is the probability that all of them spent more than \$500 on the lottery?
59. Data from the Bureau of Labor Statistics' Consumer Expenditure Survey show that annual expenditures for cellular phone services per consumer unit increased from \$210 in 2001 to \$608 in 2007. Let the standard deviation of annual cellular expenditure be \$48 in 2001 and \$132 in 2007.
- a. What is the probability that the average annual expenditure of 100 cellular customers in 2001 exceeded \$200?
- b. What is the probability that the average annual expenditure of 100 cellular customers in 2007 exceeded \$600?
60. According to a recent report, scientists in New England say they have identified a set of genetic variants that predicts extreme longevity with 77% accuracy (*The New York Times*, July 1, 2010). Assume 150 patients decide to get their genome sequenced.
- a. If the claim by scientists is accurate, what is the probability that more than 120 patients will get a correct diagnosis for extreme longevity?
- b. If the claim by scientists is accurate, what is the probability that fewer than 70% of the patients will get a correct diagnosis for extreme longevity?
61. American workers are increasingly planning to delay retirement (*U.S. News & World Report*, June 30, 2010). According to a Pew Research Center comprehensive survey, 35% of employed adults of age 62 and older say they have pushed back their retirement date.
- a. What is the probability that in a sample of 100 employed adults of age 62 and older, more than 40% have pushed back their retirement date?
- b. What is the probability that in a sample of 200 employed adults of age 62 and older, more than 40% have pushed back their retirement date?
- c. Comment on the difference between the two estimated probabilities.
62. The producer of a particular brand of soup claims that its sodium content is 50% less than that of its competitor. The food label states that the sodium content measures 410 milligrams per serving. Assume the population of sodium content is normally distributed with a standard deviation of 25 milligrams. Inspectors take periodic samples of 25 cans and measure the sodium content. The following sample means are obtained.

405	412	399	420	430	428
-----	-----	-----	-----	-----	-----

- a. Construct an  $\bar{x}$  chart. Plot the sample means on the  $\bar{x}$  chart.
- b. Can the inspectors conclude that the producer is advertising the sodium content accurately? Explain.
63. **FILE Packaging.** A variety of packaging solutions exist for products that must be kept within a specific temperature range. Cold chain distribution is particularly useful in the food and pharmaceutical industries. A packaging company strives to maintain a constant temperature for its packages. It is believed that the temperature of its packages follows a normal distribution with a mean of 5 degrees Celsius and a standard deviation of 0.3 degree Celsius. Inspectors take weekly samples for 5 weeks of eight randomly selected boxes and report the following temperatures in degrees Celsius. A portion of the data is given below.
- | Week 1 | Week 2 | Week 3 | Week 4 | Week 5 |
|--------|--------|--------|--------|--------|
| 3.98   | 5.52   | 5.79   | 3.98   | 5.14   |
| 4.99   | 5.52   | 6.42   | 5.79   | 6.25   |
| :      | :      | :      | :      | :      |
| 4.95   | 4.95   | 5.44   | 5.95   | 4.28   |
- a. Construct an  $\bar{x}$  chart for quality control. Plot the five weekly sample means on the  $\bar{x}$  chart.
- b. Are any points outside the control limits? Does it appear that the process is in control? Explain.
64. Acceptance sampling is an important quality control technique, where a batch of data is tested to determine if the proportion of units having a particular attribute exceeds a given percentage. Suppose that 10% of produced items are known to be nonconforming. Every week a batch of items is evaluated and the production machines are adjusted if the proportion of nonconforming items exceeds 15%.
- a. What is the probability that the production machines will be adjusted if the batch consists of 50 items?
- b. What is the probability that the production machines will be adjusted if the batch consists of 100 items?
65. In the previous question, suppose that the management decides to use a  $\bar{p}$  chart for the analysis. As noted earlier, 10% of produced items are known to be nonconforming. The firm analyzes a batch of production items for 6 weeks and computes the following percentages of nonconforming items.

Week	Nonconforming Percentage
1	5.5%
2	13.1%
3	16.8%
4	13.6%
5	19.8%
6	2.0%

- Suppose weekly batches consisted of 50 items. Construct a  $\bar{p}$  chart and determine if the machine needs adjustment in any of the weeks.
- Suppose weekly batches consisted of 100 items. Construct a  $\bar{p}$  chart and determine if the machine needs adjustment in any of the weeks.

## CASE STUDIES

**CASE STUDY 7.1** The significant decline of savings in the United States from the 1970s and 1980s to the 1990s and 2000s has been widely discussed by economists (<http://money.cnn.com>, June 30, 2010). According to the Bureau of Economic Analysis, the savings rate of American households, defined as a percentage of the disposable personal income, was 4.20% in 2009. The reported savings rate is not uniform across the country. A public policy institute conducts two of its own surveys to compute the savings rate in the Midwest. In the first survey, a sample of 160 households is taken and the average savings rate is found to be 4.48%. Another sample of 40 households finds an average savings rate of 4.60%. Assume that the population standard deviation is 1.4%.

In a report, use the above information to:

- Compute the probability of obtaining a sample mean that is at least as high as the one computed in each of the two surveys.
- Use these probabilities to decide which of the two samples is likely to be more representative of the United States as a whole.

**CASE STUDY 7.2** According to a report, college graduates in 2010 were likely to face better job prospects than 2009 graduates (*The New York Times*, May 24, 2010). Many employers who might have been pessimistic at the start of the 2009–2010 academic year were making more offers than expected. Despite the improvement in job prospects, the Bureau of Labor Statistics reported that the current jobless rate for college graduates under age 25 was still 8%. For high school graduates under age 25 who did not enroll in college, the current jobless rate was 24.5%. Cindy Chan works in the sales department of a trendy apparel company and has recently been relocated to a small town in Iowa. She finds that there are a total of 220 college graduates and 140 high school graduates under age 25 who live in this town. Cindy wants to gauge the demand for her products by the number of youths in this town who are employed.

In a report, use the above information to:

- Compute the expected number of college and high school graduates who are employed.
- Report the probabilities that at least 200 college graduates and at least 100 high school graduates under age 25 are employed.

**CASE STUDY 7.3** Hockey pucks used by the National Hockey League (NHL) and other professional leagues weigh an average of 163 grams (5.75 ounces). A quality inspector monitors the manufacturing process for hockey pucks. She takes eight samples of  $n = 10$ . Measured in grams, the weights appear in the following table. It is believed that puck weights are normally distributed, and when the production process is in control,  $\mu = 163$  and  $\sigma = 7.5$ . A portion of the data, measured in grams, is shown in the accompanying table.



**Data for Case Study 7.3** Hockey Puck Weights (in grams)

#1	#2	#3	#4	#5	#6	#7	#8
162.2	165.8	156.4	165.3	168.6	167.0	186.8	178.3
159.8	166.2	156.4	173.3	175.8	171.4	160.4	163.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
160.3	160.6	152.2	166.4	168.2	168.4	176.8	171.3

In a report, use the above information to:

1. Prepare a control chart that specifies a centerline as well as an upper control limit (UCL) and a lower control limit (LCL).
2. Use the control chart to determine whether the process is in control.

## APPENDIX 7.1 Derivation of the Mean and the Variance for $\bar{X}$ and $\bar{P}$

### Sample Mean, $\bar{X}$

Let the expected value and the variance of the population  $X$  be denoted by  $E(X) = \mu$  and  $Var(X) = \sigma^2$ , respectively. The sample mean  $\bar{X}$  based on a random draw of  $n$  observations,  $X_1, X_2, \dots, X_n$ , from the population is computed as  $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ .

We use the properties of the sum of random variables to derive

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{E(X_1) + E(X_2) + \dots + E(X_n)}{n} \\ &= \frac{\mu + \mu + \dots + \mu}{n} = \frac{n\mu}{n} = \mu. \end{aligned}$$

Since the sample mean is based on  $n$  independent draws from the population, the covariance terms drop out and the variance of the sample mean is thus derived as:

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n^2} Var(X_1 + X_2 + \dots + X_n) \\ &= \frac{1}{n^2} (Var(X_1) + Var(X_2) + \dots + Var(X_n)) \\ &= \frac{\sigma^2 + \sigma^2 + \dots + \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

### Sample Proportion, $\bar{P}$

Let  $X$  be a binomial random variable representing the number of successes in  $n$  trials. Recall from Chapter 5 that  $E(X) = np$  and  $Var(X) = np(1 - p)$  where  $p$  is the probability of success. For the sample proportion  $\bar{P} = \frac{X}{n}$ ,

$$\begin{aligned} E(\bar{P}) &= E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{np}{n} = p, \quad \text{and} \\ Var(\bar{P}) &= Var\left(\frac{X}{n}\right) = \frac{Var(X)}{n^2} = \frac{np(1 - p)}{n^2} = \frac{p(1 - p)}{n}. \end{aligned}$$

## APPENDIX 7.2 Properties of Point Estimators

We generally discuss the performance of an estimator in terms of its statistical properties. Some of the desirable properties of a point estimator include unbiasedness, consistency, and efficiency. An estimator is **unbiased** if, based on repeated sampling from the population, the

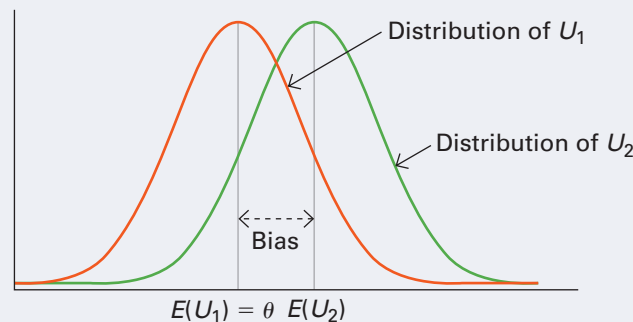


average value of the estimator equals the population parameter. In other words, for an unbiased estimator, the expected value of the point estimator equals the population parameter.

### UNBIASEDNESS

An estimator is **unbiased** if its expected value equals the unknown population parameter being estimated.

Figure A7.1 shows the sampling distributions for two estimators  $U_1$  and  $U_2$ , which are assumed to be normally distributed. Let  $\theta$  (the Greek letter read as theta) be the true parameter value of the population. Estimator  $U_1$  is unbiased because its expected value  $E(U_1)$  equals  $\theta$ . Estimator  $U_2$  is biased because  $E(U_2) \neq \theta$ ; the amount of bias is given by the difference between  $E(U_2)$  and  $\theta$ .



**FIGURE A7.1** The distributions of unbiased ( $U_1$ ) and biased ( $U_2$ ) estimators

Since  $E(\bar{X}) = \mu$  and  $E(\bar{P}) = p$ ,  $\bar{X}$  and  $\bar{P}$  are the unbiased estimators of  $\mu$  and  $p$ , respectively. This property is independent of the sample size. For instance, the expected value of the sample mean is equal to the population mean irrespective of the sample size.

We often compare the performance of the unbiased estimators in terms of their relative **efficiency**. An estimator is deemed efficient if its variability between samples is smaller than that of other unbiased estimators. Recall that the variability is often measured by the standard error of the estimator. For an unbiased estimator to be efficient, its standard error must be lower than that of other unbiased estimators. It is well documented that the estimators  $\bar{X}$  and  $\bar{P}$  are not only unbiased, but also efficient.

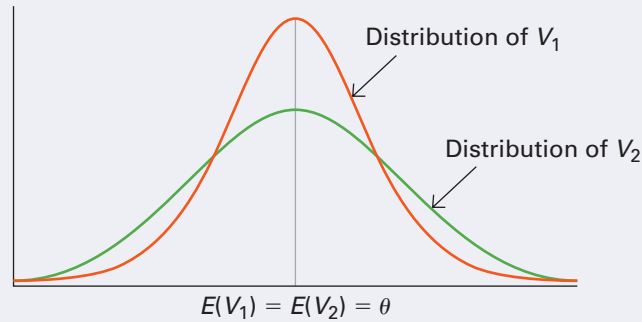
### EFFICIENCY

An unbiased estimator is **efficient** if its standard error is lower than that of other unbiased estimators.

Figure A7.2 shows the sampling distributions for two unbiased estimators  $V_1$  and  $V_2$  for the true population parameter  $\theta$ . Again, for illustration,  $V_1$  and  $V_2$  follow the normal distribution. While both  $V_1$  and  $V_2$  are unbiased ( $E(V_1) = E(V_2) = \theta$ ),  $V_1$  is more efficient because it has less variability.

Another desirable property, which is often considered a minimum requirement for an estimator, is **consistency**. An estimator is consistent if it approaches the population parameter of interest as the sample size increases. Consistency implies that we will get the inference right if we take a large enough sample. The estimators  $\bar{X}$  and  $\bar{P}$  are not only unbiased, but also consistent. For instance, the sample mean collapses to the population

**FIGURE A7.2**  
The distributions of  
efficient ( $V_1$ ) and less  
efficient ( $V_2$ ) estimators



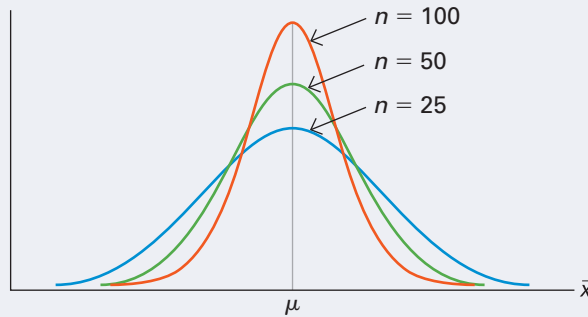
mean ( $\bar{X} \rightarrow \mu$ ) as the sample size approaches infinity ( $n \rightarrow \infty$ ). An unbiased estimator is consistent if its standard deviation, or its standard error, collapses to zero as the sample size increases.

### CONSISTENCY

An estimator is **consistent** if it approaches the unknown population parameter being estimated as the sample size grows larger.

The consistency of  $\bar{X}$  is illustrated in Figure A7.3.

**FIGURE A7.3**  
The distribution of a  
consistent estimator  $\bar{X}$   
for various sample sizes



As the sample size  $n$  increases, the variability of  $\bar{X}$  decreases. In particular as  $n \rightarrow \infty$ ,  $SD(\bar{X}) = \sigma/\sqrt{n} \rightarrow 0$ , thus implying that  $\bar{X}$  is a consistent estimator of  $\mu$ .

## APPENDIX 7.3 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor.

### Minitab

#### Generating a Random Sample

- A. (Replicating Example 7.1) From the menu choose **Calc > Random Data > Integer**.
- B. Enter 100 as the **Number of rows of data to generate**; enter C1 for **Store in column**; enter 1 for **Minimum value** and 20000 as **Maximum value**.

## Constructing an $\bar{x}$ Chart

- A. (Replicating Figure 7.A). Stack all wait times into one column.
- B. From the menu choose **Stat > Control Charts > Variables Charts for Subgroups > Xbar**.
- C. Choose **All observations for a chart are in one column**, and in the box directly under this one, select wait times. For **Subgroup sizes**, enter the number 4. Choose **Xbar Options** and enter 60 for **Mean** and 30 for **Standard deviation**.

**FILE**  
*Lux\_Hotel*

## SPSS

### Constructing an $\bar{x}$ Chart

- A. (Replicating Figure 7.A). Stack all wait times into one column. In an adjacent column, indicate how the data are grouped and label this column 'group'. For instance, the first four observations are given the value 1; the next four observations are given the value 2, and so on.
- B. From the menu select **Analyze > Quality Control > Control Charts > X-bar, R, s**.
- C. Under **Process Measurement**, select wait times, and under **Subgroups Defined by** select group. Under **Charts**, select **X-bar using standard deviation**. Choose **Options**. After **Number of Sigmas**, enter 3, and after **Minimum subgroup size**, enter 4. Choose **Statistics**. Under **Specification Limits**, enter 105 for **Upper**, 15 for **Lower**, and 60 for **Target**.

**FILE**  
*Lux\_Hotel*

## JMP

### Generating a Random Sample

- A. (Replicating Example 7.1) Right-click on the header at the top of the column in the spreadsheet view and select **Formula**. Under **Functions (grouped)**, choose **Random > Random Integer**.
- B. Put the insertion marker on the box for **n1** and click the insert button (shown as a caret ^ with the mathematical operations) until you see **n2** next to **n1**. Enter 1 for **n1** and 20000 for **n2**.

### Constructing the $\bar{x}$ Chart

- A. (Replicating Figure 7.A). Stack all wait times into one column.
- B. From the menu choose **Analyze > Quality and Process > Control Chart > X-bar**.
- C. Under **Select Columns**, select wait times, and under **Cast Columns into Roles**, Process. Under **Parameters**, select **KSigma** and enter 3. Under **Sample Size**, select **Sample Size Constant** and enter 4. Select **Specify Stats**. Enter 30 for **Sigma** and 60 for **Mean(measure)**.

**FILE**  
*Lux\_Hotel*

# 8

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 8.1 Explain an interval estimator.
- LO 8.2 Calculate a confidence interval for the population mean when the population standard deviation is known.
- LO 8.3 Describe the factors that influence the width of a confidence interval.
- LO 8.4 Discuss features of the  $t$  distribution.
- LO 8.5 Calculate a confidence interval for the population mean when the population standard deviation is not known.
- LO 8.6 Calculate a confidence interval for the population proportion.
- LO 8.7 Select a sample size to estimate the population mean and the population proportion.

# Interval Estimation

In earlier chapters, we made a distinction between the population parameters, such as the population mean and the population proportion, and the corresponding sample statistics. The sample statistics are used to make statistical inferences regarding the unknown values of the population parameters. In general, two basic methodologies emerge from the inferential branch of statistics: estimation and hypothesis testing. As discussed in Chapter 7, a point estimator uses sample data to produce a single value as an estimate for the unknown population parameter of interest. A confidence interval, on the other hand, produces a range of values that estimate the unknown population parameter. In this chapter, we develop and interpret confidence intervals for the population mean and the population proportion. Since obtaining a sample is one of the first steps in making statistical inferences, we also learn how an appropriate sample size is determined in order to achieve a certain level of precision in the estimates.



## INTRODUCTORY CASE

### Fuel Usage of “Ultra-Green” Cars

A car manufacturer advertises that its new “ultra-green” car obtains an average of 100 miles per gallon (mpg) and, based on its fuel emissions, is one of the few cars that earns an A+ rating from the Environmental Protection Agency. Jared Beane, an analyst at Pinnacle Research, records the mpg for a sample of 25 “ultra-green” cars after the cars were driven equal distances under identical conditions. Table 8.1 shows each car’s mpg.

**TABLE 8.1** MPG for a Sample of 25 “Ultra-Green” Cars

**FILE**  
**MPG**

97	117	93	79	97
87	78	83	94	96
102	98	82	96	113
113	111	90	101	99
112	89	92	96	98

Jared has already used tabular and graphical methods to summarize the data in his report. He would like to make statistical inferences regarding key population parameters. In particular, he wants to use the above sample information to:

1. Estimate the mean mpg of all ultra-green cars with 90% confidence.
2. Estimate the proportion of all ultra-green cars that obtain over 100 mpg with 90% confidence.
3. Determine the sample size that will enable him to achieve a specified level of precision in his mean and proportion estimates.

A synopsis of this case is provided at the end of Section 8.4.



## 8.1 CONFIDENCE INTERVAL FOR THE POPULATION MEAN WHEN $\sigma$ IS KNOWN

### LO 8.1

Explain an interval estimator.

Recall that a population consists of all items of interest in a statistical problem, whereas a sample is a subset of the population. Given sample data, we use the sample statistics to make inferences about the unknown population parameters, such as the population mean and the population proportion. Two basic methodologies emerge from the inferential branch of statistics: estimation and hypothesis testing. Although the sample statistics are based on a portion of the population, they contain useful information to estimate the population parameters and to conduct tests regarding the population parameters. In this chapter, we focus on estimation.

As discussed in Chapter 7, when a statistic is used to estimate a parameter, it is referred to as a point estimator, or simply an estimator. A particular value of the estimator is called a point estimate or an estimate. Recall that the sample mean  $\bar{X}$  is the estimator of the population mean  $\mu$ , and the sample proportion  $\bar{P}$  is the estimator of the population proportion  $p$ . Let us consider the introductory case where Jared Beane records the mpg for a sample of 25 ultra-green cars. We use the sample information in Table 8.1 to compute the mean mpg of the cars as  $\bar{x} = 96.52$  mpg. Similarly, since Jared is also interested in the proportion of these cars that get an mpg greater than 100 and seven of the cars in the sample satisfied this criterion, we compute the relevant sample proportion as  $\bar{p} = 7/25 = 0.28$ . Therefore, our estimate for the mean mpg of all ultra-green cars is 96.52 mpg and our estimate for the proportion of all ultra-green cars with mpg greater than 100 is 0.28.

It is important to note that the above estimates are based on a sample of 25 cars and, therefore, are likely to vary between samples. For instance, the values will change if another sample of 25 cars is used. What Jared really wishes to estimate are the mean and the proportion (parameters) of all ultra-green cars (population), not just those comprising the sample. We now examine how we can extract useful information from a single sample to make inferences about these population parameters.

So far we have only discussed estimators. Often it is more informative to provide a range of values—an **interval**—rather than a single point estimate for the unknown population parameter. This range of values is called a **confidence interval**, also referred to as an **interval estimate**, for the population parameter.

### CONFIDENCE INTERVAL

A **confidence interval** provides a range of values that, with a certain level of confidence, contains the population parameter of interest.

In order to construct a confidence interval for the population mean  $\mu$  or the population proportion  $p$ , it is essential that the sampling distributions of  $\bar{X}$  and  $\bar{P}$  follow, or approximately follow, a normal distribution. Other methods that do not require the normality condition are not discussed in this text. Recall from Chapter 7 that  $\bar{X}$  follows a normal distribution when the underlying population is normally distributed; this result holds irrespective of the sample size  $n$ . If the underlying population is not normally distributed, then by the central limit theorem, the sampling distribution of  $\bar{X}$  will be approximately normal if the sample size is sufficiently large—that is, when  $n \geq 30$ . Similarly, the sampling distribution of  $\bar{P}$  is approximately normal if the sample size is sufficiently large—that is, when  $np \geq 5$  and  $n(1 - p) \geq 5$ .

The main ingredient for developing a confidence interval is the sampling distribution of the underlying statistic. The sampling distribution of  $\bar{X}$ , for example, describes how the sample mean varies between samples. Recall that the variability between samples is measured by the standard error of  $\bar{X}$ . If the standard error is small, it implies that the sample means are not only close to one another, they are also close to the unknown population mean  $\mu$ .

A confidence interval is generally associated with a **margin of error** that accounts for the standard error of the estimator and the desired confidence level of the interval. For



estimating the population mean and the population proportion, the sampling distribution of the underlying statistic is approximately normal. The symmetry implied by the normal distribution allows us to construct a confidence interval by adding and subtracting the same margin of error to the point estimate.

It is common to construct a confidence interval as: Point estimate  $\pm$  Margin of error.

An analogy to a simple weather example is instructive. If you feel that the outside temperature is about 50 degrees, then perhaps you can, with a certain level of confidence, suggest that the actual temperature is between 40 and 60 degrees. In this example, 50 degrees is analogous to a point estimate of the actual temperature, and 10 degrees is the margin of error that is added to and subtracted from this point estimate.

We know from the introductory case study that the point estimate for the population mean mpg of all ultra-green cars is 96.52 mpg; that is,  $\bar{x} = 96.52$ . We can construct a confidence interval by using the point estimate as a base to which we add and subtract the margin of error.

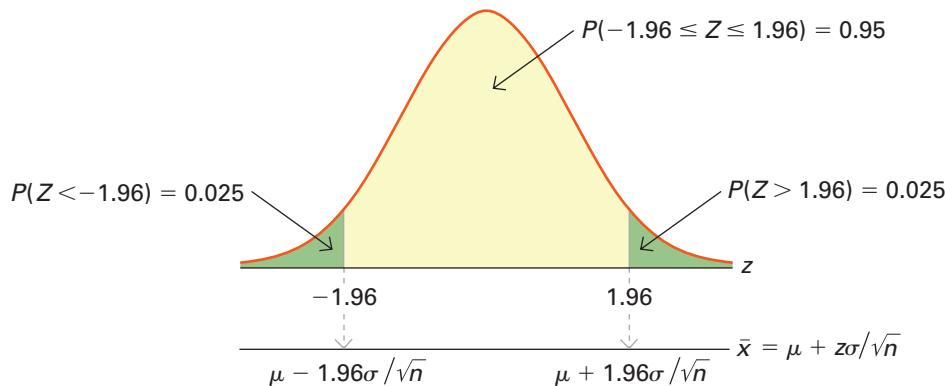
## Constructing a Confidence Interval for $\mu$ When $\sigma$ Is Known

Let us construct a 95% confidence interval for  $\mu$ . As mentioned earlier, in order to construct this interval, the sampling distribution of  $\bar{X}$  must be normal. Consider the standard normal random variable  $Z$ . Using the symmetry of  $Z$ , we can compute  $P(Z > 1.96) = P(Z < -1.96) = 0.025$ ; see Figure 8.1. Remember that  $z = 1.96$  is easily determined from the  $z$  table given the probability of 0.025 in the upper tail of the distribution. Therefore, we formulate the probability statement  $P(-1.96 \leq Z \leq 1.96) = 0.95$ .

### LO 8.2

Calculate a confidence interval for the population mean when the population standard deviation is known.

**FIGURE 8.1** Graphical depiction of  $P(Z < -1.96) = 0.025$  and  $P(Z > 1.96) = 0.025$



Since  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ , for a normal  $\bar{X}$  with mean  $\mu$  and standard error  $\sigma/\sqrt{n}$ , we get

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95.$$

We multiply by  $\sigma/\sqrt{n}$  and add  $\mu$  to obtain

$$P\left(\mu - 1.96\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 1.96\sigma/\sqrt{n}\right) = 0.95.$$

This equation (see also the lower portion of Figure 8.1) implies that there is a 0.95 probability that the sample mean  $\bar{X}$  will fall between  $\mu - 1.96\sigma/\sqrt{n}$  and  $\mu + 1.96\sigma/\sqrt{n}$ —that is, within the interval  $\mu \pm 1.96\sigma/\sqrt{n}$ . If samples of size  $n$  are drawn repeatedly from a given population, 95% of the computed sample means,  $\bar{x}$ 's, will fall within the interval and the remaining 5% will fall outside the interval.

We do not know the population mean  $\mu$ , and therefore cannot determine if a particular  $\bar{x}$  falls within the interval or not. However, we do know that  $\bar{x}$  will fall within the interval  $\mu \pm 1.96\sigma/\sqrt{n}$  if and only if  $\mu$  falls within the interval  $\bar{x} \pm 1.96\sigma/\sqrt{n}$ . This will happen 95% of the time given how the interval is constructed. Therefore, we call the interval  $\bar{x} \pm 1.96\sigma/\sqrt{n}$  a 95% confidence interval for the population mean, where  $1.96\sigma/\sqrt{n}$  is its margin of error.

Confidence intervals are often misinterpreted; you need to exercise care in characterizing them. For instance, the above 95% confidence interval does *not* imply that the probability that  $\mu$  falls in the confidence interval is 0.95. Remember that  $\mu$  is a constant, although its value is not known. It either falls in the interval (probability equals one) or does not fall in the interval (probability equals zero). The randomness comes from  $\bar{X}$ , not  $\mu$ , since many possible sample means can be derived from a population. Therefore, it is incorrect to say that the probability that  $\mu$  falls in the  $\bar{x} \pm 1.96\sigma/\sqrt{n}$  interval is 0.95. A 95% confidence interval simply implies that if numerous samples of size  $n$  are drawn from a given population, then 95% of the intervals formed by the preceding procedure (formula) will contain  $\mu$ . Keep in mind that we only use a single sample to derive the estimates. Since there are many possible samples, we will be right 95% of the time, thus giving us 95% confidence.

#### INTERPRETING A 95% CONFIDENCE INTERVAL

Technically, a 95% confidence interval for the population mean  $\mu$  implies that for 95% of the samples, the procedure (formula) produces an interval that contains  $\mu$ . Informally, we can report with 95% confidence that  $\mu$  lies in the given interval. It is not correct to say that there is a 95% chance that  $\mu$  lies in the given interval.

#### EXAMPLE 8.1

A sample of 25 cereal boxes of Granola Crunch, a generic brand of cereal, yields a mean weight of 1.02 pounds of cereal per box. Construct a 95% confidence interval for the mean weight of all cereal boxes. Assume that the weight is normally distributed with a population standard deviation of 0.03 pound.

**SOLUTION:** Note that the normality condition of  $\bar{X}$  is satisfied since the underlying population is normally distributed. A 95% confidence interval for the population mean is computed as

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 1.02 \pm 1.96 \frac{0.03}{\sqrt{25}} = 1.02 \pm 0.012.$$

With 95% confidence, we can report that the mean weight of all cereal boxes falls between 1.008 and 1.032 pounds.

While it is common to report a 95% confidence interval, in theory we can construct an interval of any level of confidence ranging from 0 to 100%. Let's now extend the analysis to include intervals of any confidence level. Let the Greek letter  $\alpha$  (alpha) denote the allowable probability of error that in Chapter 9 will define the so-called significance level. This is the probability that the estimation procedure will generate an interval that does not contain  $\mu$ . The **confidence coefficient** ( $1 - \alpha$ ) is interpreted as the probability that the estimation procedure will generate an interval that contains  $\mu$ . Thus, the probability of error  $\alpha$  is related to the confidence coefficient and the confidence level as follows:

- Confidence coefficient =  $1 - \alpha$ , and
- Confidence level =  $100(1 - \alpha)\%$ .

For example, a confidence coefficient of 0.95 implies that the probability of error  $\alpha$  equals  $1 - 0.95 = 0.05$  and the confidence level equals  $100(1 - 0.05)\% = 95\%$ . Similarly, for a 90% confidence interval, the confidence coefficient equals 0.90 and  $\alpha = 1 - 0.90 = 0.10$ . The following statement generalizes the construction of a confidence interval for  $\mu$  when  $\sigma$  is known.

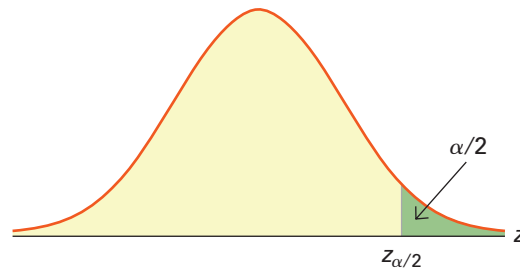
### CONFIDENCE INTERVAL FOR $\mu$ WHEN $\sigma$ IS KNOWN

A  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  when the population standard deviation  $\sigma$  is known is computed as

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \left[ \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

This formula is valid only if  $\bar{X}$  (approximately) follows a normal distribution.

The notation  $z_{\alpha/2}$  is the  $z$  value associated with the probability of  $\alpha/2$  in the upper tail of the standard normal probability distribution. In other words, if  $Z$  is a standard normal random variable and  $\alpha$  is any probability, then  $z_{\alpha/2}$  represents a  $z$  value such that the area under the  $z$  curve to the right of  $z_{\alpha/2}$  is  $\alpha/2$ —that is,  $P(Z \geq z_{\alpha/2}) = \alpha/2$ . Figure 8.2 depicts the notation  $z_{\alpha/2}$ .



**FIGURE 8.2** Graphical depiction of the notation  $z_{\alpha/2}$

As discussed earlier, for a 95% confidence interval,  $\alpha = 0.05$  and  $\alpha/2 = 0.025$ . Therefore,  $z_{\alpha/2} = z_{0.025} = 1.96$ . Similarly, using the  $z$  table or Excel's function NORM.S.INV, we can derive the following:

- For a 90% confidence interval,  $\alpha = 0.10$ ,  $\alpha/2 = 0.05$ , and  $z_{\alpha/2} = z_{0.05} = 1.645$ .
- For a 99% confidence interval,  $\alpha = 0.01$ ,  $\alpha/2 = 0.005$ , and  $z_{\alpha/2} = z_{0.005} = 2.576$ .

## The Width of a Confidence Interval

The margin of error used in the computation of the confidence interval for the population mean, when the population standard deviation is known, is  $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ . Since we are basically adding and subtracting this quantity from  $\bar{x}$ , the width of the confidence interval is two times the margin of error. In Example 8.1, the margin of error for a 95% confidence interval is 0.012 and the width of the interval is  $1.032 - 1.008 = 2(0.012) = 0.024$ . Now let's examine how the width of a confidence interval is influenced by various factors.

### LO 8.3

Describe the factors that influence the width of a confidence interval.

- For a given confidence level  $100(1 - \alpha)\%$  and sample size  $n$ , the larger the population standard deviation  $\sigma$ , the wider the confidence interval.

### EXAMPLE 8.1b

Let the standard deviation of the population in Example 8.1 be 0.05 instead of 0.03. Compute the 95% confidence interval based on the same sample information.

**SOLUTION:** We use the same formula as before, but we substitute 0.05 for the standard deviation instead of 0.03:

$$1.02 \pm 1.96 \frac{0.05}{\sqrt{25}} = 1.02 \pm 0.020.$$

The width has increased from 0.024 to  $2(0.020) = 0.040$ .

- II. For a given confidence level  $100(1 - \alpha)\%$  and population standard deviation  $\sigma$ , the smaller the sample size  $n$ , the wider the confidence interval.

### EXAMPLE 8.1c

Instead of 25 observations, let the sample in Example 8.1 be based on 16 observations. Compute the 95% confidence interval using the same sample mean of 1.02 pounds and the same population standard deviation of 0.03.

**SOLUTION:** Again we use the same formula as before, but this time we substitute 16 for  $n$  instead of 25:

$$1.02 \pm 1.96 \frac{0.03}{\sqrt{16}} = 1.02 \pm 0.015.$$

The width has increased from 0.024 to  $2(0.015) = 0.030$ .

- III. For a given sample size  $n$  and population standard deviation  $\sigma$ , the greater the confidence level  $100(1 - \alpha)\%$ , the wider the confidence interval.

### EXAMPLE 8.1d

Compute the 99%, instead of the 95%, confidence interval based on the information in Example 8.1.

**SOLUTION:** Now we use the same formula and substitute the value 2.576 for  $z_{\alpha/2}$  instead of 1.96:

$$1.02 \pm 2.576 \frac{0.03}{\sqrt{25}} = 1.02 \pm 0.015.$$

The width has increased from 0.024 to  $2(0.015) = 0.030$ .

The precision is directly linked with the width of the confidence interval—the wider the interval, the lower is its precision. Continuing with the weather analogy, a temperature estimate of 40 to 80 degrees is imprecise because the interval is too wide to be of value. We lose precision when the sample does not reveal a great deal about the population, resulting in a wide confidence interval. Examples 8.1b and 8.1c suggest that the estimate will be less precise if the variability of the underlying population is high ( $\sigma$  is high) or a small segment of the population is sampled ( $n$  is small). Example 8.1d relates the width with the confidence level. For given sample information, the only way we can gain confidence is by making the interval wider. If you are 95% confident that the outside temperature is between 40 and 60, then you can increase your confidence level to 99% only by using a wider range, say between 35 and 65. This result also helps us understand the difference between precision (width of the interval) and the confidence level. There is a trade-off between the amount of confidence we have in an interval and its width.

### EXAMPLE 8.2

IQ tests are designed to yield scores that are approximately normally distributed. A reporter is interested in estimating the average IQ of employees in a large high-tech firm in California. She gathers the IQ scores from 22 employees of this firm and records the sample mean IQ as 106. She assumes that the population standard deviation is 15.

- Compute 90% and 99% confidence intervals for the average IQ in this firm.
- Use these results to infer if the mean IQ in this firm is significantly different from the national average of 100.

**SOLUTION:**

- For the 90% confidence interval,  $z_{\alpha/2} = z_{0.05} = 1.645$ . Similarly, for the 99% confidence interval,  $z_{\alpha/2} = z_{0.005} = 2.576$ .

The 90% confidence interval is  $106 \pm 1.645 \frac{15}{\sqrt{22}} = 106 \pm 5.26$ .

The 99% confidence interval is  $106 \pm 2.576 \frac{15}{\sqrt{22}} = 106 \pm 8.24$ .

Note that the 99% interval is wider than the 90% interval.

- With 90% confidence, the reporter can infer that the average IQ of this firm's employees differs from the national average, since the value 100 falls outside the 90% confidence interval, [100.74, 111.26]. However, she cannot infer the same result with 99% confidence, since the wider range of the interval, [97.76, 114.24], includes the value 100. We will study the link between estimation and testing in more detail in the next chapter.

## Using Excel to Construct a Confidence Interval for $\mu$ When $\sigma$ Is Known

We can use Excel's functions to construct a confidence interval. These functions are particularly useful with large data sets. Consider the following example.

### EXAMPLE 8.3

Table 8.2 lists a portion of the weights (in grams) for a sample of 80 hockey pucks. Construct the 92% confidence interval for the population mean weight assuming that the population standard deviation is 7.5 grams.

**FILE**  
*Hockey\_Pucks*

**TABLE 8.2** Hockey Puck  
Weights,  $n = 80$

Weight (in grams)
162.2
159.8
⋮
171.3

**SOLUTION:** We need to compute  $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ . We are given  $\sigma = 7.5$  and  $n = 80$ . In order to find  $\bar{x}$ , we open *Hockey\_Pucks*, find an empty cell, and calculate the sample mean by inputting “=AVERAGE(A2:A81)”; Excel returns 166.71. For the 92% confidence interval,  $\alpha = 0.08$ , we need to find  $z_{\alpha/2} = z_{0.04}$ . Excel's NORM.S.INV function finds a particular  $z$  value for a given cumulative probability. Since we want the  $z$  value such that the area under the  $z$  curve to the right of  $z_{0.04}$  is 0.04, we insert “=NORM.S.INV(0.96)”. Excel returns 1.751 or, equivalently,  $z_{0.04} = 1.751$ . Inserting these values into the formula and simplifying yields:  $166.71 \pm 1.751 \frac{7.5}{\sqrt{80}} = 166.71 \pm 1.47$ . With 92% confidence, we conclude that the mean weight of all hockey pucks falls between 165.24 and 168.18 grams.

## EXERCISES 8.1

### Mechanics

- Find  $z_{\alpha/2}$  for each of the following confidence levels used in estimating the population mean.
  - 90%
  - 98%
  - 88%
- Find  $z_{\alpha/2}$  for each of the following confidence levels used in estimating the population mean.
  - 89%
  - 92%
  - 96%
- A simple random sample of 25 observations is derived from a normally distributed population with a known standard deviation of 8.2.
  - Is the condition that  $\bar{X}$  is normally distributed satisfied? Explain.
  - Compute the margin of error with 80% confidence.
  - Compute the margin of error with 90% confidence.
  - Which of the two margins of error will lead to a wider interval?
- Consider a population with a known standard deviation of 26.8. In order to compute an interval estimate for the population mean, a sample of 64 observations is drawn.
  - Is the condition that  $\bar{X}$  is normally distributed satisfied? Explain.
  - Compute the margin of error at the 95% confidence level.
  - Compute the margin of error at the 95% confidence level based on a larger sample of 225 observations.
  - Which of the two margins of error will lead to a wider confidence interval?
- Discuss the factors that influence the margin of error for the confidence interval for the population mean. What can a practitioner do to reduce the margin of error?
- An article in the *National Geographic News* ("U.S. Racking Up Huge Sleep Debt," February 24, 2005) argues that Americans are increasingly skimping on their sleep. A researcher in a small Midwestern town wants to estimate the mean weekday sleep time of its adult residents. He takes a random sample of 80 adult residents and records their weekday mean sleep time as 6.4 hours. Assume that the population standard deviation is fairly stable at 1.8 hours.
  - Calculate the 95% confidence interval for the population mean weekday sleep time of all adult residents of this Midwestern town.
  - Can we conclude with 95% confidence that the mean sleep time of all adult residents in this Midwestern town is not 7 hours?
- A family is relocating from St. Louis, Missouri, to California. Due to an increasing inventory of houses in St. Louis, it is taking longer than before to sell a house. The wife is concerned and wants to know when it is optimal to put their house on the market. Her realtor friend informs them that the last 26 houses that sold in their neighborhood took an average time of 218 days to sell. The realtor also tells them that based on her prior experience, the population standard deviation is 72 days.
  - What assumption regarding the population is necessary for making an interval estimate for the population mean?
  - Construct the 90% confidence interval for the mean sale time for all homes in the neighborhood.
- U.S. consumers are increasingly viewing debit cards as a convenient substitute for cash and checks. The average amount spent annually on a debit card is \$7,790 (*Kiplinger's*, August 2007). Assume that this average was based on a sample of 100 consumers and that the population standard deviation is \$500.
  - At 99% confidence, what is the margin of error?
  - Construct the 99% confidence interval for the population mean amount spent annually on a debit card.

### Applications

- The average life expectancy for Bostonians is 78.1 years (*The Boston Globe*, August 16, 2010). Assume that this average was based on a sample of 50 Bostonians and that the population standard deviation is 4.5 years.
  - What is the point estimate of the population mean?
  - At 90% confidence, what is the margin of error?
  - Construct the 90% confidence interval for the population average life expectancy of Bostonians.
- In order to estimate the mean 30-year fixed mortgage rate for a home loan in the United States, a random sample of 28 recent loans is taken. The average calculated from this sample is 5.25%. It can be assumed that 30-year fixed mortgage rates are normally distributed with a standard deviation of 0.50%. Compute 90% and 99% confidence intervals for the population mean 30-year fixed mortgage rate.
- A manager is interested in estimating the mean time (in minutes) required to complete a job. His assistant uses a sample of 100 observations to report the confidence interval as [14.355, 17.645]. The population standard deviation is known to be equal to 10 minutes.
  - Find the sample mean time used to compute the confidence interval.
  - Determine the confidence level used for the analysis.



13. **FILE CT\_Undergrad\_Debt.** A study reports that recent college graduates from New Hampshire face the highest average debt of \$31,048 (*The Boston Globe*, May 27, 2012). A researcher from Connecticut wants to determine how recent undergraduates from that state fare. He collects data on debt from 40 recent undergraduates. A portion of the data is shown in the accompanying table. Assume that the population standard deviation is \$5,000.

Debt
24,040
19,153
⋮
29,329

- Use Excel to construct the 95% confidence interval for the mean debt of all undergraduates from Connecticut.
  - Use the 95% confidence interval to determine if the debt of Connecticut undergraduates differs from that of New Hampshire undergraduates.
14. **FILE Hourly\_Wage.** An economist wants to estimate the mean hourly wage of all workers. She collects data on 50 hourly wage earners. A portion of the data is shown in the accompanying table. Assume that the population

standard deviation is \$6. Construct and interpret 90% and 99% confidence intervals for the mean hourly wage of all workers.

Hourly Wage (in \$)
37.85
21.72
⋮
24.18

15. **FILE Highway\_Speeds.** A safety officer is concerned about speeds on a certain section of the New Jersey Turnpike. He records the speeds of 40 cars on a Saturday afternoon. The accompanying table shows a portion of the results. Assume that the population standard deviation is 5 mph. Construct the 95% confidence interval for the mean speed of all cars on that section of the turnpike. Are the safety officer's concerns valid if the speed limit is 55 mph? Explain.

Highway Speeds (mph)
70
60
⋮
65

## 8.2 CONFIDENCE INTERVAL FOR THE POPULATION MEAN WHEN $\sigma$ IS UNKNOWN

So far we have considered confidence intervals for the population mean when the population standard deviation  $\sigma$  is known. In reality,  $\sigma$  is rarely known. Recall from Chapter 3 that the population variance and the population standard deviation are calculated as  $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$  and  $\sigma = \sqrt{\sigma^2}$ , respectively. It is highly unlikely that  $\sigma$  is known when  $\mu$  is not. However, there are instances when the population standard deviation is considered fairly stable and, therefore, can be determined from prior experience. In these cases, the population standard deviation is treated as known.

Recall that the margin of error in a confidence interval depends on the standard error of the estimator and the desired confidence level. With  $\sigma$  unknown, the standard error of  $\bar{X}$ , given by  $\sigma/\sqrt{n}$ , can be conveniently estimated by  $s/\sqrt{n}$ , where  $s$  denotes the sample standard deviation. For convenience, we denote this estimate of the standard error of  $\bar{X}$  also by  $se(\bar{X}) = s/\sqrt{n}$ .

### The $t$ Distribution

As discussed earlier, in order to derive a confidence interval for  $\mu$ , it is essential that  $\bar{X}$  be normally distributed. A normally distributed  $\bar{X}$  is standardized as  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  where  $Z$  follows the  $z$  distribution. Another standardized statistic, which uses the estimator  $S$  in place of  $\sigma$ , is computed as  $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ . The random variable  $T$  follows the **Student's  $t$  distribution**, more commonly known as the  **$t$  distribution**.<sup>1</sup>

<sup>1</sup>William S. Gossett (1876–1937) published his research concerning the  $t$  distribution under the pen name “Student” because his employer, the Guinness Brewery, did not allow employees to publish their research results.

#### LO 8.4

Discuss features of the  $t$  distribution.

### THE $t$ DISTRIBUTION

If a random sample of size  $n$  is taken from a normal population with a finite variance, then the statistic  $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$  follows the  $t$  distribution with  $(n - 1)$  degrees of freedom,  $df$ .

The  $t$  distribution is actually a family of distributions, which are similar to the  $z$  distribution in that they are all bell-shaped and symmetric around zero. However, all  $t$  distributions have slightly broader tails than the  $z$  distribution. Each  $t$  distribution is identified by the **degrees of freedom**, or simply  $df$ . The degrees of freedom determine the extent of the broadness of the tails of the distribution; the fewer the degrees of freedom, the broader the tails. Since the  $t$  distribution is defined by the degrees of freedom, it is common to refer to it as the  $t_{df}$  distribution.

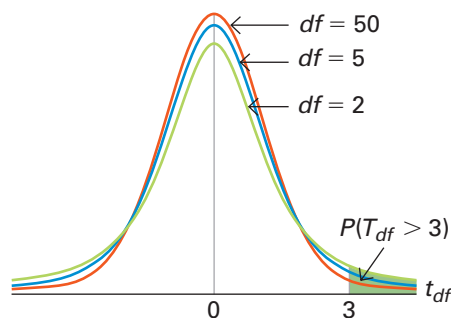
Specifically, the degrees of freedom refer to the number of independent pieces of information that go into the calculation of a given statistic and, in this sense, can be “freely chosen.” Consider the number of independent observations that enter into the calculation of the sample mean. If it is known that  $\bar{x} = 20$ ,  $n = 4$ , and three of the observations have values of  $x_1 = 16$ ,  $x_2 = 24$ , and  $x_3 = 18$ , then there is no choice but for the fourth observation to have a value of 22. In other words, three degrees of freedom are involved in computing  $\bar{x} = 20$  if  $n = 4$ ; in effect, one degree of freedom is lost.

### Summary of the $t_{df}$ Distribution

- Like the  $z$  distribution, the  $t_{df}$  distribution is bell-shaped and symmetric around 0 with asymptotic tails (the tails get closer and closer to the horizontal axis but never touch it).
- The  $t_{df}$  distribution has slightly broader tails than the  $z$  distribution.
- The  $t_{df}$  distribution consists of a family of distributions where the actual shape of each one depends on the degrees of freedom  $df$ . As  $df$  increases, the  $t_{df}$  distribution becomes similar to the  $z$  distribution; it is identical to the  $z$  distribution when  $df$  approaches infinity.

From Figure 8.3 we note that the tails of the  $t_2$  and  $t_5$  distributions are broader than the tails of the  $t_{50}$  distribution. For instance, for  $t_2$  and  $t_5$ , the area exceeding a value of 3, or  $P(T_{df} > 3)$ , is greater than that for  $t_{50}$ . In addition, the  $t_{50}$  resembles the  $z$  distribution.

**FIGURE 8.3**  
The  $t_{df}$  distribution with various degrees of freedom



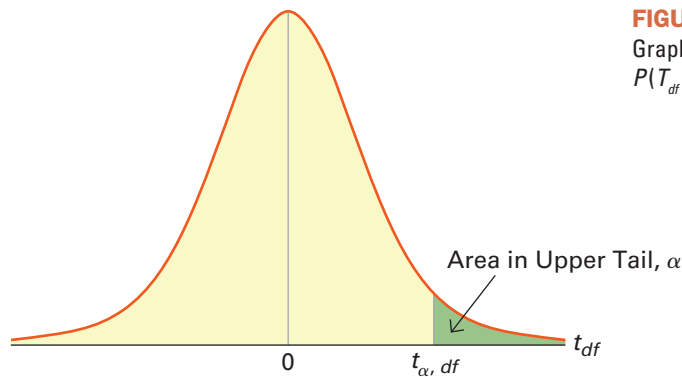
### Locating $t_{df}$ Values and Probabilities

Table 8.3 lists  $t_{df}$  values for selected upper-tail probabilities and degrees of freedom  $df$ . Table 2 of Appendix A provides a more complete table. Since the  $t_{df}$  distribution is a family of distributions identified by the  $df$  parameter, the  $t$  table is not as comprehensive as the  $z$  table. It only lists probabilities corresponding to a limited number of values. Also, unlike the cumulative probabilities in the  $z$  table, the  $t$  table provides the probabilities in the upper tail of the distribution.

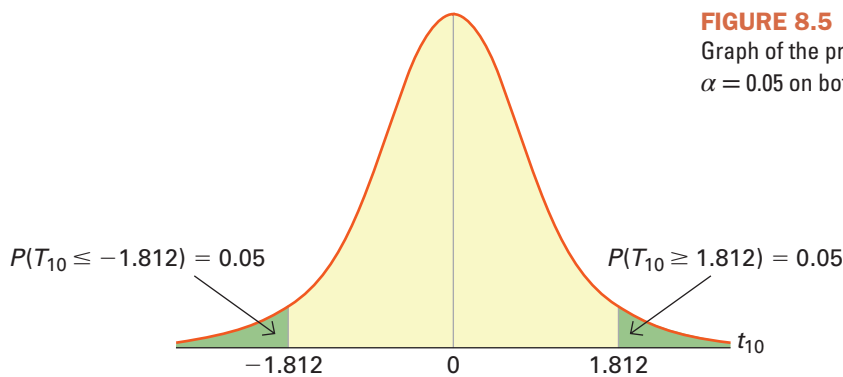
**TABLE 8.3** Portion of the *t* Table

<i>df</i>	Area in Upper Tail, $\alpha$					
	0.20	0.10	0.05	0.025	0.01	0.005
1	1.376	3.078	6.314	12.706	31.821	63.657
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	0.879	1.372	<b>1.812</b>	2.228	2.764	3.169
⋮	⋮	⋮	⋮	⋮	⋮	⋮
∞	0.842	1.282	1.645	1.960	2.326	2.576

We use the notation  $t_{\alpha,df}$  to denote a value such that the area in the upper tail equals  $\alpha$  for a given  $df$ . In other words, for a random variable  $T_{df}$ , the notation  $t_{\alpha,df}$  represents a value such that  $P(T_{df} \geq t_{\alpha,df}) = \alpha$ . Similarly,  $t_{\alpha/2,df}$  represents a value such that  $P(T_{df} \geq t_{\alpha/2,df}) = \alpha/2$ . Figure 8.4 illustrates the notation.

**FIGURE 8.4**  
Graphical depiction of  $P(T_{df} \geq t_{\alpha,df}) = \alpha$ 

When determining the value  $t_{\alpha,df}$ , we need two pieces of information: (a) the sample size  $n$ , or analogously,  $df = n - 1$ , and (b)  $\alpha$ . For instance, suppose we want to find the value  $t_{\alpha,df}$  with  $\alpha = 0.05$  and  $df = 10$ —that is,  $t_{0.05,10}$ . Using Table 8.3, we look at the first column labeled  $df$  and find the row 10. We then continue along this row until we reach the column  $\alpha = 0.05$ . The value 1.812 suggests that  $P(T_{10} \geq 1.812) = 0.05$ . Due to the symmetry of the  $t$  distribution, we also get  $P(T_{10} \leq -1.812) = 0.05$ . Figure 8.5 shows these results graphically. Also, since the area under the entire  $t_{df}$  distribution sums to one, we deduce that  $P(T_{10} < 1.812) = 1 - 0.05 = 0.95$ , which also equals  $P(T_{10} > -1.812)$ .

**FIGURE 8.5**  
Graph of the probability  $\alpha = 0.05$  on both sides of  $T_{10}$ 

Sometimes the exact probability cannot be determined from the  $t$  table. For example, given  $df = 10$ , the exact probability  $P(T_{10} \geq 1.562)$  is not included in the table. However, this probability is between 0.05 and 0.10 because the value 1.562 falls between 1.372 and 1.812. Similarly,  $P(T_{10} < 1.562)$  is between 0.90 and 0.95. We can use Excel and other statistical packages to find exact probabilities.

### EXAMPLE 8.4

Compute  $t_{\alpha,df}$  for  $\alpha = 0.025$  using 2, 5, and 50 degrees of freedom.

#### SOLUTION:

- For  $df = 2$ ,  $t_{0.025,2} = 4.303$ .
- For  $df = 5$ ,  $t_{0.025,5} = 2.571$ .
- For  $df = 50$ ,  $t_{0.025,50} = 2.009$ .

Note that the  $t_{df}$  values change with the degrees of freedom. Moreover, as  $df$  increases, the  $t_{df}$  distribution begins to resemble the  $z$  distribution. In fact, with  $df = \infty$ ,  $t_{0.025,\infty} = 1.96$ , which is identical to the corresponding  $z$  value; recall that  $P(Z \geq 1.96) = 0.025$ .

### LO 8.5

Calculate a confidence interval for the population mean when the population standard deviation is not known.

## Constructing a Confidence Interval for $\mu$ When $\sigma$ Is Unknown

We can never stress enough the importance of the requirement that  $\bar{X}$  follows a normal distribution in estimating the population mean. Recall that  $\bar{X}$  follows the normal distribution when the underlying population is normally distributed or when the sample size is sufficiently large ( $n \geq 30$ ). We still construct the confidence interval for  $\mu$  as: Point estimate  $\pm$  Margin of error. However, when the population standard deviation is unknown, we now use the  $t_{df}$  distribution to calculate the margin of error.

### CONFIDENCE INTERVAL FOR $\mu$ WHEN $\sigma$ IS NOT KNOWN

A  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  when the population standard deviation  $\sigma$  is not known is computed as

$$\bar{x} \pm t_{\alpha/2,df} \frac{s}{\sqrt{n}} \quad \text{or} \quad \left[ \bar{x} - t_{\alpha/2,df} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,df} \frac{s}{\sqrt{n}} \right],$$

where  $s$  is the sample standard deviation. This formula is valid only if  $\bar{X}$  (approximately) follows a normal distribution.

As before,  $100(1 - \alpha)\%$  is the confidence level and  $t_{\alpha/2,df}$  is the  $t_{df}$  value associated with the probability  $\alpha/2$  in the upper tail of the distribution with  $df = n - 1$ . In other words,  $P(T_{df} > t_{\alpha/2,df}) = \alpha/2$ . It is important to note that uncertainty is increased when we estimate the population standard deviation with the sample standard deviation, making the confidence interval wider, especially for smaller samples. This is appropriately captured by the wider tail of the  $t_{\alpha/2}$  distribution.

### EXAMPLE 8.5

In the introductory case of this chapter, Jared Beane wants to estimate the mean mpg for all ultra-green cars. Table 8.1 lists the mpg of a sample of 25 cars. Use this information to construct a 90% confidence interval for the population mean. Assume that mpg follows a normal distribution.

**SOLUTION:** The condition that  $\bar{X}$  follows a normal distribution is satisfied since we assumed that mpg is normally distributed. Thus, we construct the confidence interval as  $\bar{x} \pm t_{\alpha/2,df} \frac{s}{\sqrt{n}}$ . This is a classic example where a statistician has access only to sample data. Since the population standard deviation is not known, the sample standard deviation has to be computed from the sample. From the sample

data in Table 8.1, we find that  $\bar{x} = \frac{\sum x_i}{n} = \frac{2413}{25} = 96.52$  mpg and  $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{2746.24}{25-1}} = 10.70$ ; alternatively, we can use Excel to find these values. For the 90% confidence interval,  $\alpha = 0.10$ ,  $\alpha/2 = 0.05$ , and given  $n = 25$ ,  $df = 25 - 1 = 24$ . Thus,  $t_{0.05,24} = 1.711$ .

The 90% confidence interval for  $\mu$  is computed as

$$\bar{x} \pm t_{\alpha/2, df} \frac{s}{\sqrt{n}} = 96.52 \pm 1.711 \frac{10.70}{\sqrt{25}} = 96.52 \pm 3.66.$$

Thus, Jared concludes with 90% confidence that the average mpg of all ultra-green cars is between 92.86 mpg and 100.18 mpg. Note that the manufacturer's claim that the ultra-green car will average 100 mpg cannot be rejected by the sample data since the value 100 falls within the 90% confidence interval.

## Using Excel to Construct a Confidence Interval for $\mu$ When $\sigma$ Is Unknown

Again we find that Excel's functions are quite useful when constructing confidence intervals. Consider the following example.

### EXAMPLE 8.6

A recent article found that Massachusetts residents spent an average of \$860.70 on the lottery in 2010 (www.businessweek.com, March 14, 2012). In order to verify the results, a researcher at a Boston think tank surveys 100 Massachusetts residents and asks them about their annual expenditures on the lottery. Table 8.4 shows a portion of the results. Construct the 95% confidence interval for the average annual expenditures on the lottery for all Massachusetts residents. Do the results dispute the article's claim? Explain.

**TABLE 8.4** Massachusetts Residents' Annual Lottery Expenditures,  $n = 100$

**FILE**  
**Lottery**

Annual Lottery Expenditures (in \$)
790
594
⋮
759

**SOLUTION:** We need to compute  $\bar{x} \pm t_{\alpha/2, df} \frac{s}{\sqrt{n}}$ . The only value that is readily available is  $n = 100$ . In order to find  $\bar{x}$  and  $s$ , we open **Lottery**, find empty cells, and input “=AVERAGE(A2:A101)” and “=STDEV.S(A2:A101)”, respectively; Excel returns a sample mean of 841.94 and a sample standard deviation of 217.15. For the 95% confidence interval with  $\alpha = 0.05$  and  $df = n - 1 = 100 - 1 = 99$ , we need to find  $t_{\alpha/2, df} = t_{0.025, 99}$ . Excel's T.INV function finds a particular  $t_{df}$  value for a given cumulative probability. Since we want  $t_{0.025, 99}$  such that the area under the  $t_{99}$  curve to the right of  $t_{0.025, 99}$  is 0.025, we insert “=T.INV(0.975, 99).” Excel returns 1.984. Excel also provides another extremely similar function, “=TINV( $\alpha, df$ )” that directly computes  $t_{\alpha/2, df}$ . For example, “=TINV(0.05, 99)” also returns 1.984. Inserting the values returned by Excel into the formula and simplifying yields:  $841.94 \pm 1.984 \frac{217.15}{\sqrt{100}} = 841.94 \pm 43.08$ . With 95% confidence, we conclude that the average annual expenditures on the lottery for all Massachusetts residents fall between \$798.86 and \$885.02. The results do not dispute the article's claim since the interval includes the article's reported mean value of \$860.70.

## EXERCISES 8.2

### Mechanics

16. Find  $t_{\alpha, df}$  from the following information.
- $\alpha = 0.025$  and  $df = 12$
  - $\alpha = 0.10$  and  $df = 12$
  - $\alpha = 0.025$  and  $df = 25$
  - $\alpha = 0.10$  and  $df = 25$
17. We use the  $t$  distribution to construct a confidence interval for the population mean when the underlying population standard deviation is not known. Under the assumption that the population is normally distributed, find  $t_{\alpha/2, df}$  for the following scenarios.
- A 90% confidence level and a sample of 28 observations.
  - A 95% confidence level and a sample of 28 observations.
  - A 90% confidence level and a sample of 15 observations.
  - A 95% confidence level and a sample of 15 observations.
18. A random sample of 24 observations is used to estimate the population mean. The sample mean and the sample standard deviation are calculated as 104.6 and 28.8, respectively. Assume that the population is normally distributed.
- Construct the 90% confidence interval for the population mean.
  - Construct the 99% confidence interval for the population mean.
  - Use your answers to discuss the impact of the confidence level on the width of the interval.
19. Consider a normal population with an unknown population standard deviation. A random sample results in  $\bar{X} = 48.68$  and  $s^2 = 33.64$ .
- Compute the 95% confidence interval for  $\mu$  if  $\bar{X}$  and  $s^2$  were obtained from a sample of 16 observations.
  - Compute the 95% confidence interval for  $\mu$  if  $\bar{X}$  and  $s^2$  were obtained from a sample of 25 observations.
  - Use your answers to discuss the impact of the sample size on the width of the interval.
20. Let the following sample of 8 observations be drawn from a normal population with unknown mean and standard deviation: 22, 18, 14, 25, 17, 28, 15, 21.
- Calculate the sample mean and the sample standard deviation.
  - Construct the 80% confidence interval for the population mean.
  - Construct the 90% confidence interval for the population mean.
  - What happens to the margin of error as the confidence level increases from 80% to 90%?
- Assume the normal distribution for the underlying population to construct the 90% confidence interval for the population mean.
22. A popular weight loss program claims that with its recommended healthy diet regimen, users lose significant weight within a month. In order to estimate the mean weight loss of all customers, a nutritionist takes a sample of 18 dieters and records their weight loss one month after joining the program. He computes the sample mean and the standard deviation of weight loss as 12.5 pounds and 9.2 pounds, respectively. He believes that weight loss is likely to be normally distributed.
- Calculate the margin of error with 95% confidence.
  - Compute the 95% confidence interval for the population mean.
23. The manager of The Cheesecake Factory in Boston reports that on six randomly selected weekdays, the number of customers served was 120, 130, 100, 205, 185, and 220. She believes that the number of customers served on weekdays follows a normal distribution. Construct the 90% confidence interval for the average number of customers served on weekdays.
24. According to a recent survey, high school girls average 100 text messages daily (*The Boston Globe*, April 21, 2010). Assume that the survey was based on a random sample of 36 high school girls. The sample standard deviation is computed as 10 text messages daily.
- Calculate the margin of error with 99% confidence.
  - What is the 99% confidence interval for the population mean texts that all high school girls send daily?
25. The Chartered Financial Analyst (CFA®) designation is fast becoming a requirement for serious investment professionals. Although it requires a successful completion of three levels of grueling exams, it also entails promising careers with lucrative salaries. A student of finance is curious about the average salary of a CFA charterholder. He takes a random sample of 36 recent charterholders and computes a mean salary of \$158,000 with a standard deviation of \$36,000. Use this sample information to determine the 95% confidence interval for the average salary of a CFA charterholder.
26. The *Sudoku* puzzle has recently become very popular all over the world. It is based on a  $9 \times 9$  grid and the challenge is to fill in the grid so that every row, every column, and every  $3 \times 3$  box contains the digits 1 through 9. A researcher is interested in estimating the average time taken by a college student to solve the puzzle. He takes a random sample of 8 college students and records their solving times (in minutes) as 14, 7, 17, 20, 18, 15, 19, 28.
- Construct the 99% confidence interval for the average time taken by a college student to solve a *Sudoku* puzzle.
  - What assumption is necessary to make this inference?

### Applications

21. A random sample of eight drugstores shows the following prices (in \$) of a popular pain reliever:

3.50	4.00	2.00	3.00	2.50	3.50	2.50	3.00
------	------	------	------	------	------	------	------



27. Executive compensation has risen dramatically compared to the rising levels of an average worker's wage over the years. Sarah is an MBA student who decides to use her statistical skills to estimate the mean CEO compensation in 2010 for all large companies in the United States. She takes a random sample of six CEO compensations.

Firm	Compensation (in \$ millions)
Intel	8.20
Coca-Cola	2.76
Wells Fargo	6.57
Caterpillar	3.88
McDonald's	6.56
U.S. Bancorp	4.10

SOURCE: <http://finance.yahoo.com>.

- Help Sarah use the above information to construct the 90% confidence interval for the mean CEO compensation of all large companies in the United States?
  - What assumption is necessary for deriving the interval estimate?
  - How can the margin of error reported in part a be reduced?
28. As reported by [tradingeconomics.com](http://tradingeconomics.com) on September 2, 2012, the unemployment rates (in %) in major economies around the world are as follows:

Country	Unemployment Rate (in %)
Australia	5.2
China	4.1
France	10.0
Germany	6.8
India	3.8
United Kingdom	8.0
United States	8.3

- Calculate the margin of error used in the 95% confidence level for the population mean unemployment rate. Explain the assumption made for the analysis.
  - How can we reduce the margin of error for the 95% confidence interval?
29. A price-earnings ratio or P/E ratio is calculated as a firm's share price compared to the income or profit earned by the firm per share. Generally, a high P/E ratio suggests that investors are expecting higher earnings growth in the future compared to companies with a lower P/E ratio. The following table shows the P/E ratios for a sample of firms in the footwear industry:

Firm	P/E Ratio
Brown Shoe Co., Inc.	26
CROCS, Inc.	13
DSW, Inc.	21
Foot Locker, Inc.	16
Nike, Inc.	21

SOURCE: <http://biz.yahoo.com>, data retrieved September 2, 2012.

Let these ratios represent a random sample drawn from a normally distributed population. Construct the 90% confidence interval for the mean P/E ratio for the entire footwear industry.

30. The monthly closing stock prices (rounded to the nearest dollar) for Panera Bread Co. for the first six months of 2010 are reported in the following table.

Month	Closing Stock Price
January 2010	\$71
February 2010	73
March 2010	76
April 2010	78
May 2010	81
June 2010	75

SOURCE: <http://finance.yahoo.com>.

- Calculate the sample mean and the sample standard deviation.
  - Compute the 90% confidence interval for the mean stock price of Panera Bread Co., assuming that the stock price is normally distributed.
  - What happens to the margin of error if a higher confidence level is used for the interval estimate?
31. The following table shows the annual returns (in percent) for Fidelity's Electronics and Utilities funds.

Year	Electronics	Utilities
2010	17	11
2011	-8	13
2012	4	7
2013	39	21
2014	38	22

SOURCE: <http://finance.yahoo.com>, data retrieved April 3, 2015.

- Derive 99% confidence intervals for the mean return for Fidelity's Electronics and Utilities funds.
  - What did you have to assume to make the above inferences?
32. Suppose the 90% confidence interval for the mean SAT scores of applicants at a business college is given by [1690, 1810]. This confidence interval uses the sample mean and the sample standard deviation based on 25 observations. What are the sample mean and the sample standard deviation used when computing the interval?
33. A teacher wants to estimate the mean time (in minutes) that students take to go from one classroom to the next. His research assistant uses the sample time of 36 students to report the confidence interval as [8.20, 9.80].
- Find the sample mean time used to compute the confidence interval.
  - Determine the confidence level if the sample standard deviation used for the interval is 2.365.
34. In order to lure female customers, a new clothing store offers free gourmet coffee and pastry to its customers. The average daily revenue over the past five-week period

has been \$1,080 with a standard deviation of \$260. Use this sample information to construct the 95% confidence interval for the average daily revenue. The store manager believes that the coffee and pastry strategy would lead to an average daily revenue of \$1,200. Use the above 95% interval to determine if the manager is wrong.

35. **FILE Startups.** Many of today's leading companies, including Google, Microsoft, and Facebook, are based on technologies developed within universities. Lisa Fisher is a business school professor who believes that a university's research expenditure in \$ millions (Research) and the age of its technology transfer office in years (Duration) are major factors that enhance innovation. She wants to know what the average values are for the Research and the Duration variables. She collects data from 143 universities on these variables for the academic year 2008. A portion of the data is shown in the accompanying table.

Research (\$ millions)	Duration
\$145.52	23
\$237.52	23
⋮	⋮
\$154.38	9

SOURCE: Association of University Managers and National Science Foundation.

- a. Construct and interpret the 95% confidence interval for the mean research expenditure of all universities.
- b. Construct and interpret the 95% confidence interval for the mean duration of all universities.
36. **FILE Economics.** An associate dean of a university wishes to compare the means on the standardized final exams in microeconomics and macroeconomics. He has access to a random sample of 40 scores from each of these two courses. A portion of the data is shown in the accompanying table.

Micro	Macro
85	48
78	79
⋮	⋮
75	74

- a. Use Excel to construct 95% confidence intervals for the mean score in microeconomics and the mean score in macroeconomics.
- b. Explain why the widths of the two intervals are different.

37. **FILE Math\_Scores.** For decades, people have believed that boys are innately more capable than girls in math. In other words, due to the intrinsic differences in brains, boys are better suited for doing math than girls. Recent research challenges this stereotype, arguing that gender differences in math performance have more to do with culture than innate aptitude. Others argue, however, that while the average may be the same, there is more variability in math ability for boys than girls, resulting in some boys with soaring math skills. A portion of the data on math scores of boys and girls is shown in the accompanying table.

Boys	Girls
74	83
89	76
⋮	⋮
66	74

- a. Use Excel to construct 95% confidence intervals for the mean scores of boys and the mean scores of girls. Explain your assumptions.
- b. Explain why the widths of the two intervals are different.
38. **FILE Debt\_Payments.** A recent study found that consumers are making average monthly debt payments of \$983 (Experian.com, November 11, 2010). The accompanying table shows a portion of average debt payments for 26 metropolitan areas. Use Excel to construct 90% and 95% confidence intervals for the population mean. Comment on the width of the interval.

City	Debt Payments
Washington, D.C.	\$1,285
Seattle	1,135
⋮	⋮
Pittsburgh	763

SOURCE: www.Experian.com, November 11, 2010.

## 8.3 CONFIDENCE INTERVAL FOR THE POPULATION PROPORTION

### LO 8.6

Calculate a confidence interval for the population proportion.

Sometimes the parameter of interest describes a population that is qualitative rather than quantitative. Recall that while the population mean  $\mu$  and the population variance  $\sigma^2$  describe quantitative data, the population proportion  $p$  is the essential descriptive measure when the data type is qualitative. The parameter  $p$  represents the proportion of successes in the population, where success is defined by a particular outcome. Examples of population proportions include the proportion of women students at a university, the proportion of defective items in a manufacturing process, and the default probability on a mortgage loan.

As in the case of the population mean, we estimate the population proportion on the basis of its sample counterpart. In particular, we use the sample proportion  $\bar{P}$  as the point estimator of the population proportion  $p$ . Also, although the sampling distribution of  $\bar{P}$  is based on a binomial distribution, we can approximate it by a normal distribution for large samples, according to the central limit theorem. This approximation is valid when the sample size  $n$  is such that  $np \geq 5$  and  $n(1 - p) \geq 5$ .

Using the normal approximation for  $\bar{P}$  with  $E(\bar{P}) = p$  and  $se(\bar{P}) = \sqrt{p(1 - p)/n}$ , and analogous to the derivation of the confidence interval for the population mean, a  $100(1 - \alpha)\%$  confidence interval for the population proportion is

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \quad \text{or} \quad \left[ \bar{p} - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}, \bar{p} + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right].$$

This confidence interval is theoretically sound; however, it cannot be implemented because it uses  $p$  in the derivation, which is unknown. Since we always use large samples for the normal distribution approximation, we can also conveniently replace  $p$  with its estimate  $\bar{p}$  in the construction of the interval. Therefore, for  $\sqrt{\frac{p(1-p)}{n}}$ , we substitute  $\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$ . This substitution yields a feasible confidence interval for the population proportion.

#### CONFIDENCE INTERVAL FOR $p$

A  $100(1 - \alpha)\%$  confidence interval for the population proportion  $p$  is computed as

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad \text{or} \quad \left[ \bar{p} - z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \right]$$

This formula is valid only if  $\bar{P}$  (approximately) follows a normal distribution.

The normality condition is evaluated at the sample proportion  $\bar{p}$ . In other words, for constructing a confidence interval for the population proportion  $p$ , we require that  $n\bar{p} \geq 5$  and  $n(1 - \bar{p}) \geq 5$ .

#### EXAMPLE 8.7

In the introductory case of this chapter, Jared Beane wants to estimate the proportion of all ultra-green cars that obtain over 100 mpg. Use the information in Table 8.1 to construct 90% and 99% confidence intervals for the population proportion.

**SOLUTION:** As shown in Table 8.1, 7 of the 25 cars obtain over 100 mpg; thus, the point estimate of the population proportion is  $\bar{p} = 7/25 = 0.28$ . Note that the normality condition is satisfied since  $np \geq 5$  and  $n(1 - p) \geq 5$ , where  $p$  is evaluated at  $\bar{p} = 0.28$ . With a 90% confidence level,  $\alpha/2 = 0.10/2 = 0.05$ ; thus, we find  $z_{\alpha/2} = z_{0.05} = 1.645$ . Substituting the appropriate values into  $\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$  yields

$$0.28 \pm 1.645 \sqrt{\frac{0.28(1-0.28)}{25}} = 0.28 \pm 0.148.$$

With 90% confidence, Jared reports that the percentage of cars that obtain over 100 mpg is between 13.2% and 42.8%.

If Jared had wanted a 99% confidence level, we would use  $\alpha/2 = 0.01/2 = 0.005$  and  $z_{\alpha/2} = z_{0.005} = 2.576$  to obtain

$$0.28 \pm 2.576 \sqrt{\frac{0.28(1-0.28)}{25}} = 0.28 \pm 0.231.$$

At a higher confidence level of 99%, the interval for the percentage of cars that obtain over 100 mpg becomes 4.9% to 51.1%. Given the current sample size of 25 cars, Jared can gain confidence (from 90% to 99%) at the expense of precision, as the corresponding margin of error increases from 0.148 to 0.231.

## EXERCISES 8.3

### Mechanics

39. A random sample of 80 observations results in 50 successes.
- Construct the 95% confidence interval for the population proportion of successes.
  - Construct the 95% confidence interval for the population proportion of failures.
40. Assume  $\bar{p} = 0.6$  in a sample of size  $n = 50$ .
- Construct the 95% confidence interval for the population proportion.
  - What happens to the margin of error if the above sample proportion is based on  $n = 200$  instead of  $n = 50$ ?
41. A sample of 80 results in 30 successes.
- Calculate the point estimate for the population proportion of successes.
  - Construct 90% and 99% confidence intervals for the population proportion.
  - Can we conclude at 90% confidence that the population proportion differs from 0.5?
  - Can we conclude at 99% confidence that the population proportion differs from 0.5?
42. A random sample of 100 observations results in 40 successes.
- What is the point estimate for the population proportion of successes?
  - Construct 90% and 99% confidence intervals for the population proportion.
  - Can we conclude at 90% confidence that the population proportion differs from 0.5?
  - Can we conclude at 99% confidence that the population proportion differs from 0.5?
43. In a sample of 30 observations the number of successes equals 18.
- Construct the 88% confidence interval for the population proportion of successes.
  - Construct the 98% confidence interval for the population proportion of successes.
  - What happens to the margin of error as you move from an 88% confidence interval to a 98% confidence interval?
- population proportion of people who would pay off debts with an unexpected tax refund.
46. In a CNNMoney.com poll conducted on July 13, 2010, a sample of 5,324 Americans were asked about what matters most to them in a place to live. Thirty-seven percent of the respondents felt job opportunities matter most.
- Construct the 90% confidence interval for the proportion of Americans who feel that good job opportunities matter most in a place to live.
  - Construct the 99% confidence interval for the proportion of Americans who feel that good job opportunities matter most in a place to live.
  - Which of the above two intervals has a higher margin of error? Explain why.
47. An economist reports that 560 out of a sample of 1,200 middle-income American households actively participate in the stock market.
- Construct the 90% confidence interval for the proportion of middle-income Americans who actively participate in the stock market.
  - Can we conclude that the proportion of middle-income Americans who actively participate in the stock market is not 50%?
48. In an *NBC News/Wall Street Journal* poll of 1,000 American adults conducted August 5–9, 2010, 44% of respondents approved of the job that Barack Obama was doing in handling the economy.
- Compute the 90% confidence interval for the proportion of Americans who approved of Barack Obama's handling of the economy.
  - What is the resulting margin of error?
  - Compute the margin of error associated with the 99% confidence level.
49. In a recent poll of 760 homeowners in the United States, one in five homeowners reports having a home equity loan that he or she is currently paying off. Using a confidence coefficient of 0.90, derive the interval estimate for the proportion of all homeowners in the United States that hold a home equity loan.
50. Obesity is generally defined as 30 or more pounds over a healthy weight. A recent study of obesity reports 27.5% of a random sample of 400 adults in the United States to be obese.
- Use this sample information to compute the 90% confidence interval for the adult obesity rate in the United States.
  - Is it reasonable to conclude with 90% confidence that the adult obesity rate in the United States differs from 30%?
51. An accounting professor is notorious for being stingy in giving out good letter grades. In a large section of 140 students in the fall semester, she gave out only 5% A's, 23% B's, 42% C's, and 30% D's and F's. Assuming that this was

### Applications

44. A recent poll of 1,079 adults finds that 51% of Americans support Arizona's stringent new immigration enforcement law, even though it may lead to racial profiling (*The New York Times/CBS News*, April 28–May 2, 2010). Use the sample information to compute the 95% confidence interval for the population parameter of interest.
45. A survey of 1,026 people asked: "What would you do with an unexpected tax refund?" Forty-seven percent responded that they would pay off debts (*Vanity Fair*, June 2010).
- At 95% confidence, what is the margin of error?
  - Construct the 95% confidence interval for the

a representative class, compute the 95% confidence interval of the probability of getting at least a B from this professor.

52. A survey conducted by CBS News asked 1,026 respondents: "What would you do with an unexpected tax refund?" The responses are summarized in the following table.

Response	Frequency
Pay off debts	482
Put it in the bank	308
Spend it	112
I never get a refund	103
Other	21

SOURCE: *Vanity Fair*, June 2010.

- Construct the 90% confidence interval for the population proportion of those who would put the tax refund in the bank.
  - Construct the 90% confidence interval for the population proportion of those who never get a refund.
53. A recent survey asked 5,324 individuals: What's most important to you when choosing where to live? The responses are shown by the following frequency distribution.

Response	Frequency
Good jobs	1,969
Affordable homes	799
Top schools	586
Low crime	1,225
Things to do	745

SOURCE: CNNMoney.com, July 13, 2010.

- Calculate the margin of error used in the 95% confidence level for the population proportion of those who believe that low crime is most important.
  - Calculate the margin of error used in the 95% confidence level for the population proportion of those who believe that good jobs or affordable homes are most important.
  - Explain why the margins of error in parts a and b are different.
54. One in five 18-year-old Americans has not graduated from high school (*The Wall Street Journal*, April 19, 2007). A mayor of a northeastern city comments that its residents do not have the same graduation rate as the rest of the country. An analyst from the Department of Education decides to test the mayor's claim. In particular, she draws a random sample of 80 18-year-olds in the city and finds that 20 of them have not graduated from high school.
- Compute the point estimate for the proportion of 18-year-olds who have not graduated from high school in this city.
  - Use this point estimate to derive the 95% confidence interval for the population proportion.
  - Can the mayor's comment be justified at 95% confidence?

## 8.4 SELECTING THE REQUIRED SAMPLE SIZE

LO 8.7

So far we have discussed how a confidence interval provides useful information on an unknown population parameter. We compute the confidence interval by adding and subtracting the margin of error to/from the point estimate. If the margin of error is very large, the confidence interval becomes too wide to be of much value. For instance, little useful information can be gained from a confidence interval that suggests that the average annual starting salary of a business graduate is between \$16,000 and \$64,000. Similarly, an interval estimate that 10% to 60% of business students pursue an MBA is not very informative.

Statisticians like precision in their interval estimates, which is implied by a low margin of error. If we are able to increase the size of the sample, the larger  $n$  reduces the margin of error for the interval estimates. Although a larger sample size improves precision, it also entails the added cost in terms of time and money. Before getting into data collection, it is important that we first decide on the sample size that is adequate for what we wish to accomplish. In this section, we examine the required sample size, for a desired margin of error, in the confidence intervals for the population mean  $\mu$  and the population proportion  $p$ . In order to be conservative, we always round up non-integer values for the required sample size.

Select a sample size to estimate the population mean and the population proportion.

### Selecting $n$ to Estimate $\mu$

Consider a confidence interval for  $\mu$  with a known population standard deviation  $\sigma$ . In addition, let  $E$  denote the desired margin of error. In other words, you do not want the sample mean to deviate from the population mean by more than  $E$ , for a given level of confidence.

Since  $E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ , we rearrange this equation to derive the formula for the required sample size as  $n = \left( \frac{z_{\alpha/2} \sigma}{E} \right)^2$ . The sample size can be computed if we specify the population standard deviation  $\sigma$ , the value of  $z_{\alpha/2}$  based on the confidence level  $100(1 - \alpha)\%$ , and the desired margin of error  $E$ .

This formula is based on a knowledge of  $\sigma$ . However, in most cases  $\sigma$  is not known and, therefore, has to be estimated. Note that the sample standard deviation  $s$  cannot be used as an estimate for  $\sigma$  because  $s$  can be computed only after a sample of size  $n$  has been selected. In such cases, we replace  $\sigma$  with its reasonable estimate  $\hat{\sigma}$ .

#### THE REQUIRED SAMPLE SIZE WHEN ESTIMATING THE POPULATION MEAN

For a desired margin of error  $E$ , the minimum sample size  $n$  required to estimate a  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  is

$$n = \left( \frac{z_{\alpha/2} \hat{\sigma}}{E} \right)^2,$$

where  $\hat{\sigma}$  is a reasonable estimate of  $\sigma$  in the planning stage.

If  $\sigma$  is known, we replace  $\hat{\sigma}$  with  $\sigma$ . Sometimes we use the sample standard deviation from a preselected sample as  $\hat{\sigma}$  in the planning stage. Another choice for  $\hat{\sigma}$  is to use an estimate of the population standard deviation from prior studies. Finally, if the lowest and highest possible values of the population are available, a rough approximation for the population standard deviation is given by  $\hat{\sigma} = \text{range}/4$ .

#### EXAMPLE 8.8

Let us revisit Example 8.5, where Jared Beane wants to construct the 90% confidence interval for the mean mpg of all ultra-green cars. Suppose Jared would like to constrain the margin of error to within 2 mpg. Further, Jared knows that the lowest mpg in the population is 76 mpg, whereas the highest is 118 mpg. How large a sample does Jared need to compute the 90% confidence interval for the population mean?

**SOLUTION:** For the 90% confidence level, Jared computes  $z_{\alpha/2} = z_{0.05} = 1.645$ . He estimates the population standard deviation as  $\hat{\sigma} = \text{range}/4 = (118 - 76)/4 = 10.50$ . Given  $E = 2$ , the required sample size is

$$n = \left( \frac{z_{\alpha/2} \hat{\sigma}}{E} \right)^2 = \left( \frac{1.645 \times 10.50}{2} \right)^2 = 74.58,$$

which is rounded up to 75. Therefore, Jared needs a random sample of at least 75 ultra-green cars to provide a more precise interval estimate of the mean mpg.

### Selecting $n$ to Estimate $p$

The margin of error  $E$  for the confidence interval for the population proportion  $p$  is  $E = z_{\alpha/2} \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$ , where  $\bar{p}$  represents the sample proportion. By rearranging, we derive the formula for the required sample size as  $n = \left( \frac{z_{\alpha/2}}{E} \right)^2 \bar{p}(1 - \bar{p})$ . Analogous to the case of the population mean, this formula is not feasible because it uses  $\bar{p}$ , which cannot be computed unless a sample of size  $n$  has already been selected. We replace  $\bar{p}$  with a reasonable estimate  $\hat{p}$  of the population proportion  $p$ .



### THE REQUIRED SAMPLE SIZE WHEN ESTIMATING THE POPULATION PROPORTION

For a desired margin of error  $E$ , the minimum sample size  $n$  required to estimate a  $100(1 - \alpha)\%$  confidence interval for the population proportion  $p$  is

$$n = \left( \frac{z_{\alpha/2}}{E} \right)^2 \hat{p}(1 - \hat{p}),$$

where  $\hat{p}$  is a reasonable estimate of  $p$  in the planning stage.

Sometimes we use the sample proportion from a preselected sample as  $\hat{p}$  in the planning stage. Another choice for  $\hat{p}$  is to use an estimate of the population proportion from prior studies. If no other reasonable estimate of the population proportion is available, we can use  $\hat{p} = 0.5$  as a conservative estimate to derive the optimal sample size; note that the required sample is the largest when  $\hat{p} = 0.5$ .

#### EXAMPLE 8.9

Let us revisit Example 8.7, where Jared Beane wants to construct the 90% confidence interval for the proportion of all ultra-green cars that obtain over 100 mpg. Jared does not want the margin of error to be more than 0.10. How large a sample does Jared need for his analysis of the population proportion?

**SOLUTION:** For the 90% confidence level, Jared computes  $z_{\alpha/2} = z_{0.05} = 1.645$ . Since no estimate for the population proportion is readily available, Jared uses a conservative estimate of  $\hat{p} = 0.50$ . Given  $E = 0.10$ , the required sample size is

$$n = \left( \frac{z_{\alpha/2}}{E} \right)^2 \hat{p}(1 - \hat{p}) = \left( \frac{1.645}{0.10} \right)^2 0.50(1 - 0.50) = 67.65,$$

which is rounded up to 68. Therefore, Jared needs to find another random sample of at least 68 ultra-green cars to provide a more precise interval estimate for the proportion of all ultra-green cars that obtain over 100 mpg.

## SYNOPSIS OF INTRODUCTORY CASE

Jared Beane, an analyst at a research firm, prepares to write a report on the new ultra-green car that boasts an average of 100 mpg. Based on a sample of 25 cars, Jared reports with 90% confidence that the average mpg of all ultra-green cars is between 92.86 mpg and 100.18 mpg. Jared also constructs the 90% confidence interval for the proportion of cars that obtain more than 100 mpg and reports the interval between 0.132 and 0.428. Jared wishes to increase the precision of his confidence intervals by reducing the margin of error. If his desired margin of error is 2 mpg for the population mean, he must use a sample of at least 75 cars for the analysis. Jared also wants to reduce the margin of error to 0.10 for the proportion of cars that obtain more than 100 mpg. Using a conservative estimate, he calculates that a sample of at least 68 cars is needed to achieve this goal. Thus, in order to gain precision in the interval estimate for both the mean and the proportion with 90% confidence, Jared's sample must contain at least 75 cars.



## EXERCISES 8.4

### Mechanics

55. The lowest and highest observations in a population are 20 and 80, respectively. What is the minimum sample size  $n$  required to estimate  $\mu$  with 80% confidence if the desired margin of error is  $E = 2.6$ ? What happens to  $n$  if you decide to estimate  $\mu$  with 95% confidence?
56. Find the required sample size for estimating the population mean in order to be 95% confident that the sample mean is within 10 units of the population mean. Assume that the population standard deviation is 40.
57. You need to compute a 99% confidence interval for the population mean. How large a sample should you draw to ensure that the sample mean does not deviate from the population mean by more than 1.2? (Use 6.0 as an estimate of the population standard deviation from prior studies.)
58. What is the minimum sample size  $n$  required to estimate  $\mu$  with 90% confidence if the desired margin of error is  $E = 1.2$ ? The population standard deviation is estimated as  $\hat{\sigma} = 3.5$ . What happens to  $n$  if the desired margin of error decreases to  $E = 0.7$ ?
59. What is the minimum sample size  $n$  required to estimate  $p$  with 95% confidence if the desired margin of error  $E = 0.08$ ? The population proportion is estimated as  $\hat{p} = 0.36$  from prior studies. What happens to  $n$  if the desired margin of error increases to  $E = 0.12$ ?
60. In the planning stage, a sample proportion is estimated as  $\hat{p} = 40/50 = 0.80$ . Use this information to compute the minimum sample size  $n$  required to estimate  $p$  with 99% confidence if the desired margin of error  $E = 0.12$ . What happens to  $n$  if you decide to estimate  $p$  with 90% confidence?
61. You wish to compute a 95% confidence interval for the population proportion. How large a sample should you draw to ensure that the sample proportion does not deviate from the population proportion by more than 0.06? No prior estimate for the population proportion is available.
62. minimum number of gas stations that she should include in her sample if she uses the standard deviation estimate of \$0.32, as reported in the popular press?
64. An analyst would like to construct 95% confidence intervals for the mean stock returns in two industries. Industry A is a high-risk industry with a known population standard deviation of 20.6%, whereas Industry B is a low-risk industry with a known population standard deviation of 12.8%.
  - a. What is the minimum sample size required by the analyst if she wants to restrict the margin of error to 4% for Industry A?
  - b. What is the minimum sample size required by the analyst if she wants to restrict the margin of error to 4% for Industry B?
  - c. Why do the above results differ if they use the same margin of error?
65. The manager of a pizza chain in Albuquerque, New Mexico, wants to determine the average size of their advertised 16-inch pizzas. She takes a random sample of 25 pizzas and records their mean and standard deviation as 16.10 inches and 1.8 inches, respectively. She subsequently computes the 95% confidence interval of the mean size of all pizzas as [15.36, 16.84]. However, she finds this interval to be too broad to implement quality control and decides to reestimate the mean based on a bigger sample. Using the standard deviation estimate of 1.8 from her earlier analysis, how large a sample must she take if she wants the margin of error to be under 0.5 inch?
66. The manager of a newly opened Target store wants to estimate the average expenditure of his customers. From a preselected sample, the standard deviation was determined to be \$18. The manager would like to construct the 95% confidence interval for the mean customer expenditure.
  - a. Find the appropriate sample size necessary to achieve a margin of error of \$5.
  - b. Find the appropriate sample size necessary to achieve a margin of error of \$3.

### Applications

62. Mortgage lenders often use FICO® scores to check the credit worthiness of consumers applying for real estate loans. In general, FICO scores range from 300 to 850 with higher scores representing a better credit profile. A lender in a Midwestern town would like to estimate the mean credit score of its residents. What is the required number of sample FICO scores needed if the lender does not want the margin of error to exceed 20, with 95% confidence?
63. An analyst from an energy research institute in California wishes to precisely estimate the 99% confidence interval for the average price of unleaded gasoline in the state. In particular, she does not want the sample mean to deviate from the population mean by more than \$0.06. What is the
67. A budget airline wants to estimate what proportion of customers would consider paying \$12 for in-flight wireless access. Given that the airline has no prior knowledge of the proportion, how many customers would it have to sample to ensure a margin of error of no more than 0.05 for a 90% confidence interval?
68. Newscasters wish to estimate the proportion of registered voters who support the incumbent candidate in the mayoral election. In an earlier poll of 240 registered voters, 110 had supported the incumbent candidate. Find the sample size required to construct the 90% confidence interval if newscasters do not want the margin of error to exceed 0.02.

69. A survey by the AARP (*Money*, June 2007) reported that approximately 70% of people in the 50 to 64 age bracket have tried some type of alternative therapy (for instance, acupuncture or the use of nutrition supplements). Assume this survey was based on a sample of 400 people.

- Identify the relevant parameter of interest for these qualitative data and compute its point estimate as well as the margin of error with 90% confidence.
- You decide to redo the analysis with the margin of error reduced to 2%. How large a sample do you need to draw? State your assumptions in computing the required sample size.

70. Subprime lending was big business in the United States in the mid-2000s, when lenders provided mortgages to people with poor credit. However, subsequent increases

in interest rates coupled with a drop in home values necessitated many borrowers to default. Suppose a recent report finds that two in five subprime mortgages are likely to default nationally. A research economist is interested in estimating default rates in Illinois with 95% confidence. How large a sample is needed to restrict the margin of error to within 0.06, using the reported national default rate?

71. A business student is interested in estimating the 99% confidence interval for the proportion of students who bring laptops to campus. He wishes a precise estimate and is willing to draw a large sample that will keep the sample proportion within five percentage points of the population proportion. What is the minimum sample size required by this student, given that no prior estimate of the population proportion is available?

## WRITING WITH STATISTICS

Callie Fitzpatrick, a research analyst with an investment firm, has been asked to write a report summarizing the weekly stock performance of Home Depot and Lowe's. Her manager is trying to decide whether or not to include one of these stocks in a client's portfolio and the average stock performance is one of the factors influencing this decision. Callie decides to use descriptive measures to summarize stock returns in her report, as well as provide confidence intervals for the average return for Home Depot and Lowe's. She collects weekly returns for each firm for the first eight months of 2010. A portion of the return data is shown in Table 8.5.



**TABLE 8.5** Weekly Returns (in percent) for Home Depot and Lowe's

**FILE**  
Weekly\_Returns

Date	Home Depot	Lowe's
1/11/2010	-1.44	-1.59
1/19/2010	-2.98	-3.53
⋮	⋮	⋮
8/30/2010	-2.61	-3.89

SOURCE: <http://finance.yahoo.com>.

Callie would like to use the sample information to:

- Summarize weekly returns for Home Depot and Lowe's.
- Provide confidence intervals for the average weekly returns.
- Make recommendations for further analysis.

## Sample Report— Weekly Stock Performance: Home Depot vs. Lowe's

Grim news continues to distress the housing sector. On August 24, 2010, Reuters reported that the sales of previously owned U.S. homes took a record plunge in July to the slowest pace in 15 years. Combine this fact with the continued fallout from the subprime mortgage debacle, a sluggish economy, and high unemployment, and the housing sector appears quite unstable. Have these unfavorable events managed to trickle down and harm the financial performance of Home Depot and Lowe's, the two largest home improvement retailers in the United States?

One way to analyze their financial stability is to observe their stock performance during this period. In order to make valid statements concerning the reward of holding these stocks, weekly return data for each firm were gathered from January through August of 2010. Table 8.A summarizes the important descriptive statistics.

**TABLE 8.A** Descriptive Statistics for Weekly Returns of Home Depot and Lowe's ( $n = 34$ )

	Home Depot	Lowe's
Mean	0.00%	−0.33%
Median	0.76%	−0.49%
Minimum	−8.08%	−7.17%
Maximum	5.30%	7.71%
Standard deviation	3.59%	3.83%
Margin of error with 95% confidence	1.25%	1.34%

Over the past 34 weeks, Home Depot posted both a higher average return and median return of 0.00% and 0.76%, respectively. Lowe's return over the same period was negative, whether the central tendency was measured by its mean (−0.33%) or its median (−0.49%). In terms of dispersion, Lowe's return data had the higher standard deviation (3.83% > 3.59%). In terms of descriptive measures, the investment in Home Depot's stock not only provided higher returns, but also was less risky than the investment in Lowe's stock.

Table 8.A also shows the margins of error for 95% confidence intervals for the mean returns. With 95% confidence, the mean return for Home Depot fell in the range [−1.25%, 1.25%], while that for Lowe's fell in the range [−1.67%, 1.01%]. Given that these two intervals overlap, one cannot conclude that Home Depot delivered the higher reward over this period—a conclusion one may have arrived at had only the point estimates been evaluated. It is not possible to recommend one stock over the other for inclusion in a client's portfolio based solely on the mean return performance. Other factors, such as the correlation between the stock and the existing portfolio, must be analyzed before this decision can be made.

## CONCEPTUAL REVIEW

### LO 8.1 Explain an interval estimator.

The sample mean  $\bar{X}$  is the point estimator for the population mean  $\mu$ , and the sample proportion  $\bar{P}$  is the point estimator for the population proportion  $p$ . Sample values of the point estimators represent the point estimates for the population parameter of interest;  $\bar{x}$  and  $\bar{p}$  are the point estimates for  $\mu$  and  $p$ , respectively. While a point estimator provides a single value that approximates the unknown parameter, a **confidence interval**, or an

**interval estimate**, provides a range of values that, with a certain level of confidence, will contain the population parameter of interest.

Often, we construct a confidence interval as: point estimate  $\pm$  margin of error. The **margin of error** accounts for the variability of the estimator and the desired confidence level of the interval.

**LO 8.2 Calculate a confidence interval for the population mean when the population standard deviation is known.**

A  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  when the population standard deviation  $\sigma$  is known is computed as  $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ , where  $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  is the margin of error. This formula is valid only if  $\bar{X}$  (approximately) follows a normal distribution.

**LO 8.3 Describe the factors that influence the width of a confidence interval.**

The **precision** of a confidence interval is directly linked with the **width** of the interval: the wider the interval, the lower is its precision. A confidence interval is wider (a) the greater the population standard deviation  $\sigma$ , (b) the smaller the sample size  $n$ , and (c) the greater the confidence level.

**LO 8.4 Discuss features of the  $t$  distribution.**

The  **$t$  distribution** is a family of distributions that are similar to the  $z$  distribution, in that they are all symmetric and bell-shaped around zero with asymptotic tails. However, the  $t$  distribution has broader tails than does the  $z$  distribution. Each  $t$  distribution is identified by a parameter known as the **degrees of freedom  $df$** . The  $df$  determine the extent of broadness—the smaller the  $df$ , the broader the tails. Since the  $t$  distribution is defined by the degrees of freedom, it is common to refer to it as the  $t_{df}$  distribution.

**LO 8.5 Calculate a confidence interval for the population mean when the population standard deviation is not known.**

A  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  when the population standard deviation  $\sigma$  is not known is computed as  $\bar{x} \pm t_{\alpha/2, df} \frac{s}{\sqrt{n}}$ , where  $s$  is the sample standard deviation. This formula is valid only if  $\bar{X}$  (approximately) follows a normal distribution.

**LO 8.6 Calculate a confidence interval for the population proportion.**

A  $100(1 - \alpha)\%$  confidence interval for the population proportion  $p$  is computed as  $\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$ , where  $\bar{p}$  is the sample proportion. This formula is valid only if  $\bar{P}$  (approximately) follows a normal distribution.

**LO 8.7 Select a sample size to estimate the population mean and the population proportion.**

For a desired margin of error  $E$ , the minimum  $n$  required to estimate  $\mu$  with  $100(1 - \alpha)\%$  confidence is  $n = \left( \frac{z_{\alpha/2} \hat{\sigma}}{E} \right)^2$ , where  $\hat{\sigma}$  is a reasonable estimate of  $\sigma$  in the planning stage. If  $\sigma$  is known, we replace  $\hat{\sigma}$  with  $\sigma$ . Other choices for  $\hat{\sigma}$  include an estimate from a preselected sample, prior studies, or  $\hat{\sigma} = \text{range}/4$ .

For a desired margin of error  $E$ , the minimum  $n$  required to estimate  $p$  with  $100(1 - \alpha)\%$  confidence is  $n = \left( \frac{z_{\alpha/2}}{E} \right)^2 \hat{p}(1 - \hat{p})$ , where  $\hat{p}$  is a reasonable estimate of  $p$  in the planning stage. Choices for  $\hat{p}$  include an estimate from a preselected sample or prior studies; a conservative estimate of  $\hat{p} = 0.5$  is used when no other reasonable estimate is available.



# ADDITIONAL EXERCISES AND CASE STUDIES

## Exercises

72. Over a 10-year sample period, the mean and the standard deviation of annual returns on a portfolio you are analyzing were 10% and 15%, respectively. You assume that returns are normally distributed. Construct the 95% confidence interval for the population mean.
73. A hair salon in Cambridge, Massachusetts, reports that on seven randomly selected weekdays, the number of customers who visited the salon were 40, 30, 28, 22, 36, 16, and 50. It can be assumed that weekday customer visits follow a normal distribution.
- Construct the 90% confidence interval for the average number of customers who visit the salon on weekdays.
  - Construct the 99% confidence interval for the average number of customers who visit the salon on weekdays.
  - What happens to the width of the interval as the confidence level increases?
74. According to data from the Organization for Economic Cooperation and Development, the average U.S. worker takes 16 days of vacation each year (*The Wall Street Journal*, June 20, 2007). Assume that these data were based on a sample of 225 workers and that the sample standard deviation is 12 days.
- Construct the 95% confidence interval for the population mean.
  - At the 95% confidence level, can we conclude that the average U.S. worker does not take 14 days of vacation each year?
75. Recently, six single-family homes in San Luis Obispo County in California sold at the following prices (in \$1,000s): 549, 449, 705, 529, 639, 609.
- Construct the 95% confidence interval for the mean sale price in San Luis Obispo County.
  - What assumption have you made when constructing this confidence interval?
76. Students who graduated from college in 2010 owed an average of \$25,250 in student loans (*The New York Times*, November 2, 2011). An economist wants to determine if average debt has changed. She takes a sample of 40 recent graduates and finds that their average debt was \$27,500 with a standard deviation of \$9,120. Use the 90% confidence interval to determine if average debt has changed.
77. A machine that is programmed to package 1.20 pounds of cereal is being tested for its accuracy.

In a sample of 36 cereal boxes, the sample mean filling weight is calculated as 1.22 pounds. The population standard deviation is known to be 0.06 pound.

- Identify the relevant parameter of interest for these quantitative data and compute its point estimate as well as the margin of error with 95% confidence.
  - Can we conclude that the packaging machine is operating improperly?
  - How large a sample must we take if we want the margin of error to be at most 0.01 pound with 95% confidence?
78. The SAT is the most widely used test in the undergraduate admissions process. Scores on the math portion of the SAT are believed to be normally distributed and range from 200 to 800. A researcher from the admissions department at the University of New Hampshire is interested in estimating the mean math SAT scores of the incoming class with 90% confidence. How large a sample should she take to ensure that the margin of error is below 15?
79. A recent study by Allstate Insurance Co. finds that 82% of teenagers have used cell phones while driving (*The Wall Street Journal*, May 5, 2010). Suppose this study was based on a random sample of 50 teen drivers.
- Construct the 99% confidence interval for the proportion of all teenagers that have used cell phones while driving.
  - What is the margin of error with 99% confidence?
80. The following table shows the annual returns (in percent) for the Vanguard Energy Fund.

Year	Return
2010	13
2011	-2
2012	3
2013	18
2014	-14

Source: <http://finance.yahoo.com>, data retrieved April 4, 2015.

- Calculate the point estimate for  $\mu$ .
  - Construct the 95% confidence interval for  $\mu$ .
  - What assumption did you make when constructing the interval?
81. **FILE MV\_Houses.** A realtor wants to estimate the mean price of houses in Mission Viejo, California.



She collects a sample of 36 recent house sales, a portion of which is shown in the accompanying table. Assume that the population standard deviation is 100 (in \$1,000s). Construct and interpret 95% and 98% confidence intervals for the mean price of all houses in Mission Viejo, CA.

Prices (in \$1,000s)
430
520
⋮
430

82. **FILE MI\_Life\_Expectancy.** Residents of Hawaii have the longest life expectancies, averaging 81.48 years ([www.worldlifeexpectancy.com](http://www.worldlifeexpectancy.com), data retrieved June 4, 2012). A sociologist collects data on the age at death for 50 recently deceased Michigan residents. A portion of the data is shown in the accompanying table. Assume that the population standard deviation is 5 years.

Age at Death
76.4
76.0
⋮
73.6

- Use Excel to construct the 95% confidence interval for the mean life expectancy of all residents of Michigan.
  - Use the 95% confidence interval to determine if the mean life expectancy of Michigan residents differs from that for Hawaii residents.
83. **FILE Fastballs.** The manager of a minor league baseball team wants to estimate the average fastball speed of two pitchers. He clocks 50 fastballs, in miles per hour, for each pitcher. A portion of the data is shown in the accompanying table.

Pitcher 1	Pitcher 2
87	82
86	92
⋮	⋮
86	93

- Use Excel to construct 95% confidence intervals for the mean speed for each pitcher.
  - Explain why the widths of the two intervals are different.
84. **FILE Theater.** The new manager of a theater would like to offer discounts to increase the number of tickets sold for shows on Monday and Tuesday evenings. She uses a sample of 30 weeks to record the

number of tickets sold on these two days. A portion of the data is shown in the accompanying table.

Monday	Tuesday
221	208
187	199
⋮	⋮
194	180

- Use Excel to compare the margin of error for the 95% confidence intervals for the mean number of tickets sold for shows on Monday and Tuesday evenings.
  - Use Excel to construct the 95% confidence intervals for the mean number of tickets sold for shows on Monday and Tuesday evenings.
  - Determine if the population mean differs from 200 for shows on Monday and Tuesday evenings.
85. **FILE Ann Arbor\_Rental.** Real estate investment in college towns continues to promise good returns (*The Wall Street Journal*, September 24, 2010). Marcela Treisman works for an investment firm in Michigan. Her assignment is to analyze the rental market in Ann Arbor, which is home to the University of Michigan. She gathers data on monthly rents for 2011 along with the square footage of 40 homes. A portion of the data is shown in the accompanying table.

Monthly Rent	Square Footage
645	500
675	648
⋮	⋮
2400	2700

SOURCE: [www.zillow.com](http://www.zillow.com).

- Use Excel to construct 90% and 95% confidence intervals for the mean rent for all rental homes in Ann Arbor, Michigan.
  - Use Excel to construct 90% and 95% confidence intervals for the mean square footage for all rental homes in Ann Arbor, Michigan.
86. According to a survey of 1,235 businesses by IDC, a market-research concern in Framingham, Massachusetts, 12.1% of sole proprietors are engaging in e-commerce (*The Wall Street Journal*, July 26, 2007).
- With 95% confidence, what is the margin of error when estimating the proportion of sole proprietors that engage in e-commerce?
  - Construct the 95% confidence interval for the population proportion.

87. A Monster.com poll of 3,057 individuals asked: “What’s the longest vacation you plan to take this summer?” The following relative frequency distribution summarizes the results.

Response	Relative Frequency
A few days	0.21
A few long weekends	0.18
One week	0.36
Two weeks	0.22

SOURCE: *The Boston Globe*, June 12, 2007.

- Construct the 95% confidence interval for the proportion of people who plan to take a one-week vacation this summer.
  - Construct the 99% confidence interval for the proportion of people who plan to take a one-week vacation this summer.
  - Which of the two confidence intervals is wider?
88. Linda Barnes has learned from prior studies that one out of five applicants gets admitted to top MBA programs in the country. She wishes to construct her own 90% confidence interval for the acceptance rate in top MBA programs. How large a sample should she take if she does not want the acceptance rate of the sample to deviate from that of the population by more than five percentage points? State your assumptions in computing the required sample size.
89. **FILE Field Choice.** There is a declining interest among teenagers to pursue a career in science and health care (*U.S. News & World Report*, May 23, 2011). Thirty college-bound students in Portland, Oregon, are asked about the field they would like to pursue in college. The choices offered in the questionnaire are science, business, and other. The gender information also is included in the questionnaire. A portion of the data is shown below.

Field Choice	Gender
Business	Male
Other	Female
:	:
Science	Female

- Compare the 95% confidence interval for the proportion of students who would like to pursue science with the proportion who would like to pursue business.
  - Construct and interpret the 90% confidence interval for the proportion of female students who are college bound.
90. **FILE Pedestrians.** A recent study examined “sidewalk rage” in an attempt to find insight into anger’s origins and offer suggestions for anger-management treatments (*The Wall Street Journal*,

February 15, 2011). “Sidewalk ragers” tend to believe that pedestrians should behave in a certain way. One possible strategy for sidewalk ragers is to avoid walkers who are distracted by other activities such as smoking and tourism. Sample data were obtained from 50 pedestrians in Lower Manhattan. It was noted if the pedestrian was smoking (equaled 1 if smoking, 0 otherwise) or was a tourist (equaled 1 if tourist, 0 otherwise). The accompanying table shows a portion of the data.

Smoking	Tourist
0	1
0	1
:	:
0	0

- Construct and interpret the 95% confidence interval for the proportion of pedestrians in Lower Manhattan who smoke while walking.
  - Construct and interpret the 95% confidence interval for the proportion of pedestrians in Lower Manhattan who are tourists.
91. An economist would like to estimate the 95% confidence interval for the average real estate taxes collected by a small town in California. In a prior analysis, the standard deviation of real estate taxes was reported as \$1,580. What is the minimum sample size required by the economist if he wants to restrict the margin of error to \$500?
92. An employee of the Bureau of Transportation Statistics has been given the task of estimating the proportion of on-time arrivals of a budget airline. A prior study had estimated this on-time arrival rate as 78.5%. What is the minimum number of arrivals this employee must include in the sample to ensure that the margin of error for a 95% confidence interval is no more than 0.05?
93. According to a recent report by the PEW Research Center, 85% of adults under 30 feel optimistic about the economy, but the optimism is shared by only 45% of those who are over 50 (*Newsweek*, September 13, 2010). A research analyst would like to construct 95% confidence intervals for the proportion patterns in various regions of the country. She uses the reported rates by the PEW Research Center to determine the sample size that would restrict the margin of error to within 0.05.
- How large a sample is required to estimate the proportion of adults under 30 who feel optimistic about the economy?
  - How large a sample is required to estimate the proportion of adults over 50 who feel optimistic about the economy?

## CASE STUDIES

**CASE STUDY 8.1** Texas is home to more than one million undocumented immigrants, and most of them are stuck in low-paying jobs. Meanwhile, the state also suffers from a lack of skilled workers. The Texas Workforce Commission estimates that 133,000 jobs are currently unfilled, many because employers cannot find qualified applicants (*The Boston Globe*, September 29, 2011). Texas was the first state to pass a law that allows children of undocumented immigrants to pay in-state college tuition rates if they have lived in Texas for three years and plan to become permanent residents. The law passed easily back in 2001 because most legislators believed that producing college graduates and keeping them in Texas benefits the business community. In addition, since college graduates earn more money, they also provide the state with more revenue. Carol Capaldo wishes to estimate the mean hourly wage of workers with various levels of education. She collects a sample of the hourly wages of 30 Texas workers with a bachelor's degree or higher, 30 Texas workers with only a high school diploma, and 30 Texas workers who did not finish high school. A portion of the data is shown in the accompanying table.

**Data for Case Study 8.1** Hourly Wages of Texas Workers by Education Level (in \$)

Bachelor's Degree or Higher	High School Diploma	No High School Diploma
\$22.50	\$12.68	\$11.21
19.57	11.23	8.54
:	:	:
21.44	7.47	10.27

**FILE**  
Texas\_Wages

In a report, use the above information to:

1. Use descriptive statistics to compare the hourly wages for the three education levels.
2. Construct and interpret 95% confidence intervals for the mean hourly wage at each education level.

**CASE STUDY 8.2** The following table presents a portion of the annual returns for two mutual funds offered by the investment giant Fidelity. The *Fidelity Select Automotive Fund* invests primarily in companies engaged in the manufacturing, marketing, or sales of automobiles, trucks, specialty vehicles, parts, tires, and related services. The *Fidelity Gold Fund* invests primarily in companies engaged in exploration, mining, processing, or dealing in gold and, to a lesser degree, in other precious metals and minerals.

**Data for Case Study 8.2** Annual Total Return (%) History

**FILE**  
Fidelity\_Returns

Year	Annual Total Return (%) History	
	Fidelity Select Automotive Fund	Fidelity Select Gold Fund
2001	22.82	24.99
2002	-6.48	64.28
:	:	:
2009	122.28	38.00

SOURCE: <http://finance.yahoo.com>.

In a report, use the above information to:

1. Use descriptive statistics to compare the returns of the mutual funds.
2. Assess reward by constructing and interpreting 95% confidence intervals for the population mean return. What assumption did you make for the interval estimates?

**CASE STUDY 8.3** The information gathered from opinion polls and political surveys is becoming so increasingly important for candidates on the campaign trail that it is hard to imagine an election that lacks extensive polling. An NBC News/*Wall Street Journal* survey (August 5–9, 2010) of 1,000 adults asked people’s preferences on candidates and issues prior to the midterm 2010 elections. Some of the responses to the survey are shown below, as well as responses from prior surveys. (Copyright © 2010 Dow Jones & Co., Inc.)

*Question:* In general, do you approve or disapprove of the way Barack Obama is handling the aftermath of the Gulf Coast oil spill in August 2010 (and George W. Bush’s handling of Katrina in March 2006)?

	August 2010	March 2006
Approve	50%	36%
Disapprove	38%	53%
Not sure	12%	11%

*Question:* Which are more important to you in your vote for Congress this November: domestic issues such as the economy, health care, and immigration; or international issues such as Afghanistan, Iran, and terrorism?

	August 2010	September 2006
Domestic issues	73%	43%
International issues	12%	28%
Both equally important	15%	28%

In a report, construct 95% confidence intervals for the relevant population proportions to:

1. Compare the approval rates of President Obama’s handling of the Gulf Coast oil spill and President George W. Bush’s handling of the Hurricane Katrina crisis.
2. Compare the importance of domestic issues in August 2010 and in September 2006.

## APPENDIX 8.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill’s Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### Estimating $\mu$ , $\sigma$ Known

- A. (Replicating Example 8.3) From the menu choose **Stat > Basic Statistics > 1-Sample Z**.
- B. Select **Samples in columns**, and then select Weight. Choose **Options**. Enter 92.0 for **Confidence Level**.

#### Estimating $\mu$ , $\sigma$ Unknown

- A. (Replicating Example 8.6) From the menu choose **Stat > Basic Statistics > 1-Sample t**.
- B. Select **Samples in columns**, and then select Expenditures. Choose **Options**. Enter 95.0 for **Confidence Level**.

**FILE**

Hockey\_Puck

**FILE**

Lottery

## Estimating $p$

- A. (Replicating Example 8.7) From the menu choose **Stat > Basic Statistics > 1-Proportion**.
- B. Select **Summarized data**. Enter 7 for **Number of events** and 25 for **Number of trials**. Choose **Options**. Enter 90.0 for **Confidence Level** and check **Use test and interval based on normal distribution**.

## SPSS

### Estimating $\mu, \sigma$ Unknown

(Replicating Example 8.6) From the menu choose **Analyze > Compare Means > One-Sample T Test**. Under **Test Variable(s)**, select Expenditures. Choose **Options**. After **Confidence Interval Percentage** enter 95.

**FILE**  
*Lottery*

## JMP

### Estimating $\mu, \sigma$ Known

- A. (Replicating Example 8.3) From the menu choose **Analyze > Distribution**.
- B. Under **Select Columns**, select Weight, and then under **Cast Selected Columns into Roles**, select **Y, Columns**.
- C. Click on the red triangle in the output window beside Weight. Choose **Confidence Interval > Other**, and after **Enter (1-alpha for Confidence level)**, enter 0.92. Select **Use known sigma** and enter 7.5.

**FILE**  
*Hockey\_Puck*

### Estimating $\mu, \sigma$ Unknown

- A. (Replicating Example 8.6) From the menu choose **Analyze > Distribution**.
- B. Under **Select Columns**, select Expenditures, and then under **Cast Selected Columns into Roles**, select **Y, Columns**.
- C. Click on the red triangle in the output window beside Expenditures. Choose **Confidence Interval > 0.95**.

**FILE**  
*Lottery*

# 9

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 9.1 Define the null hypothesis and the alternative hypothesis.
- LO 9.2 Distinguish between Type I and Type II errors.
- LO 9.3 Conduct a hypothesis test using the  $p$ -value approach.
- LO 9.4 Conduct a hypothesis test using the critical value approach.
- LO 9.5 Differentiate between the test statistics for the population mean.
- LO 9.6 Specify the test statistic for the population proportion.

# Hypothesis Testing

In Chapter 8, we used confidence intervals to estimate an unknown population parameter of interest. In this chapter, we will focus on the second major area of statistical inference: hypothesis testing.

We use a hypothesis test to challenge the status quo, or some belief about an underlying population parameter, based on sample data. In particular, we develop hypothesis tests for the population mean and the population proportion. For instance, we may wish to test whether the average age of MBA students in the United States is less than 30 years or whether the proportion of defective items in a production process differs from 5%. In either case, since we do not have access to the entire population, we have to perform statistical inference on the basis of limited sample information. If the sample information is not consistent with the status quo, we use the hypothesis testing framework to determine if the inconsistency is real (that is, we contradict the status quo) or due to chance (that is, we do not contradict the status quo).





## INTRODUCTORY CASE

### Undergraduate Study Habits

Are today's college students studying hard or hardly studying? A recent study asserts that over the past five decades the number of hours that the average college student studies each week has been steadily dropping (*The Boston Globe*, July 4, 2010). In 1961, students invested 24 hours per week in their academic pursuits, whereas today's students study an average of 14 hours per week.

Susan Knight is a dean at a large university in California. She wonders if the study trend is reflective of the students at her university. She randomly selects 35 students and asks their average study time per week (in hours). The responses are shown in Table 9.1.

**FILE** **TABLE 9.1** Average Hours Studied per Week for a Sample of 35 College Students

Study_Hours	25	17	8	14	17	7	11
	19	16	9	15	12	17	19
	26	14	22	17	14	35	24
	11	21	6	20	27	17	6
	29	10	10	4	25	13	16

Summary measures:  $\bar{x} = 16.37$  hours and  $s = 7.22$  hours.

Susan wants to use the sample information to:

1. Determine if the mean study time of students at her university is below the 1961 national average of 24 hours per week.
2. Determine if the mean study time of students at her university differs from today's national average of 14 hours per week.

A synopsis of this case is provided at the end of Section 9.3.

Define the null hypothesis and the alternative hypothesis.

Every day people make decisions based on their beliefs about the true state of the world. They hold certain things to be true and others to be false, and then act accordingly. For example, an engineer believes that a certain steel cable has a breaking strength of 5,000 pounds or more, and then permits its use at a construction site; a manufacturer believes that a certain process yields capsules that contain precisely 100 milligrams of a drug, and then ships the capsules to a pharmacy; an agronomist believes that a new fertilizer increases soy bean production by more than 30%, and then switches to this new fertilizer; a manager believes that an incoming shipment contains less than 2% of defects, and then accepts the shipment. In these cases, and many more, the formation of these beliefs may have started as a mere conjecture, an informed guess, or a proposition tentatively advanced as true. When people formulate a belief in this way, we refer to it as a hypothesis. Sooner or later, however, every hypothesis eventually confronts evidence that either substantiates or refutes it. Determining the validity of an assumption of this nature is called hypothesis testing.

We use hypothesis testing to resolve conflicts between two competing hypotheses on a particular population parameter of interest. We refer to one hypothesis as the **null hypothesis**, denoted  $H_0$ , and the other as the **alternative hypothesis**, denoted  $H_A$ . We think of the null hypothesis as corresponding to a presumed default state of nature or status quo. The alternative hypothesis, on the other hand, contradicts the default state or status quo.

#### NULL HYPOTHESIS VERSUS ALTERNATIVE HYPOTHESIS

When constructing a hypothesis test, we define a **null hypothesis**, denoted  $H_0$ , and an **alternative hypothesis**, denoted  $H_A$ . We conduct a hypothesis test to determine whether or not sample evidence contradicts  $H_0$ .

In statistics, we use sample information to make inferences regarding the unknown population parameters of interest. In this chapter, our goal is to determine if the null hypothesis can be rejected in favor of the alternative hypothesis. An analogy can be drawn with applications in the medical and legal fields, where we can define the null hypothesis as “an individual is free of a particular disease” or “an accused is innocent.” In both cases, the verdict is based on limited evidence, which in statistics translates into making a decision based on limited sample information.

### The Decision to “Reject” or “Not Reject” the Null Hypothesis

The hypothesis testing procedure enables us to make one of two decisions. If sample evidence is inconsistent with the null hypothesis, we reject the null hypothesis. Conversely, if sample evidence is not inconsistent with the null hypothesis, then we do not reject the null hypothesis. It is not correct to conclude that “we accept the null hypothesis” because while the sample data may not be inconsistent with the null hypothesis, they do not necessarily prove that the null hypothesis is true.

On the basis of sample information, we either “**reject the null hypothesis**” or “**do not reject the null hypothesis**.”

Consider the example just referenced where the null is defined as “an individual is free of a particular disease.” Suppose a medical procedure does not detect this disease. On the basis of this limited information, we can only conclude that we are unable to detect the disease (do not reject the null hypothesis). It does not necessarily prove that the person does not have the disease (accept the null hypothesis). Similarly, in the court example where the null hypothesis

is defined as “an accused is innocent,” we can conclude that the person is guilty (reject the null hypothesis) or that there is not enough evidence to convict (do not reject the null hypothesis).

## Defining the Null and the Alternative Hypotheses

As mentioned earlier, we use a hypothesis test to contest the status quo, or some belief about an underlying population parameter, based on sample data. A very crucial step concerns the formulation of the two competing hypotheses, since the conclusion of the test depends on how the hypotheses are stated. As a general guideline, whatever we wish to establish is placed in the alternative hypothesis, whereas the null hypothesis includes the status quo. If we are unable to reject the null hypothesis, then we maintain the status quo or “business as usual.” However, if we reject the null hypothesis, this establishes that the evidence supports the alternative hypothesis, which may require that we take some kind of action. For instance, if we reject the null hypothesis that an individual is free of a particular disease, then we conclude that the person is sick, for which treatment may be prescribed. Similarly, if we reject that an accused is innocent, we conclude that the person is guilty and should be suitably punished.

In most applications, a requirement in hypothesis testing is that some form of the equality sign appears in the null hypothesis. (The justification for the equality sign will be provided later.) In general, any statement including one of the three signs “=”, “ $\leq$ ”, or “ $\geq$ ” is valid for the null hypothesis. Given that the alternative hypothesis states the opposite of the null hypothesis, the alternative hypothesis is then specified with a “ $\neq$ ”, “ $>$ ”, or “ $<$ ” sign.

As a general guideline, we use the alternative hypothesis as a vehicle to establish something new—that is, contest the status quo. In most applications, the null hypothesis regarding a particular population parameter of interest is specified with one of the following signs: =,  $\leq$ , or  $\geq$ ; the alternative hypothesis is then specified with the corresponding opposite sign:  $\neq$ ,  $>$ , or  $<$ .

A hypothesis test can be **one-tailed** or **two-tailed**. A two-tailed test is defined when the alternative hypothesis includes the sign “ $\neq$ ”. For example,  $H_0: \mu = \mu_0$  versus  $H_A: \mu \neq \mu_0$  and  $H_0: p = p_0$  versus  $H_A: p \neq p_0$  are examples of two-tailed tests, where  $\mu_0$  and  $p_0$  represent hypothesized values of the population mean and the population proportion, respectively. If the null hypothesis is rejected, it suggests that the true parameter does not equal the hypothesized value.

A one-tailed test, on the other hand, involves a null hypothesis that can only be rejected on one side of the hypothesized value. For example, consider  $H_0: \mu \leq \mu_0$  versus  $H_A: \mu > \mu_0$ . Here we can reject the null hypothesis only when there is substantial evidence that the population mean is greater than  $\mu_0$ . It is also referred to as a **right-tailed test** since rejection of the null hypothesis occurs on the right side of the hypothesized mean. Another example is a **left-tailed test**,  $H_0: \mu \geq \mu_0$  versus  $H_A: \mu < \mu_0$ , where the null hypothesis can only be rejected on the left side of the hypothesized mean. One-tailed tests for the population proportion are defined similarly.

### ONE-TAILED VERSUS TWO-TAILED HYPOTHESIS TESTS

Hypothesis tests can be **one-tailed** or **two-tailed**. In a **one-tailed test**, we can reject the null hypothesis only on one side of the hypothesized value of the population parameter. In a **two-tailed test**, we can reject the null hypothesis on either side of the hypothesized value of the population parameter.

In general, we follow three steps when formulating the competing hypotheses:

- Identify the relevant population parameter of interest.
- Determine whether it is a one- or a two-tailed test.
- Include some form of the equality sign in the null hypothesis and use the alternative hypothesis to establish a claim.

The following examples highlight one- and two-tailed tests for the population mean and the population proportion. In each example, we want to state the appropriate competing hypotheses.

### EXAMPLE 9.1

A trade group predicts that back-to-school spending will average \$606.40 per family this year. A different economic model is needed if the prediction is wrong. Specify the null and the alternative hypotheses to determine if a different economic model may be needed.

**SOLUTION:** Given that we are examining average back-to-school spending, the parameter of interest is the population mean. Since we want to be able to determine if the population mean differs from \$606.40 ( $\mu \neq 606.40$ ), we need a two-tailed test and formulate the null and alternative hypotheses as

$$H_0: \mu = 606.40$$

$$H_A: \mu \neq 606.40$$

The trade group is advised to use a different economic model if the null hypothesis is rejected.

### EXAMPLE 9.2

An advertisement for a popular weight-loss clinic suggests that participants in its new diet program lose, on average, more than 10 pounds. A consumer activist wants to determine if the advertisement's claim is valid. Specify the null and the alternative hypotheses to validate the advertisement's claim.

**SOLUTION:** The advertisement's claim concerns average weight loss; thus, the parameter of interest is again the population mean. This is an example of a one-tailed test because we want to determine if the mean weight loss is more than 10 pounds ( $\mu > 10$ ). We specify the competing hypotheses as

$$H_0: \mu \leq 10 \text{ pounds}$$

$$H_A: \mu > 10 \text{ pounds}$$

The underlying claim that the mean weight loss is more than 10 pounds is true if our decision is to reject the null hypothesis. Conversely, if we do not reject the null hypothesis, we infer that the claim is not supported by the sample data.

### EXAMPLE 9.3

A television research analyst wishes to test a claim that more than 50% of the households will tune in for a TV episode. Specify the null and the alternative hypotheses to test the claim.

**SOLUTION:** This is an example of a one-tailed test regarding the population proportion  $p$ . Given that the analyst wants to determine whether  $p > 0.50$ , this claim is placed in the alternative hypothesis, whereas the null hypothesis is just its opposite.

$$H_0: p \leq 0.50$$

$$H_A: p > 0.50$$

The claim that more than 50% of the households will tune in for a TV episode is valid only if the null hypothesis is rejected.

### EXAMPLE 9.4

It is generally believed that at least 60% of the residents in a small town in Texas are happy with their lives. A sociologist is concerned about the lingering economic crisis and wants to determine whether the crisis has adversely affected the happiness level in this town. Specify the null and the alternative hypotheses to determine if the sociologist's concern is valid.

**SOLUTION:** This is also a one-tailed test regarding the population proportion  $p$ . While the population proportion has been at least 0.60 ( $p \geq 0.60$ ), the sociologist wants to establish that the current population proportion is below 0.60 ( $p < 0.60$ ). Therefore, the hypotheses are formulated as

$$H_0: p \geq 0.60$$

$$H_A: p < 0.60$$

In this case, the sociologist's concern is valid if the null hypothesis is rejected. Nothing new is established if the null hypothesis is not rejected.

## Type I and Type II Errors

Since the decision of a hypothesis test is based on limited sample information, we are bound to make errors. Ideally, we would like to be able to reject the null hypothesis when the null hypothesis is false and not reject the null hypothesis when the null hypothesis is true. However, we may end up rejecting or not rejecting the null hypothesis erroneously. In other words, sometimes we reject the null hypothesis when we should not, or choose not to reject the null hypothesis when we should.

We consider two types of errors in the context of hypothesis testing: a **Type I error** and a **Type II error**. A Type I error is committed when we reject the null hypothesis when the null hypothesis is actually true. On the other hand, a Type II error is made when we do not reject the null hypothesis and the null hypothesis is actually false.

Table 9.2 summarizes the circumstances surrounding Type I and Type II errors. Two correct decisions are possible: not rejecting the null hypothesis when the null hypothesis is true and rejecting the null hypothesis when the null hypothesis is false. Conversely, two incorrect decisions (errors) are also possible: rejecting the null hypothesis when the null hypothesis is true (Type I error) and not rejecting the null hypothesis when the null hypothesis is false (Type II error).

### LO 9.2

Distinguish between Type I and Type II errors.

**TABLE 9.2** Type I and Type II Errors

Decision	Null hypothesis is true	Null hypothesis is false
Reject the null hypothesis	Type I error	Correct decision
Do not reject the null hypothesis	Correct decision	Type II error

### EXAMPLE 9.5

Consider the following hypotheses that relate to the medical example mentioned earlier.

$$H_0: \text{A person is free of a particular disease}$$

$$H_A: \text{A person has a particular disease}$$

Suppose a person takes a medical test that attempts to detect this disease. Discuss the consequences of a Type I error and a Type II error.



**SOLUTION:** A Type I error occurs when the medical test indicates that the person has the disease (reject  $H_0$ ), but, in reality, the person is free of the disease. We often refer to this type of result as a false positive. If the medical test shows that the person is free of the disease (do not reject  $H_0$ ), when the person actually has the disease, then a Type II error occurs. We often call this type of result a false negative. Arguably, the consequences of a Type II error in this example are more serious than those of a Type I error.

### EXAMPLE 9.6

Consider the following competing hypotheses that relate to the court of law.

$H_0$ : An accused person is innocent

$H_A$ : An accused person is guilty

Suppose the accused person is judged by a jury of her peers. Discuss the consequences of a Type I error and a Type II error.

**SOLUTION:** A Type I error is a verdict that finds that the accused is guilty (reject  $H_0$ ) when she is actually innocent. A Type II error is a verdict that finds that the accused is innocent (do not reject  $H_0$ ) when she is actually guilty. In this example, it is not clear which of the two errors is more costly to society.

As noted in Example 9.6, it is not always easy to determine which of the two errors has more serious consequences. For given evidence, there is a trade-off between these errors; by reducing the likelihood of a Type I error, we implicitly increase the likelihood of a Type II error, and vice versa. The only way we can reduce both errors is by collecting more evidence. Let us denote the probability of a Type I error by  $\alpha$ , the probability of a Type II error by  $\beta$ , and the strength of the evidence by the sample size  $n$ . Therefore, we can conclude that the only way we can lower both  $\alpha$  and  $\beta$  is by increasing  $n$ . For a given  $n$ , however, we can reduce  $\alpha$  only at the expense of a higher  $\beta$  and reduce  $\beta$  only at the expense of a higher  $\alpha$ . The optimal choice of  $\alpha$  and  $\beta$  depends on the relative cost of these two types of errors, and determining these costs is not always easy. Typically, the decision regarding the optimal level of Type I and Type II errors is made by the management of a firm where the job of a statistician is to conduct the hypothesis test for a chosen value of  $\alpha$ .

## EXERCISES 9.1

- Explain why the following hypotheses are not constructed correctly.
  - $H_0: \mu \leq 10; H_A: \mu \geq 10$
  - $H_0: \mu \neq 500; H_A: \mu = 500$
  - $H_0: p \leq 0.40; H_A: p > 0.42$
  - $H_0: \bar{X} \leq 128; H_A: \bar{X} > 128$
- Which of the following statements are valid null and alternative hypotheses? If they are invalid hypotheses, explain why.
  - $H_0: \bar{X} \leq 210; H_A: \bar{X} > 210$
  - $H_0: \mu = 120; H_A: \mu \neq 120$
  - $H_0: p \leq 0.24; H_A: p > 0.24$
  - $H_0: \mu < 252; H_A: \mu > 252$
- Explain why the following statements are not correct.
  - "With my methodological approach, I can reduce the Type I error with the given sample information without changing the Type II error."
  - "I have already decided how much of the Type I error I am going to allow. A bigger sample will not change either the Type I or Type II error."
  - "I can reduce the Type II error by making it difficult to reject the null hypothesis."
  - "By making it easy to reject the null hypothesis, I am reducing the Type I error."
- Which of the following statements are correct? Explain if incorrect.



- a. "I accept the null hypothesis since sample evidence is not inconsistent with the null hypothesis."
  - b. "Since sample evidence cannot be supported by the null hypothesis, I reject the null hypothesis."
  - c. "I can establish a given claim if sample evidence is consistent with the null hypothesis."
  - d. "I cannot establish a given claim if the null hypothesis is not rejected."
5. Construct the null and the alternative hypotheses for the following tests:
    - a. Test if the mean weight of cereal in a cereal box differs from 18 ounces.
    - b. Test if the stock price increases on more than 60% of the trading days.
    - c. Test if Americans get an average of less than seven hours of sleep.
  6. Define the consequences of Type I and Type II errors for each of the tests considered in the preceding question.
  7. Construct the null and the alternative hypotheses for the following claims:
    - a. "I am going to get the majority of the votes to win this election."
    - b. "I suspect that your 10-inch pizzas are, on average, less than 10 inches in size."
    - c. "I will have to fine the company since its tablets do not contain an average of 250 mg of ibuprofen as advertised."
  8. Discuss the consequences of Type I and Type II errors for each of the claims considered in the preceding question.
  9. A polygraph (lie detector) is an instrument used to determine if an individual is telling the truth. These tests are considered to be 95% reliable. In other words, if an individual lies, there is a 0.95 probability that the test will detect a lie. Let there also be a 0.005 probability that the test erroneously detects a lie even when the individual is actually telling the truth. Consider the null hypothesis, "the individual is telling the truth," to answer the following questions.
    - a. What is the probability of a Type I error?
    - b. What is the probability of a Type II error?
    - c. Discuss the consequences of Type I and Type II errors.
    - d. What is wrong with the statement, "I can prove that the individual is telling the truth on the basis of the polygraph result."
  10. The screening process for detecting a rare disease is not perfect. Researchers have developed a blood test that is considered fairly reliable. It gives a positive reaction in 98% of the people who have that disease. However, it erroneously gives a positive reaction in 3% of the people who do not have the disease. Consider the null hypothesis "the individual does not have the disease" to answer the following questions.
    - a. What is the probability of a Type I error?
    - b. What is the probability of a Type II error?
    - c. Discuss the consequences of Type I and Type II errors.
    - d. What is wrong with the nurse's analysis, "The blood test result has proved that the individual is free of disease."
  11. The manager of a large manufacturing firm is considering switching to new and expensive software that promises to significantly reduce its assembly costs. Before purchasing the software, the manager wants to conduct a hypothesis test to determine if the new software does significantly reduce its assembly costs.
    - a. Is the manager of the manufacturing firm more concerned about a Type I error or a Type II error? Explain.
    - b. Is the software company more concerned about a Type I error or a Type II error? Explain.
  12. A consumer group has accused a restaurant for using higher fat content than what is reported on its menu. The group has been asked to conduct a hypothesis test to substantiate its claims.
    - a. Is the manager of the restaurant more concerned about a Type I error or a Type II error? Explain.
    - b. Is the consumer group more concerned about a Type I error or a Type II error? Explain.

## 9.2 HYPOTHESIS TEST FOR THE POPULATION MEAN WHEN $\sigma$ IS KNOWN

In order to introduce the basic methodology for hypothesis testing, we first conduct a hypothesis test regarding the population mean  $\mu$  under the assumption that the population standard deviation  $\sigma$  is known. While it is true that  $\sigma$  is rarely known, there are instances when  $\sigma$  is considered fairly stable, and therefore, can be determined from prior experience. In such cases,  $\sigma$  is treated as known. Fortunately, this assumption has no bearing on the overall procedure of conducting a hypothesis test; a procedure we use throughout the remainder of the text.

A hypothesis test regarding the population mean  $\mu$  is based on the sampling distribution of the sample mean  $\bar{X}$ . In particular, it uses the fact that  $E(\bar{X}) = \mu$  and  $se(\bar{X}) = \sigma/\sqrt{n}$ . Also, in order to implement the test, it is essential that  $\bar{X}$  is normally distributed. Recall

that  $\bar{X}$  is normally distributed when the underlying population is normally distributed. If the underlying population is not normally distributed, then, by the central limit theorem,  $\bar{X}$  is approximately normally distributed if the sample size is sufficiently large—that is,  $n \geq 30$ .

The basic principle of hypothesis testing is to first assume that the null hypothesis is true and then determine if sample evidence contradicts this assumption. This principle is analogous to the scenario in the court of law where the null hypothesis is defined as “the individual is innocent” and the decision rule is best described by “innocent until proven guilty.”

We follow a four-step procedure when implementing a hypothesis test. We make a distinction between two equivalent methods—the ***p*-value approach** and the **critical value approach**—for hypothesis testing. The four-step procedure with the two approaches is valid for one-tailed and two-tailed tests regarding the population mean, the population proportion, or any other population parameter of interest.

### LO 9.3

Conduct a hypothesis test using the *p*-value approach.

## The *p*-Value Approach

Suppose a sociologist wants to establish that the mean retirement age is greater than 67 ( $\mu > 67$ ). It is assumed that retirement age is normally distributed with a known population standard deviation of 9 years ( $\sigma = 9$ ). We can investigate the sociologist’s belief by specifying the competing hypotheses as

$$H_0: \mu \leq 67$$

$$H_A: \mu > 67$$

Let a random sample of 25 retirees produce an average retirement age of 71—that is,  $\bar{x} = 71$ . This sample evidence casts doubt on the validity of the null hypothesis, since the sample mean is greater than the hypothesized value,  $\mu_0 = 67$ . However, the discrepancy between  $\bar{x}$  and  $\mu_0$  does not necessarily imply that the null hypothesis is false. Perhaps the discrepancy can be explained by pure chance. It is common to evaluate this discrepancy in terms of the appropriate test statistic.

### TEST STATISTIC FOR $\mu$ WHEN $\sigma$ IS KNOWN

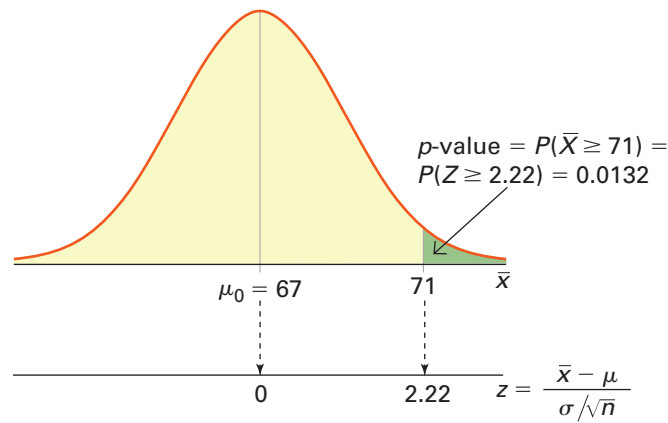
The value of the **test statistic** for the hypothesis test of the **population mean  $\mu$**  when the **population standard deviation  $\sigma$  is known** is computed as

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

where  $\mu_0$  is the hypothesized value of the population mean. This formula is valid only if  $\bar{X}$  (approximately) follows a normal distribution.

Note that the value of the test statistic  $z$  is evaluated at  $\mu = \mu_0$ , which explains why we need some form of the equality sign in the null hypothesis. Given that the population is normally distributed with a known standard deviation,  $\sigma = 9$ , we compute the value of the test statistic as  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{71 - 67}{9/\sqrt{25}} = 2.22$ . Therefore, comparing  $\bar{x} = 71$  with 67 is identical to comparing  $z = 2.22$  with 0, where 67 and 0 are the means of  $\bar{X}$  and  $Z$ , respectively.

We now compute the ***p*-value**, which is the likelihood of obtaining a sample mean that is at least as extreme as the one derived from the given sample, under the assumption that the null hypothesis is true as an equality—that is,  $\mu_0 = 67$ . Since in this example  $\bar{x} = 71$ , we define the extreme value as a sample mean of 71 or higher and use the  $z$  table to compute the *p*-value as  $P(\bar{X} \geq 71) = P(Z \geq 2.22) = 1 - 0.9868 = 0.0132$ . Figure 9.1 shows the computed *p*-value.



**FIGURE 9.1** The  $p$ -value for a right-tailed test with  $z = 2.22$

Note that when the null hypothesis is true, there is only a 1.32% chance that the sample mean will be 71 or more. This seems like a very small chance, but is it small enough to allow us to reject the null hypothesis in favor of the alternative hypothesis? Let's see how we define "small enough."

Remember that a Type I error occurs when we reject the null hypothesis when it is actually true. We define the *allowed* probability of making a Type I error as  $\alpha$ ; we refer to  $100\alpha\%$  as the **significance level**. The  $p$ -value, on the other hand, is referred to as the *observed* probability of making a Type I error. When using the  $p$ -value approach, **the decision rule is to reject the null hypothesis if the  $p$ -value  $< \alpha$  and not reject the null hypothesis if the  $p$ -value  $\geq \alpha$ .**

We generally choose a value for  $\alpha$  *before* implementing a hypothesis test; that is, we set the rules of the game before playing. Most hypothesis tests are conducted using a significance level of 1%, 5%, or 10%, using  $\alpha = 0.01$ , 0.05, or 0.10, respectively. For example,  $\alpha = 0.05$  means that we allow a 5% chance of rejecting a true null hypothesis. We can also interpret these conventional significance levels as follows:

- If we reject a null hypothesis at the 10% significance level ( $\alpha = 0.10$ ), then we have *some evidence* that the null hypothesis is false;
- If we reject a null hypothesis at the 5% significance level ( $\alpha = 0.05$ ), then we have *strong evidence* that the null hypothesis is false; and
- If we reject a null hypothesis at the 1% significance level ( $\alpha = 0.01$ ), then we have *very strong evidence* that the null hypothesis is false.

In our example, given the  $p$ -value of 0.0132, if we decide to reject the null hypothesis, then there is a 1.32% chance that our decision will be erroneous.

Suppose we had chosen  $\alpha = 0.05$  to conduct the above test. At this significance level, we reject the null hypothesis because  $0.0132 < 0.05$ . This means that the sample data support the sociologist's claim that the average retirement age is greater than 67 years old. Individuals may be working past the normal retirement age of 67 either because their savings have been depleted due to the financial crisis and/or because this generation is expected to outlive any previous generation and needs a job to pay the bills. We should note that if  $\alpha$  had been set at 0.01, then the findings would have been different. At this smaller significance level, the evidence does not allow us to reject the null hypothesis ( $0.0132 > 0.01$ ). At the 1% significance level, we cannot conclude that the mean retirement age is greater than 67.

In the retirement age example of a right-tailed test, we calculated the  $p$ -value as  $P(Z \geq z)$ . Analogously, for a left-tailed test the  $p$ -value is given by  $P(Z \leq z)$ . For a two-tailed test, the extreme values exist on both sides of the distribution of the test statistic. Given the symmetry of the  $z$  distribution, the  $p$ -value for a two-tailed test is twice that of the  $p$ -value for a one-tailed test. It is calculated as  $2P(Z \geq z)$  if  $z > 0$  or as  $2P(Z \leq z)$  if  $z < 0$ .

### THE $p$ -VALUE APPROACH

Under the assumption that  $\mu = \mu_0$ , the  $p$ -value is the likelihood of observing a sample mean that is at least as extreme as the one derived from the given sample. Its calculation depends on the specification of the alternative hypothesis.

Alternative Hypothesis	$p$ -value
$H_A: \mu > \mu_0$	Right-tail probability: $P(Z \geq z)$
$H_A: \mu < \mu_0$	Left-tail probability: $P(Z \leq z)$
$H_A: \mu \neq \mu_0$	Two-tail probability: $2P(Z \geq z)$ if $z > 0$ or $2P(Z \leq z)$ if $z < 0$

The decision rule: Reject  $H_0$  if the  $p$ -value  $< \alpha$ .

Figure 9.2 shows the three different scenarios of determining the  $p$ -value depending on the specification of the competing hypotheses.

**FIGURE 9.2** The  $p$ -values for one- and two-tailed tests

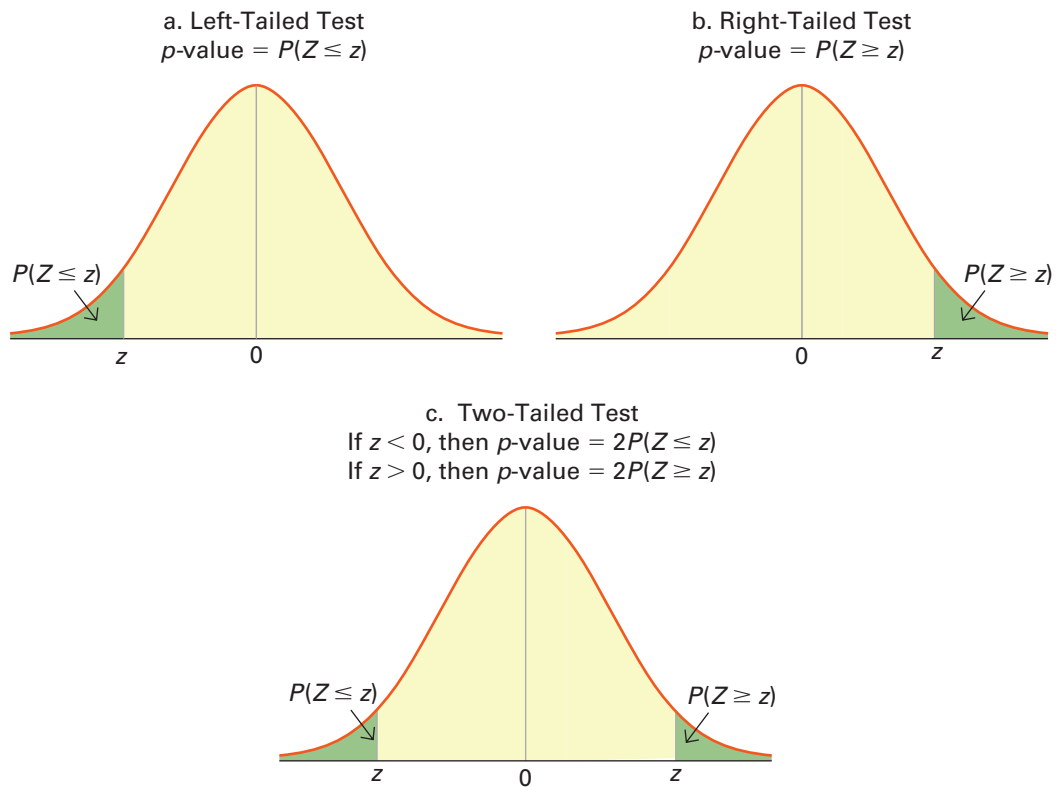


Figure 9.2a shows the  $p$ -value for a left-tailed test. Since the appropriate test statistic follows the standard normal distribution, we calculate the  $p$ -value as  $P(Z \leq z)$ . When calculating the  $p$ -value for a right-tailed test (see Figure 9.2b), we find the area to the right of the value of the test statistic  $z$  or, equivalently,  $P(Z \geq z)$ . Figure 9.2c shows the  $p$ -value for a two-tailed test, calculated as  $2P(Z \leq z)$  when  $z < 0$  or as  $2P(Z \geq z)$  when  $z > 0$ .

It is important to note that we *cannot* reject  $H_0$  for a right-tailed test if  $\bar{x} \leq \mu_0$ , or equivalently,  $z \leq 0$ . Consider, for example, a right-tailed test with the hypotheses specified as  $H_0: \mu \leq 67$  versus  $H_A: \mu > 67$ . Here, if  $\bar{x} = 65$ , there is no need for formal testing since we have no discrepancy between the sample mean and the hypothesized value of the population mean. Similarly, we *cannot* reject  $H_0$  for a left-tailed test if  $\bar{x} \geq \mu_0$  or, equivalently,  $z \geq 0$ . We will now summarize the four-step procedure using the  $p$ -value approach.

#### THE FOUR-STEP PROCEDURE USING THE $p$ -VALUE APPROACH

**Step 1. Specify the null and the alternative hypotheses.** We identify the relevant population parameter of interest, determine whether it is a one- or a two-tailed test and, most importantly, include some form of the equality sign in the null hypothesis and place whatever we wish to establish in the alternative hypothesis.

**Step 2. Specify the significance level.** Before implementing a hypothesis test, we first specify  $\alpha$ , which is the *allowed* probability of making a Type I error.

**Step 3. Calculate the value of the test statistic and the  $p$ -value.** When the population standard deviation  $\sigma$  is known, the value of the test statistic is  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ , where  $\mu_0$  is the hypothesized value of the population mean. For a right-tailed test, the  $p$ -value is  $P(Z \geq z)$ , and for a left-tailed test, the  $p$ -value is  $P(Z \leq z)$ . For a two-tailed test, the  $p$ -value is  $2P(Z \geq z)$  if  $z > 0$ , or  $2P(Z \leq z)$  if  $z < 0$ . The  $p$ -value is also referred to as the *observed* probability of making a Type I error.

**Step 4. State the conclusion and interpret results.** The decision rule is to reject the null hypothesis when  $p\text{-value} < \alpha$  and not reject the null hypothesis when  $p\text{-value} \geq \alpha$ .

#### EXAMPLE 9.7

A research analyst disputes a trade group's prediction that back-to-school spending will average \$606.40 per family this year. She believes that average back-to-school spending will significantly differ from this amount. She decides to conduct a test on the basis of a random sample of 30 households with school-age children. She calculates the sample mean as \$622.85. She also believes that back-to-school spending is normally distributed with a population standard deviation of \$65. She wants to conduct the test at the 5% significance level.

- Specify the competing hypotheses in order to test the research analyst's claim.
- In this hypothesis test, what is the allowed probability of a Type I error?
- Calculate the value of the test statistic and the  $p$ -value.
- At the 5% significance level, does average back-to-school spending differ from \$606.40?

#### SOLUTION:

- Since we want to determine if the average is different from the predicted value of \$606.40, we specify the hypotheses as

$$H_0: \mu = 606.40$$

$$H_A: \mu \neq 606.40$$

- The allowed probability of a Type I error is equivalent to the significance level of the test, which in this example is given as  $\alpha = 0.05$ .
- Note that  $\bar{X}$  is normally distributed since it is computed from a random sample drawn from a normal population. Since  $\sigma$  is known, the test statistic follows the standard normal distribution, and its value is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{622.85 - 606.40}{65/\sqrt{30}} = 1.39.$$

For a two-tailed test with a positive value for the test statistic, we compute the  $p$ -value as  $2P(Z \geq 1.39)$ . From the  $z$  table, we first find  $P(Z \geq 1.39) = 1 - 0.9177 = 0.0823$ ; so the  $p$ -value  $= 2 \times 0.0823 = 0.1646$ .

- d. The decision rule is to reject the null hypothesis if the  $p$ -value is less than  $\alpha$ . Since  $0.1646 > 0.05$ , we do not reject  $H_0$ . Therefore, at the 5% significance level, we cannot conclude that average back-to-school spending differs from \$606.40 per family this year. The sample data do not support the research analyst's claim.

#### LO 9.4

Conduct a hypothesis test using the critical value approach.

## The Critical Value Approach

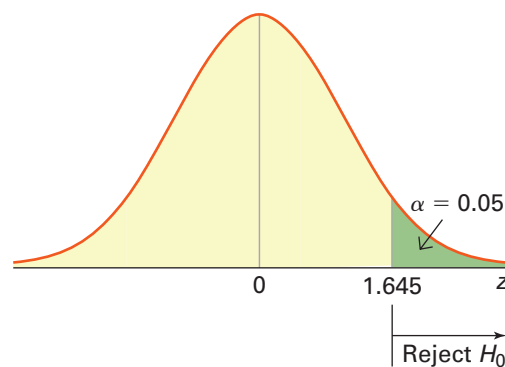
We always use sample evidence and the chosen significance level  $\alpha$  to conduct hypothesis tests. The  $p$ -value approach makes the comparison in terms of probabilities. The value of the test statistic is used to compute the  $p$ -value, which is then compared with  $\alpha$  in order to arrive at a decision. As we will see shortly, most statistical software packages report  $p$ -values, so the  $p$ -value approach to hypothesis testing tends to be favored by most researchers and practitioners. The critical value approach, on the other hand, makes the comparison directly in terms of the value of the test statistic. This approach is particularly useful when a computer is unavailable and all calculations must be done manually. Some also find the critical value approach more intuitively appealing. Both approaches, however, always lead to the same conclusion.

Earlier, we had used the  $p$ -value approach to validate a sociologist's claim that the mean retirement age in the United States is greater than 67 at the 5% significance level. In a random sample of 25 retirees, the average retirement age was 71. It was also assumed that the retirement age is normally distributed with a population standard deviation of 9 years. With the critical value approach, we still specify the competing hypotheses and calculate the value of the test statistic as we did with the  $p$ -value approach. In the retirement age example, the competing hypotheses are  $H_0: \mu \leq 67$  versus  $H_A: \mu > 67$  and the value of the test statistic is  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{71 - 67}{9/\sqrt{25}} = 2.22$ .

The critical value approach specifies a region of values, also called the **rejection region**, such that if the value of the test statistic falls into this region, then we reject the null hypothesis. The **critical value** is a point that separates the rejection region from the nonrejection region. Once again we need to make distinctions between the three types of competing hypotheses. For a right-tailed test, the critical value is  $z_\alpha$ , where  $P(Z \geq z_\alpha) = \alpha$ . The resulting rejection region includes values greater than  $z_\alpha$ .

With  $\alpha$  known, we can easily find the corresponding  $z_\alpha$  from the  $z$  table. In the retirement age example with  $\alpha = 0.05$ , we evaluate  $P(Z \geq z_\alpha) = 0.05$  to derive the critical value as  $z_\alpha = z_{0.05} = 1.645$ . Figure 9.3 shows the critical value as well as the corresponding rejection region of the test.

**FIGURE 9.3**  
The critical value for a right-tailed test with  $\alpha = 0.05$

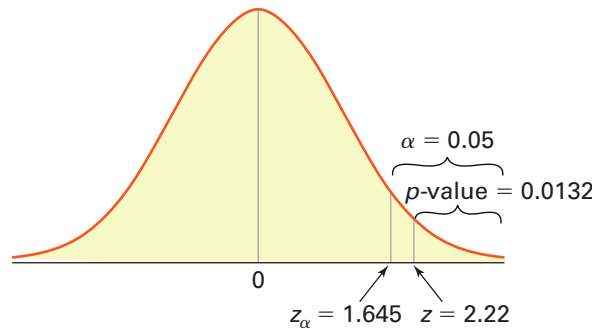


As shown in Figure 9.3, the decision rule is to reject  $H_0$  if  $z > 1.645$ . Since the value of the test statistic,  $z = 2.22$ , exceeds the critical value,  $z_\alpha = 1.645$ , we reject the null hypothesis



and conclude that the mean age is significantly greater than 67. Thus, we confirm the conclusion reached with the  $p$ -value approach.

We would like to stress that we always arrive at the same conclusion whether we use the  $p$ -value approach or the critical value approach. If  $z$  falls in the rejection region, then the  $p$ -value must be less than  $\alpha$ . Similarly, if  $z$  does not fall in the rejection region, then the  $p$ -value must be greater than  $\alpha$ . Figure 9.4 shows the equivalence of the two results in the retirement age example of a right-tailed test.



**FIGURE 9.4** Equivalent conclusions resulting from the  $p$ -value and the critical value approaches

We reject the null hypothesis because the  $p$ -value = 0.0132 is less than  $\alpha = 0.05$ , or, equivalently, because  $z = 2.22$  is greater than  $z_{\alpha} = 1.645$ .

The above example uses a right-tailed test to calculate the critical value as  $z_{\alpha}$ . Given the symmetry of the  $z$  distribution around zero, the critical value for a left-tailed test is simply  $-z_{\alpha}$ . For a two-tailed test, we split the significance level in half to determine *two* critical values  $-z_{\alpha/2}$  and  $z_{\alpha/2}$  where  $P(Z \geq z_{\alpha/2}) = \alpha/2$ .

#### THE CRITICAL VALUE APPROACH

The critical value approach specifies a region such that if the value of the test statistic falls into the region, the null hypothesis is rejected. The specification of the competing hypotheses and the significance level determine the critical value(s).

Alternative Hypothesis	Critical Value
$H_A: \mu > \mu_0$	Right-tailed critical value is $z_{\alpha}$ , where $P(Z \geq z_{\alpha}) = \alpha$ .
$H_A: \mu < \mu_0$	Left-tailed critical value is $-z_{\alpha}$ , where $P(Z \geq z_{\alpha}) = \alpha$ .
$H_A: \mu \neq \mu_0$	Two-tailed critical values $-z_{\alpha/2}$ and $z_{\alpha/2}$ , where $P(Z \geq z_{\alpha/2}) = \alpha/2$ .

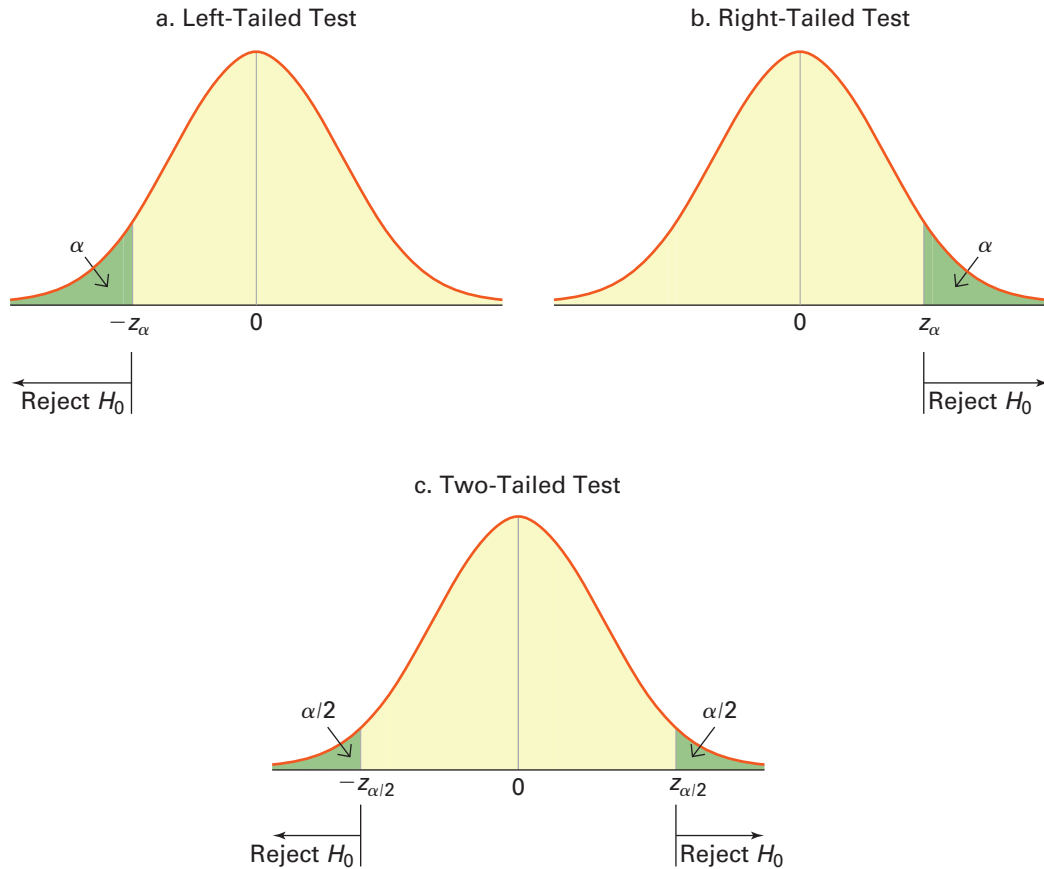
#### The decision rule:

- Reject  $H_0$  if  $z > z_{\alpha}$  for a right-tailed test.
- Reject  $H_0$  if  $z < -z_{\alpha}$  for a left-tailed test.
- Reject  $H_0$  if  $z > z_{\alpha/2}$  or  $z < -z_{\alpha/2}$  for a two-tailed test.

For a given  $\alpha$ , Figure 9.5 shows the three different scenarios of determining the critical value(s) depending on the specification of the competing hypotheses.

Figure 9.5a shows a negative critical value for a left-tailed test where we reject the null hypothesis if  $z < -z_{\alpha}$ . Similarly, Figure 9.5b shows a positive critical value for a right-tailed test where we reject the null hypothesis if  $z > z_{\alpha}$ . There are two critical values for a two-tailed test, where we reject the null hypothesis when  $z < -z_{\alpha/2}$  or when  $z > z_{\alpha/2}$  (see Figure 9.5c).

**FIGURE 9.5** Critical values for one- and two-tailed tests



We will now summarize the four-step procedure using the critical value approach.

#### THE FOUR-STEP PROCEDURE USING THE CRITICAL VALUE APPROACH

**Step 1. Specify the null and the alternative hypotheses.** We identify the relevant parameter of interest, determine whether it is a one- or a two-tailed test, and, most importantly, include some form of the equality sign in the null hypothesis and place whatever we wish to establish in the alternative hypothesis.

**Step 2. Specify the significance level and find the critical value(s).** We first specify  $\alpha$ , which is the *allowed* probability of making a Type I error. When the population standard deviation  $\sigma$  is known, the critical value for a right-tailed test is  $z_\alpha$ , where  $P(Z \geq z_\alpha) = \alpha$ , and the critical value for a left-tailed test is  $-z_\alpha$ . The critical values for a two-tailed test are  $-z_{\alpha/2}$  and  $z_{\alpha/2}$ , where  $P(Z \geq z_{\alpha/2}) = \alpha/2$ .

**Step 3. Calculate the value of the test statistic.** The value of the test statistic is  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ , where  $\mu_0$  is the hypothesized value of the population mean.

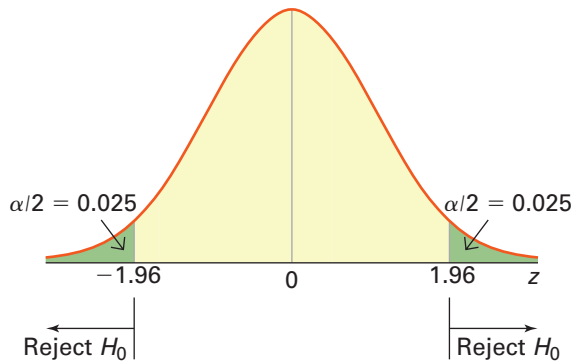
**Step 4. State the conclusion and interpret results.** If the value of the test statistic falls in the rejection region, the decision rule is to reject the null hypothesis. So for a right-tailed test, we reject the null hypothesis if  $z > z_\alpha$ ; for a left-tailed test, we reject the null hypothesis if  $z < -z_\alpha$ ; and for a two-tailed test, we reject the null hypothesis if  $z > z_{\alpha/2}$  or  $z < -z_{\alpha/2}$ .

### EXAMPLE 9.8

Repeat Example 9.7 using the critical value approach. Recall that a research analyst wishes to determine if average back-to-school spending differs from \$606.40. A random sample of 30 households, drawn from a normally distributed population with a population standard deviation of \$65, results in a sample mean of \$622.85. The test is conducted at the 5% significance level.

**SOLUTION:** The competing hypotheses and the value of the test statistic are the same; that is,  $H_0: \mu = 606.40$  versus  $H_A: \mu \neq 606.40$  and  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{622.85 - 606.40}{65/\sqrt{30}} = 1.39$ . For a two-tailed test, we split the significance level in half to determine *two* critical values, one on each side of the distribution of the test statistic. Given a 5% level of significance,  $\alpha/2 = 0.05/2 = 0.025$  is used to derive  $z_{\alpha/2} = z_{0.025}$  as 1.96. Thus, the critical values are  $-1.96$  and  $1.96$ . As shown in Figure 9.6, the decision rule is to reject  $H_0$  if  $z > 1.96$  or  $z < -1.96$ .

**FIGURE 9.6** The critical values for a two-tailed test with  $\alpha = 0.05$



Since  $z = 1.39$  does not fall in the rejection region ( $-1.96 < 1.39 < 1.96$ ), we do not reject the null hypothesis. At the 5% significance level, we cannot conclude that average back-to-school spending differs from \$606.40 per family. As always, our conclusion is consistent with that using the  $p$ -value approach.

## Confidence Intervals and Two-Tailed Hypothesis Tests

A confidence interval for the population parameter is sometimes used as an alternative method for conducting a two-tailed hypothesis test. Given that we conduct the hypothesis test at the  $\alpha$  significance level, we can use the sample data to determine a corresponding  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$ . If the confidence interval does not contain the hypothesized value of the population mean  $\mu_0$ , then we reject the null hypothesis. If the confidence interval contains  $\mu_0$ , then we do not reject the null hypothesis.

### IMPLEMENTING A TWO-TAILED TEST USING A CONFIDENCE INTERVAL

The general specification for a  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  when the population standard deviation  $\sigma$  is known is computed as

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \left[ \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

Given a hypothesized value of the population mean  $\mu_0$ , the **decision rule** is

$$\text{Reject } H_0 \text{ if } \mu_0 < \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \text{ or if } \mu_0 > \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

### EXAMPLE 9.9

Repeat Example 9.8 by constructing a confidence interval for  $\mu$ .

**SOLUTION:** We are testing  $H_0: \mu = 606.40$  versus  $H_A: \mu \neq 606.40$  at the 5% significance level. We use  $n = 30$ ,  $\bar{x} = 622.85$ , and  $\sigma = 65$ , along with  $\alpha = 0.05$ , to determine the 95% confidence interval. We find  $z_{\alpha/2} = z_{0.025} = 1.96$  and compute

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 622.85 \pm 1.96 \frac{65}{\sqrt{30}} = 622.85 \pm 23.26,$$

resulting in the interval [599.59, 646.11]. Since the hypothesized value of the population mean  $\mu_0 = 606.40$  falls within the 95% confidence interval, we do not reject  $H_0$ . Thus, we arrive at the same conclusion as with the  $p$ -value and the critical value approaches; that is, the sample data do not support the research analyst's claim that average back-to-school spending differs from \$606.40 per family this year.

As shown above, we use the confidence interval as an alternative method for conducting a two-tailed test. It is possible to adjust the confidence interval to accommodate a one-tailed test, but we do not discuss this adjustment in this text.

## Using Excel to Test $\mu$ When $\sigma$ Is Known

Excel provides several functions that simplify the steps of a hypothesis test. Here we discuss one of these functions using the following example.

### EXAMPLE 9.10

A recent report in *The New York Times* (August 7, 2010) suggests that consumers are spending less not only as a response to the economic downturn, but also due to a realization that excessive spending does not make them happier. A researcher wants to use debit card data to contradict the generally held view that the average amount spent annually on a debit card is at least \$8,000. She surveys 20 consumers and asks them how much they spend annually on their debit cards. The results are given below.

7,960	7,700	7,727	7,704	8,543	7,661	7,767	8,761	7,530	8,128
7,938	7,771	7,272	8,113	7,727	7,697	7,690	8,000	8,079	7,547

It is assumed that the population standard deviation is \$500 and that spending on debit cards is normally distributed. Test the claim at a 1% level of significance.

**SOLUTION:** The researcher would like to establish that average spending on debit cards is less than \$8,000, or, equivalently,  $\mu < 8,000$ . Thus, we formulate the hypotheses as

$$H_0: \mu \geq 8,000$$

$$H_A: \mu < 8,000$$

The normality condition of  $\bar{X}$  is satisfied since spending on debit cards is assumed to be normally distributed. Also, since the population standard deviation is known, the test statistic is assumed to follow the  $z$  distribution. Excel's function Z.TEST returns the  $p$ -value for a right-tailed test, or equivalently,  $P(Z \geq z)$ . For a left-tailed test, as in this example, we simply subtract the value that Excel returns from one. (For a two-tailed test, if the value that Excel returns for  $P(Z \geq z)$  is less than 0.50, we multiply this value by 2 to obtain the  $p$ -value; if the value that Excel returns is

**FILE**  
Debit\_Spending

greater than 0.50, we calculate the  $p$ -value as  $2 \times (1 - P(Z \geq z))$ .) In general, this function takes the form “=Z.TEST(array,  $\mu_0$ ,  $\sigma$ ),” where *array* specifies the cell locations of the relevant data,  $\mu_0$  is the hypothesized value of the population mean under the null hypothesis, and  $\sigma$  is the value of the known population standard deviation. In order to solve Example 9.10, we open the *Debit\_Spending* data file, find an empty cell, and insert “=Z.TEST(A2:A21, 8000, 500).” Excel returns the probability of 0.8851, which corresponds to a right-tailed  $p$ -value. Subtracting this value from one yields 0.1149, which is the  $p$ -value for the left-tailed test.

The hypothesis test is conducted at the 1% significance level. Thus, since 0.1149 is not less than  $\alpha = 0.01$ , we do not reject the null hypothesis. In other words, at a 1% level of significance, the researcher cannot conclude that annual spending on debit cards is less than \$8,000. Perhaps these findings can be reconciled with a report that claims that individuals are shunning their credit cards and using debit cards to avoid incurring more debt (<http://Businessweek.com>, September 8, 2010).

## One Last Remark

An important component of any well-executed statistical analysis is to clearly communicate the results. Thus, it is not sufficient to end the analysis with a conclusion that you reject the null hypothesis or you do not reject the null hypothesis. You must interpret the results, clearly reporting whether or not the claim regarding the population parameter of interest can be justified on the basis of the sample information.

## EXERCISES 9.2

### Mechanics

13. Consider the following hypotheses:

$$H_0: \mu \leq 12.6$$

$$H_A: \mu > 12.6$$

A sample of 25 observations yields a sample mean of 13.4. Assume that the sample is drawn from a normal population with a known population standard deviation of 3.2.

- Calculate the  $p$ -value.
  - What is the conclusion if  $\alpha = 0.10$ ?
  - Calculate the  $p$ -value if the above sample mean was based on a sample of 100 observations.
  - What is the conclusion if  $\alpha = 0.10$ ?
14. Redo the preceding question using the critical value approach.

15. Consider the following hypotheses:

$$H_0: \mu \geq 150$$

$$H_A: \mu < 150$$

A sample of 80 observations results in a sample mean of 144. The population standard deviation is known to be 28.

- What is the critical value for the test with  $\alpha = 0.01$  and with  $\alpha = 0.05$ ?
- Does the above sample evidence enable us to reject the null hypothesis at  $\alpha = 0.01$ ?
- Does the above sample evidence enable us to reject the null hypothesis at  $\alpha = 0.05$ ?

16. Redo the preceding question using the  $p$ -value approach.

17. Consider the following hypotheses:

$$H_0: \mu = 1800$$

$$H_A: \mu \neq 1800$$

The population is normally distributed with a population standard deviation of 440. Compute the value of the test statistic and the resulting  $p$ -value for each of the following sample results. For each sample, determine if you can reject the null hypothesis at the 10% significance level.

- $\bar{x} = 1850$ ;  $n = 110$
- $\bar{x} = 1850$ ;  $n = 280$
- $\bar{x} = 1650$ ;  $n = 32$
- $\bar{x} = 1700$ ;  $n = 32$

18. Consider the following hypothesis test:

$$H_0: \mu \leq -5$$

$$H_A: \mu > -5$$

A random sample of 25 observations yields a sample mean of  $-8$ . The population standard deviation is 10. Calculate the  $p$ -value. What is the conclusion to the test if  $\alpha = 0.05$ ?

19. Consider the following hypothesis test:

$$H_0: \mu \leq 75$$

$$H_A: \mu < 75$$

A random sample of 100 observations yields a sample mean of 80. The population standard deviation is 30. Calculate the  $p$ -value. What is the conclusion to the test if  $\alpha = 0.10$ ?

20. Consider the following hypothesis test:

$$H_0: \mu = -100$$

$$H_A: \mu \neq -100$$

A random sample of 36 observations yields a sample mean of  $-125$ . The population standard deviation is  $42$ . If  $\alpha = 0.01$ , find the critical value(s). What is the conclusion to the test?

21. Consider the following hypotheses:

$$H_0: \mu = 120$$

$$H_A: \mu \neq 120$$

The population is normally distributed with a population standard deviation of  $46$ .

- Use a  $5\%$  level of significance to determine the critical value(s) of the test.
- What is the conclusion with  $\bar{x} = 132$  and  $n = 50$ ?
- Use a  $10\%$  level of significance to determine the critical value(s) of the test.
- What is the conclusion with  $\bar{x} = 108$  and  $n = 50$ ?

## Applications

- It is advertised that the average braking distance for a small car traveling at  $65$  miles per hour equals  $120$  feet. A transportation researcher wants to determine if the statement made in the advertisement is false. She randomly test drives  $36$  small cars at  $65$  miles per hour and records the braking distance. The sample average braking distance is computed as  $114$  feet. Assume that the population standard deviation is  $22$  feet.
  - State the null and the alternative hypotheses for the test.
  - Calculate the value of the test statistic and the  $p$ -value.
  - Use  $\alpha = 0.01$  to determine if the average braking distance differs from  $120$  feet.
  - Repeat the test with the critical value approach.
- Customers at Costco spend an average of  $\$130$  per trip (*The Wall Street Journal*, October 6, 2010). One of Costco's rivals would like to determine whether its customers spend more per trip. A survey of the receipts of  $25$  customers found that the sample mean was  $\$135.25$ . Assume that the population standard deviation is  $\$10.50$  and that spending follows a normal distribution.
  - Specify the null and alternative hypotheses to test whether average spending at the rival's store is more than  $\$130$ .
  - Calculate the value of the test statistic. Calculate the  $p$ -value.
  - At the  $5\%$  significance level, what is the conclusion to the test?
  - Repeat the test using the critical value approach.
- In May 2008, CNN reported that sports utility vehicles (SUVs) are plunging toward the "endangered" list. Due to the uncertainty of oil prices and environmental concerns, consumers are replacing gas-guzzling vehicles with fuel-efficient smaller cars. As a result, there has been a big drop in the demand for new as well as used SUVs. A sales manager of a used car dealership for SUVs believes that it takes more than  $90$  days, on average, to sell an SUV. In order to test his claim, he samples  $40$  recently sold SUVs and finds that it took an average of  $95$  days to sell an SUV. He believes that the population standard deviation is fairly stable at  $20$  days.
  - State the null and the alternative hypotheses for the test.
  - What is the  $p$ -value?
  - Is the sales manager's claim justified at  $\alpha = 0.01$ ?
  - Repeat the above hypothesis test with the critical value approach.
- An article in the *National Geographic News* (February 24, 2005) reports that Americans are increasingly skimping on their sleep. A researcher wants to determine if Americans are sleeping less than the recommended  $7$  hours of sleep on weekdays. He takes a random sample of  $150$  Americans and computes the average sleep time of  $6.7$  hours on weekdays. Assume that the population is normally distributed with a known standard deviation of  $2.1$  hours.
  - Use the  $p$ -value approach to test the researcher's claim at  $\alpha = 0.01$ .
  - Use the critical value approach to test the researcher's claim at  $\alpha = 0.01$ .
- A local bottler in Hawaii wishes to ensure that an average of  $16$  ounces of passion fruit juice is used to fill each bottle. In order to analyze the accuracy of the bottling process, he takes a random sample of  $48$  bottles. The mean weight of the passion fruit juice in the sample is  $15.80$  ounces. Assume that the population standard deviation is  $0.8$  ounce.
  - State the null and the alternative hypotheses for the test.
  - Use the critical value approach to test the bottler's concern at  $\alpha = 0.05$ .
  - Make a recommendation to the bottler.
- FILE MV\_Houses.** A realtor in Mission Viejo, California, believes that the average price of a house is more than  $\$500,000$ .
  - State the null and the alternative hypotheses for the test.
  - The data accompanying this exercise show house prices. (Data are in  $\$1,000$ s.) Use Excel's  $Z.TEST$  function to calculate the  $p$ -value. Assume the population standard deviation is  $\$100$  (in  $\$1,000$ s).
  - At  $\alpha = 0.05$ , what is the conclusion to the test? Is the realtor's claim supported by the data?
- FILE Home\_Depot.** The data accompanying this exercise show the weekly stock price for Home Depot. Assume that stock prices are normally distributed with a population standard deviation of  $\$3$ .



- a. State the null and the alternative hypotheses in order to test whether or not the average weekly stock price differs from \$30.
  - b. Specify the critical value(s) of the test at the 5% significance level.
  - c. Compute the value of the test statistic.
  - d. At  $\alpha = 0.05$ , can you conclude that the average weekly stock price does not equal \$30?
29. **FILE Hourly\_Wage.** An economist wants to test if the average hourly wage is less than \$22.
- a. State the null and the alternative hypotheses for the test.
  - b. The data accompanying this exercise show hourly wages. Use Excel's Z.TEST function to calculate the  $p$ -value. Assume that the population standard deviation is \$6.
- c. At  $\alpha = 0.05$ , what is the conclusion to the test? Is the average hourly wage less than \$22?
30. **FILE CT\_Undergrad\_Debt.** On average, a college student last year graduated with \$27,200 in debt (*The Boston Globe*, May 27, 2012). A researcher collects data on debt from 40 recent undergraduates from Connecticut. Assume that the population standard deviation is \$5,000.
- a. The researcher believes that recent undergraduates from Connecticut have more debt than the national average. Specify the competing hypotheses to test this belief.
  - b. Specify the critical value(s) of the test at the 10% significance level.
  - c. Compute the value of the test statistic.
  - d. Do the data support the researcher's claim, at  $\alpha = 0.10$ ?

## 9.3 HYPOTHESIS TEST FOR THE POPULATION MEAN WHEN $\sigma$ IS UNKNOWN

LO 9.5

So far we have considered hypothesis tests for the population mean  $\mu$  under the assumption that the population standard deviation  $\sigma$  is known. In most business applications,  $\sigma$  is not known and we have to replace  $\sigma$  with the sample standard deviation  $s$  to estimate the standard error of  $\bar{X}$ .

Differentiate between the test statistics for the population mean.

### TEST STATISTIC FOR $\mu$ WHEN $\sigma$ IS UNKNOWN

The value of the **test statistic** for the hypothesis test of the **population mean  $\mu$**  when the **population standard deviation  $\sigma$  is unknown** is computed as

$$t_{df} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where  $\mu_0$  is the hypothesized value of the population mean and the degrees of freedom  $df = n - 1$ . This formula is valid only if  $\bar{X}$  (approximately) follows a normal distribution.

The next two examples show how we use the four-step procedure for hypothesis testing when we are testing the population mean  $\mu$  and the population standard deviation  $\sigma$  is unknown. When conducting this test manually, it turns out that the critical value approach is slightly easier to implement since the exact  $p$ -value may not be available from the  $t$  table. So, here we choose to show the critical value approach first.

### EXAMPLE 9.11

In the introductory case to this chapter, the dean at a large university in California wonders if students at her university study less than the 1961 national average of 24 hours per week. She randomly selects 35 students and asks their average study time per week (in hours). From their responses (see Table 9.1), she calculates a sample mean of 16.37 hours and a sample standard deviation of 7.22 hours.

- a. Specify the competing hypotheses to test the dean's concern.
- b. At the 5% significance level, specify the critical value(s).

- c. Calculate the value of the test statistic.
- d. What is the conclusion to the hypothesis test?

**SOLUTION:**

- a. This is an example of a one-tailed test where we would like to determine if the mean hours studied is less than 24—that is,  $\mu < 24$ . We formulate the competing hypotheses as

$$H_0: \mu \geq 24 \text{ hours}$$

$$H_A: \mu < 24 \text{ hours}$$

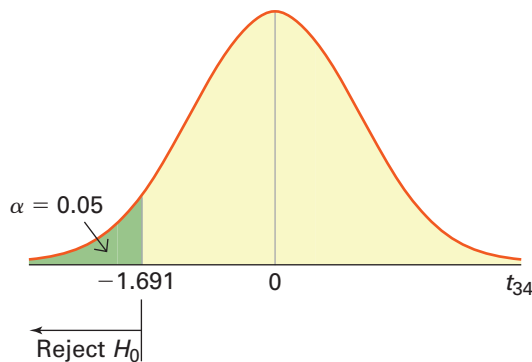
- b. Recall that for any statistical inference regarding the population mean, it is essential that the sample mean  $\bar{X}$  is normally distributed. This condition is satisfied because the sample size is greater than 30, specifically  $n = 35$ . Since we have a left-tailed test, the critical value is given by  $-t_{\alpha, df}$  where  $P(T_{df} \geq t_{\alpha, df}) = \alpha$ . Referencing the  $t$  table with  $\alpha = 0.05$  and  $df = n - 1 = 34$ , we first find  $t_{\alpha, df} = t_{0.05, 34} = 1.691$ . Therefore, the critical value is  $-t_{0.05, 34} = -1.691$ . As shown in Figure 9.7, the decision rule is to reject the null hypothesis if the value of the test statistic is less than  $-1.691$ .

- c. Given  $\bar{x} = 16.37$  and  $s = 7.22$ , we compute the value of the test statistic as

$$t_{34} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{16.37 - 24}{7.22/\sqrt{35}} = -6.25.$$

- d. Since  $t_{34} = -6.25$  is less than  $-1.691$ , we reject the null hypothesis. At the 5% significance level, we conclude that average study time at the university is less than the 1961 average of 24 hours per week.

**FIGURE 9.7** The critical value for a left-tailed test with  $\alpha = 0.05$  and  $df = 34$



**EXAMPLE 9.12**

As the introductory case to this chapter mentions, recent research finds that today's undergraduates study an average of 14 hours per week. Using the sample data from Table 9.1, the dean would also like to test if the mean study time of students at her university differs from today's national average of 14 hours per week.

- a. Formulate the competing hypotheses for this test.
- b. Calculate the value of the test statistic.
- c. Approximate the  $p$ -value.
- d. At the 5% significance level, what is the conclusion to this test?

**SOLUTION:**

- a. The dean would like to test if the mean study time of students at her university differs from 14 hours per week. Therefore, we formulate the hypotheses for this two-tailed test as

$$H_0: \mu = 14 \text{ hours}$$

$$H_A: \mu \neq 14 \text{ hours}$$

- b. Given  $n = 35$ ,  $\bar{x} = 16.37$ , and  $s = 7.22$ , we calculate the value of the test statistic as

$$t_{34} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{16.37 - 14}{7.22/\sqrt{35}} = 1.94.$$

- c. Since  $t_{34} = 1.94 > 0$ , the  $p$ -value for a two-tailed test is  $2P(T_{34} \geq t_{34})$ . Referencing the  $t$  table for  $df = 34$ , we find that the exact probability  $P(T_{34} \geq 1.94)$  cannot be determined. Table 9.3 shows a portion of the  $t$  table where we see that  $t_{34} = 1.94$  lies between 1.691 and 2.032. This means that  $P(T_{34} \geq 1.94)$  is strictly between  $P(T_{34} \geq 2.032) = 0.025$  and  $P(T_{34} \geq 1.691) = 0.05$ —that is,  $0.025 < P(T_{34} \geq 1.94) < 0.05$ . Multiplying this double inequality by 2 results in  $0.05 < p\text{-value} < 0.10$ . In Example 9.13, we will show how to use Excel to find exact  $p$ -values.

**TABLE 9.3** Portion of the  $t$  Table

df	Area in Upper Tail, $\alpha$					
	0.20	0.10	0.05	0.025	0.01	0.005
1	1.376	3.078	6.341	12.706	31.821	63.657
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
34	0.852	1.307	1.691	2.032	2.441	2.728

- d. Since the  $p$ -value satisfies  $0.05 < p\text{-value} < 0.10$ , it must be greater than  $\alpha = 0.05$ ; we do not reject the null hypothesis. Therefore, the mean study time of students at the university is not significantly different from today's national average of 14 hours per week.

## Using Excel to Test $\mu$ When $\sigma$ Is Unknown

Again we find that Excel's functions are quite useful when calculating the value of the test statistic as well as calculating the exact  $p$ -value. Consider the following example.

### EXAMPLE 9.13

Residents of Hawaii have the longest life expectancies, averaging 81.48 years ([www.worldlifeexpectancy.com](http://www.worldlifeexpectancy.com); data retrieved June 4, 2012). A sociologist collects data on the age at death for 50 recently deceased Michigan residents. Table 9.4 shows a portion of the data.

- The sociologist believes that the life expectancies of Michigan residents are significantly less than those of Hawaii residents. Specify the competing hypotheses to test this belief.
- Calculate the value of the test statistic and the exact  $p$ -value.
- At the 1% significance level, do the data support the sociologist's belief?

**TABLE 9.4** Michigan Residents' Age at Death,  $n = 50$

**FILE**  
*MI\_Life\_Expectancy*

Age at Death
76.4
76.0
$\vdots$
73.6

#### SOLUTION:

- a. In order to determine whether Michigan residents have shorter life expectancies than Hawaii residents, we set up the following competing hypotheses

$$H_0: \mu \geq 81.48$$

$$H_A: \mu < 81.48$$

- b. As we saw in earlier chapters, Excel has all the necessary built-in functions to calculate the value of the test statistic,  $t_{df} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ . We open **MI\_Life\_Expectancy**, find an empty cell, and input “=(AVERAGE(A2:A51) – 81.48)/(STDEV.S(A2:A51)/sqrt(50)).” Excel returns a value of  $-4.75$ , so  $t_{49} = -4.75$ . Excel offers three different functions to find probabilities for the  $t$  distribution. The choice of the appropriate function depends on the specification of the competing hypotheses. In order to obtain the  $p$ -value for a right-tailed test, we use the function “=T.DIST.RT( $t_{df}$ ,  $df$ ),” where  $t_{df}$  is the value of the test statistic and  $df$  is the relevant degrees of freedom. The function “=T.DIST( $t_{df}$ ,  $df$ , 1)” returns the  $p$ -value for a left-tailed test, where the argument 1 prompts Excel to return a cumulative probability. The function “=T.DIST.2T( $|t_{df}|$ ,  $df$ )” returns the  $p$ -value for a two-tailed test, but note that we use the absolute value of the test statistic as the first argument. In order to find the exact  $p$ -value for this example—that is,  $P(T_{49} \leq -4.75)$ —we input “=T.DIST( $-4.75$ , 49, 1).” Excel returns  $9.06 \times 10^{-6}$ ; this indicates that  $P(T_{49} \leq -4.75) = 9.06 \times 10^{-6} \approx 0$ .
- c. Since the  $p$ -value is less than 0.01, we reject the null hypothesis. At the 1% significance level, the data suggest that Michigan residents have shorter life spans than Hawaii residents. These results support the claim of the sociologist.

## SYNOPSIS OF INTRODUCTORY CASE



A recent report claims that undergraduates are studying far less today as compared to five decades ago (*The Boston Globe*, July 4, 2010). The report finds that in 1961 students invested 24 hours per week in their academic pursuits, whereas today's students study an average of 14 hours per week. In an attempt to determine whether or not this national trend is present at a large university in California, 35 students are randomly selected and asked their average study time per week (in hours). The sample produces a mean of 16.37 hours with a standard deviation of 7.22 hours. Two hypothesis tests are conducted. The first test examines whether the mean study time of students at this university is below the 1961 national average of 24 hours per

week. At the 5% significance level, the sample data suggest that the mean is significantly less than 24 hours per week. The second test investigates whether the mean study time of students at this university differs from today's national average of 14 hours per week. At the 5% significance level, the results suggest that the mean study time is not significantly different from 14 hours per week. Thus, the sample results support the overall findings of the report: undergraduates study, on average, 14 hours per week, far below the 1961 average of 24 hours per week. The present analysis, however, does not explain why that might be the case. For instance, it cannot be determined whether students have just become lazier, or if with the advent of the computer, they can access information in less time.

## EXERCISES 9.3

### Mechanics

31. Consider the following hypotheses:

$$H_0: \mu \leq 210$$

$$H_A: \mu > 210$$

Approximate the  $p$ -value for this test based on the following sample information.

- $\bar{x} = 216; s = 26; n = 40$
- $\bar{x} = 216; s = 26; n = 80$
- $\bar{x} = 216; s = 16; n = 40$
- $\bar{x} = 214; s = 16; n = 40$

32. Which of the sample information in the preceding question enables us to reject the null hypothesis at  $\alpha = 0.01$  and at  $\alpha = 0.10$ ?

33. Consider the following hypotheses:

$$H_0: \mu = 12$$

$$H_A: \mu \neq 12$$

Approximate the  $p$ -value for this test based on the following sample information.

- $\bar{x} = 11; s = 3.2; n = 36$
- $\bar{x} = 13; s = 3.2; n = 36$
- $\bar{x} = 11; s = 2.8; n = 36$
- $\bar{x} = 11; s = 2.8; n = 49$

34. Which of the sample information in the preceding question enables us to reject the null hypothesis at  $\alpha = 0.01$  and at  $\alpha = 0.10$ ?

35. Determine the critical values for the following tests for the population mean with an unknown population standard deviation. The analysis is based on 18 observations drawn from a normally distributed population at a 1% level of significance.

- $H_0: \mu \leq 52$  versus  $H_A: \mu > 52$
- $H_0: \mu = 9.2$  versus  $H_A: \mu \neq 9.2$
- $H_0: \mu \geq 5.6$  versus  $H_A: \mu < 5.6$
- $H_0: \mu = 10$  versus  $H_A: \mu \neq 10$

36. In order to test if the population mean differs from 16, you draw a random sample of 32 observations and compute the sample mean and the sample standard deviation as 15.2 and 0.6, respectively. Use (a) the  $p$ -value approach and (b) the critical value approach to implement the test at a 1% level of significance.

37. In order to conduct a hypothesis test for the population mean, a random sample of 24 observations is drawn from a normally distributed population. The resulting sample mean and sample standard deviation are calculated as 4.8 and 0.8, respectively. Use the  $p$ -value approach to conduct the following tests at  $\alpha = 0.05$ .

- $H_0: \mu \leq 4.5$  against  $H_A: \mu > 4.5$
- $H_0: \mu = 4.5$  against  $H_A: \mu \neq 4.5$

38. Use the critical value approach to conduct the same two tests in the preceding question at  $\alpha = 0.05$ .

39. Consider the following hypotheses:

$$H_0: \mu = 8$$

$$H_A: \mu \neq 8$$

The population is normally distributed. A sample produces the following observations:

6	9	8	7	7	11	10
---	---	---	---	---	----	----

Use the  $p$ -value approach to conduct the test at a 5% level of significance.

40. Consider the following hypotheses:

$$H_0: \mu \geq 100$$

$$H_A: \mu < 100$$

The population is normally distributed. A sample produces the following observations:

95	99	85	80	98	97
----	----	----	----	----	----

Use the critical value approach to conduct the test at a 1% level of significance.

### Applications

41. A machine that is programmed to package 1.20 pounds of cereal in each cereal box is being tested for its accuracy. In a sample of 36 cereal boxes, the mean and standard deviation are calculated as 1.22 pounds and 0.06 pound, respectively.

- Set up the null and the alternative hypotheses to determine if the machine is working improperly—that is, it is either underfilling or overfilling the cereal boxes.
- Calculate the value of the test statistic.
- Approximate the  $p$ -value. At a 5% level of significance, can you conclude that the machine is working improperly? Explain.
- Repeat the exercise using the critical value approach.

42. The manager of a small convenience store does not want her customers standing in line for too long prior to a purchase. In particular, she is willing to hire an employee for another cash register if the average wait time of the customers is more than five minutes. She randomly observes the wait time (in minutes) of customers during the day as:

3.5	5.8	7.2	1.9	6.8	8.1	5.4
-----	-----	-----	-----	-----	-----	-----

- Set up the null and the alternative hypotheses to determine if the manager needs to hire another employee.
- Calculate the value of the test statistic. What assumption regarding the population is necessary to implement this step?
- Use the critical value approach to decide whether the manager needs to hire another employee at  $\alpha = 0.10$ .
- Repeat the above analysis with the  $p$ -value approach.

43. Small, energy-efficient, Internet-centric, new computers are increasingly gaining popularity (*The New York Times*, July 20, 2008). These computers, often called netbooks, have scant onboard memory and are intended largely for surfing websites and checking e-mail. Some of the biggest companies are wary of the new breed of computers because their low price could threaten PC makers' already thin profit margins. An analyst comments that the larger companies have a cause for concern since the mean price of these small computers has fallen below \$350. She examines six popular brands of these small computers and records their retail prices as:

\$322	\$269	\$373	\$412	\$299	\$389
-------	-------	-------	-------	-------	-------

- What assumption regarding the distribution of the price of small computers is necessary to test the analyst's claim?
  - Specify the appropriate null and alternative hypotheses to test the analyst's claim.
  - Calculate the value of the test statistic.
  - At the 5% significance level, specify the critical value(s). What is the conclusion to the test? Should the larger computer companies be concerned?
44. A local brewery wishes to ensure that an average of 12 ounces of beer is used to fill each bottle. In order to analyze the accuracy of the bottling process, the bottler takes a random sample of 48 bottles. The sample mean weight and the sample standard deviation of the bottles are 11.80 ounces and 0.8 ounce, respectively.
- State the null and the alternative hypotheses for the test.
  - Do you need to make any assumption regarding the population for testing?
  - At  $\alpha = 0.05$ , specify the critical value(s). What is the decision rule?
  - Make a recommendation to the bottler.
45. Based on the average predictions of 47 members of the National Association of Business Economists (NABE), the U.S. gross domestic product (GDP) will expand by 3.2% in 2011 (*The Wall Street Journal*, May 23, 2010). Suppose the sample standard deviation of their predictions was 1%. At a 5% significance level, test if the mean forecast GDP of all NABE members is greater than 3%.
46. A car manufacturer is trying to develop a new sports car. Engineers are hoping that the average amount of time that the car takes to go from 0 to 60 miles per hour is below 6 seconds. The car company tested 12 of the cars and clocked their performance times. Three of the cars clocked in at 5.8 seconds, 5 cars at 5.9 seconds, 3 cars at 6.0 seconds, and 1 car at 6.1 seconds. At a 5% level of significance, test if the new sports car is meeting its goal to go from 0 to 60 miles per hour in less than 6 seconds. Assume a normal distribution for the analysis.

47. In September 2007, U.S. home prices fell at a record pace, and price declines in Los Angeles and Orange counties in California outpaced other major metropolitan areas (*Los Angeles Times*, November 28, 2007). The report was based on the Standard & Poor's/Case-Shiller index that measures the value of single-family homes based on their sales histories. According to this index, the prices in San Diego dropped by an average of 9.6% from a year earlier. Assume that the survey was based on recent sales of 34 houses in San Diego that also resulted in a standard deviation of 5.2%. Can we conclude that the mean drop of all home prices in San Diego is greater than the 7% drop in Los Angeles? Use a 1% level of significance for the analysis.

48. A mortgage specialist would like to analyze the average mortgage rates for Atlanta, Georgia. He studies the following sample APR quotes. These are the annual percentage rates (APR) for 30-year fixed loans. If he is willing to assume that these rates are randomly drawn from a normally distributed population, can he conclude that the mean mortgage rate for the population exceeds 4.2%? Test the hypothesis at a 10% level of significance using (a) the  $p$ -value approach and (b) the critical value approach.

Financial Institution	APR
G Squared Financial	4.125%
Best Possible Mortgage	4.250
Hersch Financial Group	4.250
Total Mortgages Services	4.375
Wells Fargo	4.375
Quicken Loans	4.500
Amerisave	4.750

SOURCE: MSN Money.com; data retrieved October 1, 2010.

49. (Use Excel) One of the consequences of the economic meltdown has been a free fall of the stock market's average price/earnings ratio, or P/E ratio (*The Wall Street Journal*, August 30, 2010). Generally, a high P/E ratio suggests that investors are expecting higher earnings growth in the future compared to companies with a lower P/E ratio. An analyst wants to determine if the P/E ratio of firms in the footwear industry is different from the overall average of 14.9. The table below shows the P/E ratios for a sample of seven firms in the footwear industry:

Firm	P/E Ratio
Brown Shoe Co., Inc.	20.54
Collective Brands, Inc.	9.33
Crocs, Inc.	22.63
DSW, Inc.	14.42
Nike, Inc.	18.68
Skechers USA, Inc.	9.35
Timberland Co.	14.93

SOURCE: <http://biz.yahoo.com>, data retrieved August 23, 2010.



- a. State the null and the alternative hypotheses in order to test whether the P/E ratio of firms in the footwear industry differs from the overall average of 14.9.
  - b. What assumption regarding the population is necessary?
  - c. Use Excel to calculate the value of the test statistic and the exact  $p$ -value.
  - d. At  $\alpha = 0.10$ , does the P/E ratio of firms in the footwear industry differ from the overall average of 14.9? Explain.
50. **FILE MPG.** The data accompanying this exercise show miles per gallon (MPG).
- a. State the null and the alternative hypotheses in order to test whether the average MPG differs from 95.
  - b. Use Excel to calculate the value of the test statistic and the exact  $p$ -value.
  - c. At  $\alpha = 0.05$ , can you conclude that the average MPG differs from 95?
51. **FILE Debt Payments.** A recent study found that consumers are making average monthly debt payments of \$983 (Experian.com, November 11, 2010). The data accompanying this exercise show the average debt payments for 26 metropolitan areas; a portion of the data is shown in the following table.

City	Debt Payments
Washington, D.C.	\$1,285
Seattle	1,135
:	:
Pittsburgh	763

SOURCE: www.Experian.com, November 11, 2010.

- a. State the null and the alternative hypotheses in order to test whether average monthly debt payments are greater than \$900.
  - b. What assumption regarding the population is necessary to implement this step?
  - c. Use Excel to calculate the value of the test statistic and the exact  $p$ -value.
  - d. At  $\alpha = 0.05$ , are average monthly debt payments greater than \$900? Explain.
52. **FILE Highway Speeds.** A police officer is concerned about speeds on a certain section of Interstate 95. The data accompanying this exercise show the speeds of 40 cars on a Saturday afternoon.
- a. The speed limit on this portion of Interstate 95 is 65 mph. Specify the competing hypotheses in order to determine if the average speed is greater than the speed limit.
  - b. Specify the critical value(s) for the test at the 1% significance level.
  - c. Compute the value of the test statistic.
  - d. At  $\alpha = 0.01$ , are the officer's concerns warranted? Explain.
53. **FILE Lottery.** A recent article found that Massachusetts residents spent an average of \$860.70 on the lottery in 2010, more than three times the U.S. average (www.businessweek.com, March 14, 2012). A researcher at a Boston think tank believes that Massachusetts residents spend significantly less than this amount. He surveys 100 Massachusetts residents and asks them about their annual expenditures on the lottery.
- a. Specify the competing hypotheses to test the researcher's claim.
  - b. Specify the critical value(s) of the test at the 10% significance level.
  - c. Compute the value of the test statistic.
  - d. At the 10% significance level, do the data support the researcher's claim? Explain.

## 9.4 HYPOTHESIS TEST FOR THE POPULATION PROPORTION

### LO 9.6

As discussed earlier, sometimes the variable of interest is *qualitative* rather than *quantitative*. While the population mean  $\mu$  describes quantitative data, the population proportion  $p$  is the essential descriptive measure when the data type is qualitative. The parameter  $p$  represents the proportion of observations with a particular attribute.

As in the case for the population mean, we estimate the population proportion on the basis of its sample counterpart. In particular, we use the sample proportion  $\bar{P}$  to estimate the population proportion  $p$ . Recall that although  $\bar{P}$  is based on a binomial distribution, it can be approximated by a normal distribution in large samples. This approximation is considered valid when  $np \geq 5$  and  $n(1 - p) \geq 5$ . Since  $p$  is not known, we typically test the sample size requirement under the hypothesized value of the population proportion  $p_0$ . In most applications, the sample size is large and the normal distribution approximation is justified. However, when the sample size is not deemed large enough, the statistical methods suggested here for inference regarding the population proportion are no longer valid.

Specify the test statistic for the population proportion.

Recall from Chapter 7 that the mean and the standard error of the sample proportion  $\bar{P}$  are given by  $E(\bar{P}) = p$  and  $se(\bar{P}) = \sqrt{p(1-p)/n}$ , respectively. The test statistic for  $p$  is defined as follows.

#### TEST STATISTIC FOR $p$

The value of the **test statistic** for the hypothesis test of the **population proportion  $p$**  is computed as

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}},$$

where  $p_0$  is the hypothesized value of the population proportion. This formula is valid only if  $\bar{P}$  (approximately) follows a normal distribution.

The following examples elaborate on the four-step procedure for a hypothesis test for the population proportion.

#### EXAMPLE 9.14

A popular weekly magazine asserts that fewer than 40% of households in the United States have changed their lifestyles because of escalating gas prices. A recent survey of 180 households finds that 67 households have made lifestyle changes due to escalating gas prices.

- Specify the competing hypotheses to test the magazine's claim.
- Calculate the value of the test statistic and the corresponding  $p$ -value.
- At a 10% level of significance, what is the conclusion to the test?

#### SOLUTION:

- We wish to establish that the population proportion is less than 0.40—that is,  $p < 0.40$ . Thus, we construct the competing hypotheses as

$$H_0: p \geq 0.40$$

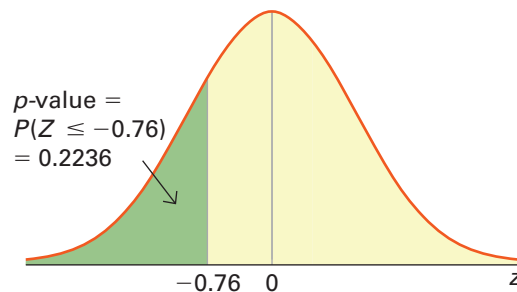
$$H_A: p < 0.40$$

- We first ensure that the normality condition is satisfied. Since both  $np_0$  and  $n(1-p_0)$  exceed 5, the normal approximation is justified. We use the sample proportion,  $\bar{p} = 67/180 = 0.3722$ , to compute the value of the test statistic as

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.3722 - 0.40}{\sqrt{0.40(1-0.40)/180}} = -0.76.$$

Since this is a left-tailed test for the population proportion, we compute the  $p$ -value as  $P(Z \leq z) = P(Z \leq -0.76) = 0.2236$ . Figure 9.8 shows the value of the test statistic and the corresponding  $p$ -value.

**FIGURE 9.8** The  $p$ -value for a left-tailed test with  $z = -0.76$



- c. The  $p$ -value of 0.2236 is greater than the chosen  $\alpha = 0.10$ . Therefore, we do not reject the null hypothesis. This means that the magazine's claim that fewer than 40% of households in the United States have changed their lifestyles because of escalating gas prices is not justified by the sample data. Such a conclusion may be welcomed by firms that have invested in alternative energy.

### EXAMPLE 9.15

Nearly one in three children and teens in the United States is obese or overweight (*Health*, October 2010). A health practitioner in the Midwest collects data on 200 children and teens and finds that 84 of them are either obese or overweight.

- The health practitioner believes that the proportion of obese and overweight children in the Midwest is not representative of the national proportion. Specify the competing hypotheses to test her claim.
- At the 1% significance level, specify the critical value(s).
- Calculate the value of the test statistic.
- Do the sample data support the health practitioner's belief?

#### SOLUTION:

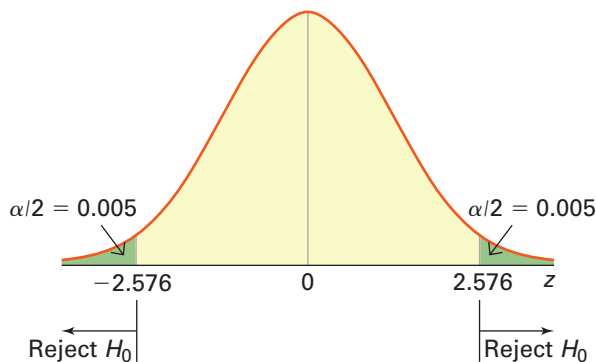
- a. The parameter of interest is again the population proportion  $p$ . The health practitioner wants to test if the population proportion of obese or overweight children in the Midwest differs from the national proportion of  $1/3 \approx 0.33$ . We construct the hypotheses as

$$H_0: p = 0.33$$

$$H_A: p \neq 0.33$$

- b. When evaluated at  $p_0 = 0.33$  with  $n = 200$ , the normality requirement that  $np \geq 5$  and  $n(1 - p) \geq 5$  is easily satisfied. Given a 1% level of significance and a two-tailed test,  $\alpha/2 = 0.01/2 = 0.005$  is used to find  $z_{\alpha/2} = z_{0.005} = 2.576$ . As shown in Figure 9.9, the critical values are  $-2.576$  and  $2.576$ .

**FIGURE 9.9** The critical values for a two-tailed test with  $\alpha = 0.01$



- c. We use  $\bar{p} = 84/200 = 0.42$  to calculate the value of the test statistic as

$$z = \frac{0.42 - 0.33}{\sqrt{0.33(1 - 0.33)/200}} = 2.71.$$

- d. The decision rule is to reject  $H_0$  if  $z < -2.576$  or if  $z > 2.576$ . Since the value of the test statistic,  $z = 2.71$ , is greater than 2.576, the appropriate decision is to reject the null hypothesis. Therefore, at the 1% significance level, the practitioner concludes that the proportion of obese or overweight children in the Midwest is not the same as the national proportion of 0.33. Given that the test statistic fell in the right side of the distribution, the practitioner can conduct further analysis to determine whether or not the proportion of obese or overweight children in the Midwest is significantly greater than the national proportion.

## EXERCISES 9.4

### Mechanics

54. Consider the following hypotheses:

$$H_0: p \geq 0.38$$

$$H_A: p < 0.38$$

Compute the  $p$ -value based on the following sample information.

- $x = 22$ ;  $n = 74$
  - $x = 110$ ;  $n = 300$
  - $\bar{p} = 0.34$ ;  $n = 50$
  - $\bar{p} = 0.34$ ;  $n = 400$
55. Which sample information in the preceding question enables us to reject the null hypothesis at  $\alpha = 0.01$  and at  $\alpha = 0.10$ ?

56. Consider the following hypotheses:

$$H_0: p = 0.32$$

$$H_A: p \neq 0.32$$

Compute the  $p$ -value based on the following sample information

- $x = 20$ ;  $n = 66$
  - $x = 100$ ;  $n = 264$
  - $\bar{p} = 0.40$ ;  $n = 40$
  - $\bar{p} = 0.38$ ;  $n = 180$
57. Which sample information in the preceding question enables us to reject the null hypothesis at  $\alpha = 0.05$  and at  $\alpha = 0.10$ ?
58. Specify the critical value(s) for the following tests for the population proportion. The analysis is conducted at a 5% level of significance.
- $H_0: p \leq 0.22$ ;  $H_A: p > 0.22$
  - $H_0: p = 0.69$ ;  $H_A: p \neq 0.69$
  - $H_0: p \geq 0.56$ ;  $H_A: p < 0.56$
59. In order to conduct a hypothesis test for the population proportion, you sample 320 observations that result in 128 successes. Use the  $p$ -value approach to conduct the following tests at  $\alpha = 0.05$ .

a.  $H_0: p \geq 0.45$ ;  $H_A: p < 0.45$

b.  $H_0: p = 0.45$ ;  $H_A: p \neq 0.45$

60. Repeat the preceding exercise using the critical value approach at  $\alpha = 0.01$ .
61. You would like to determine if the population probability of success differs from 0.70. You find 62 successes in 80 binomial trials. Implement the test at a 1% level of significance.
62. You would like to determine if more than 50% of the observations in a population are below 10. At  $\alpha = 0.05$ , conduct the test on the basis of the following 20 sample observations:

8	12	5	9	14	11	9	3	7	8
12	6	8	9	2	6	11	4	13	10

### Applications

63. A recent study by Allstate Insurance Co. finds that 82% of teenagers have used cell phones while driving (*The Wall Street Journal*, May 5, 2010). In October 2010, Massachusetts enacted a law that forbids cell phone use by drivers under the age of 18. A policy analyst would like to determine whether the law has decreased the proportion of drivers under the age of 18 who use a cell phone.
- State the null and the alternative hypotheses to test the policy analyst's objective.
  - Suppose a sample of 200 drivers under the age of 18 results in 150 who still use a cell phone while driving. What is the value of the test statistic? What is the  $p$ -value?
  - At  $\alpha = 0.05$ , has the law been effective?
  - Repeat this exercise using the critical value approach with  $\alpha = 0.05$ .
64. In order to endure financial hardships such as unemployment and medical emergencies, Americans have increasingly been raiding their already fragile retirement accounts (*MSN Money*, July 16, 2008). It is reported that between 1998 and 2004, about 12% of

families with 401(k) plans borrowed from them. An economist is concerned that this percentage now exceeds 20%. He randomly surveys 190 households with 401(k) plans and finds that 50 are borrowing against them.

- a. Set up the null and the alternative hypotheses to test the economist's concern.
  - b. Compute the value of the test statistic.
  - c. Use the  $p$ -value approach to test if the economist's concern is justifiable at  $\alpha = 0.05$ .
65. The margarita is one of the most common tequila-based cocktails, made with tequila mixed with Triple Sec and lime or lemon juice, often served with salt on the glass rim. A common ratio for a margarita is 2:1:1, which includes 50% tequila, 25% Triple Sec, and 25% fresh lime or lemon juice. A manager at a local bar is concerned that the bartender uses incorrect proportions in more than 50% of margaritas. He secretly observes the bartender and finds that he used the correct proportions in only 10 out of 30 margaritas. Use the critical value approach to test if the manager's suspicion is justified at  $\alpha = 0.05$ .
66. A movie production company is releasing a movie with the hopes of many viewers returning to see the movie in the theater for a second time. Their target is to have 30 million viewers, and they want more than 30% of the viewers to return to see the movie again. They show the movie to a test audience of 200 people, and after the movie they asked them if they would see the movie in theaters again. Of the test audience, 68 people said they would see the movie again.
- a. At a 5% level of significance, test if more than 30% of the viewers will return to see the movie again.
  - b. Repeat the analysis at a 10% level of significance.
  - c. Interpret your results.
67. Recent research commissioned by Vodafone suggests that older workers are the happiest employees (*BBC News*, July 21, 2008). The report documents that 70% of older workers in England feel fulfilled, compared with just 50% of younger workers. A demographer believes that an identical pattern does not exist in Asia. A survey of 120 older workers in Asia finds that 75 feel fulfilled. A similar survey finds that 58% of 210 younger workers feel fulfilled.
- a. At a 5% level of significance, test if older workers in Asia feel less fulfilled than their British counterparts.
  - b. At a 5% level of significance, test if younger workers in Asia feel more fulfilled than their British counterparts.
68. A politician claims that he is supported by a clear majority of voters. In a recent survey, 24 out of 40 randomly

selected voters indicated that they would vote for the politician. Is the politician's claim justified at a 5% level of significance?

69. New research shows that many banks are unwittingly training their online customers to take risks with their passwords and other sensitive account information, leaving them more vulnerable to fraud (Yahoo.com, July 23, 2008). Even web-savvy surfers could find themselves the victims of identity theft because they have been conditioned to ignore potential signs about whether the banking site they are visiting is real or a bogus site served up by hackers. Researchers at the University of Michigan found design flaws in 78% of the 214 U.S. financial institution websites they studied. Is the above sample evidence sufficient to conclude that more than three out of four financial institutions that offer online banking facilities are prone to fraud? Use a 5% significance level for the test.
70. The Social Security Administration is not expected to provide any increases in Social Security benefits for the second straight year (*U.S. News & World Report*, October 4, 2010). With increasing medical prices, it is claimed that more than 60% of seniors are likely to make serious adjustments to their lifestyle. Test this claim at a 1% level of significance if in a survey of 140 seniors, 90 reported that they have made serious adjustments to their lifestyle.
71. **FILE Silicon Valley.** According to a report on workforce diversity, about 60% of the employees in high-tech firms in Silicon Valley are white and about 20% are Asian (<http://moneycnn.com>, November 9, 2011). Women, along with blacks and Hispanics, are highly underrepresented. Just about 30% of all employees are women, with blacks and Hispanics accounting for only about 15% of the workforce. Tara Jones is a recent college graduate, working for a large high-tech firm in Silicon Valley. She wants to determine if her firm faces the same diversity as in the report. She collects gender and ethnicity information on 50 employees in her firm. A portion of the data is shown in the accompanying table.

Gender	Ethnicity
Female	White
Male	White
:	:
Male	Nonwhite

- a. At the 5% level of significance, determine if the proportion of women in Tara's firm is different from 30%.
- b. At the 5% level of significance, determine if the proportion of whites in Tara's firm is more than 50%.

WRITING WITH STATISTICS



The Associated Press reports that income inequality is at record levels in the United States (September 28, 2010). Over the years, the rich have become richer while working-class wages have stagnated. A local Latino politician has been vocal regarding his concern about the welfare of Latinos, especially given the recent downturn of the U.S. economy. In various speeches, he has stated that the mean salary of Latino households in his county has fallen below the 2008 mean of \$49,000. He has also stated that the proportion of Latino households making less than \$30,000 has risen above the 2008 level of 20%. Both of his statements are based on income data for 36 Latino households in the county, as shown in Table 9.5.

**TABLE 9.5** Representative Sample of Latino Household Incomes in 2010

FILE

Latino\_Income

22	36	78	103	38	43
62	53	26	28	25	31
62	44	51	38	77	37
29	38	46	52	61	57
20	72	41	73	16	32
52	28	69	27	53	46

Incomes are measured in \$1,000s and have been adjusted for inflation.

Trevor Jones is a newspaper reporter who is interested in verifying the concerns of the local politician.

Trevor wants to use the sample information to:

1. Determine if the mean income of Latino households has fallen below the 2008 level of \$49,000.
2. Determine if the percentage of Latino households making less than \$30,000 has risen above 20%.

Sample  
Report—  
Income  
Inequality in  
the United  
States

One of the hotly debated topics in the United States is that of growing income inequality. Market forces such as increased trade and technological advances have made highly skilled and well-educated workers more productive, thus increasing their pay. Institutional forces, such as deregulation, the decline of unions, and the stagnation of the minimum wage, have contributed to income inequality. Arguably, this income inequality has been felt by minorities, especially African Americans and Latinos, since a very high proportion of both groups is working class. The condition has been further exacerbated by the Great Recession.

A sample of 36 Latino households resulted in a mean household income of \$46,278 with a standard deviation of \$19,524. The sample mean is below the 2008 level of \$49,000. In addition, nine Latino households, or 25%, make less than \$30,000; the corresponding percentage in 2008 was 20%. Based on these results, a politician concludes that current market conditions continue to negatively impact the welfare of Latinos. However, it is essential to provide statistically significant evidence to substantiate these claims. Toward this end, formal tests of hypotheses regarding the population mean and the population proportion are conducted. The results of the tests are summarized in Table 9.A.



**TABLE 9.A** Test Statistic Values and  $p$ -Values for Hypothesis Tests

Hypotheses	Test Statistic Value	$p$ -value
$H_0: \mu \geq 49,000$ $H_A: \mu < 49,000$	$t_{35} = \frac{46,278 - 49,000}{19,524 / \sqrt{36}} = -0.84$	0.2033
$H_0: p \leq 0.20$ $H_A: p > 0.20$	$z = \frac{0.25 - 0.20}{\sqrt{\frac{(0.20)(0.80)}{36}}} = 0.75$	0.2266

When testing whether the mean income of Latino households has fallen below the 2008 level of \$49,000, a test statistic value of  $-0.84$  is obtained. Given a  $p$ -value of 0.2033, the null hypothesis regarding the population mean, specified in Table 9.A, cannot be rejected at any reasonable level of significance. Similarly, given a  $p$ -value of 0.2266, the null hypothesis regarding the population proportion cannot be rejected. Therefore, sample evidence does not support the claims that the mean income of Latino households has fallen below \$49,000 or that the proportion of Latino households making less than \$30,000 has risen above 20%. Perhaps the politician's remarks were based on a cursory look at the sample statistics and not on a thorough statistical analysis.

## CONCEPTUAL REVIEW

### LO 9.1 Define the null hypothesis and the alternative hypothesis.

Every hypothesis test contains two competing hypotheses: the **null hypothesis**, denoted  $H_0$ , and the **alternative hypothesis**, denoted  $H_A$ . We can think of the null hypothesis as corresponding to a presumed default state of nature or status quo, whereas the alternative hypothesis contradicts the default state or status quo.

On the basis of sample information, we either **reject  $H_0$**  or **do not reject  $H_0$** . As a general guideline, whatever we wish to establish is placed in the alternative hypothesis. If we reject the null hypothesis, we are able to conclude that the alternative hypothesis is true.

Hypothesis tests can be **one-tailed** or **two-tailed**. A one-tailed test allows the rejection of the null hypothesis only on one side of the hypothesized value of the population parameter. In a two-tailed test, the null hypothesis can be rejected on both sides of the hypothesized value of the population parameter.

### LO 9.2 Distinguish between Type I and Type II errors.

Since the statistical conclusion of a hypothesis test relies on sample data, there are two types of errors that may occur: a **Type I error** or a **Type II error**. A Type I error is committed when we reject the null hypothesis when it is actually true. On the other hand, a Type II error is made when we do not reject the null hypothesis when it is actually false. We denote the probability of a Type I error by  $\alpha$  and the probability of a Type II error by  $\beta$ . For a given sample size  $n$ , a decrease (increase) in  $\alpha$  will increase (decrease)  $\beta$ . However, both  $\alpha$  and  $\beta$  will decrease if the sample size  $n$  increases.

### LO 9.3 Conduct a hypothesis test using the $p$ -value approach.

Every hypothesis test can be implemented by following a four-step procedure. There are two equivalent approaches, namely the  **$p$ -value approach** and the **critical value approach**. For the  $p$ -value approach, we follow these four steps:

**Step 1. Specify the null and the alternative hypotheses.** We identify the relevant population parameter of interest, determine whether it is a one- or a two-tailed test and, most importantly, include some form of the equality sign in the null hypothesis and place whatever we wish to establish in the alternative hypothesis.

**Step 2. Specify the significance level.** Before implementing a hypothesis test, we first specify  $\alpha$ , which is the *allowed* probability of making a Type I error.

**Step 3. Calculate the value of the test statistic and the  $p$ -value.** We derive the value of the test statistic by converting the estimate of the relevant population parameter into its corresponding standardized value, either  $z$  or  $t_{df}$ .

The  $p$ -value is the probability that the test statistic is as extreme as its value computed from the given sample. We can also interpret it as the *observed* probability of making a Type I error. If the test statistic follows the  $z$  distribution, then the  $p$ -value is calculated as

- $P(Z \geq z)$  for a right-tailed test,
- $P(Z \leq z)$  for a left-tailed test, or
- $2P(Z \geq z)$  if  $z > 0$  or  $2P(Z \leq z)$  if  $z < 0$  for a two-tailed test.

$Z$  and  $z$  are replaced with  $T_{df}$  and  $t_{df}$  if the test statistic follows the  $t_{df}$  distribution with degrees of freedom  $df = n - 1$ .

**Step 4. State the conclusion and interpret results.** The decision rule is to reject the null hypothesis if the  $p$ -value  $< \alpha$ , where  $\alpha$  is the chosen significance level.

#### **LO 9.4 Conduct a hypothesis test using the critical value approach.**

For the critical value approach, we follow these four steps:

**Step 1** is the same as the  $p$ -value approach; that is, we specify the competing hypotheses.

**Step 2. Specify the significance level and find the critical value(s).** We first specify  $\alpha$ , which is the *allowed* probability of making a Type I error. The critical value(s) is a point that separates the rejection region from the nonrejection region. If the test statistic follows the  $z$  distribution, then for a given  $\alpha$ , we find the critical value(s) as

- $z_\alpha$  where  $P(Z \geq z_\alpha) = \alpha$ , for a right-tailed test;
- $-z_\alpha$  where  $P(Z \leq -z_\alpha) = \alpha$ , for a left-tailed test; or
- $-z_{\alpha/2}$  and  $z_{\alpha/2}$ , where  $P(Z \geq z_{\alpha/2}) = \alpha/2$  for a two-tailed test.

$Z$  and  $z_\alpha$  are replaced with  $T_{df}$  and  $t_{\alpha,df}$  if the test statistic follows the  $t_{df}$  distribution with  $n - 1$  degrees of freedom.

**Step 3. Calculate the value of the test statistic.** We derive the value of the test statistic by converting the estimate of the relevant population parameter into its corresponding standardized value, either  $z$  or  $t_{df}$ .

**Step 4. State the conclusion and interpret results.** The decision rule with the critical value approach is to reject the null hypothesis if the test statistic falls in the rejection region, or,

- For a right-tailed test, reject  $H_0$  if  $z > z_\alpha$ ;
- For a left-tailed test, reject  $H_0$  if  $z < -z_\alpha$ ; or
- For a two-tailed test, reject  $H_0$  if  $z < -z_{\alpha/2}$  or if  $z > z_{\alpha/2}$ .

$z$  is replaced by  $t_{df}$  if the assumed test statistic follows the  $t_{df}$  distribution with degrees of freedom  $df = n - 1$ .

#### **LO 9.5 Differentiate between the test statistics for the population mean.**

The value of the test statistic for the hypothesis test for the **population mean  $\mu$  when the population standard deviation  $\sigma$  is known** is computed as  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ , where  $\mu_0$  is the hypothesized value of the population mean. The value of the test statistic for the hypothesis test for the **population mean  $\mu$  when the population standard deviation  $\sigma$  is unknown** is computed as  $t_{df} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ , where  $\mu_0$  is the hypothesized value of the population mean and degrees of freedom  $df = n - 1$ .

**LO 9.6 Specify the test statistic for the population proportion.**

The value of the test statistic for the hypothesis test for the **population proportion  $p$**  is computed as  $z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$ , where  $p_0$  is the hypothesized value of the population proportion.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

72. A pharmaceutical company has developed a new drug for depression. There is a concern, however, that the drug also raises the blood pressure of its users. A researcher wants to conduct a test to validate this claim. Is the manager of the pharmaceutical company more concerned about a Type I error or a Type II error? Explain.
73. A company has developed a new diet that it claims will lower one's weight by more than 10 pounds. Health officials decide to conduct a test to validate this claim.
  - a. Is the manager of the company more concerned about a Type I error or a Type II error? Explain.
  - b. Should the consumers be more concerned about a Type I error or a Type II error? Explain.
74. An advertisement for a popular weight loss clinic suggests that participants in its new diet program lose, on average, more than 10 pounds. A consumer activist decides to test the authenticity of the claim. She follows the progress of 18 women who recently joined the weight reduction program. She calculates the mean weight loss of these participants as 10.8 pounds with a standard deviation of 2.4 pounds.
  - a. Set up the competing hypotheses to test the advertisement's claim.
  - b. At the 5% significance level, specify the critical value(s). What is the decision rule?
  - c. Calculate the value of the test statistic.
  - d. What does the consumer activist conclude?
75. A phone manufacturer wants to compete in the touch screen phone market. He understands that the lead product has a battery life of just 5 hours. The manufacturer claims that while the new touch phone is more expensive, its battery life is more than twice as long as that of the leading product. In order to test the claim, a researcher samples 45 units of the new phone and finds that the sample battery life averages 10.5 hours with a sample standard deviation of 1.8 hours.
  - a. Set up the competing hypotheses to test the manufacturer's claim.
  - b. Compute the value of the test statistic.
  - c. Use the critical value approach to test the phone manufacturer's claim at  $\alpha = 0.05$ .
  - d. Repeat the analysis with the  $p$ -value approach.
76. A city council is deciding whether or not to spend additional money to reduce the amount of traffic. The council decides that it will increase the transportation budget if the amount of waiting time for drivers exceeds 20 minutes. A sample of 32 main roads results in a mean waiting time of 22.08 minutes with a standard deviation of 5.42 minutes. Conduct a hypothesis test at a 1% level of significance to determine whether or not the city should increase its transportation budget.
77. Rates on 30-year fixed mortgages continue to be at historic lows (*Chron Business News*, September 23, 2010). According to Freddie Mac, the average rate for 30-year fixed loans for the week was 4.37%. An economist wants to test if there is any change in the mortgage rates in the following week. She searches the Internet for 30-year fixed loans in the following week and reports the rates offered by seven banks as 4.25%, 4.125%, 4.375%, 4.50%, 4.75%, 4.375%, and 4.875%. Assume that rates are normally distributed.
  - a. State the hypotheses to test if the average mortgage rate differs from 4.37%.
  - b. Specify the critical value(s) with  $\alpha = 0.05$ .
  - c. What is the value of the test statistic?
  - d. At the 5% significance level, does the average mortgage rate differ from 4.37%? Explain.
78. The Great Recession cost America trillions of dollars in lost wealth and also levied a heavy toll on the national psyche (*The Wall Street Journal*, December 21, 2009). According to a recent poll, just 33% of those surveyed said America was headed in the right direction. Suppose this poll was based on a sample of 1,000 people. Does the sample evidence suggest that the proportion of Americans who feel that America is headed in the right direction is below 35%? Use a 5% level of significance for the analysis. What if the sample size was 2,000?
79. A retailer is looking to evaluate its customer service. Management has determined that if the retailer wants to stay competitive, then it will have to have at least a 90% satisfaction rate among its

customers. Management will take corrective actions if the satisfaction rate falls below 90%. A survey of 1,200 customers showed that 1,068 were satisfied with their customer service.

- a. State the hypotheses to test if the retailer needs to improve its services.
  - b. What is the value of the test statistic?
  - c. Compute the  $p$ -value.
  - d. Interpret the results at  $\alpha = 0.05$ .
80. A national survey found that 33% of high school students said they texted or e-mailed while driving (*The Boston Globe*, June 8, 2012). These findings came a day after a Massachusetts teenager was convicted for causing a fatal crash while texting. A researcher wonders whether texting or e-mailing while driving is more prevalent among Massachusetts teens. He surveys 100 teens and 42% of them admitted that they texted or e-mailed while behind the wheel. Can he conclude at the 1% significance level that Massachusetts teens engage in this behavior at a rate greater than the national rate?
81. A television network is deciding whether or not to give its newest television show a spot during prime viewing time at night. For this to happen, it will have to move one of its most viewed shows to another slot. The network conducts a survey asking its viewers which show they would rather watch. The network will keep its current lineup of shows unless the majority of the customers want to watch the new show. The network receives 827 responses, of which 428 indicate that they would like to see the new show in the lineup.
- a. Set up the hypotheses to test if the television network should give its newest television show a spot during prime viewing time at night.
  - b. Compute the value of the test statistic.
  - c. Define the rejection region(s) at  $\alpha = 0.01$ .
  - d. What should the television network do?
82. A Pew Research study finds that 23% of Americans use only a cell phone, and no land line, for making phone calls (*The Wall Street Journal*, October 14, 2010). A year later, a researcher samples 200 Americans and finds that 51 of them use only cell phones for making phone calls.
- a. Set up the hypotheses in order to determine whether the proportion of Americans who solely use cell phones to make phone calls differs from 23%.
  - b. Compute the value of the test statistic and the corresponding  $p$ -value.
  - c. At  $\alpha = 0.05$ , are the sample data inconsistent with Pew Research's findings of 2010? What do the sample data suggest?
83. **FILE Metals.** Using data from the past 25 years, an investor wants to test whether the average return of Vanguard's Precious Metals and Mining Fund is greater than 12%. Assume returns are normally distributed with a population standard deviation of 30%.
- a. State the null and the alternative hypotheses for the test.
  - b. Use Excel's Z.TEST function to calculate the  $p$ -value.
  - c. At  $\alpha = 0.05$ , what is the conclusion? Is the return on Vanguard's Precious Metals and Mining Fund greater than 12%?
84. **FILE Midwest Drivers.** On average, Americans drive 13,500 miles per year (*The Boston Globe*, June 7, 2012). An economist gathers data on the driving habits of 50 residents in the Midwest.
- a. The economist believes that the average number of miles driven annually by Midwesterners is different from the U.S. average. Specify the competing hypotheses to test the economist's claim.
  - b. Use Excel to calculate the value of the test statistic and the exact  $p$ -value.
  - c. At the 10% significance level, do the data support the researcher's claim? Explain.
85. **FILE Convenience Stores.** An entrepreneur examines monthly sales (in \$1,000s) for 40 convenience stores in Rhode Island.
- a. State the null and the alternative hypotheses in order to test whether average sales differ from \$130,000.
  - b. Use Excel to calculate the value of the test statistic and the exact  $p$ -value.
  - c. At  $\alpha = 0.05$ , what is your conclusion to the test? Do average sales differ from \$130,000?
86. **FILE DJIA Volume.** The euro-zone crisis has wreaked havoc on U.S. stock markets (*The Wall Street Journal*, June 8, 2012). A portfolio analyst wonders if the average trading volume on the Dow Jones Industrial Average (DJIA) has increased since the beginning of the year. She gathers data on daily trading volumes for 30 days.
- a. She finds that the average trading volume in the beginning of the year was about 4,000 shares (in millions). Specify the competing hypotheses to test her claim.
  - b. Specify the critical value(s) at the 5% significance level.
  - c. Compute the value of the test statistic.
  - d. At the 5% significance level, does it appear that the trading volume has increased since the beginning of the year?
87. **FILE Study Hard.** A recent report suggests that business majors spend the least amount of time on course work than do all other college students (*The New York Times*, November 17, 2011). A provost of

a university conducts a survey of 50 business and 50 nonbusiness students. Students are asked if they study hard, defined as spending at least 20 hours per week on course work. The response shows “yes” if they study hard or “no” otherwise; a portion is shown in the following table.

Business Majors	Nonbusiness Majors
Yes	No
No	Yes
⋮	⋮
Yes	Yes

- At the 5% level of significance, determine if the percentage of business majors who study hard is less than 20%.
- At the 5% level of significance, determine if the percentage of nonbusiness majors who study hard is more than 20%.

## CASE STUDIES

**CASE STUDY 9.1** Harvard University has recently revolutionized its financial aid policies, aimed at easing the financial strain on middle and upper-middle income families (*Newsweek*, August 18–25, 2008). The expected contribution of students who are admitted to Harvard has been greatly reduced. Many other elite private colleges are following suit to compete for top students. The motivation for these policy changes stems from competition from public universities as well as political pressure.

A spokesman from an elite college claims that elite colleges have been very responsive to financial hardships faced by families due to the rising costs of education. Now, he says, families with an income of \$40,000 will have to spend less than \$6,500 to send their children to prestigious colleges. Similarly, families with incomes of \$80,000 and \$120,000 will have to spend less than \$20,000 and \$35,000, respectively, for their children’s education.

Although in general, the cost of attendance has gone down at each family-income level, it still varies by thousands of dollars among prestigious schools. The accompanying table shows information on the cost of attendance by family income for 10 prestigious schools.

**Data for Case Study 9.1** Cost of Attendance to Schools by Family Income

School	Family Income		
	\$40,000	\$80,000	\$120,000
Amherst College	\$ 5,302	\$19,731	\$37,558
Bowdoin College	5,502	19,931	37,758
Columbia University	4,500	12,800	36,845
Davidson College	5,702	20,131	37,958
Harvard University	3,700	8,000	16,000
Northwestern University	6,311	26,120	44,146
Pomona College	5,516	19,655	37,283
Princeton University	3,887	11,055	17,792
Univ. of California system	10,306	19,828	25,039
Yale University	4,300	6,048	13,946

SOURCE: *Newsweek*, August 18–25, 2008.

**FILE**  
Family\_Income

In a report, use the sample information to:

- Determine whether families with income of \$40,000 will have to spend less than \$6,500 to send their children to prestigious colleges. (Use  $\alpha = 0.05$ .)
- Repeat the hypothesis test from part 1 by testing the spokesman’s claims concerning college costs for families with incomes of \$80,000 and \$120,000, respectively. (Use  $\alpha = 0.05$ .)
- Assess the validity of the spokesman’s claims.



**CASE STUDY 9.2** The effort to reward city students for passing Advanced Placement tests is part of a growing trend nationally and internationally. Financial incentives are offered in order to lift attendance and achievement rates. One such program in Dallas, Texas, offers \$100 for every Advanced Placement test on which a student scores a three or higher (Reuters, September 20, 2010). A wealthy entrepreneur decides to experiment with the same idea of rewarding students to enhance performance, but in Chicago. He offers monetary incentives to students at an inner-city high school. Due to this incentive, 122 students take the Advancement Placement tests. Twelve tests are scored at 5, the highest possible score. There are 49 tests with scores of 3 and 4, and 61 tests with failing scores of 1 and 2. Historically, about 100 of these tests are taken at this school each year, where 8% score 5, 38% score 3 and 4, and the remaining are failing scores of 1 and 2.

In a report, use the sample information to:

1. Provide a descriptive analysis of student achievement on Advanced Placement before and after the monetary incentive is offered.
2. Conduct a hypothesis test that determines, at the 5% significance level, whether the monetary incentive has resulted in a higher proportion of scores of 5, the highest possible score.
3. At the 5% significance level, has the monetary incentive decreased the proportion of failing scores of 1 and 2?
4. Assess the effectiveness of monetary incentives in improving student achievement.

**CASE STUDY 9.3** The Gallup-Healthways Well-Being Index ([www.well-beingindex.com](http://www.well-beingindex.com)) provides an assessment measure of health and well-being of U.S. residents. By collecting periodic data on life evaluation, physical health, emotional health, healthy behavior, work environment, and basic access, this assessment measure is of immense value to researchers in diverse fields such as business, medical sciences, and journalism. The overall composite score, as well as a score in each of the above six categories, is calculated on a scale from 0 to 100, where 100 represents fully realized well-being. In 2009, the overall well-being index score of American residents was reported as 65.9. Let the following table represent the overall well-being score of a random sample of 35 residents in Hawaii.

**Data for Case Study 9.3** Overall Well-being of Hawaiians,  $n = 35$

20	40	40	100	60	20	40
90	90	60	60	90	90	90
80	100	90	80	80	80	100
70	90	80	100	20	70	90
80	30	80	90	90	80	30

In a report, use the sample information to:

1. Determine whether the well-being score of Hawaiians is more than the national average of 65.9 at the 5% significance level.
2. Determine if fewer than 40% of Hawaiians report a score below 50 at the 5% significance level.
3. Use your results to comment on the well-being of Hawaiians.

**FILE**  
Hawaiians

## APPENDIX 9.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.



## Minitab

### Testing $\mu$ , $\sigma$ Known

- A. (Replicating Example 9.10). From the menu choose **Stat > Basic Statistics > 1-Sample Z**.
- B. Select **One or more samples, each in a column** and select Debit Spending. Enter 500 for **Known standard deviation**. Select **Perform hypothesis test** and enter 8000 after **Hypothesized mean**. Choose **Options**. After **Alternative hypothesis**, select 'Mean < hypothesized mean.'

FILE

*Debit\_Spending*

### Testing $\mu$ , $\sigma$ Unknown

- A. (Replicating Example 9.13) From the menu choose **Stat > Basic Statistics > 1-Sample t**.
- B. Select **One or more samples, each in a column** and select Age. Select **Perform hypothesis test** and enter 81.48 after **Hypothesized mean**. Choose **Options**. After **Alternative hypothesis**, select 'Mean < hypothesized mean.'

FILE

*MI\_Life\_Expectancy*

### Testing $p$

- A. (Replicating Example 9.14) From the menu choose **Stat > Basic Statistics > 1-Proportion**.
- B. Choose **Summarized data** and then enter 67 after **Number of events** and 180 after **Number of trials**. Select **Perform hypothesis test** and enter 0.40 for **Hypothesized proportion**. Choose **Options**. After **Alternative hypothesis**, select 'Proportion < hypothesized proportion' and after **Method** select 'Normal approximation.'

## SPSS

### Testing $\mu$ , $\sigma$ Unknown

(Replicating Example 9.13). From the menu choose **Analyze > Compare Means > One-Sample T Test**. Under **Test Variable(s)**, select Age. After **Test Value**, enter 81.48.

FILE

*MI\_Life\_Expectancy*

## JMP

### Testing $\mu$ , $\sigma$ Known

- A. (Replicating Example 9.10). From the menu choose **Analyze > Distribution**.
- B. Under **Select Columns**, select Debit Spending, and then under **Cast Selected Columns into Roles**, select **Y, Columns**.
- C. Click on the red triangle in the output window beside Debit Spending. Choose **Test Mean**. After **Specify Hypothesized Mean**, enter 8000, and after **Enter true standard deviation to do z-test rather than t test**, enter 500.

FILE

*Debit\_Spending*

### Testing $\mu$ , $\sigma$ Unknown

- A. (Replicating Example 9.13). From the menu choose **Analyze > Distribution**.
- B. Under **Select Columns**, select Age, and then under **Cast Selected Columns into Roles**, select **Y, Columns**.
- C. Click on the red triangle in the output window beside Age. Choose **Test Mean**. After **Specify Hypothesized Mean**, enter 81.48.

FILE

*MI\_Life\_Expectancy*

# 10

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 10.1 Make inferences about the difference between two population means based on independent sampling.
- LO 10.2 Make inferences about the mean difference based on matched-pairs sampling.
- LO 10.3 Make inferences about the difference between two population proportions based on independent sampling.

# Statistical Inference Concerning Two Populations

In the preceding two chapters, we used estimation and hypothesis testing to analyze a single parameter, such as the population mean and the population proportion. In this chapter, we extend our discussion from the analysis of a single population to the comparison of two populations. We first analyze differences between two population means. For instance, an economist may be interested in analyzing the salary difference between male and female employees. Similarly, a marketing researcher might want to compare the operating lives of two popular brands of batteries. In these examples, we use independent sampling for the analysis. We will also consider the mean difference of two populations based on matched-pairs sampling. An example would be a consumer group activist wanting to analyze the mean weight of customers before and after they enroll in a new diet program. Finally, we look at qualitative data and compare the difference between two population proportions. For instance, marketing executives and advertisers are often interested in the different preferences between males and females when determining where to target advertising dollars. In each of the statistical inferences concerning two populations, we first develop the procedure for estimation and then follow with hypothesis testing.



## INTRODUCTORY CASE

### Effectiveness of Mandatory Caloric Postings

The federal health-care law enacted in March 2010 requires chain restaurants with 20 locations or more to post caloric information on their menus. The government wants calorie listings posted to make it easier for consumers to select healthier options. New York City pioneered the requirement of caloric information on menus in 2008, but research has shown mixed results on whether this requirement has prompted consumers to select healthier foods (*The Wall Street Journal*, August 31, 2010). Molly Hosler, a nutritionist in San Mateo, California, would like to study the effects of a recent local menu ordinance requiring caloric postings. She obtains transaction data for 40 Starbucks cardholders around the time that San Mateo implemented the ordinance. The average drink and food calories were recorded for each customer prior to the ordinance and then after the ordinance. Table 10.1 shows a portion of the data.

**TABLE 10.1** Average Caloric Intake Before and After Menu-Labeling Ordinance

**FILE**  
Drink\_Calories  
Food\_Calories

Customer	Drink Calories		Food Calories	
	Before	After	Before	After
1	141	142	395	378
2	137	140	404	392
⋮	⋮	⋮	⋮	⋮
40	147	141	406	400

Molly wants to use the sample information to:

1. Determine whether average calories of purchased drinks declined after the passage of the ordinance.
2. Determine whether average calories of purchased food declined after the passage of the ordinance.
3. Assess the implications of caloric postings for Starbucks and other chains.

A synopsis of this case is provided at the end of Section 10.2.

## 10.1 INFERENCE CONCERNING THE DIFFERENCE BETWEEN TWO MEANS

### LO 10.1

Make inferences about the difference between two population means based on independent sampling.

In this section, we consider statistical inference about the difference between two population means based on **independent random samples**. Independent random samples are samples that are completely unrelated to one another. Consider the example where we are interested in the difference between male and female salaries. For one sample, we collect data from the male population, while for the other sample we gather data from the female population. The two samples are considered to be independent because the selection of one is in no way influenced by the selection of the other. Similarly, in a comparison of battery lives between Brand A and Brand B, one sample comes from the Brand A population, while the other sample comes from the Brand B population. Again, both samples can be considered to be drawn independently.

#### INDEPENDENT RANDOM SAMPLES

Two (or more) random samples are considered independent if the process that generates one sample is completely separate from the process that generates the other sample. The samples are clearly delineated.

### Confidence Interval for $\mu_1 - \mu_2$

As discussed earlier, we use sample statistics to estimate the population parameter of interest. For example, the sample mean  $\bar{X}$  is the point estimator for the population mean  $\mu$ . In a similar vein, the difference between the two sample means  $\bar{X}_1 - \bar{X}_2$  is a point estimator for the difference between two population means  $\mu_1 - \mu_2$ , where  $\mu_1$  is the mean of the first population and  $\mu_2$  is the mean of the second population. The values of the sample means  $\bar{x}_1$  and  $\bar{x}_2$  are computed from two independent random samples with  $n_1$  and  $n_2$  observations, respectively.

Let's first discuss the sampling distribution of  $\bar{X}_1 - \bar{X}_2$ . As in the case of a single population mean, this estimator is unbiased; that is,  $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$ . Moreover, recall that the statistical inference regarding the population mean  $\mu$  is based on the condition that the sample mean  $\bar{X}$  is normally distributed. Similarly, for statistical inference regarding  $\mu_1 - \mu_2$ , it is imperative that the sampling distribution of  $\bar{X}_1 - \bar{X}_2$  is normal. Therefore, we generally assume that the two sample means are derived from two independent, normally distributed populations because a linear combination of normally distributed random variables is also normally distributed. If the underlying populations cannot be assumed to be normal, then by the central limit theorem, the sampling distribution of  $\bar{X}_1 - \bar{X}_2$  is approximately normal only if both sample sizes are sufficiently large; that is,  $n_1 \geq 30$  and  $n_2 \geq 30$ .

As in the case of a single population mean, we consider two scenarios. If we know the variances of the two populations  $\sigma_1^2$  and  $\sigma_2^2$  (or the standard deviations  $\sigma_1$  and  $\sigma_2$ ), we use the  $z$  distribution for the statistical inference. A more common case is to use the  $t_{df}$  distribution, where the sample variances,  $s_1^2$  and  $s_2^2$ , are used in place of the unknown population variances. When  $\sigma_1^2$  and  $\sigma_2^2$  are not known, we will examine two cases: (a) they can be assumed equal ( $\sigma_1^2 = \sigma_2^2$ ) or (b) they cannot be assumed equal ( $\sigma_1^2 \neq \sigma_2^2$ ).

The confidence interval for the difference in means is based on the same procedure outlined in Chapter 8. In particular, the formula for the confidence interval will follow the standard format given by: Point estimate  $\pm$  Margin of error.

We use sample data to calculate the point estimate for  $\mu_1 - \mu_2$  as the difference between the two sample means  $\bar{x}_1 - \bar{x}_2$ . The margin of error equals the standard error  $se(\bar{X}_1 - \bar{X}_2)$  multiplied by  $z_{\alpha/2}$  or  $t_{\alpha/2, df}$ , depending on whether or not the population variances are known.

### CONFIDENCE INTERVAL FOR $\mu_1 - \mu_2$

A  $100(1 - \alpha)\%$  confidence interval for the difference between two population means  $\mu_1 - \mu_2$  is given by

1.  $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ , if the population variances,  $\sigma_1^2$  and  $\sigma_2^2$ , are known.
2.  $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, df} \sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$ , if  $\sigma_1^2$  and  $\sigma_2^2$  are unknown but assumed equal. A pooled estimate of the common variance is  $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ , where  $s_1^2$  and  $s_2^2$  are the corresponding sample variances and the degrees of freedom  $df = n_1 + n_2 - 2$ .
3.  $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ , if  $\sigma_1^2$  and  $\sigma_2^2$  are unknown and cannot be assumed equal. The degrees of freedom  $df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$ . Since the resultant value for  $df$  is rarely an integer, we generally round the value down to obtain the appropriate  $t$  value from the  $t$  table.

These formulas are valid only if  $\bar{X}_1 - \bar{X}_2$  (approximately) follows a normal distribution.

Note that in the case when we construct a confidence interval for  $\mu_1 - \mu_2$  where  $\sigma_1^2$  and  $\sigma_2^2$  are unknown but assumed equal, we calculate a pooled estimate of the common variance  $s_p^2$ . In other words, because the two populations are assumed to have the same population variance, the two sample variances  $s_1^2$  and  $s_2^2$  are simply two separate estimates of this population variance. We estimate the population variance by a *weighted* average of  $s_1^2$  and  $s_2^2$ , where the weights applied are their respective degrees of freedom relative to the total number of degrees of freedom. In the case when  $\sigma_1^2$  and  $\sigma_2^2$  are unknown and cannot be assumed equal, we cannot calculate a pooled estimate of the population variance.

### EXAMPLE 10.1

A consumer advocate analyzes the nicotine content in two brands of cigarettes. A sample of 20 cigarettes of Brand A resulted in an average nicotine content of 1.68 milligrams with a standard deviation of 0.22 milligram; 25 cigarettes of Brand B yielded an average nicotine content of 1.95 milligrams with a standard deviation of 0.24 milligram.

Brand A	Brand B
$\bar{x}_1 = 1.68$ mg	$\bar{x}_2 = 1.95$ mg
$s_1 = 0.22$ mg	$s_2 = 0.24$ mg
$n_1 = 20$	$n_2 = 25$

Construct the 95% confidence interval for the difference between the two population means. Nicotine content is assumed to be normally distributed. In addition, the population variances are unknown but assumed equal.

**SOLUTION:** We wish to construct a confidence interval for  $\mu_1 - \mu_2$  where  $\mu_1$  is the mean nicotine level for Brand A and  $\mu_2$  is the mean nicotine level for Brand B. Since the population variances are unknown but assumed equal, we use the formula

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, df} \sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

We calculate the point estimate  $\bar{x}_1 - \bar{x}_2 = 1.68 - 1.95 = -0.27$ . In order to find  $t_{\alpha/2, df}$ , we determine  $df = n_1 + n_2 - 2 = 20 + 25 - 2 = 43$ . For a 95% confidence interval ( $\alpha = 0.05$ ), we reference the  $t$  table to find  $t_{0.025, 43} = 2.017$ .



We then calculate the pooled estimate of the population variance as

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(20 - 1)(0.22)^2 + (25 - 1)(0.24)^2}{20 + 25 - 2} = 0.0535.$$

Inserting the appropriate values into the formula, we have

$$-0.27 \pm 2.017 \sqrt{0.0535 \left( \frac{1}{20} + \frac{1}{25} \right)} = -0.27 \pm 0.14.$$

In other words, the 95% confidence interval for the difference between the two means ranges from  $-0.41$  to  $-0.13$ . Shortly, we will use this interval to conduct a two-tailed hypothesis test.

## Hypothesis Test for $\mu_1 - \mu_2$

As always, when we specify the null hypothesis and the alternative hypothesis, it is important that we identify the relevant population parameter of interest, determine whether we conduct a one- or a two-tailed test, and finally include some form of the equality sign in the null hypothesis and use the alternative hypothesis to establish a claim. When conducting hypothesis tests concerning the parameter  $\mu_1 - \mu_2$ , the competing hypotheses will take one of the following general forms:

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: \mu_1 - \mu_2 = d_0$	$H_0: \mu_1 - \mu_2 \leq d_0$	$H_0: \mu_1 - \mu_2 \geq d_0$
$H_A: \mu_1 - \mu_2 \neq d_0$	$H_A: \mu_1 - \mu_2 > d_0$	$H_A: \mu_1 - \mu_2 < d_0$

In most applications, the hypothesized difference  $d_0$  between two population means  $\mu_1$  and  $\mu_2$  is zero. In this scenario, a two-tailed test determines whether the two means differ from one another; a right-tailed test determines whether  $\mu_1$  is greater than  $\mu_2$ ; and a left-tailed test determines whether  $\mu_1$  is less than  $\mu_2$ .

We can also construct hypotheses where the hypothesized difference  $d_0$  is a value other than zero. For example, if we wish to determine if the mean return of an emerging market fund is more than two percentage points higher than that of a developed market fund, the resulting hypotheses are  $H_0: \mu_1 - \mu_2 \leq 2$  versus  $H_A: \mu_1 - \mu_2 > 2$ .

### EXAMPLE 10.2

Revisit Example 10.1.

- Specify the competing hypotheses in order to determine whether the average nicotine levels differ between Brand A and Brand B.
- Using the 95% confidence interval, what is the conclusion to the test?

#### SOLUTION:

- We want to determine if the average nicotine levels differ between the two brands, or  $\mu_1 \neq \mu_2$ , so we formulate a two-tailed hypothesis test as

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0$$

- In Example 10.1, we calculated the 95% confidence interval for the difference between the two means as  $-0.27 \pm 0.14$ , or equivalently, the confidence



interval ranges from  $-0.41$  to  $-0.13$ . This interval does not contain zero, the value hypothesized under the null hypothesis. This information allows us to reject  $H_0$ ; the sample data support the conclusion that average nicotine levels between the two brands differ at the 5% significance level.

While it is true that we can use confidence intervals to conduct two-tailed hypothesis tests, the four-step procedure outlined in Chapter 9 can be implemented to conduct one- or two-tailed hypothesis tests. (It is possible to adjust the confidence interval to accommodate a one-tailed test, but we do not discuss this modification.) The only real change in the process is the specification of the test statistic. We convert the point estimate  $\bar{x}_1 - \bar{x}_2$  into the value  $z$  or  $t_{df}$  of the corresponding test statistic by dividing the difference of  $(\bar{x}_1 - \bar{x}_2) - d_0$  by the standard error of the estimator  $(\bar{X}_1 - \bar{X}_2)$ .

### TEST STATISTIC FOR TESTING $\mu_1 - \mu_2$

The value of the test statistic for a hypothesis test concerning the difference between two population means,  $\mu_1 - \mu_2$ , is computed using one of the following three formulas:

1. If  $\sigma_1^2$  and  $\sigma_2^2$  are known, then the value of the test statistic is computed as

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

2. If  $\sigma_1^2$  and  $\sigma_2^2$  are unknown but assumed equal, then the value of the test statistic

$$\text{is computed as } t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}, \text{ where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \text{ and}$$

$$df = n_1 + n_2 - 2.$$

3. If  $\sigma_1^2$  and  $\sigma_2^2$  are unknown and cannot be assumed equal, then the value

$$\text{of the test statistic is computed as } t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \text{ where}$$

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} \text{ is rounded down to the nearest integer.}$$

These formulas are valid only if  $\bar{X}_1 - \bar{X}_2$  (approximately) follows a normal distribution.

### EXAMPLE 10.3

An economist claims that average weekly food expenditure for households in City 1 is more than the average weekly food expenditure for households in City 2. She surveys 35 households in City 1 and obtains an average weekly food expenditure of \$164. A sample of 30 households in City 2 yields an average weekly food expenditure of \$159. Prior studies suggest that the population standard deviation for City 1 and City 2 are \$12.50 and \$9.25, respectively.

City 1	City 2
$\bar{x}_1 = \$164$	$\bar{x}_2 = \$159$
$\sigma_1 = \$12.50$	$\sigma_2 = \$9.25$
$n_1 = 35$	$n_2 = 30$

- a. Specify the competing hypotheses to test the economist's claim.
- b. Calculate the value of the test statistic and its associated  $p$ -value.
- c. At the 5% significance level, is the economist's claim supported by the data?

**SOLUTION:**

- a. The relevant parameter of interest is  $\mu_1 - \mu_2$ , where  $\mu_1$  is the mean weekly food expenditure for City 1 and  $\mu_2$  is the mean weekly food expenditure for City 2. The economist wishes to determine if the mean weekly food expenditure in City 1 is more than that of City 2; that is,  $\mu_1 > \mu_2$ . This is an example of a right-tailed test where the appropriate hypotheses are

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_A: \mu_1 - \mu_2 > 0$$

- b. Since the population standard deviations are known, we compute the value of the test statistic as

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(164 - 159) - 0}{\sqrt{\frac{(12.50)^2}{35} + \frac{(9.25)^2}{30}}} = \frac{5}{2.70} = 1.85.$$

The  $p$ -value of the above right-tailed test is computed as  $p\text{-value} = P(Z \geq 1.85) = 1 - 0.9678 = 0.0322$ .

- c. We reject the null hypothesis since the  $p$ -value of 0.0322 is less than the chosen  $\alpha = 0.05$ . Therefore, at the 5% significance level, the economist concludes that average weekly food expenditure in City 1 is more than that of City 2.

## Using Excel for Testing Hypotheses about $\mu_1 - \mu_2$

Excel provides several options that simplify the steps when conducting a hypothesis test that compares two means. This is especially useful when we are given raw sample data and we first have to compute the sample means and the sample standard deviations for the test. Here, we discuss one of the options using the next example.

### EXAMPLE 10.4

Table 10.2 shows annual return data for 10 firms in the gold industry and 10 firms in the oil industry. Can we conclude at the 5% significance level that the average returns in the two industries differ? Here we assume that the sample data are drawn independently from normal populations with unequal population variances. The assumption concerning the population variances is reasonable since variance is a common measure of risk when analyzing financial returns; we cannot assume that the risk from investing in the gold industry is the same as the risk from investing in the oil industry.

**TABLE 10.2** Annual Returns (in percent)

**FILE**  
*Gold\_Oil*

Gold	Oil
6	-3
15	15
19	28
26	18
2	32
16	31
31	15
14	12
15	10
16	15

**SOLUTION:** We let  $\mu_1$  denote the mean return for the gold industry and  $\mu_2$  denote the mean return for the oil industry. Since we wish to test whether the mean returns differ, we set up the null and alternative hypotheses as

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0$$

Given that we are testing the difference between two means when the population variances are unknown and not equal, we need to calculate  $t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ .

Recall that the calculation for the degrees of freedom for the corresponding test statistic is rather involved. Using one command on Excel, we are provided with the value of the test statistic, the degrees of freedom, and the  $p$ -value, as well as the relevant critical values. We follow these steps.

- Open the *Gold\_Oil* data file.
- Choose **Data > Data Analysis > t-Test: Two-Sample Assuming Unequal Variances > OK**. (Note: Excel provides two other options when we want to test the difference between two population means from independent samples and we have access to the raw data. If the population variances are known, we use the option **z-Test: Two-Sample for Means**. If the population variances are unknown but assumed equal, we use the option **t-Test: Two-Sample Assuming Equal Variances**.)
- See Figure 10.1. In the dialog box, choose *Variable 1 Range* and select the gold data. Then, choose *Variable 2 Range* and select the oil data. Enter a *Hypothesized Mean Difference* of 0 since  $d_0 = 0$ , check the *Labels* box if you include Gold and Oil as headings, and enter an  $\alpha$  value of 0.05 since the test is conducted at the 5% significance level. Choose an output range and click **OK**.

**FIGURE 10.1** Excel's dialog box for  $t$  test with unequal variances

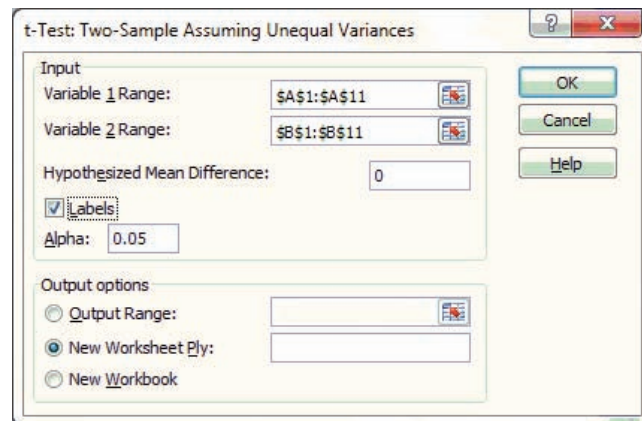


Table 10.3 shows the relevant output.

The output from Excel allows us to conduct the hypothesis test using either the  $p$ -value approach or the critical value approach. Given that we have a two-tailed hypothesis test, the relevant  $p$ -value is 0.7661 (see **P(T ≤ t) two-tail** in Table 10.3). At the 5% significance level, we cannot reject  $H_0$  since the  $p$ -value is greater than 0.05. While average returns in the oil industry seem to slightly outperform average returns in the gold industry ( $\bar{x}_2 = 17.3\% > 16.0\% = \bar{x}_1$ ), the difference is not statistically significant.

**TABLE 10.3** Excel's Output for  $t$  Test with Unequal Variances

	Gold	Oil
Mean	16	17.3
Variance	70.6667	114.2333
Observations	10	10
Hypothesized Mean Difference	0	
Df	17	
<b>t Stat</b>	<b>-0.3023</b>	
P(T ≤ t) one-tail	0.3830	
t Critical one-tail	1.7396	
<b>P(T ≤ t) two-tail</b>	<b>0.7661</b>	
<b>t Critical two-tail</b>	<b>2.1098</b>	

We now show that we reach the same conclusion concerning the mean return of these two industries using the critical value approach as well as the confidence interval approach. Given  $\alpha = 0.05$ , the relevant critical values for this two-tailed test are  $-2.1098$  and  $2.1098$  (see **t Critical two-tail** in Table 10.3). The decision rule is to reject  $H_0$  if  $t_{17} > 2.1098$  or  $t_{17} < -2.1098$ . The value of the test statistic is  $t_{17} = -0.3023$  (see **t Stat** in Table 10.3). Since the value of the test statistic is between the two critical values,  $-2.1098 < -0.3023 < 2.1098$ , we cannot reject the null hypothesis. As always, our conclusion is consistent with that of the  $p$ -value approach.

Finally, given the information in Table 10.3, it is also possible to calculate the corresponding 95% confidence interval for  $\mu_1 - \mu_2$  as

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (16.0 - 17.3) \pm 2.1098 \sqrt{\frac{70.6667}{10} + \frac{114.2333}{10}}$$

$$= -1.3 \pm 9.07$$

That is, the 95% confidence interval for the difference between the two means ranges from  $-10.37$  to  $7.77$ . We note that this interval contains zero, the value hypothesized under the null hypothesis. Using a 95% confidence interval, we again cannot support the conclusion that the population mean returns differ at the 5% significance level.

### A Note on the Assumption of Normality

In Example 10.4, we may have made a strong assumption that the populations were normally distributed. We could not invoke the central limit theorem, as we had small sample sizes. In Chapter 12, we will explore tests that check for normality. If we wish to draw inferences about  $\mu_1 - \mu_2$  from nonnormal populations, we can use the nonparametric Wilcoxon rank-sum test for independent samples, discussed in Chapter 20.

## EXERCISES 10.1

### Mechanics

- Consider the following data drawn independently from normally distributed populations:

$$\begin{array}{ll} \bar{x}_1 = 25.7 & \bar{x}_2 = 30.6 \\ \sigma_1^2 = 98.2 & \sigma_2^2 = 87.4 \\ n_1 = 20 & n_2 = 25 \end{array}$$

- Construct the 95% confidence interval for the difference between the population means.
- Specify the competing hypotheses in order to determine whether or not the population means differ.
- Using the confidence interval from part a, can you reject the null hypothesis? Explain.

2. Consider the following data drawn independently from normally distributed populations:

$$\begin{aligned}\bar{x}_1 &= -10.5 & \bar{x}_2 &= -16.8 \\ s_1^2 &= 7.9 & s_2^2 &= 9.3 \\ n_1 &= 15 & n_2 &= 20\end{aligned}$$

- Construct the 95% confidence interval for the difference between the population means. Assume the population variances are unknown but equal.
  - Specify the competing hypotheses in order to determine whether or not the population means differ.
  - Using the confidence interval from part a, can you reject the null hypothesis? Explain.
3. Consider the following competing hypotheses and accompanying sample data drawn independently from normally distributed populations.

$$\begin{aligned}H_0: \mu_1 - \mu_2 &= 0 \\ H_A: \mu_1 - \mu_2 &\neq 0\end{aligned}$$

$$\begin{aligned}\bar{x}_1 &= 57 & \bar{x}_2 &= 63 \\ \sigma_1 &= 11.5 & \sigma_2 &= 15.2 \\ n_1 &= 20 & n_2 &= 20\end{aligned}$$

- Using the  $p$ -value approach, test whether the population means differ at the 5% significance level.
  - Repeat the hypothesis test using the critical value approach.
4. Consider the following competing hypotheses and accompanying sample data. The two populations are known to be normally distributed.

$$\begin{aligned}H_0: \mu_1 - \mu_2 &\leq 0 \\ H_A: \mu_1 - \mu_2 &> 0\end{aligned}$$

$$\begin{aligned}\bar{x}_1 &= 20.2 & \bar{x}_2 &= 17.5 \\ s_1 &= 2.5 & s_2 &= 4.4 \\ n_1 &= 10 & n_2 &= 12\end{aligned}$$

- Implement the test at the 5% significance level under the assumption that the population variances are unknown but equal.
  - Repeat the analysis at the 10% significance level.
5. Consider the following competing hypotheses and accompanying sample data drawn independently from normally distributed populations.

$$\begin{aligned}H_0: \mu_1 - \mu_2 &\geq 0 \\ H_A: \mu_1 - \mu_2 &< 0\end{aligned}$$

$$\begin{aligned}\bar{x}_1 &= 249 & \bar{x}_2 &= 272 \\ s_1 &= 35 & s_2 &= 23 \\ n_1 &= 10 & n_2 &= 10\end{aligned}$$

- Implement the test at the 5% significance level under the assumption that the population variances are unknown but equal.

- Implement the test at the 5% significance level under the assumption that the population variances are unknown and are not equal.

6. Consider the following competing hypotheses and accompanying sample data.

$$\begin{aligned}H_0: \mu_1 - \mu_2 &= 5 \\ H_A: \mu_1 - \mu_2 &\neq 5\end{aligned}$$

$$\begin{aligned}\bar{x}_1 &= 57 & \bar{x}_2 &= 43 \\ s_1 &= 21.5 & s_2 &= 15.2 \\ n_1 &= 22 & n_2 &= 18\end{aligned}$$

Assume that the populations are normally distributed with unknown but equal variances.

- Calculate the value of the test statistic.
  - Using the  $p$ -value approach, test the above hypotheses at the 5% significance level.
  - Repeat the analysis using the critical value approach.
7. Consider the following sample data drawn independently from normally distributed populations with equal population variances.

Sample 1	Sample 2
12.1	8.9
9.5	10.9
7.3	11.2
10.2	10.6
8.9	9.8
9.8	9.8
7.2	11.2
10.2	12.1

- Construct the relevant hypotheses to test if the mean of the second population is greater than the mean of the first population.
  - What is the inference of the test at a 1% level of significance?
  - What is the inference of the test at a 10% level of significance?
8. Consider the following sample data drawn independently from normally distributed populations with unequal population variances.

Sample 1	Sample 2
88	98
110	114
102	118
96	128
74	102
120	110

- Construct the relevant hypothesis to test if the means of the two populations differ.

- What is the value of the test statistic?
- Approximate the  $p$ -value.
- What is the inference of the test at a 10% level of significance?

## Applications

- According to a new Health of Boston report, female residents in Boston have a higher average life expectancy as compared to male residents (*The Boston Globe*, August 16, 2010). You collect the following sample data to verify the results of the report. You also use the historical (population) standard deviation of 8.2 years for females and 8.6 years for males.

Female	Male
$\bar{x}_1 = 81.1$ years	$\bar{x}_2 = 74.8$ years
$n_1 = 32$	$n_2 = 32$

- Set up the hypotheses to test whether the average life expectancy of female Bostonians is higher than that of male Bostonians.
  - Calculate the value of the test statistic and its  $p$ -value.
  - At the 10% significance level, can we conclude that female Bostonians live longer than male Bostonians?
  - Repeat the hypothesis test using the critical value approach.
- A joint project of the U.S. Census Bureau and the National Science Foundation shows that people with a bachelor's degree who transferred from a community college earn less than those who start at a four-year school (*USA TODAY*, March 17, 2009). Previous studies referred to this occurrence as a "community college penalty." Lucille Barnes wonders if a similar pattern applies to her university. The accompanying table shows the average salary of 100 graduates with an associate degree and the average salary of 100 graduates with no associate degree. Lucille believes that the population standard deviation is \$4,400 for graduates with an associate degree and \$1,500 for graduates with no associate degree.

Bachelor's Degree with Associate Degree	Bachelor's Degree with No Associate Degree
$\bar{x}_1 = \$52,000$	$\bar{x}_2 = \$54,700$
$n_1 = 100$	$n_2 = 100$

- Set up the hypotheses to test if the report's conclusion also applies to Lucille's university.
  - Calculate the value of the test statistic and its  $p$ -value.
  - At the 5% significance level, can we conclude that there is a "community college penalty" at Lucille's university?
- The Chartered Financial Analyst (CFA®) designation is fast becoming a requirement for serious investment professionals. It is an attractive alternative to getting

an MBA for students wanting a career in investment. A student of finance is curious to know if a CFA designation is a more lucrative option than an MBA. He collects data on 38 recent CFAs with a mean salary of \$138,000 and a standard deviation of \$34,000. A sample of 80 MBAs results in a mean salary of \$130,000 with a standard deviation of \$46,000.

- Use the  $p$ -value approach to test if a CFA designation is more lucrative than an MBA at the 5% significance level. Do not assume that the population variances are equal. Make sure to state the competing hypotheses.
  - Repeat the analysis with the critical value approach.
- An entrepreneur owns some land that he wishes to develop. He identifies two development options: build condominiums or build apartment buildings. Accordingly, he reviews public records and derives the following summary measures concerning annual profitability based on a random sample of 30 for each such local business venture. For the analysis, he uses a historical (population) standard deviation of \$22,500 for condominiums and \$20,000 for apartment buildings.

Condominiums	Apartment Buildings
$\bar{x}_1 = \$244,200$	$\bar{x}_2 = \$235,800$
$n_1 = 30$	$n_2 = 30$

- Set up the hypotheses to test whether the mean profitability differs between condominiums and apartment buildings.
  - Compute the value of the test statistic and the corresponding  $p$ -value.
  - At the 5% significance level, what is the conclusion to the test? What if the significance level is 10%?
- David Anderson has been working as a lecturer at Michigan State University for the last three years. He teaches two large sections of introductory accounting every semester. While he uses the same lecture notes in both sections, his students in the first section outperform those in the second section. He believes that students in the first section not only tend to get higher scores, they also tend to have lower variability in scores. David decides to carry out a formal test to validate his hunch regarding the difference in average scores. In a random sample of 18 students in the first section, he computes a mean and a standard deviation of 77.4 and 10.8, respectively. In the second section, a random sample of 14 students results in a mean of 74.1 and a standard deviation of 12.2.
- Construct the null and the alternative hypotheses to test David's hunch.
  - Compute the value of the test statistic. What assumption regarding the populations is necessary to implement this step?



- c. Implement the test at  $\alpha = 0.01$  and interpret your results.
14. A design engineer at Sperling Manufacturing, a supplier of high-quality ball bearings, claims a new machining process can result in a higher daily output rate. Accordingly, the production group is conducting an experiment to determine if this claim can be substantiated. The mean and standard deviation of bearings in a sample of 8 days' output using the new process equal 2,613.63 and 90.78, respectively. A similar sample of 10 days' output using the old process yield the mean and standard deviation of 2,485.10 and 148.22, respectively.
- Set up the hypotheses to test whether the mean output rate of the new process exceeds that of the old process. Assume normal populations and equal population variances for each process. Use the critical value approach for the analysis.
  - Compute the value of the test statistic.
  - At the 5% significance level, what is the conclusion of the experiment?
  - At the 1% significance level, what is the conclusion of the experiment?
15. A phone manufacturer wants to compete in the touch screen phone market. Management understands that the leading product has a less than desirable battery life. They aim to compete with a new touch phone that is guaranteed to have a battery life more than two hours longer than the leading product. A recent sample of 120 units of the leading product provides a mean battery life of 5 hours and 40 minutes with a standard deviation of 30 minutes. A similar analysis of 100 units of the new product results in a mean battery life of 8 hours and 5 minutes and a standard deviation of 55 minutes. It is not reasonable to assume that the population variances of the two products are equal.
- Set up the hypotheses to test if the new product has a battery life more than two hours longer than the leading product.
  - Implement the test at the 5% significance level using the critical value approach.
16. In May 2008, CNN reported that sports utility vehicles (SUVs) are plunging toward the "endangered" list. Due to soaring oil prices and environmental concerns, consumers are replacing gas-guzzling vehicles with fuel-efficient smaller cars. As a result, there has been a big drop in the demand for new as well as used SUVs. A sales manager of a used car dealership believes that it takes 30 days longer to sell an SUV as compared to a small car. In the last two months, he sold 18 SUVs that took an average of 95 days to sell with a standard deviation of 32 days. He also sold 38 small cars with an average of 48 days to sell and a standard deviation of 24 days.
- Construct the null and the alternative hypotheses to contradict the manager's claim.
  - Compute the value of the test statistic under the assumption that the populations are normally distributed and that the variability of selling time for the SUVs and the small cars is the same.
  - Implement the test at  $\alpha = 0.10$  and interpret your results.
17. **FILE Refrigerator Longevity.** A consumer advocate researches the length of life between two brands of refrigerators, Brand A and Brand B. He collects data (measured in years) on the longevity of 40 refrigerators for Brand A and repeats the sampling for Brand B.
- Specify the competing hypotheses to test whether the average length of life differs between the two brands.
  - Using the appropriate commands in Excel, find the value of the test statistic. Assume that  $\sigma_A^2 = 4.4$  and  $\sigma_B^2 = 5.2$ . What is the  $p$ -value?
  - At the 5% significance level, what is the conclusion?
18. **FILE Website Searches.** "The See Me" marketing agency wants to determine if time of day for a television advertisement influences website searches for a product. They have extracted the number of website searches occurring during a one-hour period after an advertisement was aired for a random sample of 30 day and 30 evening advertisements. A portion of the data is shown in the accompanying table.
- | Day Searches | Evening Searches |
|--------------|------------------|
| 96670        | 118379           |
| 97855        | 111005           |
| :            | :                |
| 95103        | 114721           |
- Set up the hypotheses to test whether the mean number of website searches differs between the day and evening advertisements.
  - Using the appropriate commands in Excel, find the value of the test statistic. Assume the population variances are unknown but equal. What are the critical value(s) and the rejection rule?
  - At the 5% significance level, what is the conclusion?
19. **FILE Different Diets.** (Use Excel) According to a study published in the *New England Journal of Medicine*, overweight people on low-carbohydrate and Mediterranean diets lost more weight and got greater cardiovascular benefits than people on a conventional low-fat diet (*The Boston Globe*, July 17, 2008). A nutritionist wishes to verify these results and documents the weight loss (in pounds) of 30 dieters on the low-carbohydrate and Mediterranean diets and 30 dieters on the low-fat diet.

- Set up the hypotheses to test the claim that the mean weight loss for those on low-carbohydrate or Mediterranean diets is greater than the mean weight loss for those on a conventional low-fat diet.
- Using the appropriate commands in Excel, find the value of the test statistic. Assume that the population variances are equal and that the test is conducted at the 5% significance level. Specify the critical value(s) and the decision rule.
- At the 5% significance level, can the nutritionist conclude that people on low-carbohydrate or Mediterranean diets lost more weight than people on a conventional low-fat diet?

20. **FILE Tractor Times.** The production department at Greenside Corporation, a manufacturer of lawn equipment, has devised a new manual assembly method for its lawn tractors. Now it wishes to determine if it is reasonable to conclude the mean assembly time of the new method is less than the old method. Accordingly, they have randomly sampled assembly times from 40 tractors using the old method and 32 tractors using the new method. A portion of the data is shown in the accompanying table.

Old Method Times (Minutes)	New Method Times (Minutes)
32	30
36	32
⋮	⋮

- Set up the hypotheses to test the claim that the mean assembly time using the new method is less than the old method.
  - Using the appropriate commands in Excel, find the value of the test statistic and the  $p$ -value. Assume the population variances are unknown and not equal.
  - At the 5% significance level, what is the conclusion? What if the significance level is 10%?
21. **FILE Nicknames.** Baseball has always been a favorite pastime in America and is rife with statistics and theories. In a recent paper, researchers showed that major league players who have nicknames live  $2\frac{1}{2}$  years longer than those without them (*The Wall Street Journal*, July 16, 2009). You do not believe in this result and decide to collect data on the lifespan of 30 baseball players along with a nickname variable that equals 1 if the player had a nickname and 0 otherwise. A portion of the data is shown in the accompanying table.

Years	Nickname
74	1
62	1
⋮	⋮
64	0

- Create two subsamples consisting of players with and without nicknames. Calculate the average longevity for each subsample.
- Specify the hypotheses to contradict the claim made by the researchers.
- State the conclusion of the test using a 5% level of significance. Assume the population variances are unknown but equal.

22. **FILE Starting Salaries.** Recent evidence suggests that graduating from college during bad economic times can impact the graduate's earning power for a long time (*Financial Times*, June 1, 2012). An associate dean at a prestigious college wants to determine if the starting salary of his college graduates has declined from 2008 to 2010. He expects the variance of the salaries to be different between these two years. A portion of the data is shown in the accompanying table.

Salary 2008 (\$)	Salary 2010 (\$)
35,000	34,000
56,000	62,000
⋮	⋮
47,000	54,000

Use the  $p$ -value approach, at the 5% significance level, to determine if the mean starting salary has decreased from 2008 to 2010. Describe your steps clearly.

23. **FILE Spending Gender.** Researchers at The Wharton School of Business have found that men and women shop for different reasons. While women enjoy the shopping experience, men are on a mission to get the job done. Men do not shop as frequently, but when they do, they make big purchases like expensive electronics. The accompanying table shows a portion of the amount spent over the weekend by 40 men and 60 women at a local mall. Assume the population variances are unknown, but equal.

Spending by Men (\$)	Spending by Women (\$)
85	90
102	79
⋮	⋮

At the 1% significance level, use the critical value approach to determine if the mean amount spent by men is more than that by women. Describe your steps clearly.

## 10.2 INFERENCE CONCERNING MEAN DIFFERENCES

### LO 10.2

One of the crucial assumptions in Section 10.1 concerning differences between two population means is that the samples are drawn independently. As mentioned earlier, two samples are independent if the selection of one is not influenced by the selection of the other. When we want to conduct tests on two population means based on samples that we believe are not independent, we need to employ a different methodology.

Make inferences about the mean difference based on matched-pairs sampling.

A common case of dependent sampling, commonly referred to as **matched-pairs sampling**, is when the samples are paired or matched in some way. Such samples are useful in evaluating strategies because the comparison is made between “apples” and “apples.” For instance, an effective way to assess the benefits of a new medical treatment is by evaluating the same patients before and after the treatment. If, however, one group of people is given the treatment and another group is not, then it is not clear if the observed differences are due to the treatment or due to other important differences between the groups.

For matched-pairs sampling, the parameter of interest is referred to as the mean difference  $\mu_D$  where  $D = X_1 - X_2$ , and the random variables  $X_1$  and  $X_2$  are matched in a pair. The statistical inference regarding  $\mu_D$  is based on the estimator  $\bar{D}$ , representing the sample mean difference. It requires that  $X_1 - X_2$  is normally distributed or that the sample size is sufficiently large ( $n \geq 30$ ).

### Recognizing a Matched-Pairs Experiment

It is important to be able to determine whether a particular experiment uses independent or matched-pairs sampling. In general, two types of matched-pairs sampling occur:

1. The first type of matched-pairs sample is characterized by a measurement, an intervention of some type, and then another measurement. We generally refer to these experiments as “before” and “after” studies. For example, an operation manager of a production facility wants to determine whether a new workstation layout improves productivity at her plant. She first measures output of employees before the layout change. Then she measures output of the same employees after the change. Another classic before-and-after example concerns weight loss of clients at a diet center. In these examples, the same individual gets sampled before and after the experiment.
2. The second type of matched-pairs sample is characterized by a pairing of observations, where it is not the same individual who gets sampled twice. Suppose an agronomist wishes to switch to an organic fertilizer but is unsure what the effects might be on his crop yield. It is important to the agronomist that the yields be similar. He matches 20 adjacent plots of land using the nonorganic fertilizer on one half of the plot and the organic fertilizer on the other.

In order to recognize a matched-pairs experiment, we watch for a natural pairing between one observation in the first sample and one observation in the second sample. If a natural pairing exists, then the experiment involves matched samples.

### Confidence Interval for $\mu_D$

When constructing a confidence interval for the mean difference  $\mu_D$ , we follow the same general format of point estimate  $\pm$  margin of error.

#### CONFIDENCE INTERVAL FOR $\mu_D$

A  $100(1 - \alpha)\%$  confidence interval for the mean difference  $\mu_D$  is given by

$$\bar{d} \pm t_{\alpha/2, df} s_D / \sqrt{n},$$

where  $\bar{d}$  and  $s_D$  are the mean and the standard deviation, respectively, of the  $n$  sample differences and  $df = n - 1$ . This formula is valid only if  $\bar{D}$  (approximately) follows a normal distribution.

In the next example, the values for  $\bar{d}$  and  $s_D$  are explicitly given; we will outline the calculations when we discuss hypothesis testing.

### EXAMPLE 10.5

A manager is interested in improving productivity at a plant by changing the layout of the workstation. She measures the variable representing productivity of 10 workers before the change and again after the change. She calculates the following summary statistics for the sample productivity difference:  $\bar{d} = 8.5$ ,  $s_D = 11.38$ , and  $n = 10$ . Construct the 95% confidence interval for the mean difference, assuming that the productivity variable, before and after, is normally distributed.

**SOLUTION:** In order to construct the 95% confidence interval for the mean difference, we use  $\bar{d} \pm t_{\alpha/2, df} s_D / \sqrt{n}$ . With  $df = n - 1 = 10 - 1 = 9$  and  $\alpha = 0.05$ , we find  $t_{\alpha/2, df} = t_{0.025, 9} = 2.262$ . Plugging the relevant values into the formula, we calculate  $8.5 \pm 2.262(11.38 / \sqrt{10}) = 8.5 \pm 8.14$ . That is, the 95% confidence interval for the mean difference ranges from 0.36 to 16.64. This represents a fairly wide interval, caused by the high standard deviation  $s_D$  of the 10 sample differences.

## Hypothesis Test for $\mu_D$

As before, we generally want to test whether the mean difference  $\mu_D$  is equal to, greater than, or less than a given hypothesized mean difference  $d_0$ , or:

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: \mu_D = d_0$	$H_0: \mu_D \leq d_0$	$H_0: \mu_D \geq d_0$
$H_A: \mu_D \neq d_0$	$H_A: \mu_D > d_0$	$H_A: \mu_D < d_0$

In practice, the competing hypotheses tend to be based on  $d_0 = 0$ . For example, when testing if the mean difference differs from zero, we use a two-tailed test with the competing hypotheses defined as  $H_0: \mu_D = 0$  versus  $H_A: \mu_D \neq 0$ . If, on the other hand, we wish to determine whether or not the mean difference differs by some amount, say by 5 units, we set  $d_0 = 5$  and define the competing hypotheses as  $H_0: \mu_D = 5$  versus  $H_A: \mu_D \neq 5$ . One-tailed tests are defined similarly.

### EXAMPLE 10.6

Using the information from Example 10.5, can the manager conclude at the 5% significance level that there has been a change in productivity since the adoption of the new workstation?

**SOLUTION:** In order to determine whether or not there has been a change in the mean difference, we formulate the null and the alternative hypotheses as

$$H_0: \mu_D = 0$$

$$H_A: \mu_D \neq 0$$

In Example 10.5, we found that the 95% confidence interval for the mean difference ranges from 0.36 to 16.64. Although the interval is very wide, the entire range is above the hypothesized value of zero. Therefore, at the 5% significance level the sample data suggest that the mean difference differs from zero. In other words, there has been a change in productivity due to the different layout in the workstation.

We now examine the four-step procedure to conduct one- or two-tailed hypothesis tests concerning the mean difference. We again convert the sample mean difference into its

corresponding  $t_{df}$  statistic by dividing the difference between the sample mean difference and the hypothesized mean difference by the standard error of the estimator.

#### TEST STATISTIC FOR TESTING $\mu_D$

The value of the test statistic for a hypothesis test concerning the population mean difference  $\mu_D$  is computed as  $t_{df} = \frac{\bar{d} - d_0}{s_D / \sqrt{n}}$ , where  $df = n - 1$ ,  $\bar{d}$  and  $s_D$  are the mean and the standard deviation, respectively, of the  $n$  sample differences, and  $d_0$  is the hypothesized mean difference. This formula is valid only if  $\bar{D}$  (approximately) follows a normal distribution.

**FILE**  
Drink\_Calories

### EXAMPLE 10.7

Let's revisit the chapter's introductory case. Recall that a local ordinance requires chain restaurants to post caloric information on their menus. A nutritionist wants to examine whether average drink calories declined at Starbucks after the passage of the ordinance. The nutritionist obtains transaction data for 40 Starbucks cardholders and records their average drink calories prior to the ordinance and then after the ordinance. A portion of the data is shown in Table 10.4. Using the critical value approach, can she conclude at the 5% significance level that the ordinance reduced average drink calories?

**SOLUTION:** We first note that this is a matched-pairs experiment; specifically, it conforms to a “before” and “after” type of study. Moreover, we want to find out whether average drink calories consumed prior to the ordinance are significantly greater than average drink calories consumed after passage of the ordinance. Thus, we want to test if the mean difference  $\mu_D$  is greater than zero, where  $D = X_1 - X_2$ ,  $X_1$  denotes drink calories before the ordinance, and  $X_2$  denotes drink calories after the ordinance for a randomly selected Starbucks customer. We specify the competing hypotheses as

$$H_0: \mu_D \leq 0$$

$$H_A: \mu_D > 0$$

The normality condition for the test is satisfied since the sample size  $n \geq 30$ . The value of the test statistic is calculated as  $t_{df} = \frac{\bar{d} - d_0}{s_D / \sqrt{n}}$  where  $d_0$  equals 0. In order to determine  $\bar{d}$  and  $s_D$ , we first calculate the difference  $d_i$  for each  $i$ -th consumer. For instance, consumer 1 consumes 141 calories prior to the ordinance and 142 calories after the ordinance, for a difference of  $d_1 = 141 - 142 = -1$ . The differences for a portion of the other consumers appear in the fourth column of Table 10.4.

**Table 10.4** Data and Calculations for Example 10.7,  $n = 40$

Customer	Drink Calories		$d_i$	$(d_i - \bar{d})^2$
	Before	After		
1	141	142	-1	$(-1 - 2.1)^2 = 9.61$
2	137	140	-3	$(-3 - 2.1)^2 = 26.01$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
40	147	141	6	$(6 - 2.1)^2 = 15.21$
			$\Sigma d_i = 84$	$\Sigma (d_i - \bar{d})^2 = 2593.60$

We obtain the sample mean as

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{84}{40} = 2.10.$$

Similarly, in the fifth column of Table 10.4, we square the differences between  $d_i$  and  $\bar{d}$ . Summing these squared differences yields the numerator in the formula for the sample variance  $s_D^2$ . The denominator is simply  $n - 1$ , so:

$$s_D^2 = \frac{\sum (d_i - \bar{d})^2}{n - 1} = \frac{2593.60}{40 - 1} = 66.50.$$

As usual, the standard deviation is the positive square root of the sample variance—that is,  $s_D = \sqrt{66.50} = 8.15$ . We compute the value of the  $t_{df}$  test statistic with  $df = n - 1 = 40 - 1 = 39$  as

$$t_{39} = \frac{\bar{d} - d_0}{s_D / \sqrt{n}} = \frac{2.10 - 0}{8.15 / \sqrt{40}} = 1.63.$$

Given a right-tailed hypothesis test with  $df = 39$ , the relevant critical value with  $\alpha = 0.05$  is found as  $t_{\alpha, df} = t_{0.05, 39} = 1.685$ . Thus, the decision rule is to reject  $H_0$  if  $t_{39} > 1.685$ . Since  $t_{39} = 1.63$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that the posting of nutritional information decreases average drink calories.

We should note that once we have calculated the mean difference and the standard deviation of the mean difference, the hypothesis test essentially reduces to a one-sample  $t$  test for the population mean.

## Using Excel for Testing Hypotheses about $\mu_D$

Excel provides an option that simplifies the calculations for a hypothesis test concerning  $\mu_D$ . Example 10.8 illustrates the procedure.

### FILE

*Food\_Calories*

### EXAMPLE 10.8

The nutritionist from Example 10.7 also wants to use the data from the 40 Starbucks cardholders in order to determine if the posting of caloric information has reduced the intake of average food calories. This test is also conducted at the 5% significance level.

**SOLUTION:** We set up the same competing hypotheses as in Example 10.7 since we want to know if food caloric intake was greater before the ordinance as compared to after the ordinance.

$$H_0: \mu_D \leq 0$$

$$H_A: \mu_D > 0$$

If we follow these steps, then Excel provides the sample value of the test statistic, the  $p$ -value, and the critical value(s).

- Open the *Food\_Calories* data file.
- Choose **Data > Data Analysis > t-Test: Paired Two Sample for Means > OK**.
- See Figure 10.2. In the dialog box, choose *Variable 1 Range* and select the data in the Before column. Choose *Variable 2 Range* and select the data in the After column. Enter a *Hypothesized Mean Difference* of 0 since  $d_0 = 0$ , check the *Labels* box if you include Before and After as headings, and enter an  $\alpha$  value of 0.05 since the test is conducted at the 5% significance level. Choose an output range and click **OK**.



**FIGURE 10.2** Excel's dialog box for  $t$  test with paired sample

Table 10.5 shows the relevant output.

**TABLE 10.5** Excel's Output for  $t$  Test with Paired Sample

	Before	After
Mean	400.275	391.475
Variance	49.94808	42.3583
Observations	40	40
Pearson Correlation	0.27080	
Hypothesized Mean Difference	0	
Df	39	
<b>t Stat</b>	<b>6.7795</b>	
<b>P(T ≤ t) one-tail</b>	<b>2.15E-08</b>	
<b>t Critical one-tail</b>	<b>1.6849</b>	
P(T ≤ t) two-tail	4.31E-08	
t Critical two-tail	2.0227	

The output from Excel allows us to conduct the hypothesis test using either the  $p$ -value approach or the critical value approach. Given that we have a one-tailed hypothesis test, the relevant  $p$ -value is 2.15E-08—that is, virtually zero. At the 5% significance level, we can reject  $H_0$  because the  $p$ -value is less than 0.05.

Given degrees of freedom of 39 and  $\alpha = 0.05$ , the relevant critical value for this one-tailed test is  $t_{\alpha, df} = t_{0.05, 39} = 1.6849$  (see **t Critical one-tail** in Table 10.5). Since the value of the test statistic is greater than the critical value,  $6.7795 > 1.6849$ , we reach the same decision to reject the null hypothesis. Thus, at the 5% significance level we can conclude that average food caloric intake has declined after the ordinance.

### One Last Note on the Matched-Pairs Experiment

Similar to our remarks in the last section, when making inferences concerning  $\mu_D$ , we require that  $\bar{D}$  (approximately) follows a normal distribution. If  $\bar{D}$  is not normally distributed, we can use the nonparametric Wilcoxon signed-rank test for matched pairs, discussed in Chapter 20.

## SYNOPSIS OF INTRODUCTORY CASE



In an effort to make it easier for consumers to select healthier options, the government wants chain restaurants to post caloric information on their menus. A nutritionist studies the effects of a recent local menu ordinance requiring caloric postings at a Starbucks in San Mateo, California. She obtains transaction data for 40 Starbucks cardholders and records their average drink and food calories prior to the ordinance and then after the ordinance. Two hypothesis tests are conducted. The first test examines whether drink caloric intake is less since the passage of the ordinance. After conducting a test on the mean difference at the 5% significance level, the nutritionist

infers that the ordinance did not prompt consumers to reduce their drink caloric intake. The second test investigates whether food caloric intake is less since the passage of the ordinance. At the 5% significance level, the sample data suggest that consumers have reduced their food caloric intake since the passage of the ordinance. In sum, while the government is trying to ensure that customers process the calorie information as they are ordering, the results are consistent with research that has shown mixed results on whether mandatory caloric postings are prompting consumers to select healthier foods.

## EXERCISES 10.2

### Mechanics

24. A sample of 20 paired observations generates the following data:  $\bar{d} = 1.3$  and  $s_D^2 = 2.6$ . Assume a normal distribution.
- Construct the 90% confidence interval for the mean difference  $\mu_D$ .
  - Using the confidence interval, test whether the mean difference differs from zero. Explain.
25. The following table contains information on matched sample values whose differences are normally distributed.

Number	Sample 1	Sample 2
1	18	21
2	12	11
3	21	23
4	22	20
5	16	20
6	14	17
7	17	17
8	18	22

- Construct the 95% confidence interval for the mean difference  $\mu_D$ .
- Specify the competing hypotheses in order to test whether the mean difference differs from zero.
- Using the confidence interval from part a, are you able to reject  $H_0$ ? Explain.

26. Consider the following competing hypotheses and accompanying results from matched samples:

$$H_0: \mu_D \geq 0; H_A: \mu_D < 0$$

$$\bar{d} = -2.8, s_D = 5.7, n = 12$$

- At the 5% significance level, find the critical value(s).
  - Calculate the value of the test statistic under the assumption that the sample difference is normally distributed.
  - What is the conclusion to the hypothesis test?
27. Consider the following competing hypotheses and accompanying results from matched samples:

$$H_0: \mu_D \leq 2; H_A: \mu_D > 2$$

$$\bar{d} = 5.6, s_D = 6.2, n = 10$$

- Calculate the value of the test statistic and approximate the  $p$ -value assuming that the sample difference is normally distributed.
  - Use the 1% significance level to make a conclusion.
28. A sample of 35 paired observations generates the following results:  $\bar{d} = 1.2$  and  $s_D = 3.8$ .
- Specify the appropriate hypotheses to test if the mean difference is greater than zero.
  - Compute the value of the test statistic and approximate the  $p$ -value.
  - At the 5% significance level, can you conclude that the mean difference is greater than zero? Explain.

- d. Repeat the hypothesis test using the critical value approach.

29. Consider the following matched samples representing observations before and after an experiment. Assume that the sample differences are normally distributed.

Before	2.5	1.8	1.4	-2.9	1.2	-1.9	-3.1	2.5
After	2.9	3.1	3.9	-1.8	0.2	0.6	-2.5	2.9

- Construct the competing hypotheses to determine if the experiment increases the magnitude of the observations.
- Implement the test at a 5% significance level.
- Do the results change if we implement the test at a 1% significance level?

## Applications

30. A manager of an industrial plant asserts that workers on average do not complete a job using Method A in the same amount of time as they would using Method B. Seven workers are randomly selected. Each worker's completion time (in minutes) is recorded by the use of Method A and Method B.

Worker	Method A	Method B
1	15	16
2	21	25
3	16	18
4	18	22
5	19	23
6	22	20
7	20	20

- Specify the null and alternative hypotheses to test the manager's assertion.
  - At the 10% significance level, specify the critical value(s) and the decision rule.
  - Assuming that the completion time difference is normally distributed, calculate the value of the test statistic.
  - Is the manager's assertion supported by the data?
31. A diet center claims that it has the most effective weight loss program in the region. Its advertisements say, "Participants in our program lose more than 5 pounds within a month." Six clients of this program are weighed on the first day of the diet and then one month later.

Client	Weight on First Day of Diet	Weight One Month Later
1	158	151
2	205	200
3	170	169
4	189	179
5	149	144
6	135	129

- Specify the null and alternative hypotheses that test the diet center's claim.
- Assuming that weight loss is normally distributed, calculate the value of the test statistic.
- Approximate the  $p$ -value.
- At the 5% significance level, do the data support the diet center's claim?

32. A bank employs two appraisers. When approving borrowers for mortgages, it is imperative that the appraisers value the same types of properties consistently. To make sure that this is the case, the bank examines six properties that the appraisers had valued recently.

Property	Value from Appraiser 1	Value from Appraiser 2
1	\$235,000	\$239,000
2	195,000	190,000
3	264,000	271,000
4	315,000	310,000
5	435,000	437,000
6	515,000	525,000

- Specify the competing hypotheses that determine whether there is any difference between the values estimated by Appraiser 1 and Appraiser 2.
  - At the 5% significance level, find the critical value(s).
  - Assuming that the value difference is normally distributed, calculate the value of the test statistic.
  - Is there sufficient evidence to conclude that the appraisers are inconsistent in their estimates? Explain.
33. The quality department at ElectroTech is examining which of two microscope brands (Brand A or Brand B) to purchase. They have hired someone to inspect six circuit boards using both microscopes. Below are the results in terms of the number of defects (e.g., solder voids, misaligned components) found using each microscope.

Circuit Board	Number of defects with Brand A	Number of defects with Brand B
1	12	14
2	8	9
3	16	16
4	14	12
5	9	8
6	13	15

- Specify the null and alternative hypotheses to test for differences in the defects found between the microscope brands.

- b. At the 5% significance level, find the critical value(s) of the test. What is the decision rule?
- c. Assuming that the difference in defects is normally distributed, calculate the value of the test statistic.
- d. Based on the preceding results, is there a difference between the microscope brands?
34. A computer technology firm wishes to check whether the speed of a new processor exceeds that of an existing processor when used in one of its popular laptop computer models. Accordingly, it measures the time required to complete seven common tasks on two otherwise identical computers, one with the new processor and one with the existing processor. The time required is as follows:

Task	Time Required with New Processor (seconds)	Time Required with Existing Processor (seconds)
1	1.47	1.68
2	2.59	2.99
3	5.21	5.69
4	3.49	3.75
5	3.99	4.25
6	3.10	2.99
7	5.75	6.19

- a. Specify the null and alternative hypotheses to test whether the time required for the new processor is less than the existing processor.
- b. At the 5% significance level, what is the critical value(s)? What is the decision rule?
- c. Assuming that the difference in time is normally distributed, calculate the value of the test statistic.
- d. Based on the above results, is the new processor faster than the old processor?
35. **FILE Mock\_SAT.** A recent report criticizes SAT-test-preparation providers for promising big score gains without any hard data to back up such claims (*The Wall Street Journal*, May 20, 2009). Suppose eight college-bound students take a mock SAT, complete a three-month test-prep course, and then take the real SAT.

Student	Score on Mock SAT	Score on Real SAT
1	1830	1840
2	1760	1800
3	2000	2010
4	2150	2190
5	1630	1620
6	1840	1960
7	1930	1890
8	1710	1780

- a. Specify the competing hypotheses that determine whether completion of the test-prep course increases a student's score on the real SAT.
- b. Using the appropriate commands in Excel, calculate the value of the test statistic and the  $p$ -value. Assume that the SAT scores difference is normally distributed.
- c. At the 5% significance level, do the sample data support the test-prep providers' claims?
36. **FILE Insurance\_Premiums.** The marketing department at Insure-Me, a large insurance company, wants to advertise that customers can save, on average, more than \$100 on their annual automotive insurance policies (relative to their closest competitor) by switching their policies to Insure-Me. However, to avoid potential litigation for false advertising, they select a random sample of 50 policyholders and compare their premiums to those of their closest competitor. A portion of the data is presented below.

Policyholder	Competitor's Premium	"Insure-Me" Premium
1	958	1086
2	1034	366
⋮	⋮	⋮
50	1161	964

- a. Specify the competing hypotheses to determine whether the mean difference between the competitor's premium and Insure-Me's premium is over \$100.
- b. Using the appropriate commands in Excel, find the value of the test statistic and the  $p$ -value.
- c. What is the conclusion at the 5% significance level? What is the conclusion at the 10% significance level?
37. **FILE Electronic\_Utilities.** The following table shows the annual returns (in percent) for Fidelity's Select Electronic and Select Utilities mutual funds.

Year	Electronic	Utilities
2001	-14.23	-21.89
2002	-50.54	-30.40
2003	71.89	26.42
2004	-9.81	24.22
2005	15.75	9.36
2006	0.30	30.08
2007	4.67	18.13
2008	-49.87	-36.00
2009	84.99	14.39

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com)

- a. Set up the hypotheses to test the claim that the mean return for the Electronic mutual fund differs from the mean return for the Utilities mutual fund.
- b. Using the appropriate commands in Excel, find the value of the test statistic. What is the  $p$ -value?
- c. At the 5% significance level, do the mean returns differ?

38. **FILE Labor Costs.** The labor quotation department at Excabar, a large manufacturing company, wants to verify the accuracy of their labor bidding process (estimated cost per unit versus actual cost per unit). They have randomly chosen 35 product quotations that subsequently were successful (meaning the company won the contract for the product). A portion of the data is shown in the accompanying table.

Product	Actual cost/unit	Estimated cost/unit
1	12.90	13.90
2	15.80	18.80
⋮	⋮	⋮
35	14.80	17.80

- Specify the competing hypotheses to determine whether there is a difference between the estimated cost and the actual cost.
  - Using the appropriate commands in Excel, find the value of the test statistic and the  $p$ -value.
  - At the 1% significance level, what is the conclusion?
39. **FILE Smoking Weight.** It is fairly common for people to put on weight when they quit smoking. While a small weight gain is normal, excessive weight gain can create new health concerns that erode the benefits of not smoking. The accompanying table shows a portion of the weight data for 50 women before quitting and six months after quitting.

Weight after Quitting	Weight before Quitting
155	140
142	144
⋮	⋮
147	135

- Construct and interpret the 95% confidence interval for the mean gain in weight.

- Use the preceding confidence interval to determine whether or not the mean gain in weight is 5 pounds.

40. **FILE Shift.** When faced with a power hitter, many baseball teams utilize a defensive shift. A shift usually involves putting three infielders on one side of second base against pull hitters. Many believe that a power hitter's batting average is lower when he faces a shift defense as compared to when he faces a standard defense. Consider the following batting averages of 10 power hitters over the 2010 and 2011 seasons when they faced a shift defense versus when they faced a standard defense.

Player	Average When Shift	Average When No Shift
Jack Cust	0.239	0.270
Adam Dunn	0.189	0.230
Prince Fielder	0.150	0.263
Adrian Gonzalez	0.186	0.251
Ryan Howard	0.177	0.317
Brian McCann	0.321	0.250
David Ortiz	0.245	0.232
Carlos Pena	0.243	0.191
Mark Teixeira	0.168	0.182
Jim Thome	0.211	0.205

SOURCE: *The Fielding Bible*-Volume III, March 2012

- Specify the competing hypotheses to determine whether the use of the defensive shift lowers a power hitter's batting average.
- Using the appropriate commands in Excel, calculate the value of the test statistic and the  $p$ -value. Assume that the batting average difference is normally distributed.
- At the 5% significance level, is the defensive shift effective in lowering a power hitter's batting average?

## 10.3 INFERENCE CONCERNING THE DIFFERENCE BETWEEN TWO PROPORTIONS

**LO 10.3**

In the preceding two sections, we focused on quantitative data, where we compared means of two populations. Now we turn our attention to qualitative data, where we provide statistical inference concerning the difference between two population proportions. This technique has many practical applications. For instance, an investor may want to determine whether the bankruptcy rate is the same for firms in the technology industry as compared to firms in the construction industry. The resulting analysis will help determine the relative risk of investing in these two industries. Or perhaps a marketing executive maintains that the proportion of women who buy a firm's product is significantly greater than the proportion of men who buy the product. If this claim is supported by the data, it provides information as to where the firm should advertise. In another case, a consumer advocacy group may state that the proportion of young adults (aged 18 to 35 years old) who carry health insurance is less than the proportion of older adults (aged 36 years or older). Health and government officials

Make inferences about the difference between two population proportions based on independent sampling.

might be particularly interested in this type of information. All of these examples deal with comparing two population proportions. Our parameter of interest is  $p_1 - p_2$ , where  $p_1$  and  $p_2$  denote the proportions in the first and second populations, respectively. The estimator for the difference between two population proportions is  $\bar{P}_1 - \bar{P}_2$ .

## Confidence Interval for $p_1 - p_2$

Since the population proportions  $p_1$  and  $p_2$  are unknown, we estimate them by  $\bar{p}_1$  and  $\bar{p}_2$ , respectively. The first sample proportion is computed as  $\bar{p}_1 = x_1/n_1$  where  $x_1$  denotes the number of successes in  $n_1$  observations drawn from population 1. Similarly,  $\bar{p}_2 = x_2/n_2$  is the sample proportion derived from population 2 where  $x_2$  is the number of successes in  $n_2$ . The difference  $\bar{p}_1 - \bar{p}_2$  is a point estimate of  $p_1 - p_2$ . Recall from Chapter 7 that the standard errors for the estimators  $\bar{P}_1$  and  $\bar{P}_2$  are  $se(\bar{P}_1) = \sqrt{\frac{p_1(1-p_1)}{n_1}}$  and  $se(\bar{P}_2) = \sqrt{\frac{p_2(1-p_2)}{n_2}}$ , respectively. Therefore, for two independently drawn samples, the standard error,  $se(\bar{P}_1 - \bar{P}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ . Since  $p_1$  and  $p_2$  are unknown, we estimate the standard error by  $\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$ . Finally, when both  $n_1$  and  $n_2$  are sufficiently large, the sampling distribution of  $\bar{P}_1 - \bar{P}_2$  can be approximated by the normal distribution. We construct a confidence interval for the difference between two population proportions using the following formula.

### CONFIDENCE INTERVAL FOR $P_1 - P_2$

A  $100(1 - \alpha)\%$  confidence interval for the difference between two population proportions  $p_1 - p_2$  is given by:

$$(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}.$$

As noted, the above formula is valid only when the two samples are sufficiently large; the general guideline is that  $n_1 p_1$ ,  $n_1(1 - p_1)$ ,  $n_2 p_2$ , and  $n_2(1 - p_2)$  must all be greater than or equal to 5, where  $p_1$  and  $p_2$  are evaluated at  $\bar{p}_1$  and  $\bar{p}_2$ , respectively.

### EXAMPLE 10.9

Despite his inexperience, candidate A appears to have gained support among the electorate. Three months ago, in a survey of 120 registered voters, 55 said that they would vote for Candidate A. Today, 41 registered voters in a sample of 80 said that they would vote for Candidate A. Construct the 95% confidence interval for the difference between the two population proportions.

**SOLUTION:** Let  $p_1$  and  $p_2$  represent the population proportion of the electorate who support the candidate today and three months ago, respectively. In order to calculate the 95% confidence interval for  $p_1 - p_2$ , we use the formula  $(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$ . We compute the sample proportions as

$$\bar{p}_1 = x_1/n_1 = 41/80 = 0.5125 \quad \text{and} \quad \bar{p}_2 = x_2/n_2 = 55/120 = 0.4583.$$

Note that the normality condition is satisfied because  $n_1 \bar{p}_1$ ,  $n_1(1 - \bar{p}_1)$ ,  $n_2 \bar{p}_2$ , and  $n_2(1 - \bar{p}_2)$  all exceed 5. For the 95% confidence interval, we use the  $z$  table to find  $z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$ . Substituting the values into the formula, we find



$$(0.5125 - 0.4583) \pm 1.96 \sqrt{\frac{0.5125(1 - 0.5125)}{80} + \frac{0.4583(1 - 0.4583)}{120}}$$

$$= 0.0542 \pm 0.1412 \text{ or } [-0.0870, 0.1954].$$

With 95% confidence, we can report that the percentage change of support for the candidate is between  $-8.70\%$  and  $19.54\%$ .

## Hypothesis Test for $p_1 - p_2$

The null and alternative hypotheses for testing the difference between two population proportions under independent sampling will take one of the following forms:

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: p_1 - p_2 = d_0$	$H_0: p_1 - p_2 \leq d_0$	$H_0: p_1 - p_2 \geq d_0$
$H_A: p_1 - p_2 \neq d_0$	$H_A: p_1 - p_2 > d_0$	$H_A: p_1 - p_2 < d_0$

We use the symbol  $d_0$  to denote a given hypothesized difference between the unknown population proportions  $p_1$  and  $p_2$ . In most cases,  $d_0$  is set to zero. For example, when testing if the population proportions differ—that is, if  $p_1 \neq p_2$ —we use a two-tailed test with the competing hypotheses defined as  $H_0: p_1 - p_2 = 0$  versus  $H_A: p_1 - p_2 \neq 0$ . If, on the other hand, we wish to determine whether or not the proportions differ by some amount, say 20%, we set  $d_0 = 0.20$  and define the competing hypotheses as  $H_0: p_1 - p_2 = 0.20$  versus  $H_A: p_1 - p_2 \neq 0.20$ . One-tailed tests are defined similarly.

### EXAMPLE 10.10

Let's revisit Example 10.9. Specify the competing hypotheses in order to determine whether the proportion of those who favor Candidate A has changed over the three-month period. Using the 95% confidence interval, what is the conclusion to the test? Explain.

**SOLUTION:** In essence, we would like to determine whether  $p_1 \neq p_2$ , where  $p_1$  and  $p_2$  represent the population proportion of the electorate who support the candidate today and three months ago, respectively. We formulate the competing hypotheses as:

$$H_0: p_1 - p_2 = 0$$

$$H_A: p_1 - p_2 \neq 0$$

In the previous example, we constructed the 95% confidence interval for the difference between the population proportions as  $[-0.0870, 0.1954]$ . We note that the interval contains zero, the value hypothesized under the null hypothesis. Therefore, we are unable to reject the null hypothesis. In other words, from the given sample data, we cannot conclude at the 5% significance level that the support for candidate A has changed.

We now introduce the standard four-step procedure for conducting one- or two-tailed hypothesis tests concerning the difference between two proportions  $p_1 - p_2$ . We transform its estimator  $\bar{P}_1 - \bar{P}_2$  into a corresponding  $z$  statistic by subtracting the hypothesized difference  $d_0$  from this estimator and dividing by the standard error of the estimator  $se(\bar{P}_1 - \bar{P}_2)$ . When we developed the confidence interval for

$p_1 - p_2$ , we assumed  $se(\bar{P}_1 - \bar{P}_2) = \sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}$ . However, if  $d_0$  is zero—that is,  $H_0: p_1 = p_2$ —both  $\bar{p}_1$  and  $\bar{p}_2$  are essentially the estimates of the same unknown population proportion. For this reason, the standard error can be improved upon by computing the pooled estimate  $\bar{p} = (x_1 + x_2)/(n_1 + n_2)$  for the unknown population proportion, which is now based on a larger sample.

### TEST STATISTIC FOR TESTING $p_1 - p_2$

The value of the test statistic for a hypothesis test concerning the difference between two proportions  $p_1 - p_2$  is computed using one of the following two formulas:

1. If the hypothesized difference  $d_0$  is zero, then the value of the test statistic is

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

$$\text{where } \bar{p}_1 = \frac{x_1}{n_1}, \bar{p}_2 = \frac{x_2}{n_2}, \text{ and } \bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2}.$$

2. If the hypothesized difference  $d_0$  is not zero, then the value of the test statistic is

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - d_0}{\sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}}.$$

As in the case of the confidence interval, the above formulas are valid only when the two samples are sufficiently large.

### EXAMPLE 10.11

Recent research by analysts and retailers claims significant gender differences when it comes to online shopping (*The Wall Street Journal*, March 13, 2008). A survey revealed that 5,400 of 6,000 men said they “regularly” or “occasionally” make purchases online, compared with 8,600 of 10,000 women surveyed. At the 5% significance level, test whether the proportion of all men who regularly or occasionally make purchases online is greater than the proportion of all women.

**SOLUTION:** We use the critical value approach to conduct this test. Let  $p_1$  and  $p_2$  denote the population proportions of men and of women who make online purchases, respectively. We wish to test whether the proportion of men who make purchases online is greater than the proportion of women; that is,  $p_1 - p_2 > 0$ . Therefore, we construct the competing hypotheses as

$$H_0: p_1 - p_2 \leq 0$$

$$H_A: p_1 - p_2 > 0$$

For a right-tailed test with  $\alpha = 0.05$ , the appropriate critical value is  $z_\alpha = z_{0.05} = 1.645$ . Since the hypothesized difference is zero, or  $d_0 = 0$ , we compute the value of the test statistic as  $z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ . We first compute the sample proportions

$\bar{p}_1 = x_1/n_1 = 5400/6000 = 0.90$  and  $\bar{p}_2 = x_2/n_2 = 8600/10000 = 0.86$ . The normality condition is satisfied since  $n_1\bar{p}_1$ ,  $n_1(1 - \bar{p}_1)$ ,  $n_2\bar{p}_2$ , and  $n_2(1 - \bar{p}_2)$  all exceed 5. Next we calculate  $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{5400 + 8600}{6000 + 10000} = 0.875$ . Thus,

$$z = \frac{(0.90 - 0.86)}{\sqrt{0.875(1 - 0.875)\left(\frac{1}{6000} + \frac{1}{10000}\right)}} = \frac{0.04}{0.0054} = 7.41.$$

The decision rule is to reject  $H_0$  if  $z > 1.645$ . Since  $7.41 > 1.645$ , we reject  $H_0$ . The proportion of men who shop online either regularly or occasionally is greater than the proportion of women at the 5% significance level. Our results are consistent with the recent decision by so many retailers to redesign their websites to attract male customers.

### EXAMPLE 10.12

While we expect relatively expensive wines to have more desirable characteristics than relatively inexpensive wines, people are often confused in their assessment of the quality of wine in a blind test (*The New York Times*, December 16, 2010). In a recent experiment at a local winery, the same wine is served to two groups of people but with different price information. In the first group, 60 people are told that they are tasting a \$25 wine, of which 48 like the wine. In the second group, only 20 of 50 people like the wine when they are told that it is a \$10 wine. The experiment is conducted to determine if the proportion of people who like the wine in the first group is more than 20 percentage points higher than in the second group. Conduct this test at the 5% significance level using the  $p$ -value approach.

**SOLUTION:** Let  $p_1$  and  $p_2$  denote the proportions of people who like the wine in groups 1 and 2, respectively. We want to test if the proportion of people who like the wine in the first group is more than 20 percentage points higher than in the second group. Thus, we construct the competing hypotheses as

$$H_0: p_1 - p_2 \leq 0.20$$

$$H_A: p_1 - p_2 > 0.20$$

We first compute the sample proportions as  $\bar{p}_1 = x_1/n_1 = 48/60 = 0.80$  and  $\bar{p}_2 = x_2/n_2 = 20/50 = 0.40$  and note that the normality condition is satisfied since  $n_1\bar{p}_1$ ,  $n_1(1 - \bar{p}_1)$ ,  $n_2\bar{p}_2$ , and  $n_2(1 - \bar{p}_2)$  all exceed 5.

Since  $d_0 = 0.20$ , the value of the test statistic is computed as

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - d_0}{\sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}} = \frac{(0.80 - 0.40) - 0.20}{\sqrt{\frac{0.80(1 - 0.80)}{60} + \frac{0.40(1 - 0.40)}{50}}} = 2.31.$$

For this right-tailed test, we compute the  $p$ -value as  $P(Z \geq 2.31) = 1 - 0.9896 = 0.0104$ . Since the  $p$ -value  $< \alpha = 0.05$ , we reject the null hypothesis. At the 5% significance level, we conclude that the proportion of people who like the wine in the first group is more than 20 percentage points higher than in the second group. Overall, this result is consistent with scientific research, which has demonstrated the power of suggestion in wine tasting.

## EXERCISES 10.3

### Mechanics

41. Given  $\bar{p}_1 = 0.85$ ,  $n_1 = 400$ ,  $\bar{p}_2 = 0.90$ ,  $n_2 = 350$ , construct the 90% confidence interval for the difference between the population proportions. Is there a difference between the population proportions at the 10% significance level? Explain.
42. Given  $x_1 = 50$ ,  $n_1 = 200$ ,  $x_2 = 70$ ,  $n_2 = 250$ , construct the 95% confidence interval for the difference between the population proportions. Is there a difference between the population proportions at the 5% significance level? Explain.

43. Consider the following competing hypotheses and accompanying sample data.

$$H_0: p_1 - p_2 \geq 0$$

$$H_A: p_1 - p_2 < 0$$

$$x_1 = 250 \quad x_2 = 275$$

$$n_1 = 400 \quad n_2 = 400$$

- a. At the 5% significance level, find the critical value(s).
- b. Calculate the value of the test statistic.
- c. What is the conclusion to the test? Is  $p_1$  significantly less than  $p_2$ ?

44. Consider the following competing hypotheses and accompanying sample data.

$$H_0: p_1 - p_2 = 0$$

$$H_A: p_1 - p_2 \neq 0$$

$$x_1 = 100 \quad x_2 = 172$$

$$n_1 = 250 \quad n_2 = 400$$

- Calculate the value of the test statistic.
  - Calculate the  $p$ -value.
  - At the 5% significance level, what is the conclusion to the test? Do the population proportions differ?
  - Repeat the analysis with the critical value approach.
45. Consider the following competing hypotheses and accompanying sample data.

$$H_0: p_1 - p_2 = 0$$

$$H_A: p_1 - p_2 \neq 0$$

$$x_1 = 300 \quad x_2 = 325$$

$$n_1 = 600 \quad n_2 = 500$$

- Calculate the value of the test statistic.
  - Calculate the  $p$ -value.
  - At the 5% significance level, what is the conclusion to the test? Do the population proportions differ?
46. Consider the following competing hypotheses and accompanying sample data.

$$H_0: p_1 - p_2 = 0.20$$

$$H_A: p_1 - p_2 \neq 0.20$$

$$x_1 = 150 \quad x_2 = 130$$

$$n_1 = 250 \quad n_2 = 400$$

- Calculate the value of the test statistic.
- Calculate the  $p$ -value.
- At the 5% significance level, what is the conclusion to the test? Can you conclude that the difference between the population proportions differs from 0.20?
- Repeat the analysis with the critical value approach.

## Applications

47. A recent study claims that girls and boys do not do equally well on math tests taken from the 2nd to 11th grades (*Chicago Tribune*, July 25, 2008). Suppose in a representative sample, 344 of 430 girls and 369 of 450 boys score at proficient or advanced levels on a standardized math test.
- Construct the 95% confidence interval for the difference between the population proportions of girls and boys who score at proficient or advanced levels.
  - Develop the appropriate null and alternative hypotheses to test whether the proportion of girls who score at proficient or advanced levels differs from the proportion of boys.

- At the 5% significance level, what is the conclusion to the test? Do the results support the study's claim?

48. Reducing scrap of 4-foot planks of hardwood is an important factor in reducing cost at a wood-flooring manufacturing company. Accordingly, engineers at Lumberworks are investigating a potential new cutting method involving lateral sawing that may reduce the scrap rate. To examine its viability, samples of 500 and 400 planks, respectively, were examined under the old and new methods. Sixty-two of the 500 planks were scrapped under the old method, whereas 36 of the 400 planks were scrapped under the new method.

- Construct the 95% confidence interval for the difference between the population scrap rates between the old and new methods, respectively.
- Specify the null and alternative hypotheses to test for differences in the population scrap rates between the old and new cutting methods, respectively.
- Using the results from part (a), can we conclude at the 5% significance level that the scrap rate of the new method is different than the old method?

49. According to the Pew report, 14.6% of newly married couples in 2008 reported that their spouse was of another race or ethnicity (*CNNLiving*, June 7, 2010). In a similar survey in 1980, only 6.8% of newlywed couples reported marrying outside their race or ethnicity. Suppose both of these surveys were conducted on 120 newly married couples.

- Specify the competing hypotheses to test the claim that there is an increase in the proportion of people who marry outside their race or ethnicity.
- Calculate the value of the test statistic and the  $p$ -value.
- At the 5% level of significance, what is the conclusion to the test?

50. Research by Harvard Medical School experts suggests that boys are more likely than girls to grow out of childhood asthma when they hit their teenage years (*BBC News*, August 15, 2008). Scientists followed over 1,000 children between the ages of 5 and 12, all of whom had mild to moderate asthma. By the age of 18, 14% of the girls and 27% of the boys seemed to have grown out of asthma. Suppose their analysis was based on 500 girls and 500 boys.

- Develop the hypotheses to test whether the proportion of boys who grow out of asthma in their teenage years is more than that of girls.
- Use the  $p$ -value approach to test the assertion in part (a) at the 5% significance level.
- Does the sample data suggest that the proportion of boys who grow out of asthma in their teenage years is more than 0.10 than that of girls? Use the critical value approach to test this assertion at the 5% significance level.

51. More people are using social media to network, rather than phone calls or e-mails (U.S. *News & World Report*, October 20, 2010). From an employment perspective, jobseekers are no longer calling up friends for help with job placement, as they can now get help online. In a recent survey of 150 jobseekers, 67 said they used LinkedIn to search for jobs. A similar survey of 140 jobseekers, conducted three years ago, had found that 58 jobseekers had used LinkedIn for their job search. Is there sufficient evidence to suggest that more people are now using LinkedIn to search for jobs as compared to three years ago? Use a 5% level of significance for the analysis.

52. The director of housekeeping at Elegante, a luxury resort hotel with two locations (*Seaside* and *Oceanfront*), wants to evaluate housekeeping performance at those two locations. Random samples of 100 rooms were inspected at each location for defects (e.g., missing towels, missing soap, dirty floors or showers, dusty tables) after being cleaned. It was found that 21 of the rooms at *Seaside* had some housekeeping defect(s), and 28 rooms at *Oceanfront* had some housekeeping defect(s).

- Develop the hypotheses to test whether the proportion of housekeeping defects differs between the two hotel locations.
- What is the value of the test statistic and the associated  $p$ -value?
- Do the results suggest that the proportion of housekeeping defects differs between the two hotel locations at the 5% significance level?
- Construct the 95% confidence interval for the difference between the population housekeeping defect rates between the two hotel locations. How can this confidence interval be used to reach the same conclusion as in part (c)?

53. Due to late delivery problems with an existing supplier, the director of procurement at ElectroTech began to place orders for electrical switches with a new supplier as part of a “dual-source” (two-supplier) strategy. Now she wants to revert to a “single-source” (i.e., one supplier) strategy to simplify purchasing activities. She wishes to conduct a test to infer whether the new supplier will continue to outperform the old supplier. Based on recent sample data, she found that 27 of 150 orders placed with the old supplier arrived late, whereas 6 of 75 orders placed with the new supplier arrived late.

- Specify the null and alternative hypotheses to test for whether the proportion of late deliveries with the new supplier is less than that of the old supplier.
- Compute the value of the test statistic.
- Use the critical value approach to test the hypotheses at the 5% significance level. What is the conclusion?

54. According to a recent report, 32.2% of American adults (aged 20 and older) are obese (*The New York Times*, August 15, 2008). Among ethnic groups in general, African-American women are more overweight than Caucasian women, but African-American men are less obese than Caucasian men. Sarah Weber, a recent college graduate, is curious to determine if the same pattern also exists in her hometown on the West Coast. She randomly selects 220 African Americans and 300 Caucasian adults for the analysis. The following table contains the sample information.

Race	Gender	Obese	Not Obese
African Americans	Males	36	94
	Females	35	55
Caucasians	Males	62	118
	Females	31	89

- Use the  $p$ -value approach to test if the proportion of obese African-American men is less than the proportion of obese Caucasian men at  $\alpha = 0.05$ .
- Use the critical value approach to test if the proportion of obese African-American women is more than the proportion of obese Caucasian women at  $\alpha = 0.05$ .
- Use the critical value approach to test if the proportion of obese African Americans differs from the proportion of obese Caucasian adults at the 5% significance level.

55. Only 26% of psychology majors are “satisfied” or “very satisfied” with their career paths as compared to 50% of accounting majors (*The Wall Street Journal*, October 11, 2010). Suppose these results were obtained from a survey of 300 psychology majors and 350 accounting majors.

- Develop the appropriate null and alternative hypotheses to test whether the proportion of accounting majors satisfied with their career paths differs from psychology majors by more than 20 percentage points.
- Calculate the value of the test statistic and the  $p$ -value.
- At the 5% significance level, what is the conclusion?

56. In an effort to make children’s toys safer and more tamperproof, toy packaging has become cumbersome for parents to remove in many cases. Accordingly, the director of marketing at Toys4Tots, a large toy manufacturer, wants to evaluate the effectiveness of a new packaging design which engineers claim will reduce customer complaints by more than 10 percentage points. Customer satisfaction surveys were sent to 250 parents who registered toys packaged under the old design and 250 parents who registered toys packaged under the new design. Of these, 85 parents expressed dissatisfaction with packaging of the old design, and 40 parents



expressed dissatisfaction with packaging of the new design.

- a. Specify the null and alternative hypotheses to test whether customer complaints have been reduced by more than 10 percentage points under the new packaging design.
  - b. What is the value of the test statistic and the associated  $p$ -value?
  - c. At the 5% significance level, do the results support the engineers' claim?
  - d. At the 1% significance level, do the results support the engineers' claim?
57. A recent report suggests that business majors spend the least amount of time on course work than all other college students (*The New York Times*, November 17, 2011). A provost of a university decides to conduct a survey where students are asked if they study hard, defined as spending at least 20 hours per week on course work. Of 120 business majors included in the survey, 20 said that they

studied hard, as compared to 48 out of 150 nonbusiness majors who said that they studied hard. At the 5% significance level, can we conclude that the proportion of business majors who study hard is less than that of nonmajors? Provide the details.

58. It has generally been believed that it is not feasible for men and women to be just friends (*The New York Times*, April 12, 2012). Others argue that this belief may not be true anymore since gone are the days when men worked and women stayed at home and the only way they could get together was for romance. In a recent survey, 186 heterosexual college students were asked if it was feasible for male and female students to be just friends. Thirty-two percent of females and 57% of males reported that it was not feasible for men and women to be just friends. Suppose the study consisted of 100 female and 86 male students. At the 5% significance level, can we conclude that there is a greater than 10 percentage point difference between the proportion of male and female students with this view? Provide the details.

## WRITING WITH STATISTICS

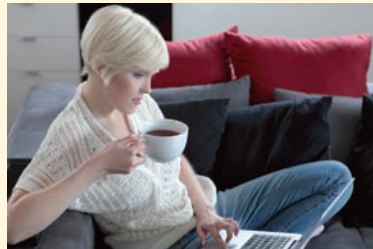


The recent phenomenon of online dating has made it as likely for would-be couples to meet via e-mail or other virtual matchmaking services as through friends and family (*CNN*, February 6, 2012). Studies that have looked at gender differences in mate selection have found that women put greater emphasis on the race and financial stability of a partner, while men mostly look for physical attractiveness. Recent survey results reported in *USA Today* (February 2, 2012) showed that 13% of women and 8% of men want their partner to be of the same ethnic background. The same survey also reported that 36% of women and 13% of men would like to meet someone who makes as much money as they do.

Anka Wilder, working for a small matchmaking service in Cincinnati, Ohio, wants to know if a similar pattern also exists with her customers. She has access to the preferences of 160 women and 120 men customers. In this sample, she finds that 28 women and 12 men customers want their partner to be of the same ethnicity. Also, 50 women and 10 men want their partner to make as much money as they do.

Anka wants to use this sample information to:

1. Determine whether the proportion of women who want their partner to be of the same ethnic background is significantly greater than that of men.
2. Determine whether the proportion of women who want their partner to make as much money as they do is more than 20 percentage points greater than that of men.



### Sample Report— Online Dating Preferences

With the advent of the Internet, there has been a surge in online dating services that connect individuals with similar interests, religious, and cultural backgrounds for personal relationships. In 1992, when the Internet was still in its infancy, less than 1% of Americans met their partners through online dating services. By 2009, about 22% of heterosexual couples and 61% of same-sex couples reported meeting online (*CNN*, February 6, 2012).



A recent survey suggested that a higher proportion of women than men would like to meet someone with a similar ethnic background. Also, the difference between the proportion of women and men who would like to meet someone who makes as much money as they do is greater than 20%.

A couple of hypothesis tests were performed to determine if similar gender differences existed for online dating customers in Cincinnati, Ohio. The sample consisted of responses from 160 women and 120 men. The summary of the test results is presented in Table 10.A.

**TABLE 10.A** Test Statistics and  $p$ -values for Hypothesis Tests

Hypotheses	Test Statistic	$p$ -value
$H_0: p_1 - p_2 \leq 0$ $H_A: p_1 - p_2 > 0$	$z = \frac{0.175 - 0.10}{\sqrt{0.1429(1 - 0.1429) \left( \frac{1}{160} + \frac{1}{120} \right)}} = 1.77$	0.0384
$H_0: p_1 - p_2 \leq 0.20$ $H_A: p_1 - p_2 > 0.20$	$z = \frac{0.3125 - 0.0833 - 0.20}{\sqrt{\frac{0.3125(1 - 0.3125)}{160} + \frac{0.0833(1 - 0.0833)}{120}}} = 0.66$	0.2546

First, it was tested if the proportion of women, denoted  $p_1$ , who want their partner to be of the same ethnicity is significantly greater than that of men, denoted  $p_2$ . It was found that 28 out of 160 women valued this trait, yielding a sample proportion of  $\bar{p}_1 = 28/160 = 0.175$ ; a similar proportion for men was calculated as  $\bar{p}_2 = 12/120 = 0.10$ . The first row of Table 10. A shows the competing hypotheses, the value of the test statistic, and the  $p$ -value for this test. At the 5% significance level, the proportion of women who want the same ethnicity was greater than that of men. In the second test,  $p_1$  and  $p_2$  denoted the proportion of women and men, respectively, who would like their partner to make as much money as they do; here  $\bar{p}_1 = 50/160 = 0.3125$  and  $\bar{p}_2 = 10/120 = 0.0833$ . The second row of Table 10. A shows the competing hypotheses, the value of the test statistic, and the  $p$ -value for this test. At the 5% significance level, the proportion of women who want their partner to make as much income as they do is not more than 20 percentage points greater than that of men. Online dating is a relatively new market and any such information is important for individuals looking for relationships as well as for service providers.

## CONCEPTUAL REVIEW

### **LO 10.1** Make inferences about the difference between two population means based on independent sampling.

**Independent samples** are samples that are completely unrelated to one another.

A  $100(1 - \alpha)\%$  **confidence interval for the difference between two population means  $\mu_1 - \mu_2$** , based on independent samples, is

- $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ , if  $\sigma_1^2$  and  $\sigma_2^2$  are known.
- $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, df} \sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$ , if  $\sigma_1^2$  and  $\sigma_2^2$  are unknown but assumed equal. The pooled sample variance is  $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ , and  $df = n_1 + n_2 - 2$ .
- $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, df} \sqrt{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}$ , if  $\sigma_1^2$  and  $\sigma_2^2$  are unknown and assumed unequal. The degrees of freedom are calculated as  $df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$ , and are rounded down to the nearest integer.

When conducting **hypothesis tests about the difference between two means**  $\mu_1 - \mu_2$ , based on independent samples, the value of the **test statistic** is:

- $z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ , if  $\sigma_1^2$  and  $\sigma_2^2$  are known.
- $t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$ , if  $\sigma_1^2$  and  $\sigma_2^2$  are unknown but assumed equal.
- $t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}}$ , if  $\sigma_1^2$  and  $\sigma_2^2$  are unknown and assumed unequal.

The degrees of freedom for the last two tests are the same as the ones defined for the corresponding confidence intervals. The formulas for estimation and testing are valid only if  $\bar{X}_1 - \bar{X}_2$  (approximately) follows a normal distribution.

### LO 10.2 Make inferences about the mean difference based on matched-pairs sampling.

A common case of dependent sampling, commonly referred to as **matched-pairs sampling**, is when the samples are paired or matched in some way.

For matched-pairs sampling, the population parameter of interest is referred to as the mean difference  $\mu_D$  where  $D = X_1 - X_2$ , and the random variables  $X_1$  and  $X_2$  are matched in a pair. A  $100(1 - \alpha)\%$  **confidence interval for the mean difference  $\mu_D$** , based on a matched-pairs sample, is given by  $\bar{d} \pm t_{\alpha/2, df} s_D / \sqrt{n}$ , where  $\bar{d}$  and  $s_D$  are the mean and the standard deviation, respectively, of  $D$ , and  $df = n - 1$ .

When conducting a **hypothesis test about  $\mu_D$**  the value of the **test statistic** is calculated as  $t_{df} = \frac{\bar{d} - d_0}{s_D / \sqrt{n}}$ , where  $d_0$  is a hypothesized mean difference and  $df = n - 1$ .

### LO 10.3 Make inferences about the difference between two population proportions based on independent sampling.

A  $100(1 - \alpha)\%$  **confidence interval for the difference between two population proportions  $p_1 - p_2$**  is given by  $(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}$ .

When conducting **hypothesis tests about the difference between two proportions  $p_1 - p_2$** , the value of the **test statistic** is calculated as:

- $z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1 - \bar{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$ , if the hypothesized difference  $d_0$  between  $p_1$  and  $p_2$  is zero.  
The pooled sample proportion is  $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$ .
- $z = \frac{(\bar{p}_1 - \bar{p}_2) - d_0}{\sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}}$ , if the hypothesized difference  $d_0$  between  $p_1$  and  $p_2$  is not zero.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

59. A new study has found that, on average, 6- to 12-year-old children are spending less time on household chores today compared to 1981 levels (*The Wall Street Journal*, August 27, 2008). Suppose two samples representative of the study's results

report the following summary statistics for the two periods:

1981 Levels	2008 Levels
$\bar{x}_1 = 30$ minutes	$\bar{x}_2 = 24$ minutes
$s_1 = 4.2$ minutes	$s_1 = 3.9$ minutes
$n_1 = 30$	$n_2 = 30$

- a. Specify the competing hypotheses to test the study's claim that children today spend less time on household chores as compared to children in 1981.
  - b. Calculate the value of the test statistic assuming that the unknown population variances are equal.
  - c. Approximate the  $p$ -value.
  - d. At the 5% significance level, do the data support the study's claim? Explain.
  - e. Repeat the hypothesis test using the critical value approach.
60. Do men really spend more money on St. Patrick's Day as compared to women? A recent survey found that men spend an average of \$43.87 while women spend an average of \$29.54 (*USA Today*, March 17, 2009). Assume that these data were based on a sample of 100 men and 100 women and the population standard deviations of spending for men and women are \$32 and \$25, respectively.
- a. Specify the competing hypotheses to determine whether men spend more money on St. Patrick's Day as compared to women.
  - b. Calculate the value of the test statistic.
  - c. Calculate the  $p$ -value.
  - d. At the 1% significance level, do men spend more money on St. Patrick's Day as compared to women? Explain.
61. **FILE Balanced\_European.** The accompanying table shows annual return data from 2001–2009 for Vanguard's Balanced Index and European Stock Index mutual funds.

Year	Balanced Index	European Stock Index
2001	–3.02%	–20.30%
2002	–9.52	–17.95
2003	19.87	38.70
2004	9.33	20.86
2005	4.65	9.26
2006	11.02	33.42
2007	6.16	13.82
2008	–22.21	–44.73
2009	20.05	31.91

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com)

- a. Set up the hypotheses to test whether the mean returns of the two funds differ. (*Hint:* This is a matched-pairs comparison.)
- b. Using the appropriate commands in Excel, calculate the value of the test statistic and the  $p$ -value. Assume that the return difference is normally distributed.
- c. At the 5% significance level, what is the conclusion?

62. **FILE Cholesterol Levels.** It is well documented that cholesterol over 200 is a risk factor in developing heart disease for both men and women (<http://Livingstrong.com>, January 11, 2011). Younger men are known to have higher cholesterol levels than younger women; however, beyond age 55, women are more likely to have higher cholesterol levels. A recent college graduate working at a local blood lab has access to the cholesterol data of 50 men and 50 women in the 20–40 age group. The accompanying table shows a portion of the data.

Men	Women
181	178
199	193
⋮	⋮
190	182

Use the critical value approach, at the 1% significance level, to determine if there are any differences in the mean cholesterol levels for men and women in the age group. It is fair to assume that the population variances for men and women are equal. Describe your steps clearly.

63. **FILE Pregnancy Weight.** It is important for women to gain the right amount of weight during pregnancy by eating a healthy, balanced diet (<http://webmd.com>). It is recommended that a woman of average weight before pregnancy should gain 25 to 35 pounds during pregnancy. The accompanying table shows a portion of the weight data for 40 women before and after pregnancy.

Weight after Pregnancy	Weight before Pregnancy
168	114
161	107
⋮	⋮
157	136

- a. At the 5% level of significance, determine if the mean weight gain of women due to pregnancy is more than 30 pounds.
  - b. At the 5% level of significance, determine if the mean weight gain of women due to pregnancy is more than 35 pounds.
64. A farmer is concerned that a change in fertilizer to an organic variant might change his crop yield. He subdivides 6 lots and uses the old fertilizer on one half of each lot and the new fertilizer on the other half. The following table shows the results.

Lot	Crop Yield Using Old Fertilizer	Crop Yield Using New Fertilizer
1	10	12
2	11	10
3	10	13
4	9	9
5	12	11
6	11	12

- Specify the competing hypotheses that determine whether there is any difference between the average crop yields from the use of the different fertilizers.
  - Assuming that crop yields are normally distributed, calculate the value of the test statistic.
  - At the 5% significance level, find the critical value(s).
  - Is there sufficient evidence to conclude that the crop yields are different? Should the farmer be concerned?
65. A recent Health of Boston report suggests that 14% of female residents suffer from asthma as opposed to 6% of males (*The Boston Globe*, August 16, 2010). Suppose 250 females and 200 males responded to the study.
- Develop the appropriate null and alternative hypotheses to test whether the proportion of females suffering from asthma is greater than the proportion of males.
  - Calculate the value of the test statistic and its associated  $p$ -value.
  - At the 5% significance level, what is the conclusion? Do the data suggest that females suffer more from asthma than males?
66. Depression engulfs millions of Americans every day. A new federal study reported that 10.9% of adults aged 18–24 identified with some level of depression versus 6.8% of adults aged 65 or older (*The Boston Globe*, October 18, 2010). Suppose 250 young adults (18–24 years old) and 200 older adults (65 years old and older) responded to the study.
- Develop the appropriate null and alternative hypotheses to test whether the proportion of young adults suffering from depression is greater than the proportion of older adults suffering from depression.
  - Calculate the value of the test statistic and the  $p$ -value.
  - At the 5% significance level, what is the conclusion? Do the sample data suggest that young adults suffer more from depression than older adults?

67. **FILE SAT\_Writing.** (Use Excel) The SAT is required of most students applying for college admission in the United States. This standardized test has gone through many revisions over the years. In 2005, a new writing section was introduced that includes a direct writing measure in the form of an essay. People argue that female students generally do worse on math tests but better on writing tests. Therefore, the new section may help reduce the usual male lead on the overall average SAT score (*The Washington Post*, August 30, 2006). Consider the following scores on the writing component of the test of eight male and eight female students.

Males	620	570	540	580	590	580	480	620
Females	660	590	540	560	610	590	610	650

- Construct the null and the alternative hypotheses to test if females outscore males on writing tests.
  - Assuming that the difference in scores is normally distributed, calculate the value of the test statistic. Do not assume that the population variances are equal.
  - Implement the test at  $\alpha = 0.01$  and interpret your results.
68. Fresh numbers from the U.S. Department of Transportation suggest that fewer flights in the U.S. arrive on time than before. The explanations offered for the lackluster performance are understaffed airlines, a high volume of travelers, and overtaxed air traffic control. A transportation analyst is interested in comparing the performance at two major international airports, namely Kennedy International (JFK) in New York and O'Hare International in Chicago. She finds that 70% of the flights were on time at JFK compared with 63% at O'Hare. Suppose these proportions were based on 200 flights at each of these two airports. The analyst believes that the proportion of on-time flights at JFK is more than 5 percentage points higher than that of O'Hare.
- Develop the competing hypotheses to test the transportation analyst's belief.
  - Compute the value of the test statistic.
  - Use the  $p$ -value approach to test the above assertion.
  - Repeat the analysis with the critical value approach.
69. **FILE Safety\_Program.** An engineer wants to determine the effectiveness of a safety program. He collects annual loss of hours due to accidents in 12 plants "before and after" the program was put into operation.

Plant	Before	After	Plant	Before	After
1	100	98	7	88	90
2	90	88	8	75	70
3	94	90	9	65	62
4	85	86	10	58	60
5	70	67	11	67	60
6	83	80	12	104	98

- Specify the competing hypotheses that determine whether the safety program was effective.
- Using the appropriate commands in Excel, calculate the value of the test statistic. Assume that the hours difference is normally distributed.
- At the 5% significance level, specify the critical value(s).
- Is there sufficient evidence to conclude that the safety program was effective? Explain.

## CASE STUDIES

**CASE STUDY 10.1** Chad Perrone is a financial analyst in Boston studying the annual return data for the health and information technology industries. He randomly samples 20 firms in each industry and notes each firm's annual return. A portion of the data is shown in the accompanying table.

**Data for Case Study 10.1** Annual Returns for Firms in Health and Information Technology Industries

Health	Information Technology
10.29%	4.77%
32.17	1.14
:	:
13.21	22.61

**FILE**  
Health\_Info

In a report, use the sample information to:

- Provide descriptive statistics and comment on the reward and risk in each industry.
- Determine whether the average returns in each industry differ at the 5% significance level. Assume that the population variances are unequal.

**CASE STUDY 10.2** The Speedo LZR Racer Suit is a high-end, body-length swimsuit that was launched on February 13, 2008. When 17 world records fell at the December 2008 European Short Course Championships in Croatia, many believed a modification in the rules surrounding swimsuits was necessary. The FINA Congress, the international governing board for swimming, banned the LZR Racer and all other body-length swimsuits from competition effective January 2010. In a statement to the public, FINA defended its position with the following statement: "FINA wishes to recall the main and core principle that swimming is a sport essentially based on the physical performance of the athlete" (*BBC Sport*, March 14, 2009).

Luke Johnson, a freelance journalist, wonders if the decision made by FINA has statistical backing. He conducts an experiment with the local university's Division I swim team. He times 10 of the swimmers swimming the 50-meter breaststroke in his/her bathing suit and then retests them while wearing the LZR Racer. A portion of the results are shown in the accompanying table.

**Data for Case Study 10.2** 50-Meter Breaststroke Times (in seconds)

Swimmer	Time in Bathing Suit	Time in LZR Racer
1	27.64	27.45
2	27.97	28.06
:	:	:
10	38.08	37.93

**FILE**  
LZR\_Racer

In a report, use the sample information to:

1. Determine whether the LZR Racer significantly improves swimmers' times at the 5% significance level. Assume that the time difference is normally distributed.
2. Comment on whether the data appear to support FINA's decision.

**CASE STUDY 10.3** Paige Thomsen is about to graduate from college at a local university in San Francisco. Her options are to look for a job in San Francisco or go home to Denver and search for work there. Recent data report that average starting salaries for college graduates is \$48,900 in San Francisco and \$40,900 in Denver (*Forbes*, June 26, 2008). Suppose these data were based on 100 recent graduates in each city where the population standard deviation is \$16,000 in San Francisco and \$14,500 in Denver. For social reasons, Paige is also interested in the percent of the population who are in their 20s. The same report states that 20% of the population are in their 20s in San Francisco; the corresponding percentage in Denver is 22%.

In a report, use the sample information to:

1. Determine whether the average starting salary in San Francisco is greater than Denver's average starting salary at the 5% significance level.
2. Determine whether the proportion of the population in their 20s differs in these two cities at the 5% significance level.

## APPENDIX 10.1 Guidelines for Other Software Packages

The following section provides brief commands for specific software packages: Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. All three packages have an option to perform a paired-comparisons test. However, it is also possible to conduct this test by finding the differences between the paired items, and then using the one-sample *t*-test discussed in Chapter 9.

Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### Testing $\mu_1 - \mu_2$

**FILE**  
*Gold\_Oil*

- A. (Replicating Example 10.4) From the menu choose **Stat > Basic Statistics > 2-Sample t**. Choose **Each sample is in its own column**, and after **Sample 1** select Gold and after **Sample 2** select Oil.
- B. Choose **Options**. After **Alternative hypothesis**, select "Difference  $\neq$  hypothesized difference".

#### Testing $\mu_D$

**FILE**  
*Food\_Calories*

- A. (Replicating Example 10.8) From the menu choose **Stat > Basic Statistics > Paired t**. **Each sample is in its own column**, and after **Sample 1** select Before and after **Sample 2** select After.
- B. Choose **Options**. After **Alternative hypothesis**, select "Difference  $>$  hypothesized difference".

#### Testing $p_1 - p_2$

- A. (Replicating Example 10.11) From the menu choose **Stat > Basic Statistics > 2 Proportions**. Choose **Summarized data**. Under **Sample 1**, enter 5400 for **Number of events** and 6000 for **Number of trials**. Under **Sample 2**, enter 8600 for **Number of events** and 10000 for **Number of trials**.



- B. Choose **Options**. After **Alternative hypothesis**, select “Difference > hypothesized difference”. After **Test method**, select “Use the pooled estimate for the proportion.”

## SPSS

### Testing $\mu_1 - \mu_2$

- A. (Replicating Example 10.4) Pool all **Gold\_Oil** data in one column and label Pooled. In adjacent column (labeled Group), denote all Gold values with 0 and all Oil values with 1.
- B. From the menu choose **Analyze > Compare Means > Independent-Samples T Test**.
- C. Select Pooled as **Test Variable(s)** and Group as **Grouping Variable**. Select **Define Groups** and enter 0 for **Group 1** and 1 for **Group 2**.

**FILE**  
*Gold\_Oil*

### Testing $\mu_D$

- A. (Replicating Example 10.8) From the menu choose **Analyze > Compare Means > Independent-Samples T Test**.
- B. Select Before as **Variable1** and After as **Variable2**.

**FILE**  
*Food\_Calories*

## JMP

### Testing $\mu_1 - \mu_2$

- A. (Replicating Example 10.4) Pool all **Gold\_Oil** data in one column and label it Pooled. In adjacent column (labeled Group and read as nominal data), denote all Gold values with 0 and all Oil values with 1.
- B. From the menu choose **Analyze > Fit Y by X**.
- C. Select Pooled as **Y, Response** and Group as **X, Factor**.
- D. Click on the red triangle next to the header that reads **Oneway Analysis of Column 1 by Column 2** and select **t test** (to use a pooled variance, select **Means/Anova/Pooled t**).

**FILE**  
*Gold\_Oil*

### Testing $\mu_D$

- A. (Replicating Example 10.8) From the menu choose **Analyze > Matched Pairs**.
- B. Choose Before and After as **Y, Paired Response**.

**FILE**  
*Food\_Calories*

# 11

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 11.1 Discuss features of the  $\chi^2$  distribution.
- LO 11.2 Construct a confidence interval for the population variance.
- LO 11.3 Conduct a hypothesis test for the population variance.
- LO 11.4 Discuss features of the  $F$  distribution.
- LO 11.5 Construct a confidence interval for the ratio of two population variances.
- LO 11.6 Conduct a hypothesis test for the ratio of two population variances.

# Statistical Inference Concerning Variance

So far, when conducting statistical inference concerning quantitative data, we have restricted our attention to the population mean. The mean is a basic measure of central location, but in many instances we are also interested in making inferences about the measures of variability or dispersion. For instance, quality-control studies use the population variance to measure the variability of the weight, size, or volume of a product. The population variance is also the most widely used quantitative measure of risk in investments. In this chapter, we study statistical inference with respect to the population variance as well as the ratio of two population variances. In order to construct confidence intervals or conduct hypothesis tests regarding the population variance, we use a new distribution called the  $\chi^2$  (chi-square) distribution. We then turn our attention to analyzing the ratio of two population variances. In order to construct confidence intervals or conduct hypothesis tests concerning this ratio, we use another new distribution called the  $F$  distribution.



## INTRODUCTORY CASE

### Assessing the Risk of Mutual Fund Returns

In Chapter 3, investment counselor Rebecca Johnson examined annual return data for two top-performing mutual funds from the last decade: Vanguard's Precious Metals and Mining fund (henceforth, Metals) and Fidelity's Strategic Income fund (henceforth, Income). Table 11.1 shows relevant descriptive statistics for the two mutual funds for the years 2000–2009. Rebecca knows that the reward of investing is measured by its average return, while the standard deviation is the most widely used measure of risk. A client of Rebecca's has specific questions related to the risk of investing. He would like to invest a portion of his money in the Metals fund so long as the risk does not exceed 25%. He also wonders if the risk of investing in the Income fund differs from 8.5%, which is the risk inherent in similar funds. Rebecca is familiar with making statistical inference with respect to the population mean and a comparison of the population means; however, she is not clear on how to implement these techniques with respect to the population standard deviation.

**TABLE 11.1** Descriptive Measures for the Metals and the Income Funds,  $n = 10$

	Metals	Income
Mean	24.65%	8.51%
Standard Deviation	37.13%	11.07%

Rebecca would like to use the above sample information to:

1. Compare the investment risks of the Metals and the Income funds by using the appropriate confidence intervals and hypothesis tests.
2. Determine whether the standard deviation of the Metals fund exceeds 25%.
3. Determine whether the standard deviation of the Income fund differs from 8.5%.

A synopsis of this case is provided at the end of Section 11.2.

## 11.1 INFERENCE CONCERNING THE POPULATION VARIANCE

The population variance is used in quality-control studies to measure the variability of the weight, size, or volume of a product. Consider, for example, a bottler who wishes its production line to fill a certain amount of beverage in each bottle. It is important not only to get the desired average amount filled in the bottles, but also to keep the variability of the amount filled below some tolerance limit. Similarly, with variance used as a quantitative measure of risk, an investor may want to evaluate his/her risk in a particular investment. Other examples for the relevance of making inference regarding the population variance include evaluating the consistency of an athlete or a team, the variability of speeds on a highway, and the variability of repair costs of a certain automobile.

Recall that we use the sample mean  $\bar{X}$  as the estimator of the population mean  $\mu$ . Similarly, we use the sample variance  $S^2$  as an estimator of the population variance  $\sigma^2$ . Using a random sample of  $n$  observations drawn from the population, we compute  $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$  as an estimate of  $\sigma^2$ . In order to examine the techniques for statistical inferences regarding  $\sigma^2$ , we first need to analyze the sampling distribution of  $S^2$ .

### LO 11.1

Discuss features of the  $\chi^2$  distribution.

### Sampling Distribution of $S^2$

Statistical inferences regarding  $\sigma^2$  are based on the  $\chi^2$  or **chi-square** distribution. Like the  $t$  distribution, the  $\chi^2$  distribution is characterized by a family of distributions, where each distribution depends on its particular degrees of freedom  $df$ . It is common, therefore, to refer to it as the  $\chi^2_{df}$  distribution.

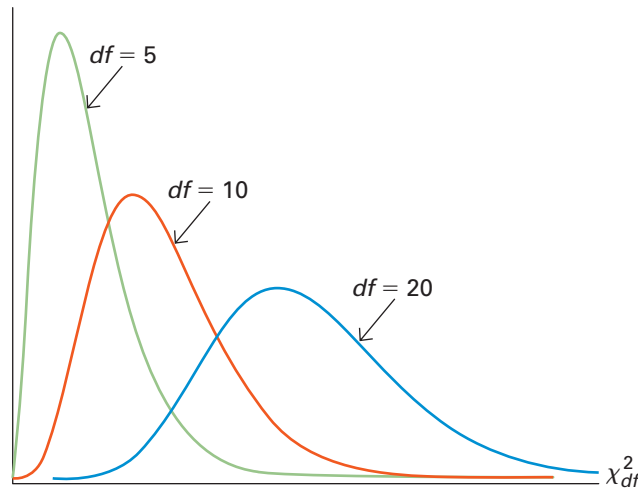
In general, the  $\chi^2_{df}$  distribution is the probability distribution of the sum of several independent squared standard normal random variables. Here  $df$  is defined as the number of squared standard normal random variables included in the summation. Recall that the estimator  $S^2$  of the population variance is based on the squared differences between the sample values and the sample mean. If  $S^2$  is computed from a random sample of  $n$  observations drawn from an underlying normal population, then we can define the  $\chi^2_{df}$  variable as  $\frac{(n-1)S^2}{\sigma^2}$ .

#### THE SAMPLING DISTRIBUTION OF $\frac{(n-1)S^2}{\sigma^2}$

If a sample of size  $n$  is taken from a normal population with a finite variance, then the statistic  $\chi^2_{df} = \frac{(n-1)S^2}{\sigma^2}$  follows the  $\chi^2_{df}$  distribution with  $df = n - 1$ .

In earlier chapters, we denoted the random variables by uppercase letters and particular outcomes of the random variables by the corresponding lowercase letters. For instance, the statistics  $Z$  and  $T_{df}$  are random variables and their values are given by  $z$  and  $t_{df}$ , respectively. It is cumbersome to continue with the distinction between the random variable and its value in this chapter. Here, we use the notation  $\chi^2_{df}$  to represent a random variable as well as its value. Similarly, for the  $F_{(df_1, df_2)}$  distribution introduced in Section 11.2, we will use  $F_{(df_1, df_2)}$  to represent both a random variable and its value.

From Figure 11.1, we note that the  $\chi^2_{df}$  distributions are positively skewed, where the extent of skewness depends on the degrees of freedom. As the  $df$  grow larger, the  $\chi^2_{df}$  distribution tends to the normal distribution. For instance, in the figure, the  $\chi^2_{20}$  distribution resembles the shape of the normal distribution.



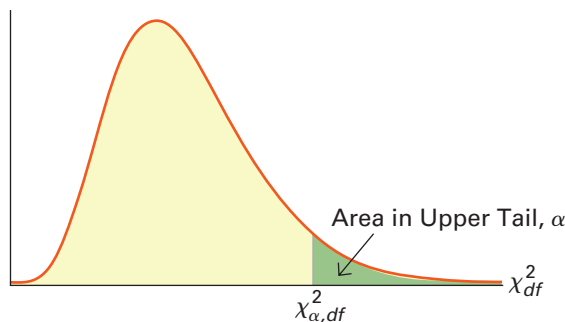
**FIGURE 11.1**  
The  $\chi^2_{df}$  distribution with various degrees of freedom

#### SUMMARY OF THE $\chi^2_{df}$ DISTRIBUTION

- The  $\chi^2_{df}$  distribution is characterized by a family of distributions, where each distribution depends on its particular degrees of freedom  $df$ .
- The values of the  $\chi^2_{df}$  distribution range from zero to infinity.
- The  $\chi^2_{df}$  distribution is positively skewed, and the extent of skewness depends on the  $df$ . As the  $df$  grow larger, the  $\chi^2_{df}$  distribution approaches the normal distribution.

### Locating $\chi^2_{df}$ Values and Probabilities

For a  $\chi^2_{df}$  distributed random variable, we use the notation  $\chi^2_{\alpha,df}$  to represent a value such that the area in the upper (right) tail of the distribution is  $\alpha$ . In other words,  $P(\chi^2_{df} \geq \chi^2_{\alpha,df}) = \alpha$ . Figure 11.2 illustrates the notation  $\chi^2_{\alpha,df}$ , which we use to locate  $\chi^2_{df}$  values and probabilities from the  $\chi^2$  (chi-square) table.



**FIGURE 11.2**  
Graphical depiction of  
 $P(\chi^2_{df} \geq \chi^2_{\alpha,df}) = \alpha$

A portion of the upper tail areas and the corresponding values for the  $\chi^2_{df}$  distributions are given in Table 11.2. Table 3 of Appendix A provides a more complete table.

Suppose we want to find the  $\chi^2_{\alpha,df}$  with  $\alpha = 0.05$  and  $df = 10$ —that is,  $\chi^2_{0.05,10}$ . Using Table 11.2, we look at the first column labeled  $df$  and find the value 10. We then continue along this row until we reach the column 0.050. Here we see the value  $\chi^2_{0.05,10} = 18.307$  such that  $P(\chi^2_{10} \geq 18.307) = 0.05$ .

**TABLE 11.2** Portion of the  $\chi^2$  table

df	Area in Upper Tail, $\alpha$									
	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
:	:	:	:	:	:	:	:	:	:	:
10	2.156	2.558	3.247	<b>3.940</b>	4.865	15.987	<b>18.307</b>	20.483	23.209	25.188
:	:	:	:	:	:	:	:	:	:	:
100	67.328	70.065	74.222	77.929	82.358	118.342	124.342	129.561	135.807	140.170

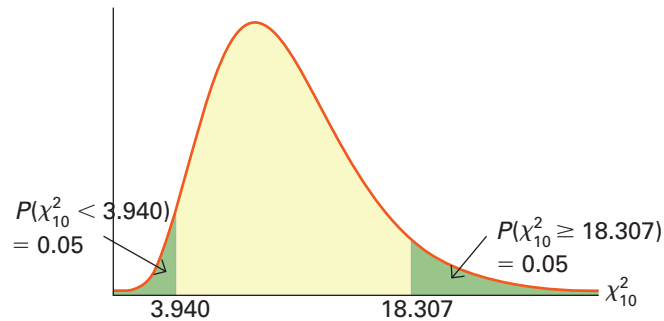
Sometimes we need to derive values in the lower (left) tail of the distribution. Given that the area under any probability distribution equals one, if the area to the left of a given value equals  $\alpha$ , then the area to the right must equal  $1 - \alpha$ . In other words, the relevant value on the lower tail of the distribution is  $\chi^2_{1-\alpha, df}$  where  $P(\chi^2_{df} \geq \chi^2_{1-\alpha, df}) = 1 - \alpha$ .

#### LOCATING $\chi^2_{df}$ VALUES ON THE LOWER TAIL

For a  $\chi^2_{df}$  distributed random variable,  $\chi^2_{1-\alpha, df}$  represents a value such that  $P(\chi^2_{df} < \chi^2_{1-\alpha, df}) = \alpha$ , or equivalently,  $P(\chi^2_{df} \geq \chi^2_{1-\alpha, df}) = 1 - \alpha$ .

Suppose that we want to find the value such that the area to the left of the  $\chi^2_{10}$  variable equals 0.05. Given that the area to the left of this value is 0.05, we know that the area to its right is  $1 - 0.05 = 0.95$ ; thus, we need to find  $\chi^2_{1-0.05, 10} = \chi^2_{0.95, 10}$ . Again, we find  $df = 10$  in the first column and follow this row until we intersect the column 0.95 and find the value 3.940. This is the value such that  $P(\chi^2_{10} \geq 3.940) = 0.95$  or  $P(\chi^2_{10} < 3.940) = 0.05$ . Figure 11.3 graphically depicts the probability  $\alpha = 0.05$  on both sides of the  $\chi^2_{10}$  distribution and the corresponding  $\chi^2_{10}$  values.

**FIGURE 11.3**  
Graph of the  
probability  $\alpha = 0.05$  on  
both sides of  $\chi^2_{10}$



#### EXAMPLE 11.1

Find the value  $x$  for which:

- $P(\chi^2_5 \geq x) = 0.025$
- $P(\chi^2_8 < x) = 0.025$

#### SOLUTION

- We find the value  $x$  such that the area in the upper tail of the distribution equals 0.025. Referencing Table 3 in Appendix A, we find  $df = 5$  in the first column and follow this row until we intersect the column 0.025 and find the value 12.833; therefore,  $x = 12.833$ .



- b. We find the value  $x$  such that the area in the lower tail of the distribution equals 0.025. We solve this problem as  $P(\chi_8^2 \geq x) = 1 - 0.025 = 0.975$ . Again referencing Table 3 in Appendix A, we find  $df = 8$  in the first column and follow this row until we intersect the column 0.975 and find  $x = 2.180$ . This is equivalent to  $P(\chi_8^2 < 2.180) = 0.025$ .

## Confidence Interval for the Population Variance

LO 11.2

Consider a  $\chi_{df}^2$  distributed random variable. We can use the notation that we just introduced to make the following probability statement concerning this random variable:

$$P(\chi_{1-\alpha/2,df}^2 \leq \chi_{df}^2 \leq \chi_{\alpha/2,df}^2) = 1 - \alpha.$$

This indicates that the probability that  $\chi_{df}^2$  falls between  $\chi_{1-\alpha/2,df}^2$  and  $\chi_{\alpha/2,df}^2$  is equal to  $1 - \alpha$ , where  $1 - \alpha$  is the familiar confidence coefficient. Substituting  $\chi_{df}^2 = \frac{(n-1)S^2}{\sigma^2}$  into the probability statement yields

$$P\left(\chi_{1-\alpha/2,df}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2,df}^2\right) = 1 - \alpha.$$

After manipulating this equation algebraically, we arrive at the formula for the confidence interval for  $\sigma^2$ .

### CONFIDENCE INTERVAL FOR $\sigma^2$

A  $100(1 - \alpha)\%$  confidence interval for the population variance  $\sigma^2$  is computed as

$$\left[ \frac{(n-1)s^2}{\chi_{\alpha/2,df}^2}, \frac{(n-1)s^2}{\chi_{1-\alpha/2,df}^2} \right],$$

where  $df = n - 1$ .

This formula is valid only when the random sample is drawn from a normally distributed population. Note that the confidence interval is not in the usual format of point estimate  $\pm$  margin of error. Since the confidence intervals for the population mean and the population proportion are based on the  $z$  or the  $t_{df}$  distributions, the symmetry of these distributions leads to the same margin of error that is added to and subtracted from the point estimate. However, for a nonsymmetric  $\chi_{df}^2$  distribution, what is added to and subtracted from the point estimate of the population variance is not the same. Finally, since the standard deviation is just the positive square root of the variance, a  $100(1 - \alpha)\%$  confidence interval for the population standard deviation is computed as

$$\left[ \sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2,df}^2}}, \sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2,df}^2}} \right].$$

### EXAMPLE 11.2

Compute 95% confidence intervals for the population standard deviation for the Metals fund and the Income fund using the data from Table 11.1 in the introductory case. Assume that returns are normally distributed.

**SOLUTION:** For the years 2000–2009 ( $n = 10$ ), the sample standard deviation for the Metals fund is  $s = 37.13\%$ , while the sample standard deviation for the Income fund is  $s = 11.07\%$ .

Construct a confidence interval for the population variance.

We first determine the 95% confidence interval for the population variance for the Metals fund. Given  $n = 10$ ,  $df = 10 - 1 = 9$ . For a 95% confidence interval,  $\alpha = 0.05$  and  $\alpha/2 = 0.025$ . Thus, we find  $\chi^2_{\alpha/2, df} = \chi^2_{0.025, 9} = 19.023$  and  $\chi^2_{1-\alpha/2, df} = \chi^2_{0.975, 9} = 2.700$ . The 95% confidence interval for the population variance is

$$\left[ \frac{(n-1)s^2}{\chi^2_{\alpha/2, df}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, df}} \right] = \left[ \frac{(10-1)(37.13)^2}{19.023}, \frac{(10-1)(37.13)^2}{2.700} \right] \\ = [652.25(\%)^2, 4,595.46(\%)^2].$$

Taking the positive square root of the limits of this interval, we find the corresponding 95% confidence interval for the population standard deviation as [25.54%, 67.79%]. With 95% confidence, we report that the standard deviation of the return for the Metals fund is between 25.54% and 67.79%. Similarly, for the Income fund, we compute the 95% confidence interval for the population standard deviation as [7.61%, 20.21%].

### LO 11.3

Conduct a hypothesis test for the population variance.

## Hypothesis Test for the Population Variance

Let's now develop the four-step procedure for conducting a hypothesis test concerning the population variance. Following the methodology used in the last two chapters, we specify the null and the alternative hypotheses as:

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \sigma^2 = \sigma_0^2$	$H_0: \sigma^2 \leq \sigma_0^2$	$H_0: \sigma^2 \geq \sigma_0^2$
$H_A: \sigma^2 \neq \sigma_0^2$	$H_A: \sigma^2 > \sigma_0^2$	$H_A: \sigma^2 < \sigma_0^2$

Here  $\sigma_0^2$  is the hypothesized value of the population variance  $\sigma^2$ . As before, we can use confidence intervals to implement two-tailed hypothesis tests; however, for one-tailed tests concerning the population variance, we implement the four-step procedure using either the critical value approach or the  $p$ -value approach.

### TEST STATISTIC FOR $\sigma^2$

The value of the **test statistic** for the hypothesis test for the **population variance**  $\sigma^2$  is computed as

$$\chi^2_{df} = \frac{(n-1)s^2}{\sigma_0^2},$$

where  $df = n - 1$ ,  $s^2$  is the sample variance, and  $\sigma_0^2$  is the hypothesized value of the population variance. This formula is valid only if the underlying population is normally distributed.

### EXAMPLE 11.3

We again consider the introductory case. Rebecca Johnson's client asks if the standard deviation of returns for the Metals fund is significantly greater than 25%. This is equivalent to testing whether or not the variance is significantly greater than  $625(\%)^2$ . We conduct this test at the 5% significance level, based on the sample information provided in Table 11.1. We implement the four-step procedure using the critical value approach and assume that returns are normally distributed.

**SOLUTION:** In this example, the relevant parameter of interest is the population variance  $\sigma^2$ . Since we wish to determine whether the variance is greater than  $625(\%)^2$ , we specify the competing hypotheses as

$$H_0: \sigma^2 \leq 625$$

$$H_A: \sigma^2 > 625$$

The  $\chi^2$  test is valid because the underlying population is assumed to be normally distributed. For this right-tailed test, the critical value  $\chi^2_{\alpha, df}$  is derived from  $P(\chi^2_{df} \geq \chi^2_{\alpha, df}) = \alpha$ . Referencing Table 3 in Appendix A with  $\alpha = 0.05$  and  $df = n - 1 = 9$ , we find the critical value  $\chi^2_{0.05, 9}$  as 16.919. The decision rule is to reject the null hypothesis if  $\chi^2_9$  exceeds 16.919.

Given that  $n = 10$  and  $s = 37.13$ , we compute the value of the test statistic as

$$\chi^2_{df} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(10-1)(37.13)^2}{625} = 19.85.$$

We reject the null hypothesis because the value of the test statistic falls in the rejection region ( $\chi^2_9 = 19.85$  exceeds  $\chi^2_{0.05, 9} = 16.919$ ). At the 5% significance level, the variance of the Metals fund is significantly greater than  $625(\%)^2$ . Analogously, the standard deviation is significantly greater than 25%, implying that the risk associated with this investment is more than the client wants to accept.

### EXAMPLE 11.4

Rebecca Johnson's client from the introductory case also wonders if the standard deviation of returns for the Income fund differs from 8.5%. This is equivalent to testing whether or not the variance differs from  $72.25(\%)^2$ . Use the  $p$ -value approach to conduct this test at the 5% significance level.

**SOLUTION:** This is an example of a two-tailed test for the population variance  $\sigma^2$ . We specify the competing hypotheses for the two-tailed test for the population variance as

$$H_0: \sigma^2 = 72.25$$

$$H_A: \sigma^2 \neq 72.25$$

Given that  $n = 10$  and  $s = 11.07$ , we compute the value of the test statistic as

$$\chi^2_{df} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(10-1)(11.07)^2}{72.25} = 15.27.$$

Recall that the  $p$ -value is the probability of obtaining a value of the test statistic that is at least as extreme as the one that we actually observed, given that the null hypothesis is true as an equality. For a two-tailed test, we double the probability that is considered extreme. For example, for a two-tailed test for the population mean  $\mu$  with a  $z$  statistic, the  $p$ -value is computed as  $2P(Z \geq z)$  if  $z > 0$  or  $2P(Z \leq z)$  if  $z < 0$ . Note that  $z > 0$  if  $\bar{x} > \mu_0$  and  $z < 0$  if  $\bar{x} < \mu_0$ . Similarly, for a two-tailed test for the population variance  $\sigma^2$ , the  $p$ -value is computed as two times the upper tail area if  $s^2 > \sigma_0^2$  or two times the lower tail area if  $s^2 < \sigma_0^2$ .

For the Income fund, since  $s^2 > \sigma_0^2$  ( $122.54 > 72.25$ ), we compute the  $p$ -value as  $2P(\chi^2_{df} \geq 15.27)$ . For  $df = 9$ , since 15.27 lies between 14.684 and 16.919 (see Table 3 in Appendix A),  $P(\chi^2_9 \geq 15.27)$  lies between 0.05 and 0.10. Multiplying this probability by two results in a  $p$ -value between 0.10 and 0.20.

At the 5% significance level, we cannot reject  $H_0$  because the  $p$ -value is greater than  $\alpha = 0.05$ . Therefore, we cannot conclude that the risk, measured by the variance of the return, differs from  $72.25(\%)^2$ ; or equivalently, we cannot conclude that the standard deviation differs from 8.5%.

## Using Excel to Calculate $p$ -Values

We can easily find exact probabilities using Excel's CHISQ.DIST.RT function. To find the upper tail area we find an empty cell and input “=CHISQ.DIST.RT( $\chi^2_{df}$ ,  $df$ ),” where  $\chi^2_{df}$  is the value of the test statistic and  $df$  are the respective degrees of freedom. For Example 11.4, we input “=CHISQ.DIST.RT(15.27, 9)” and Excel returns 0.0838. The exact  $p$ -value for the two-tailed test is  $2(0.0838)$ , or 0.1676. Again, given a significance level of 5%, we are unable to reject the null hypothesis that the population variance equals  $72.25$  ( $\sigma_0^2$ ). If  $s^2 < \sigma_0^2$  such that we need a lower tail area, we can use Excel's CHISQ.DIST function or simply subtract the probability obtained from using Excel's CHISQ.DIST.RT function from 1.

### EXERCISES 11.1

#### Concepts

- Find the value  $x$  for which:
  - $P(\chi^2_8 \geq x) = 0.025$
  - $P(\chi^2_8 \geq x) = 0.05$
  - $P(\chi^2_8 < x) = 0.025$
  - $P(\chi^2_8 < x) = 0.05$
- Find the value  $x$  for which:
  - $P(\chi^2_{20} \geq x) = 0.005$
  - $P(\chi^2_{20} \geq x) = 0.01$
  - $P(\chi^2_{20} < x) = 0.005$
  - $P(\chi^2_{20} < x) = 0.01$
- In order to construct a confidence interval for the population variance, a random sample of  $n$  observations is drawn from a normal population. Use this information to find  $\chi^2_{\alpha/2, df}$  and  $\chi^2_{1-\alpha/2, df}$  under the following scenarios.
  - A 95% confidence level with  $n = 18$
  - A 95% confidence level with  $n = 30$
  - A 99% confidence level with  $n = 18$
  - A 99% confidence level with  $n = 30$
- A random sample of 25 observations is used to estimate the population variance. The sample mean and sample standard deviation are calculated as 52.5 and 3.8, respectively. Assume that the population is normally distributed.
  - Construct the 90% interval estimate for the population variance.
  - Construct the 99% interval estimate for the population variance.
  - Use your answers to discuss the impact of the confidence level on the width of the interval.
- The following values are drawn from a normal population.

20	29	32	27	34	25	30	31
----	----	----	----	----	----	----	----

- In order to conduct a hypothesis test for the population variance, you compute  $s^2 = 75$  from a sample of 21 observations drawn from a normally distributed population. Use the critical value approach to conduct the following tests at  $\alpha = 0.10$ .
  - $H_0: \sigma^2 \leq 50$ ;  $H_A: \sigma^2 > 50$
  - $H_0: \sigma^2 = 50$ ;  $H_A: \sigma^2 \neq 50$
- Consider the following hypotheses:
$$H_0: \sigma^2 = 200$$
$$H_A: \sigma^2 \neq 200$$
Approximate the  $p$ -value based on the following sample information, where the sample is drawn from a normally distributed population.
  - $s^2 = 300$ ;  $n = 25$
  - $s^2 = 100$ ;  $n = 25$
  - Which of the above sample information enables us to reject the null hypothesis at  $\alpha = 0.05$ ?
- You would like to test the claim that the variance of a normally distributed population is more than 2 squared units. You draw a random sample of 10 observations as 2, 4, 1, 3, 2, 5, 2, 6, 1, 4. At  $\alpha = 0.10$ , test the claim using (a) the  $p$ -value approach and (b) the critical value approach.

#### Applications

- A research analyst is examining a stock for possible inclusion in his client's portfolio. Over a 10-year period, the sample mean and the sample standard deviation of annual returns on the stock were 20% and 15%, respectively. The client wants to know if the risk, as measured by the standard deviation, differs from 18%.
  - Construct the 95% confidence intervals for the population variance and the population standard deviation.
  - What assumption did you make in constructing the confidence interval?
  - Based on the results in part (a), does the risk differ from 18%?

10. A replacement part for a machine must be produced within close specifications in order for it to be acceptable to customers. A random sample of 20 parts drawn from a normally distributed population yields a sample variance of  $s^2 = 0.03$ .
- Construct the 95% confidence interval for the population variance.
  - Production specifications call for the variance in the lengths of the parts to be exactly 0.05. Comment on whether or not the specifications are being violated.

11. A consumer advocacy group is concerned about the variability in the cost of prescription medication. The group surveys eight local pharmacies and obtains the following prices for a particular brand of medication:

\$25.50	32.00	33.50	28.75	29.50	35.00	27.00	29.00
---------	-------	-------	-------	-------	-------	-------	-------

- Calculate the point estimate for the population variance.
  - The group assumes that the prices represent a random sample drawn from a normally distributed population. Construct the 90% interval estimate for the population variance.
  - The group decides to begin a lobbying effort on its members' behalf if the variance in the price does not equal 4. What should the group do?
12. The following table shows the annual returns (in percent) for the Vanguard Energy Fund from 2005 through 2009.

Year	Energy
2005	44.60
2006	19.68
2007	37.00
2008	-42.87
2009	38.36

Source: finance.yahoo.com.

- Calculate the point estimate of  $\sigma$ .
  - Construct the 95% confidence interval for  $\sigma$ . Assume that returns are normally distributed.
13. The manager of a supermarket would like the variance of the waiting times of the customers not to exceed 3 minutes-squared. She would add a new cash register if the variance exceeds this threshold. She regularly checks the waiting times of the customers to ensure that the variance does not rise above the allowed level. In a recent random sample of 28 customer waiting times, she computes the sample variance as 4.2 minutes-squared. She believes that the waiting times are normally distributed.
- State the null and the alternative hypotheses to test if the threshold has been crossed.
  - Use the  $p$ -value approach to conduct the test at  $\alpha = 0.05$ .
  - Repeat the analysis with the critical value approach.
  - What should the manager do?

14. A restaurant owner is concerned about the consistency of business. He wants to determine if the standard deviation of the profits for each week is less than \$300. The profits from last week are listed below (in dollars). Assume that profits are normally distributed.

1,825	1,642	1,675	1,553	1,925	2,037	1,902
-------	-------	-------	-------	-------	-------	-------

- State the appropriate null and alternative hypotheses for the test.
  - Use the critical value approach to test the owner's concern at  $\alpha = 0.01$ .
  - Calculate the value of the test statistic.
  - Repeat the test at  $\alpha = 0.10$ .
15. India Fund, Inc. (IFN) is a close-ended equity mutual fund launched by the Blackstone Group, Asset Management Arm. Although it promises impressive returns, it does so at the cost of greater risk. An analyst would like to test if the variance of the returns for IFN is greater than  $1,000(\%)^2$ . He uses the following sample data to test his claim at a 5% level of significance.

2001	2002	2003	2004	2005	2006	2007
-28.33%	7.87%	140.63%	30.91%	63.35%	-2.78%	58.30%

- State the competing hypotheses.
  - Given  $\alpha = 0.05$ , specify the critical value(s).
  - Calculate the value of the test statistic. What assumption regarding the IFN returns did you make?
  - Is the variance of returns greater than  $1,000(\%)^2$ ?
16. Metalworks, a supplier of machine parts, fabricates bearings for use in aeronautical applications in which the standard deviation of the bearing diameter must be within 0.002 inch maximum. Otherwise, problems with fit will occur. The engineering department is conducting an experiment to investigate adherence to this requirement. A sample of 25 bearings has revealed a sample standard deviation of 0.0024 inch.
- State the appropriate null and alternative hypotheses to test if the requirement has been violated.
  - Compute the value of the test statistic. What assumption did you make regarding the bearing diameters?
  - Use the  $p$ -value approach to conduct the test at  $\alpha = 0.05$ . What is your conclusion?
  - Would your conclusion change at the 10% significance level?
17. Some transportation experts claim that it is the variability of speeds, rather than the level of speeds, that is a critical factor in determining the likelihood of an accident occurring (*Update*, Virginia Department of Transportation, Winter 2000). One of the experts claims that driving conditions are dangerous if the variance of speeds

- exceeds 80 (mph)<sup>2</sup>. On a heavily traveled highway, a random sample of 61 cars revealed a mean and a variance of speeds of 57.5 mph and 88.7 (mph)<sup>2</sup>, respectively.
- Set up the competing hypotheses to test the expert's claim.
  - At the 5% significance level, find the critical value(s) and state the decision rule.
  - Can you conclude that driving conditions are dangerous on this highway? Explain.
18. **FILE Sewing.** To maintain high consistency in its manual sewing operations, a custom manufacturer of high-quality fashion clothing has a goal in which all sewing employees should score within a standard deviation of 9 on a sewing dexterity test. To test adherence to this goal, a random sample of 30 employees was subjected to a needle-board dexterity test.
- State the hypotheses to test whether the standard deviation of the dexterity test scores exceeds 9.
  - Calculate the value of the test statistic. Assume that dexterity scores are normally distributed.
  - Use Excel's function (CHISQ.DIST.RT or CHISQ.DIST) to compute the  $p$ -value.
  - Make a conclusion at each of the following significance levels: 1%, 5%, and 10%.
19. **FILE Hourly\_Wage.** An economist is interested in the variability of hourly wages at a production plant. She collects data on 50 hourly wage earners.
- Set up the competing hypotheses to test whether the variance of hourly wages exceeds 35 ( $\$^2$ ).
  - Calculate the value of the test statistic.
  - Use Excel's function (CHISQ.DIST.RT or CHISQ.DIST) to calculate the  $p$ -value.
  - At the 5% significance level, does the variance of hourly wages exceed 35 ( $\$^2$ )? Explain.
20. **FILE MV\_Houses.** A realtor in Mission Viejo, California, believes that the standard deviation of house prices is more than 100 units, where each unit equals \$1,000. Assume house prices are normally distributed.
- State the null and the alternative hypotheses for the test.
  - Calculate the value of the test statistic.
  - Use Excel's function (CHISQ.DIST.RT or CHISQ.DIST) to calculate the  $p$ -value.
  - At  $\alpha = 0.05$ , what is the conclusion? Is the realtor's claim supported by the data?
21. **FILE MPG.** The data accompanying this exercise show miles per gallon (mpg) for 25 cars.
- State the null and the alternative hypotheses in order to test whether the variance differs from 62 mpg<sup>2</sup>.
  - Assuming that mpg are normally distributed, calculate the value of the test statistic.
  - Use Excel's function (CHISQ.DIST.RT or CHISQ.DIST) to calculate the  $p$ -value.
  - Make a conclusion at  $\alpha = 0.01$ .
22. **FILE Rentals.** Real estate investment in college towns continues to promise good returns (*The Wall Street Journal*, September 24, 2010). With rents holding up, this is good news for investors but the same cannot be said for students. There also tends to be significant variability in rents. Consider monthly rents of two bedroom apartments in two campus towns: Ann Arbor, Michigan, and Davis, California. A portion of the data is shown in the accompanying table.

Ann Arbor Rent	Davis Rent
\$850	\$744
929	850
:	:
1450	1810

SOURCE: www.zillow.com.

- Use Excel to calculate the standard deviation of rent for Ann Arbor, Michigan, and Davis, California.
- Construct and interpret 95% confidence intervals for the standard deviation of rent for both Ann Arbor, Michigan, and Davis, California.
- For each campus town, determine if the standard deviation of rent differs from \$200; use  $\alpha = 0.05$ .

## 11.2 INFERENCE CONCERNING THE RATIO OF TWO POPULATION VARIANCES

In this section, we turn our attention to comparing two population variances  $\sigma_1^2$  and  $\sigma_2^2$ . We may want to compare products on the basis of the relative variability of their weight, size, or volume. For example, a bottler may want to compare two production facilities based on the relative variability of the amount of beverage filled at each facility. Similarly, with variance used as a quantitative measure of risk in investments, an investor may want to



compare the relative risk of two investment strategies. Other examples for the relevance of comparing two population variances include comparing the consistency of athletes or teams, the relative variability of speeds on highways, and the relative variability of repair costs of different makes of automobiles.

We specify the parameter of interest as the ratio of the population variances  $\sigma_1^2/\sigma_2^2$  rather than their difference  $\sigma_1^2 - \sigma_2^2$ . Note that the condition  $\sigma_1^2 = \sigma_2^2$  is equivalent to  $\sigma_1^2 - \sigma_2^2 = 0$  as well as  $\sigma_1^2/\sigma_2^2 = 1$ . We use the ratio of the sample variances  $S_1^2/S_2^2$  as an estimator of  $\sigma_1^2/\sigma_2^2$ , where the sample variances are computed from independent random samples drawn from two normally distributed populations. In order to examine the techniques for statistical inference, we first need to analyze the sampling distribution of  $S_1^2/S_2^2$ .

## Sampling Distribution of $S_1^2/S_2^2$

### LO 11.4

Discuss features of the  $F$  distribution.

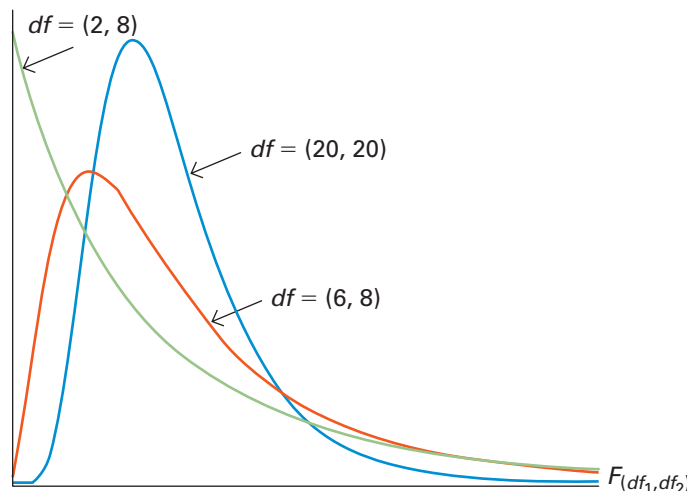
We use the sampling distribution of  $S_1^2/S_2^2$  to define a new distribution, called the  **$F$  distribution**.<sup>1</sup> Like the  $t_{df}$  and  $\chi_{df}^2$  distributions, the  $F$  distribution is characterized by a family of distributions; however, each distribution depends on *two* degrees of freedom: the numerator degrees of freedom  $df_1$  and the denominator degrees of freedom  $df_2$ . It is common to refer to it as the  $F_{(df_1, df_2)}$  distribution. As mentioned earlier, we will use  $F_{(df_1, df_2)}$  to represent both a random variable and its value.

In general, the  $F_{(df_1, df_2)}$  distribution is the probability distribution of the ratio of two independent chi-square variables, where each variable is divided by its own degrees of freedom; that is,  $F_{(df_1, df_2)} = \frac{\chi_{df_1}^2/df_1}{\chi_{df_2}^2/df_2}$ .

### THE SAMPLING DISTRIBUTION OF $S_1^2/S_2^2$ WHEN $\sigma_1^2 = \sigma_2^2$

If independent samples of size  $n_1$  and  $n_2$  are drawn from normal populations with equal variances, then the statistic  $F_{(df_1, df_2)} = S_1^2/S_2^2$  follows the  $F_{(df_1, df_2)}$  distribution with  $df_1 = n_1 - 1$  and  $df_2 = n_2 - 1$ .

Like the  $\chi_{df}^2$  distribution, the  $F_{(df_1, df_2)}$  distribution is positively skewed with values ranging from zero to infinity, but becomes increasingly symmetric as  $df_1$  and  $df_2$  increase. Figure 11.4 shows the  $F_{(df_1, df_2)}$  distribution with various degrees of freedom. Note that each  $F_{(df_1, df_2)}$  distribution is positively skewed; as  $df_1$  and  $df_2$  grow larger, the  $F_{(df_1, df_2)}$  distribution becomes less skewed and tends to the normal distribution. For instance,  $F_{(20, 20)}$  is relatively less skewed and more bell-shaped as compared to  $F_{(2, 8)}$  or  $F_{(6, 8)}$ .



**FIGURE 11.4**  
The  $F_{(df_1, df_2)}$  distribution  
with various degrees of  
freedom

<sup>1</sup>The  $F$  distribution is named in honor of Sir Ronald Fisher, who discovered the distribution in 1922.

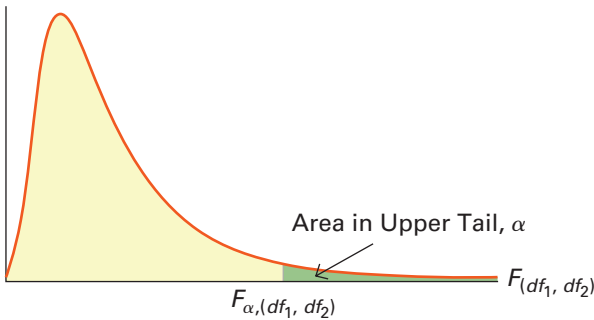
### SUMMARY OF THE $F_{(df_1, df_2)}$ DISTRIBUTION

- The  $F_{(df_1, df_2)}$  distribution is characterized by a family of distributions, where each distribution depends on two degrees of freedom,  $df_1$  and  $df_2$ .
- The values of the  $F_{(df_1, df_2)}$  distribution range from zero to infinity.
- The  $F_{(df_1, df_2)}$  distribution is positively skewed, where the extent of skewness depends on  $df_1$  and  $df_2$ . As  $df_1$  and  $df_2$  grow larger, the  $F_{(df_1, df_2)}$  distribution approaches the normal distribution.

### Locating $F_{(df_1, df_2)}$ Values and Probabilities

As with the  $\chi^2_{df}$  distribution, we use the notation  $F_{\alpha, (df_1, df_2)}$  to represent a value such that the area in the upper (right) tail of the distribution is  $\alpha$ . In other words,  $P(F_{(df_1, df_2)} \geq F_{\alpha, (df_1, df_2)}) = \alpha$ . Figure 11.5 illustrates this notation.

**FIGURE 11.5**  
Graphical depiction of  
 $P(F_{(df_1, df_2)} \geq F_{\alpha, (df_1, df_2)}) = \alpha$



A portion of the upper tail areas and the corresponding values for the  $F_{(df_1, df_2)}$  distribution are given in Table 11.3. Table 4 of Appendix A provides a more complete table.

**TABLE 11.3** Portion of the  $F$  Table

$df_2$	Area in Upper Tail, $\alpha$	$df_1$		
		6	7	8
6	0.10	3.05	3.01	2.98
	0.05	4.28	4.21	4.15
	0.025	5.82	5.70	5.60
	0.01	8.47	8.26	8.10
7	0.10	2.83	2.78	2.75
	0.05	3.87	3.79	3.73
	0.025	5.12	4.99	4.90
	0.01	7.19	6.99	6.84
8	0.10	2.67	2.62	2.59
	0.05	<b>3.58</b>	3.50	3.44
	0.025	4.65	4.53	4.43
	0.01	<b>6.37</b>	6.18	6.03

Consider the degrees of freedom given by  $df_1 = 6$  and  $df_2 = 8$ . With  $df_1 = 6$  (read from the top row) and  $df_2 = 8$  (read from the first column), we can easily determine the area in the upper tail as  $P(F_{(6,8)} \geq 3.58) = 0.05$  and  $P(F_{(6,8)} \geq 6.37) = 0.01$ . The  $F$  table is not very comprehensive and lists probabilities corresponding to a limited number of values in the upper tail of the distribution. For instance, the exact probability  $P(F_{(6,8)} \geq 3.92)$  cannot be

determined from the table and we have to rely on approximate values. All we can say is the area to the right of 3.92 is between 0.025 and 0.05. Shortly, we will use Excel to find exact probabilities.

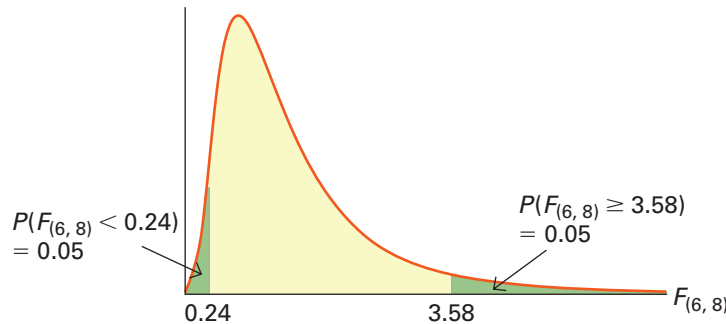
Sometimes we need to derive values such that the area to the left of a given value is equal to  $\alpha$ . Given that the area under any distribution equals one, the area to the right of the given value must equal  $1 - \alpha$ . As in the case of the  $\chi^2_{df}$  distribution, we let  $F_{1-\alpha, (df_1, df_2)}$  denote the value such that the area to its right equals  $1 - \alpha$  and thus the area to its left equals  $\alpha$ . It is convenient, however, to find  $F_{1-\alpha, (df_1, df_2)}$  using a simple rule that  $F_{1-\alpha, (df_1, df_2)} = \frac{1}{F_{\alpha, (df_2, df_1)}}$ . Note that the rule reverses the order of the numerator and the denominator degrees of freedom.

#### LOCATING $F_{(df_1, df_2)}$ VALUES IN THE LOWER TAIL

For a  $F_{(df_1, df_2)}$  distributed random variable,  $F_{1-\alpha, (df_1, df_2)}$  represents a value such that the area to the left of  $F_{1-\alpha, (df_1, df_2)}$  is  $\alpha$ —that is,  $P(F_{(df_1, df_2)} < F_{1-\alpha, (df_1, df_2)}) = \alpha$ . It is convenient to reverse the order of degrees of freedom and determine  $F_{1-\alpha, (df_1, df_2)}$  as

$$F_{1-\alpha, (df_1, df_2)} = \frac{1}{F_{\alpha, (df_2, df_1)}}.$$

Suppose we need to find  $F_{1-\alpha, (df_1, df_2)}$  where  $\alpha = 0.05$ ,  $df_1 = 6$ , and  $df_2 = 8$ . We find  $F_{0.95, (6, 8)} = \frac{1}{F_{0.05, (8, 6)}} = \frac{1}{4.15} = 0.24$ . In other words, the lower (left) tail area is  $P(F_{(6, 8)} < 0.24) = 0.05$ . Figure 11.6 graphically depicts  $P(F_{(6, 8)} \geq 3.58) = 0.05$  and  $P(F_{(6, 8)} < 0.24) = 0.05$ .



**FIGURE 11.6**  
Graph of the probability  
 $\alpha = 0.05$  on both sides  
of  $F_{(6, 8)}$

#### EXAMPLE 11.5

Find the value  $x$  for which:

- $P(F_{(7, 10)} \geq x) = 0.025$
- $P(F_{(7, 10)} < x) = 0.05$

#### SOLUTION:

- We find the value  $x$  such that the area in the upper tail of the distribution equals 0.025. Referencing Table 4 in Appendix A, we follow the column corresponding to  $df_1 = 7$  until it intersects with the row corresponding to  $df_2 = 10$  and  $\alpha = 0.025$ ; we find the value 3.95. Therefore,  $P(F_{(7, 10)} \geq 3.95) = 0.025$ .
- We find the value  $x$  such that the area in the lower tail of the distribution equals 0.05, or equivalently, the area in the upper tail of the distribution equals 0.95. We have  $F_{0.95, (7, 10)} = \frac{1}{F_{0.05, (10, 7)}} = \frac{1}{3.64} = 0.27$ . In other words,  $P(F_{(7, 10)} < 0.27) = 0.05$ .

**LO 11.5**

Construct a confidence interval for the ratio of two population variances.

## Confidence Interval for the Ratio of Two Population Variances

The formula for a confidence interval for the ratio of the population variances  $\sigma_1^2/\sigma_2^2$  is derived in a manner analogous to previous confidence intervals. Here, we simply show the end result.

### CONFIDENCE INTERVAL FOR $\sigma_1^2/\sigma_2^2$

A  $100(1 - \alpha)\%$  confidence interval for the ratio of the population variances  $\sigma_1^2/\sigma_2^2$  is computed as

$$\left[ \left( \frac{s_1^2}{s_2^2} \right) \frac{1}{F_{\alpha/2, (df_1, df_2)}}, \left( \frac{s_1^2}{s_2^2} \right) F_{\alpha/2, (df_2, df_1)} \right],$$

where for samples of size  $n_1$  and  $n_2$ ,  $df_1 = n_1 - 1$  and  $df_2 = n_2 - 1$ .

This formula is valid if the sample variances are computed from independently drawn samples from two normally distributed populations.

### EXAMPLE 11.6

Students of two sections of a statistics course took a common final examination. A professor examines the variability in scores between the two sections. Random samples of  $n_1 = 11$  and  $n_2 = 16$  yield sample variances of  $s_1^2 = 182.25$  and  $s_2^2 = 457.96$ . Construct the 95% confidence interval for the ratio of the population variances. Assume that the samples are independently drawn from two normally distributed populations.

**SOLUTION:** In order to construct the 95% confidence interval for the ratio of the population variances, we determine  $\left[ \left( \frac{s_1^2}{s_2^2} \right) \frac{1}{F_{\alpha/2, (df_1, df_2)}}, \left( \frac{s_1^2}{s_2^2} \right) F_{\alpha/2, (df_2, df_1)} \right]$ . We find the degrees of freedom as  $df_1 = n_1 - 1 = 11 - 1 = 10$  and  $df_2 = n_2 - 1 = 16 - 1 = 15$ . From the  $F$  table and given  $\alpha = .05$ , we find

$$F_{\alpha/2, (df_1, df_2)} = F_{0.025, (10, 15)} = 3.06 \quad \text{and} \quad F_{\alpha/2, (df_2, df_1)} = F_{0.025, (15, 10)} = 3.52.$$

The confidence interval is

$$\left[ \left( \frac{182.25}{457.96} \right) \frac{1}{3.06}, \left( \frac{182.25}{457.96} \right) 3.52 \right] = [0.13, 1.40].$$

Therefore, the 95% confidence interval for the ratio of the population variances ranges from 0.13 to 1.40. In other words, the variance of scores in the first section is between 13% and 140% of the variance of scores in the second section.

As we have done in earlier chapters, we will be able to use this confidence interval to conduct a two-tailed hypothesis test.

## Hypothesis Test for the Ratio of Two Population Variances

**LO 11.6**

Conduct a hypothesis test for the ratio of two population variances.

When comparing two population parameters  $\sigma_1^2$  and  $\sigma_2^2$ , the competing hypotheses will take one of the following forms:

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \sigma_1^2/\sigma_2^2 = 1$	$H_0: \sigma_1^2/\sigma_2^2 \leq 1$	$H_0: \sigma_1^2/\sigma_2^2 \geq 1$
$H_A: \sigma_1^2/\sigma_2^2 \neq 1$	$H_A: \sigma_1^2/\sigma_2^2 > 1$	$H_A: \sigma_1^2/\sigma_2^2 < 1$

A two-tailed test determines whether the two population variances are different. As noted earlier, the condition  $\sigma_1^2 = \sigma_2^2$  is equivalent to  $\sigma_1^2/\sigma_2^2 = 1$ . A right-tailed test examines whether  $\sigma_1^2$  is greater than  $\sigma_2^2$ , whereas a left-tailed test examines whether  $\sigma_1^2$  is less than  $\sigma_2^2$ .

### EXAMPLE 11.7

Let's revisit Example 11.6.

- Specify the competing hypotheses in order to determine whether or not the variances in the two statistics sections differ.
- Using the 95% confidence interval, what is the conclusion to the test?

#### SOLUTION:

- Since we want to determine if the variances differ between the two sections, we formulate a two-tailed hypothesis test as

$$H_0: \sigma_1^2/\sigma_2^2 = 1$$

$$H_A: \sigma_1^2/\sigma_2^2 \neq 1$$

- We calculated the 95% confidence interval for the ratio of the two variances that ranged from 0.13 to 1.40. We note that this interval contains the value one; thus, we do not reject  $H_0$ . The sample data do not suggest that the variances between the two statistics sections differ at the 5% significance level.

Now we use the four-step procedure to implement one- or two-tailed hypothesis tests. We use the ratio of the values of the sample variances  $s_1^2/s_2^2$  to conduct hypothesis tests regarding the ratio of the population variances  $\sigma_1^2/\sigma_2^2$ . The resulting  $F_{(df_1, df_2)}$  test is valid if the sample variances are computed from independently drawn samples from normally distributed populations.

#### TEST STATISTIC FOR $\sigma_1^2/\sigma_2^2$

The value of the **test statistic** for the hypothesis test for the **ratio of two population variances**  $\sigma_1^2/\sigma_2^2$  is computed as

$$F_{(df_1, df_2)} = s_1^2/s_2^2,$$

where for samples of size  $n_1$  and  $n_2$ ,  $df_1 = n_1 - 1$  and  $df_2 = n_2 - 1$ .

We should point out that a left-tailed test can easily be converted into a right-tailed test by interchanging the variances of the two populations. For instance, we can convert  $H_0: \sigma_1^2/\sigma_2^2 \geq 1$  versus  $H_A: \sigma_1^2/\sigma_2^2 < 1$  into  $H_0: \sigma_2^2/\sigma_1^2 \leq 1$  versus  $H_A: \sigma_2^2/\sigma_1^2 > 1$ .

#### PLACING THE LARGER SAMPLE VARIANCE IN THE NUMERATOR

It is preferable to place the larger sample variance in the numerator of the  $F_{(df_1, df_2)}$  statistic. The resulting value allows us to focus only on the upper tail of the distribution.

In other words, we define the hypotheses such that the resulting test statistic is computed as  $s_1^2/s_2^2$  when  $s_1^2 > s_2^2$  and as  $s_2^2/s_1^2$  when  $s_2^2 > s_1^2$ ; the degrees of freedom are adjusted accordingly. This saves us the additional work required to find the area in the lower tail of the  $F_{(df_1, df_2)}$  distribution.

### EXAMPLE 11.8

Let's again visit the case introduced at the beginning of this chapter. Investment counselor Rebecca Johnson wonders if the Metals fund is significantly riskier than the Income fund. We assume that returns are normally distributed to implement the test at the 5% significance level using the critical value approach. For reference, we repeat the sample descriptive measures for the two funds:

Metals fund:  $\bar{x}_1 = 24.65\%$ ,  $s_1 = 37.13\%$ , and  $n_1 = 10$

Income fund:  $\bar{x}_2 = 8.51\%$ ,  $s_2 = 11.07\%$ , and  $n_2 = 10$

**SOLUTION:** We define the population variance as the measure of risk and let  $\sigma_1^2$  and  $\sigma_2^2$  denote the population variances of the Metals and the Income funds, respectively. Since we wish to determine whether the variance of the Metals fund is greater than that of the Income fund, we specify the competing hypotheses as

$$H_0: \sigma_1^2 / \sigma_2^2 \leq 1$$

$$H_A: \sigma_1^2 / \sigma_2^2 > 1$$

Note that this specification is appropriate since  $s_1 = 37.13\%$  is greater than  $s_2 = 11.07\%$ . Had  $s_2$  been greater, we would have specified the hypotheses in terms of  $\sigma_2^2 / \sigma_1^2$  instead of  $\sigma_1^2 / \sigma_2^2$ .

The  $F$  test is valid because the underlying populations are assumed to be normally distributed. Using  $\alpha = 0.05$ , we find the critical value of this right-tailed test as  $F_{\alpha, (df_1, df_2)} = F_{0.05, (9, 9)} = 3.18$ . The decision rule is to reject the null hypothesis if  $F_{(df_1, df_2)} = F_{(9, 9)}$  exceeds 3.18. (For a two-tailed test, the critical value would be  $F_{\alpha/2, (df_1, df_2)} = F_{0.025, (9, 9)} = 4.03$ .)

We compute the value of the test statistic as

$$F_{(df_1, df_2)} = F_{(9, 9)} = \frac{s_1^2}{s_2^2} = \frac{(37.13)^2}{(11.07)^2} = 11.25.$$

We reject the null hypothesis because the value of the test statistic falls in the rejection region ( $F_{(9, 9)} = 11.25$  exceeds the critical value  $F_{0.05, (9, 9)} = 3.18$ ). At the 5% significance level, the variance of the Metals fund is significantly greater than the variance of the Income fund. Therefore, we can conclude that the Metals fund is riskier than the Income fund.

## Using Excel to Calculate the $p$ -Value for the $F_{(df_1, df_2)}$ Test Statistic

### Excel's F.DIST.RT Function

In Example 11.8, we calculated the value of the test statistic as  $F_{(9, 9)} = 11.25$ . Therefore, the  $p$ -value of this right-tailed test is given by  $P(F_{(9, 9)} \geq 11.25)$ . Since the value 11.25 is not listed in the  $F$  table, we can only state that this probability is less than 0.01.

As in the case of the  $t_{df}$  and the  $\chi_{df}^2$  distributions, we can easily get exact probabilities for the  $F_{(df_1, df_2)}$  distribution using Excel's F.DIST.RT function. In order to find the area in the upper tail of the  $F_{(df_1, df_2)}$  distribution, we find an empty cell and input “=F.DIST.RT( $F_{(df_1, df_2)}$ ,  $df_1$ ,  $df_2$ ),” where  $F_{(df_1, df_2)}$  is the value of the test statistic,  $df_1$  is the numerator degrees of freedom, and  $df_2$  is the denominator degrees of freedom. For Example 11.8, we input “=F.DIST.RT(11.25, 9, 9),” and Excel returns 0.0007. (For a two-tailed test, we would multiply this probability by two.) Again, given a 5% significance level, we reject the null hypothesis and conclude that the variance of the Metals fund is greater than the variance of the Income fund.



## Excel's F.TEST Function

If we have access to the raw data, rather than summary statistics, then Excel's F.TEST function returns the  $p$ -value for a two-tailed test. Table 11.4 shows the annual total return data for the Metals and Income funds for the years 2000–2009. We open **Fund\_Returns**, find an empty cell and input “=F.Test(Array1, Array2),” where we select the data for the Metals fund for Array 1, and then select the data for Income fund for Array 2.

Excel returns 0.0013. This is the  $p$ -value for a two-tailed test. Since we conducted a one-tailed test, we divide this value by two ( $0.0013/2 = 0.0007$ ), arriving at the  $p$ -value that we found earlier.

**TABLE 11.4** Annual Total Returns (in percent) for the Metals and Income Funds

Year	Metals	Income
2000	−7.34	4.07
2001	18.33	6.52
2002	33.35	9.38
2003	59.45	18.62
2004	8.09	9.44
2005	43.79	3.12
2006	34.30	8.15
2007	36.13	5.44
2008	−56.02	−11.37
2009	76.46	31.77

SOURCE: finance.yahoo.com

**FILE**  
**Fund\_Returns**

## SYNOPSIS OF INTRODUCTORY CASE

Vanguard's Precious Metals and Mining fund (Metals) and Fidelity's Strategic Income fund (Income) were two top-performing mutual funds for the years 2000 through 2009. At first glance, the Metals fund seems attractive since its average return is greater than the average return for the Income fund ( $24.65\% > 8.51\%$ ); however, the average return does not incorporate the risk of investing. Variance and standard deviation tend to be the most common measures of risk with financial data. An analysis of the variance and standard deviation of the returns for these funds provides additional relevant information. For the Metals fund, the 95% confidence interval for the population standard deviation of the return is between 25.54% and 67.79%, while the corresponding interval for the Income fund is between 7.61% and 20.21%. Since the intervals do not overlap, we may infer that the risk for the two funds is different with 95% confidence. Formal testing reveals that the risk of the Metals fund is greater than the risk of the Income fund at the 5% significance level.

Two more hypothesis tests are also conducted. The first test examines whether the standard deviation of the Metals fund exceeds 25%. At the 5% significance level, the sample data suggest that the standard deviation is significantly greater than 25%. The second test investigates whether or not the standard deviation of the Income fund differs from 8.5%, the risk inherent in similar funds. At the 5% significance level, the results suggest that the standard deviation is not significantly different from 8.5%. These results stress the importance of analyzing the variance and the standard deviation of the returns of an asset—an examination of only the average return of the two funds would be incomplete.



## EXERCISES 11.2

### Concepts

23. Find the value  $x$  for which:
- $P(F_{(4,8)} \geq x) = 0.025$
  - $P(F_{(4,8)} \geq x) = 0.05$
  - $P(F_{(4,8)} < x) = 0.025$
  - $P(F_{(4,8)} < x) = 0.05$
24. Use the  $F$  table to approximate the following probabilities.
- $P(F_{(10,8)} \geq 3.35)$
  - $P(F_{(10,8)} < 0.42)$
  - $P(F_{(10,8)} \geq 4.30)$
  - $P(F_{(10,8)} < 0.26)$
25. Construct the 90% interval estimate for the ratio of the population variances using the following results from two independently drawn samples from normally distributed populations.
- Sample 1:  $\bar{x}_1 = 157$ ,  $s_1^2 = 23.2$ , and  $n_1 = 9$   
 Sample 2:  $\bar{x}_2 = 148$ ,  $s_2^2 = 19.9$ , and  $n_2 = 8$
26. Consider the following measures based on independently drawn samples from normally distributed populations:
- Sample 1:  $s_1^2 = 220$ , and  $n_1 = 20$   
 Sample 2:  $s_2^2 = 196$ , and  $n_2 = 15$
- Construct the 95% interval estimate for the ratio of the population variances.
  - Using the confidence interval from part (a), test if the ratio of the population variances differs from one at the 5% significance level. Explain.
27. Consider the following competing hypotheses and relevant summary statistics:
- $$H_0: \sigma_1^2/\sigma_2^2 = 1$$
- $$H_A: \sigma_1^2/\sigma_2^2 \neq 1$$
- Sample 1:  $\bar{x}_1 = 48.5$ ,  $s_1^2 = 18.7$ , and  $n_1 = 10$   
 Sample 2:  $\bar{x}_2 = 50.2$ ,  $s_2^2 = 12.9$ , and  $n_2 = 8$
- Assume that the two populations are normally distributed.
- Using the  $p$ -value approach, conduct this hypothesis test at the 5% significance level.
  - Confirm your conclusions by using the critical value approach.
28. Consider the following competing hypotheses and relevant summary statistics:
- $$H_0: \sigma_1^2/\sigma_2^2 \leq 1$$
- $$H_A: \sigma_1^2/\sigma_2^2 > 1$$
- Sample 1:  $s_1^2 = 935$  and  $n_1 = 14$   
 Sample 2:  $s_2^2 = 812$  and  $n_2 = 11$
- Use the critical value approach to conduct this hypothesis test at the 5% significance level. State your assumptions.

29. Consider the following competing hypotheses and relevant summary statistics:

$$H_0: \sigma_1^2/\sigma_2^2 \geq 1$$

$$H_A: \sigma_1^2/\sigma_2^2 < 1$$

Sample 1:  $s_1^2 = 1,315$  and  $n_1 = 17$

Sample 2:  $s_2^2 = 1,523$  and  $n_2 = 19$

Conduct this hypothesis test at the 5% significance level. State your assumptions. (*Hint:* You may want to first convert the above left-tailed test into a right-tailed test by switching the two variances.)

### Applications

30. A firm has just developed a new cost-reducing technology for producing a certain replacement part for automobiles. Since a replacement part must be produced within close specifications in order for it to be acceptable to customers, the new technology's specifications must not deviate drastically from the older version. Suppose the sample variance for 15 parts produced using the older version is  $s_1^2 = 0.35$ , while the sample variance for 15 parts produced using the new technology is  $s_2^2 = 0.48$ . Assume that the two samples are drawn independently from normally distributed populations.
- Develop the hypotheses to test whether the population variances differ.
  - Calculate the value of the test statistic.
  - Using the critical value approach, determine the decision rule at the 5% significance level.
  - Can you conclude that the variances are different? Given that all other criteria are satisfied, should the company adopt the new technology?
31. Two basketball players on a school team are working hard on consistency of their performance. In particular, they are hoping to bring down the variance of their scores. The coach believes that the players are not equally consistent in their games. Over a 10-game period, the scores of these two players are shown below. Assume that the two samples are drawn independently from normally distributed populations.
- |          |    |    |    |    |    |    |    |    |    |    |
|----------|----|----|----|----|----|----|----|----|----|----|
| Player 1 | 13 | 15 | 12 | 18 | 14 | 15 | 11 | 13 | 11 | 16 |
| Player 2 | 11 | 21 | 18 | 9  | 20 | 11 | 13 | 11 | 19 | 8  |
- Develop the hypotheses to test whether the players differ in consistency.
  - Use the critical value approach to test the coach's claim at  $\alpha = 0.05$ .
32. The following table shows the annual returns (in percent) for Fidelity's Electronic and Utilities funds.

Year	Electronic	Utilities
2010	17	11
2011	-8	13
2012	4	7
2013	39	21
2014	38	22

SOURCE: www.finance.yahoo.com; data retrieved April 3, 2015.

Using the critical value approach, test if the population variances differ at the 5% significance level. State your assumptions.

33. Nike's total revenues (in millions of \$) for the Asian and Latin American regions for the years 2005 through 2009 are as follows:

	2005	2006	2007	2008	2009
Asia	1,897	2,054	2,296	2,888	3,322
Latin America	696	905	967	1,165	1,285

SOURCE: Nike Online Annual Reports.

- Specify the competing hypothesis in order to test whether the variance in revenues is greater in Asia than in Latin America.
  - Calculate the value of the test statistic. Assume that the revenue distributions are normally distributed.
  - Use Excel's F.DIST.RT function to calculate the  $p$ -value.
  - At  $\alpha = 0.01$ , what is your conclusion?
34. The quality manager at a battery manufacturing company wants to determine if lithium-ion batteries have less variability in discharge time than nickel-cadmium batteries. Using products with similar power draws, he has measured the time until discharge (in hours) for random samples of 16 lithium-ion batteries and 26 nickel-cadmium batteries. The sample variance of the discharge times is 0.44 hours<sup>2</sup> for the lithium-ion batteries and 0.89 hours<sup>2</sup> for the nickel-cadmium batteries.
- State the appropriate hypotheses to test whether the variance in discharge time for the lithium-ion batteries is less than the nickel-cadmium batteries.
  - Compute the value of the test statistic. What assumption did you make?
  - Use Excel's F. DIST.RT function to calculate the  $p$ -value.
  - Make a conclusion at the 5% significance level.
  - Would your conclusion change at the 10% significance level?
35. **FILE Purchase Amounts.** A marketing analyst is studying the variability in customer purchase amounts between shopping mall stores and "big box" discount stores. She suspects the variability is different between those stores due to the nature of the customers

involved. To investigate this issue in detail, she compiled two random samples each consisting of 26 purchase amounts at shopping mall stores and discount stores.

- State the appropriate hypotheses to test whether the variance of the purchase amounts differs between the two types of stores.
  - Construct the 95% confidence interval for the ratio of the population variances. Assume the purchase amount distributions are normally distributed.
  - Use the computed confidence interval to test whether the variance of the purchase amounts differs between the two stores at the 5% significance level.
  - Confirm your conclusion using Excel's F.TEST function to calculate the  $p$ -value.
36. **FILE Monthly Stock Prices.** A portion of the monthly stock prices (rounded to the nearest dollar) for Starbucks Corp. and Panera Bread Co. from 2010 to 2013 are reported in the following table.

Date	Starbucks Corp.	Panera Bread Co.
January 2010	\$22	\$71
February 2010	23	73
:	:	:
December 2013	78	177

SOURCE: finance.yahoo.com.

- State the null and the alternative hypotheses in order to determine if the variance of price differs for the two firms.
  - What assumption regarding the population is necessary to implement this step?
  - Use Excel's F.TEST function to calculate the  $p$ -value.
  - At  $\alpha = 0.05$ , what is your conclusion?
37. **FILE Packaging.** A variety of packaging solutions exist for products that must be kept within a specific temperature range. Cold chain distribution is particularly useful in the food and pharmaceutical industries. A packaging company is trying out a new packaging material that might reduce the variation of temperatures in the box. It is believed that the temperature in the box follows a normal distribution with both packaging materials. Inspectors randomly select 16 boxes of new and old packages, 24 hours after they are sealed for shipment, and report the temperatures in degrees Celsius. A portion of the data is shown in the accompanying table. Assume that the two samples are drawn independently from normally distributed populations.

New Package	Old Package
3.98	5.79
4.99	6.42
⋮	⋮
4.95	5.95

- a. State the appropriate hypotheses to test whether the new packaging material reduces the variation of temperatures in the box.

- b. Use Excel's F.TEST function to calculate the  $p$ -value.  
c. Make a conclusion at the 5% significance level.

38. **FILE Rentals.** The data accompanying this exercise include monthly rents for a two-bedroom apartment in two campus towns: Ann Arbor, Michigan, and Davis, California. Davis, California is known to have higher rents than Ann Arbor, Michigan; however, it is not clear if it also has higher variability in rents. At a 5% significance level, determine if the variance of rent in Davis, California is more than that of Ann Arbor, Michigan. State your assumptions clearly.

## WRITING WITH STATISTICS



Many environmental groups and politicians are suggesting a return to the federal 55-mile-per-hour speed limit on America's highways. They argue that a lower national speed limit will improve traffic safety, save fuel, and reduce greenhouse emissions. Elizabeth Connolly believes that more focus should be put on the variability of speed limits as opposed to average speed limits. She points to recent research that suggests that increases in speed variability decrease overall safety. Specifically, Elizabeth feels that traffic accidents are more likely to occur when the standard deviation of speeds exceeds 5 mph. She records the speeds of 40 cars from a highway with a speed limit of 55 mph (Highway 1) and the speeds

of 40 cars from a highway with a speed limit of 65 mph (Highway 2). A portion of the data is shown in Table 11.5.

**TABLE 11.5** Speeds of Cars from Highway 1 and Highway 2

Highway 1 (55-mph limit)	Highway 2 (65-mph limit)
60	70
55	65
⋮	⋮
52	65

**FILE**

### Highway\_Speeds

Elizabeth would like to use the above sample information to:

1. Determine, at the 5% significance level, whether the standard deviation on the 55-mph highway exceeds 5 mph.
2. Determine, at the 5% significance level, whether the variability on the 55-mph highway is more than the variability on the 65-mph highway.

## Sample Report—Traffic Safety and the Variation in Speed

Increasing greenhouse emissions are prompting conservationists to lobby for a return to the federal 55-mile-per-hour (mph) speed limit on America's highways. In addition, advocates point to potential money and fuel savings, noting that fuel efficiency worsens at speeds above 60 mph. It is not clear, however, if the return to 55 mph will increase traffic safety. Many believe that traffic safety is based on the variability of the speed rather than the average speed that people are driving—the more variation in speed, the more dangerous the roads.

In this report, the variability of speeds on two highways is compared. The sample consists of the speeds of 40 cars recorded on a highway with a 55-mph speed limit

(Highway 1) and the speeds of 40 cars recorded on a highway with a 65-mph speed limit (Highway 2). Table 11.A shows the most relevant descriptive measures for the analysis.

**TABLE 11.A** Summary Measures for Highway 1 and Highway 2

	Highway 1 (55-mph speed limit)	Highway 2 (65-mph speed limit)
Mean	56.60	66.00
Standard deviation	6.98	3.00
Number of cars	40	40

While it is true that cars travel at a slower speed, on average, on Highway 1 (56.60 mph < 66.00 mph), the variability of speeds is greater on Highway 1 as measured by the standard deviation (6.98 mph > 3.00 mph).

Two hypothesis tests are conducted. The first test examines whether or not the standard deviation on Highway 1 is greater than 5 mph at the 5% significance level, or alternatively  $\sigma^2 > 5^2$ . The second test analyzes whether the standard deviation on Highway 1 is significantly greater than the standard deviation on Highway 2, or alternatively  $\sigma_1^2/\sigma_2^2 > 1$ . The results of the tests are summarized in Table 11.B.

**TABLE 11.B** Competing Hypotheses, Test Statistics, and  $p$ -values

Hypotheses	Test Statistic	$p$ -value
$H_0: \sigma^2 \leq 5^2$ $H_A: \sigma^2 > 5^2$	$\chi^2_{39} = \frac{(n-1)s^2}{\sigma^2} = \frac{(40-1)(6.98)^2}{(5)^2} = 76.00$	$\approx 0.00$
$H_0: \sigma_1^2/\sigma_2^2 \leq 1$ $H_A: \sigma_1^2/\sigma_2^2 > 1$	$F_{(39,39)} = \frac{s_1^2}{s_2^2} = \frac{(6.98)^2}{(3.00)^2} = 5.41$	$\approx 0.00$

When testing whether or not the standard deviation is greater than 5 mph on Highway 1, a test statistic of 76.00 is obtained. Given its  $p$ -value that is approximately equal to zero, the null hypothesis regarding the population variance is rejected at any reasonable level of significance. In other words, the sample data suggest that the standard deviation is significantly greater than 5 mph on Highway 1. With a test statistic of 5.41 and a corresponding  $p$ -value that is approximately equal to zero, the second hypothesis test reveals that the variance for Highway 1 is significantly greater than the variance for Highway 2.

American drivers love to drive fast, which explains why safety advocates and conservationists are losing the long-running debate over lowering highway speed limits. While a 55-mph speed limit will save fuel and reduce greenhouse emissions, it is still an open question as to whether it will also enhance safety. If traffic safety is based on the variability of the speeds that people are driving rather than the average speed, then the data suggest that a return to a federal 55-mph speed limit may not necessarily enhance safety.

## CONCEPTUAL REVIEW

### LO 11.1 Discuss features of the $\chi^2$ distribution.

The  $\chi^2$  **distribution** is characterized by a family of distributions, where each distribution depends on its particular degrees of freedom  $df$ . It is common, therefore, to refer to it as the  $\chi^2_{df}$  distribution. It is positively skewed with values ranging from zero to infinity. As the  $df$  grow larger, the  $\chi^2_{df}$  distribution tends to the normal distribution.



**LO 11.2 Construct a confidence interval for the population variance.**

The sample variance  $S^2$  is a **point estimator** of the population variance  $\sigma^2$ . Statistical inferences regarding  $\sigma^2$  are based on the  $\chi^2_{df}$  distribution. A  $100(1 - \alpha)\%$  **confidence interval** for  $\sigma^2$  is computed as  $\left[ \frac{(n-1)s^2}{\chi^2_{\alpha/2, df}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, df}} \right]$ . This formula is based on the requirement that  $s^2$  is computed using a random sample drawn from a normally distributed population.

**LO 11.3 Conduct a hypothesis test for the population variance.**

The value of the **test statistic** for the **hypothesis test of  $\sigma^2$**  is computed as  $\chi^2_{df} = \frac{(n-1)s^2}{\sigma_0^2}$ , where  $\sigma_0^2$  is the hypothesized value of the population variance. We apply the four-step procedure to conduct hypothesis tests using the  $p$ -value approach or the critical value approach.

**LO 11.4 Discuss features of the  $F$  distribution.**

The  **$F$  distribution** is also characterized by a family of distributions; however, each distribution depends on *two* degrees of freedom: the numerator degrees of freedom  $df_1$  and the denominator degrees of freedom  $df_2$ . It is common to refer to it as the  $F_{(df_1, df_2)}$  distribution. The  $F_{(df_1, df_2)}$  distribution is positively skewed with values ranging from zero to infinity, but becomes increasingly symmetric as  $df_1$  and  $df_2$  increase.

**LO 11.5 Construct a confidence interval for the ratio of two population variances.**

The ratio of the sample variances  $S_1^2/S_2^2$  is a **point estimator** for the ratio of the population variances  $\sigma_1^2/\sigma_2^2$ . Statistical inferences regarding  $\sigma_1^2/\sigma_2^2$  are based on the  $F_{(df_1, df_2)}$  distribution.

A  $100(1 - \alpha)\%$  **confidence interval** for  $\sigma_1^2/\sigma_2^2$  is computed as  $\left[ \left( \frac{s_1^2}{s_2^2} \right) \frac{1}{F_{\alpha/2, (df_1, df_2)}}, \left( \frac{s_1^2}{s_2^2} \right) F_{\alpha/2, (df_2, df_1)} \right]$ .

This formula is based on the assumption that  $s_1^2$  and  $s_2^2$  are computed using independently drawn samples from two normally distributed populations.

**LO 11.6 Conduct a hypothesis test for the ratio of two population variances.**

The value of the **test statistic** for the **hypothesis test of  $\sigma_1^2/\sigma_2^2$**  is computed as  $F_{(df_1, df_2)} = s_1^2/s_2^2$ , with  $df_1 = n_1 - 1$  and  $df_2 = n_2 - 1$ . It is assumed that  $s_1^2$  and  $s_2^2$  are based on independently drawn samples from two normally distributed populations with  $\sigma_1^2/\sigma_2^2 = 1$ . We apply the four-step procedure to conduct hypothesis tests using the  $p$ -value approach or the critical value approach.

It is preferable to define the hypotheses such that the resulting test statistic is computed as  $F_{(df_1, df_2)} = s_1^2/s_2^2$  when  $s_1^2 > s_2^2$  and as  $F_{(df_2, df_1)} = s_2^2/s_1^2$  when  $s_2^2 > s_1^2$ . This saves us the additional work required to calculate the probability in the lower tail of the  $F_{(df_1, df_2)}$  distribution.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

39. A replacement part for a machine must be produced within close specifications in order for it to be acceptable to customers. A production process is considered to be working properly as long as the variance in the lengths of the parts does not exceed 0.05 squared-units. Suppose the sample variance computed from 30 parts turns out to be  $s^2 = 0.07$ . Use this sample evidence to test if the production specification is not being met at a 5% level of significance.
40. A consumer advocacy group is concerned about the variability in the cost of a generic drug. There is cause for concern if the variance of the cost exceeds 5 (\$)<sup>2</sup>. The group surveys seven local pharmacies and obtains the following prices (in \$) for a particular generic drug: 32, 36, 38, 32, 40, 31, 34.



- a. Use the  $p$ -value approach to test if there is a cause for concern for the consumer group at a 1% significance level.
  - b. What assumption regarding the generic drug prices was made in this analysis?
41. A financial analyst maintains that the risk, measured by the variance, of investing in emerging markets is more than  $280(\%)^2$ . Data on 20 stocks from emerging markets revealed the following sample results:  $\bar{x} = 12.1(\%)$  and  $s^2 = 361(\%)^2$ . Assume that the returns are normally distributed.
- a. Specify the competing hypotheses to test the analyst's claim.
  - b. At  $\alpha = 0.01$ , specify the critical value(s).
  - c. What is the value of the test statistic?
  - d. Is the financial analyst's claim supported by the data?
42. Fizzco, a beverage manufacturing company, is interested in determining whether the standard deviation of their dispensing process has changed from a required level of 5 milliliters. They have taken a random sample of 12 bottles and have measured the amount of beverage dispensed into each bottle (in milliliters) as shown below:

352	345	347	357	341	356
349	343	346	351	348	361

- a. Calculate the point estimate of  $\sigma$ .
  - b. Construct the 95% confidence interval for the population standard deviation. Assume the amount of beverage dispensed follows a normal distribution.
  - c. At the 5% significance level, can we conclude that the standard deviation of the amount of beverage dispensed differs from the required level of 5 milliliters?
43. **FILE Checkout Arrivals.** For staffing purposes, a retail store manager would like to standardize the number of checkout lanes to keep open on a particular shift. She believes that if the standard deviation of the hourly customer arrival rates is 8 customers or less, then a fixed number of checkout lanes can be staffed without excessive customer waiting time or excessive clerk idle time. However, before determining how many checkout lanes (and thus clerks) to use, she must verify that the standard deviation of the arrival rates does not exceed 8. Accordingly, a sample of 25 hourly customer arrival rates was compiled for that shift over the past week.
- a. State the hypotheses to test whether the standard deviation of the customer arrival rates exceeds 8.

- b. Calculate the value of the test statistic. Assume that customer arrival rates are normally distributed.
  - c. Use Excel's function (CHISQ.DIST.RT or CHISQ.DIST) to compute the  $p$ -value.
  - d. At  $\alpha = 0.05$ , what is your conclusion? Would your conclusion change at the 1% significance level?
44. **FILE Automotive.** The following table presents a portion of the annual returns for Fidelity's Select Automotive Fund (in percent). This mutual fund invests primarily in companies engaged in the manufacturing, the marketing, or the sales of automobiles, trucks, specialty vehicles, parts, tires, and related services.

Year	Fidelity Select Automotive Fund
1987	6.54
1988	20.06
⋮	⋮
2013	46.67

SOURCE: <http://biz.yahoo.com>.

- a. State the null and the alternative hypotheses in order to test whether the standard deviation is greater than 35%.
  - b. What assumption regarding the population is necessary to implement this step?
  - c. Calculate the value of the test statistic.
  - d. Use Excel's function (CHISQ.DIST.RT or CHISQ.DIST) to calculate the  $p$ -value.
  - e. At  $\alpha = 0.05$ , what is your conclusion?
45. John Daum and Chris Yin are star swimmers at a local college. They are preparing to compete at the NCAA Division II national championship meet, where they both have a good shot at earning a medal in the men's 100-meter freestyle event. The coach feels that Chris is not as consistent as John, even though they clock about the same average time. In order to determine if the coach's concern is valid, you clock their time in the last 20 runs and compute a standard deviation of 0.85 seconds for John and 1.20 seconds for Chris. It is fair to assume that clock time is normally distributed for both John and Chris.
- a. Specify the hypotheses to test if the variance of time for John is smaller than that of Chris.
  - b. Carry out the test at the 10% level of significance.
  - c. Who has a better likelihood of breaking the record at the meet? Explain.
46. Annual growth rates for individual firms in the toy industry tend to fluctuate dramatically, depending on consumers' tastes and current fads. Consider the following growth rates (in percent) for two companies in this industry, Hasbro and Mattel.

Year	2005	2006	2007	2008	2009
Hasbro	3.0	2.1	21.8	4.8	1.2
Mattel	1.5	9.1	5.7	-0.1	-8.2

SOURCE: Annual Reports for Hasbro, Inc. and Mattel, Inc.

- State the null and the alternative hypotheses in order to determine if the variance of growth rates differs for the two firms.
  - What assumption regarding the population is necessary to implement this step?
  - Specify the critical value(s) at  $\alpha = 0.05$ .
  - What is your conclusion?
47. At BurgerJoint, consistency in product and service is the new motto. Accordingly, management has invested in an automated French-fry dispenser to replace the manual dispensing method. The goal is to standardize the number of fries provided. The accompanying table shows the number of fries dispensed in a large-sized container based on samples of 10 orders before and after the process change. Assume these samples were drawn randomly and independently from normally distributed populations.

Manual Method (Sample 1)	27	39	29	31	30	37	41	29	38	31
Automated Method (Sample 2)	30	35	37	38	34	35	35	32	35	32

- Develop the hypotheses to test whether the variance of the new automated dispensing method is lower than the previous manual method.
  - Use Excel's F.TEST function to calculate the  $p$ -value.
  - Would your conclusion change at the 1% significance level?
48. **FILE Safety\_Stock.** An automotive parts distributor wants to standardize its safety stock level for two parts, A and B. ("Safety stock" is the excess inventory carried above the expected demand level to provide protection against demand variability.) Consequently, the distributor has randomly sampled daily demand for each part over the past 30 days.
- State the appropriate null and alternative hypotheses to test if the variances of the daily demand values for the two parts are different.
  - Calculate the value of the test statistic. Assume demand is normally distributed.
  - Use Excel's F.TEST function to calculate the  $p$ -value.
  - Make a conclusion at the 5% significance level.
  - Is your conclusion sensitive to the choice of the significance level,  $\alpha$ ? Explain.

49. **FILE Wait\_Times.** Barbara Dwyer, the manager at Lux Hotel, makes every effort to ensure that customers attempting to make phone reservations do not have to wait too long to speak with a reservation specialist. Since the hotel accepts phone reservations 24 hours a day, Barbara is especially interested in maintaining consistency in service. Barbara wants to determine if the variance of wait time in the early morning shift differs from that in the late morning shift. She uses the following independently drawn samples of wait time for phone reservations for both shifts for the analysis. Assume that wait times are normally distributed.

Shift	Wait Time (in seconds)							
Early Morning Shift: 12:00 am–6:00 am	67	48	52	71	83	59	49	66
	57	68	60	66	82	63	64	83
	37	41	60	41	87	53	66	69
Late Morning Shift: 6:00 am–12:00 pm	98	100	122	108	100	123	102	90
	125	121	120	128	123	94	128	113
	116	104	96	111	107	105	113	106

- Specify the appropriate hypotheses to test if the variance of wait time in the early morning shift differs from that in the late morning shift.
  - Use Excel's F.TEST function to conduct the test at the 1% level of significance.
  - Does the variance of wait time in the early morning shift differ from that in the late morning shift?
50. **FILE Adidas\_Revenues.** Adidas revenues (in millions of €) in Asia and Latin America for the years 2005 through 2009 are shown in the accompanying table. Assume revenues are normally distributed.

	2005	2006	2007	2008	2009
Asia	1,523	2,020	2,254	2,662	2,614
Latin America	319	499	657	893	1,006

SOURCE: Adidas Online Annual Reports.

- Specify the competing hypothesis in order to test whether the variance in revenues is greater in Asia than in Latin America.
- Use Excel to calculate the value of the test statistic.
- Use Excel's F.TEST function to calculate the  $p$ -value.
- At  $\alpha = 0.05$ , what is your conclusion?

## CASE STUDIES

**CASE STUDY 11.1** Due to environmental concerns and the never-ending volatility of gas prices, drivers are becoming more concerned with their cars' gasoline consumption. Cameron White, a research analyst at a nonprofit organization, shares these concerns and wonders whether his car's gas consumption is as efficient as it was when he first bought the new car five years ago. Despite his best intentions, he has been a bit lax in his upkeep of the car and feels that this may adversely influence its performance. At the time he purchased the car, he was told that his car would average 29 miles per gallon (mpg) on highways with a standard deviation of 1 mpg. He records his car's mpg from the last 20 fill-ups and obtains the following values.

**Data for Case Study 11.1** Gasoline Consumption: Miles per Gallon

26	28	25	29	27	28	30	27	29	28
26	28	29	27	26	27	28	25	28	27

**FILE**

*Gasoline\_Consumption*

In a report, use the above information to:

1. Construct the 95% confidence interval for the population standard deviation. Discuss any assumptions you made for the analysis.
2. Determine whether the variability has significantly increased from the original standard deviation of 1 mpg at a 5% level of significance.

**CASE STUDY 11.2** Nicholas Grammas is an investment analyst examining the performance of two mutual funds with Janus Capital Group: The Janus Balanced Fund and the Janus Overseas Fund.

- The Janus Balanced Fund: This “core” fund consists of stocks and bonds and its goal is diversification. It has historically produced solid long-term returns through different market cycles.
- The Janus Overseas Fund: This fund invests in overseas companies based on their individual merits instead of their geography or industry sector.

The following table reports the annual returns (in percent) of these two funds over the past 10 years.

**Data for Case Study 11.2** Annual Total Return (%) History

Year	Janus Balanced Fund	Janus Overseas Fund
2000	−2.16	−18.57
2001	−5.04	−23.11
2002	−6.56	−23.89
2003	13.74	36.79
2004	8.71	18.58
2005	7.75	32.39
2006	10.56	47.21
2007	10.15	27.76
2008	−15.22	−52.75
2009	24.28	78.12

**FILE**

*Janus\_Returns*

SOURCE: finance.yahoo.com.

In a report, use the above information to:

1. Describe the similarities and differences in these two funds' returns.
2. Examine whether the risk of one fund is different from the risk of the other fund at the 5% significance level. Discuss the assumptions made for the analysis.

**CASE STUDY 11.3** For decades, people have believed that boys are innately more capable than girls in math. In other words, due to the intrinsic differences in brains, boys are better suited for doing math than girls. Recent research challenges this stereotype, arguing that gender differences in math performance have more to do with culture than innate aptitude. In the U.S., for example, girls perform just as well on standardized math tests as boys. Others argue, however, that while the average may be the same, there is more variability in math ability for boys than girls, resulting in some boys with soaring math skills. A portion of representative data on math scores for boys and girls is shown in the accompanying table.

**FILE**  
*Math\_Scores*

**Data for Case Study 11.3** Math Scores for Boys and Girls

Boys	Girls
74	83
89	76
⋮	⋮
66	74

In a report, use the above information to:

1. Construct and interpret the 95% confidence interval for the ratio of the variance of math scores for boys and for girls. Discuss the assumptions made for the analysis.
2. Determine at the 5% significance level if boys have more variability in math scores than girls.

## APPENDIX 11.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab and JMP; SPSS does not provide applications suitable to this chapter. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### Confidence Interval for $\sigma^2$

- A. (Replicating Example 11.2) From the menu choose **Stat > Basic Statistics > 1 Variance**.
- B. Choose "One or more samples in each column." Select Metals and Income. Choose Options. Enter 95.0 for **Confidence Interval**.

#### Testing $\sigma^2$

- A. (Replicating Example 11.4) From the menu choose **Stat > Basic Statistics > 1 Variance**. Select **Perform hypothesis test**, select **Hypothesized variance**, and enter the value 72.25.
- B. Choose "One or more samples in each column." Select Income.
- C. Choose Options. Select "Variance  $\neq$  hypothesized variance."

#### Confidence Interval for $\sigma_1^2/\sigma_2^2$

- A. (Replicating Example 11.6) From the menu choose **Stat > Basic Statistics > 2 Variances**.

**FILE**  
*Fund\_Returns*

- B. Choose “Sample variances,” and under **Sample 1** enter 11 for **Sample size** and 182.25 for **Variance**. Under **Sample 2** enter 16 for **Sample size** and 457.96 for **Variance**. Choose **Options**. Enter 95.0 for **Confidence Interval**.

### Testing $\sigma_1^2/\sigma_2^2$

- A. (Replicating Example 11.8) From the menu choose **Stat > Basic Statistics > 2 Variances**.
- B. Choose “Sample standard deviations,” and under **Sample 1** enter 10 for **Sample size** and 37.13 for **Standard deviation**. Under **Sample 2** enter 10 for **Sample size** and 11.07 for **Standard deviation**.
- C. Choose **Options**. Select “Ratio > hypothesized ratio.”

## JMP

### Confidence Interval for $\sigma^2$

- A. (Replicating Example 11.2) From the menu choose **Analyze > Distribution**.
- B. Under **Select Columns**, select **Metals** and **Income**, and then under **Cast Selected Columns into Roles**, select **Y, Columns**.
- C. Click the red triangle in the output window next to **Metals** and **Income**, select **Confidence Interval > 0.95**.

**FILE**

*Fund\_Returns*

### Testing $\sigma^2$

- A. (Replicating Example 11.4) From the menu choose **Analyze > Distribution**.
- B. Under **Select Columns**, select **Income**, then under **Cast Selected Columns into Roles**, select **Y, Columns**.
- C. Click the red triangle in the output window next to **Income**, and select **Test Std Dev**. After **Specify Hypothesized Standard Deviation**, enter 8.5 ( $\sigma = \sqrt{72.25} = 8.5$ ).

# 12

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 12.1 Conduct a goodness-of-fit test for a multinomial experiment.
- LO 12.2 Conduct a test for independence.
- LO 12.3 Conduct a goodness-of-fit test for normality.
- LO 12.4 Conduct the Jarque-Bera test.

# Chi-Square Tests

In this chapter, we focus on the  $\chi^2$  (chi-square) distribution to develop statistical tests that compare observed data with what we would expect from a population with a specific distribution.

Generally, the chi-square tests are used to assess two types of comparison. First, a *goodness-of-fit test* is commonly used with a frequency distribution representing sample data of a qualitative variable. For instance, we may want to substantiate a claim that market shares in the automotive industry have changed dramatically over the past 10 years. Whereas a goodness-of-fit test focuses on a single qualitative variable, a *test for independence* is used to compare two qualitative variables. For example, we may want to determine whether a person's gender influences his/her purchase of a product. We can also extend the goodness-of-fit test to determine whether it is reasonable to assume that sample data are drawn from a normal population. Since we use the normal distribution with quantitative data, we first convert the raw data into a frequency distribution. Finally, we introduce the Jarque-Bera test, which allows us to test for normality using the data in their raw form.





## INTRODUCTORY CASE

### Sportswear Brands

In the introductory case to Chapter 4, Annabel Gonzalez, chief retail analyst at a marketing firm, studies the relationship between the brand name of compression garments in the sport-apparel industry and the age of the consumer. Specifically, she wants to know whether the age of the consumer influences the brand name purchased.

Her initial feeling is that the Under Armour brand attracts a younger customer, whereas the more established companies, Nike and Adidas, draw an older clientele. She believes this information is relevant to advertisers and retailers in the sporting-goods industry, as well as to some in the financial community. Suppose she collects data on 600 recent purchases in the compression-gear market. Table 12.1 summarizes the results of the sample using a contingency table, cross-classified by age and brand name.

**TABLE 12.1** Purchases of Compression Garments Based on Age and Brand Name

Age Group	Brand Name		
	Under Armour	Nike	Adidas
Under 35 years	174	132	90
35 years or older	54	72	78

Annabel wants to use the above sample information to:

1. Determine whether the two variables (Age Group and Brand Name) are related at the 5% significance level.
2. Discuss how the findings from the test for independence can be used.

A synopsis of this case will be provided at the end of Section 12.2.

## 12.1 GOODNESS-OF-FIT TEST FOR A MULTINOMIAL EXPERIMENT

### LO 12.1

Conduct a goodness-of-fit test for a multinomial experiment.

In this section, we examine whether two or more population proportions equal each other or any predetermined (hypothesized) set of values. There are many instances where we may want to make inferences of this type. For instance, in a heavily concentrated industry consisting of four firms, we may want to determine whether each firm has an equal market share. Or, in a political contest, we may want to determine whether Candidates A, B, and C will receive 70%, 20%, and 10% of the vote, respectively. Before conducting a test of this type, we must first ensure that the random experiment satisfies the conditions of a **multinomial experiment**, which is simply a generalization of the binomial experiment first introduced in Chapter 5.

Recall that a Bernoulli process, also referred to as a binomial experiment, is a series of  $n$  independent and identical trials of an experiment, where each trial has only two possible outcomes, conventionally labeled “success” and “failure.” For the binomial experiment, we generally denote the probability of success as  $p$  and the probability of failure as  $1 - p$ . Alternatively, we could let  $p_1$  and  $p_2$  represent these probabilities, where  $p_1 + p_2 = 1$ . Now let us assume that the number of outcomes per trial is  $k$  where  $k \geq 2$ .

### A MULTINOMIAL EXPERIMENT

A **multinomial experiment** consists of a series of  $n$  independent and identical trials, such that for each trial

- There are  $k$  possible outcomes called categories.
- The probability  $p_i$  associated with the  $i$ th category remains the same.
- The sum of the probabilities is one; that is,  $p_1 + p_2 + \cdots + p_k = 1$ .

Note that when  $k = 2$ , the multinomial experiment specializes to a binomial experiment.

Numerous experiments fit the conditions of a multinomial experiment. For instance,

- As compared from the previous day, a stockbroker records whether the price of a stock rises, falls, or stays the same. This example has three possible categories ( $k = 3$ ).
- A consumer rates service at a restaurant as excellent, good, fair, or poor ( $k = 4$ ).
- The admissions office records which of the six business concentrations a student picks ( $k = 6$ ).

When setting up the competing hypotheses for a multinomial experiment, we have essentially two choices. We can set all population proportions equal to the same specific value or, equivalently, equal to one another. For instance, if we want to judge on the basis of sample data whether the proportion of voters who favor four different candidates is the same, the competing hypotheses would take the following form:

$$H_0: p_1 = p_2 = p_3 = p_4 = 0.25$$

$$H_A: \text{Not all population proportions are equal to 0.25.}$$

Note that the hypothesized value under the null hypothesis is 0.25 because the population proportions must sum to one. We can also set each population proportion equal to a different predetermined (hypothesized) value. Suppose we want to determine whether 40% of the voters favor Candidate 1, 30% favor Candidate 2, 20% favor Candidate 3, and 10% favor Candidate 4. The competing hypotheses are formulated as

$$H_0: p_1 = 0.40, p_2 = 0.30, p_3 = 0.20, \text{ and } p_4 = 0.10$$

$$H_A: \text{Not all population proportions equal their hypothesized values.}$$

When conducting a test, we take a random sample and determine whether the sample proportions are close enough to the hypothesized population proportions. For this reason,

this type of test is called a **goodness-of-fit test**. Under the usual assumption that the null hypothesis is true, we derive the expected frequencies of the categories in a multinomial experiment and compare them with observed frequencies. The objective is to determine whether we can reject the null hypothesis in favor of the alternative hypothesis. To see how to conduct a goodness-of-fit test, consider the following example.

One year ago, the management at a restaurant chain surveyed its patrons to determine whether changes should be made to the menu. One question on the survey asked patrons to rate the quality of the restaurant's entrées. The percentages of the patrons responding Excellent, Good, Fair, or Poor are listed in the following table:

Excellent	Good	Fair	Poor
15%	30%	45%	10%

Based on responses to the overall survey, management decided to revamp the menu. Recently, the same question concerning the quality of entrées was asked of a random sample of 250 patrons. Their responses are shown below:

Excellent	Good	Fair	Poor
46	83	105	16

At the 5% significance level, we want to determine whether there has been any change in the population proportions calculated one year ago.

Since we want to determine whether the responses of the 250 patrons are consistent with the earlier proportions, we let the earlier population proportions denote the hypothesized proportions for the test. Thus, we use  $p_1$ ,  $p_2$ ,  $p_3$ , and  $p_4$  to denote the population proportions of those that responded Excellent, Good, Fair, or Poor, respectively, and construct the following competing hypotheses.

$$H_0: p_1 = 0.15, p_2 = 0.30, p_3 = 0.45, \text{ and } p_4 = 0.10$$

$$H_A: \text{Not all population proportions equal their hypothesized values.}$$

The first step in calculating the value of the test statistic is to calculate the expected frequency for each category. That is, we need to estimate the frequencies that we would expect to get if the null hypothesis is true. In general, in order to calculate the expected frequency  $e_i$  for category  $i$ , we multiply the sample size  $n$  by the respective hypothesized value of the population proportion  $p_i$ . For example, consider the category Excellent. If  $H_0$  is true, then we expect that 15% ( $p_1 = 0.15$ ) of 250 patrons will find the quality of entrées to be excellent. Therefore, the expected frequency of Excellent responses is 37.5 ( $= 250 \times 0.15$ ), whereas the corresponding observed frequency is 46. Expected frequencies for other responses are found similarly. Ultimately, when computing the value of the test statistic, we compare these expected frequencies to the frequencies we actually observe. The test statistic follows the  $\chi^2$  (chi-square) distribution that was discussed in Chapter 11. Because the distribution is characterized by a family of distributions, where each distribution depends on its particular degrees of freedom,  $df$ , it is common to make reference to it using the notation  $\chi^2_{df}$ .

#### TEST STATISTIC FOR GOODNESS-OF-FIT TEST

For a multinomial experiment with  $k$  categories, the value of the test statistic is calculated as

$$\chi^2_{df} = \sum \frac{(o_i - e_i)^2}{e_i},$$

where  $df = k - 1$ , and  $o_i$  and  $e_i = np_i$  are the observed frequency and the expected frequency in the  $i$ th category, respectively.

**Note:** The test is valid when the expected frequencies for each category are five or more.

Table 12.2 shows the expected frequency  $e_i$  for each category. The condition that each expected frequency  $e_i$  must equal five or more is satisfied here. As we will see shortly, sometimes it is necessary to combine data from two or more categories to achieve this result.

**TABLE 12.2** Calculation of Expected Frequency for Restaurant Example

	Hypothesized Proportion, $p_i$	Expected Frequency, $e_i = np_i$
Excellent	0.15	$250 \times 0.15 = 37.5$
Good	0.30	$250 \times 0.30 = 75.0$
Fair	0.45	$250 \times 0.45 = 112.5$
Poor	0.10	$250 \times 0.10 = 25.0$
		$\Sigma e_i = 250$

As a check on the calculations, the sum of the expected frequencies  $\Sigma e_i$  must equal the sample size  $n$ , which in this example equals 250. Once the expected frequencies are estimated, we are ready to calculate the value of the test statistic.

The  $\chi^2_{df}$  statistic measures how much the observed frequencies differ from the expected frequencies. In particular,  $\chi^2_{df}$  is computed as the sum of the standardized squared deviations. The smallest value that  $\chi^2_{df}$  can assume is zero—this occurs when each observed frequency equals its expected frequency. Rejection of the null hypothesis occurs when  $\chi^2_{df}$  is significantly greater than zero. As a result, these tests of hypotheses regarding multiple population proportions ( $p_1, p_2, p_3, \dots$ ) are always implemented as right-tailed tests. However, since the alternative hypothesis states that not all population proportions equal their hypothesized values, rejection of the null hypothesis does not indicate which proportions differ from these values.

In this example, there are four categories ( $k = 4$ ), so  $df = k - 1 = 3$ . Since a goodness-of-fit test is a right-tailed test, the critical value with  $\alpha = 0.05$  is found from the  $\chi^2$  table (Appendix A, Table 3) as  $\chi^2_{\alpha, df} = \chi^2_{0.05, 3} = 7.815$ ; we show a portion of the  $\chi^2$  table in Table 12.3. The value of the test statistic is calculated as

$$\begin{aligned}
 \chi^2_{df} &= \chi^2_3 = \Sigma \frac{(o_i - e_i)^2}{e_i} \\
 &= \frac{(46 - 37.5)^2}{37.5} + \frac{(83 - 75)^2}{75} + \frac{(105 - 112.5)^2}{112.5} + \frac{(16 - 25)^2}{25} \\
 &= 1.93 + 0.85 + 0.50 + 3.24 = 6.52.
 \end{aligned}$$

The decision rule is to reject  $H_0$  if  $\chi^2_3 > 7.815$ . Since  $\chi^2_3 = 6.52 < 7.815$ , we do not reject  $H_0$ . We cannot conclude that the proportions differ from the ones from one year ago at the 5% significance level. Management may find this news disappointing in that the goal of the menu change was to improve customer satisfaction. Responses to other questions on the survey may shed more light on whether the goals of the menu change met or fell short of expectations.

**TABLE 12.3** Portion of the  $\chi^2$  table

df	Area in Upper Tail, $\alpha$									
	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	<b>7.815</b>	9.348	11.345	12.838

## Using Excel to Calculate $p$ -Values

As usual, we can conduct the above hypothesis test by using the  $p$ -value approach to hypothesis testing rather than the critical value approach, where the  $p$ -value is derived as

$P(\chi_3^2 \geq 6.52)$ . If we refer to Table 12.3, we see that 6.52 falls between 6.251 and 7.815, which allows us to conclude that the  $p$ -value is somewhere between 0.05 and 0.10. We can use Excel's CHISQ.DIST.RT function to calculate the exact  $p$ -value. In general, to find the right-tailed area under the chi-square distribution, we find an empty cell and input “=CHISQ.DIST.RT ( $\chi_{df}^2$ ,  $df$ ),” where  $\chi_{df}^2$  is the value of the test statistic and  $df$  are the respective degrees of freedom. In the survey example, we input “=CHISQ.DIST.RT(6.52, 3).” Excel returns the probability of 0.0889, which is the  $p$ -value. Again, given a significance level of 5%, we are unable to reject the null hypothesis.

### EXAMPLE 12.1

Table 12.4 lists the market shares in 2010 of the five firms that manufacture a particular product. A marketing analyst wonders whether the market shares have changed since 2010. He surveys 200 customers. The last column of Table 12.4 shows the number of customers who recently purchased the product at each firm.

**TABLE 12.4** Market Share of Five Firms

Firm	Market Share in 2010	Number of Recent Customers
1	0.40	70
2	0.32	60
3	0.24	54
4	0.02	10
5	0.02	6

- Specify the competing hypotheses to test whether the market shares have changed since 2010.
- Calculate the value of the test statistic.
- Use  $\alpha = 0.05$  to determine if the market shares have changed since 2010.

#### SOLUTION:

- Let  $p_i$  denote the market share for the  $i$ th firm. In order to test whether the market shares have changed since 2010, we *initially* set up the competing hypotheses as

$$H_0: p_1 = 0.40, p_2 = 0.32, p_3 = 0.24, p_4 = 0.02, \text{ and } p_5 = 0.02$$

$$H_A: \text{Not all market shares equal their hypothesized values.}$$

- The value of the test statistic is calculated as  $\chi_{df}^2 = \sum \frac{(o_i - e_i)^2}{e_i}$ . The last column of Table 12.4 shows each firm's observed frequency  $o_i$ , so before applying the formula, we first calculate each firm's expected frequency  $e_i$ .

$$e_1 = 200 \times 0.40 = 80$$

$$e_2 = 200 \times 0.32 = 64$$

$$e_3 = 200 \times 0.24 = 48$$

$$e_4 = 200 \times 0.02 = 4$$

$$e_5 = 200 \times 0.02 = 4 \left. \vphantom{e_4} \right\} 8$$

We note that the expected frequencies for firms 4 and 5 are less than five. The test is valid so long as the expected frequencies in each category are five or more. In order to achieve this result, we combine the expected frequencies for firms 4 and 5 to obtain a combined frequency of eight ( $e_4 + e_5 = 8$ ). We could have made other combinations, say  $e_4$  with  $e_1$  and  $e_5$  with  $e_2$ , but we preferred to maintain a category for the less dominant firms. After making this combination, we now respecify the competing hypotheses as



$$H_0: p_1 = 0.40, p_2 = 0.32, p_3 = 0.24, \text{ and } p_4 = 0.04$$

$H_A$ : Not all market shares equal their hypothesized values.

With  $df = k - 1 = 3$ , we calculate the value of the test statistic as

$$\begin{aligned}\chi^2_3 &= \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(70 - 80)^2}{80} + \frac{(60 - 64)^2}{64} + \frac{(54 - 48)^2}{48} + \frac{(16 - 8)^2}{8} \\ &= 1.25 + 0.25 + 0.75 + 8 = 10.25.\end{aligned}$$

- c. Given  $\alpha = 0.05$ , we find the critical value as  $\chi^2_{0.05,3} = 7.815$ . The decision rule is to reject  $H_0$  if  $\chi^2_3 > 7.815$ . Since 10.25 is greater than 7.815, we reject the null hypothesis and conclude that some market shares changed since 2010.

As mentioned earlier, one limitation of this type of chi-square test is that we cannot tell which proportions differ from their hypothesized values. However, given the divergence between the observed and expected frequencies for the less dominant firms, it appears that they may be making some headway in this industry. Further testing can be conducted to see if this is the case.

## EXERCISES 12.1

### Mechanics

1. Consider a multinomial experiment with  $n = 250$  and  $k = 4$ . The null hypothesis to be tested is  $H_0: p_1 = p_2 = p_3 = p_4 = 0.25$ . The observed frequencies resulting from the experiment are

Category	1	2	3	4
Frequency	70	42	72	66

- Specify the alternative hypothesis.
  - Find the critical value at the 5% significance level.
  - Calculate the value of the test statistic.
  - What is the conclusion to the hypothesis test?
2. Consider a multinomial experiment with  $n = 400$  and  $k = 3$ . The null hypothesis is  $H_0: p_1 = 0.60, p_2 = 0.25$ , and  $p_3 = 0.15$ . The observed frequencies resulting from the experiment are

Category	1	2	3
Frequency	250	94	56

- Define the alternative hypothesis.
  - Calculate the value of the test statistic and approximate the  $p$ -value for the test.
  - At the 5% significance level, what is the conclusion to the hypothesis test?
3. A multinomial experiment produced the following results:

Category	1	2	3	4	5
Frequency	57	63	70	55	55

Can we conclude at the 1% significance level that not all population proportions are equal to 0.20?

4. A multinomial experiment produced the following results:

Category	1	2	3
Frequency	128	87	185

At the 1% significance level, can we reject  $H_0: p_1 = 0.30, p_2 = 0.20$ , and  $p_3 = 0.50$ ?

### Applications

5. You suspect that an unscrupulous employee at a casino has tampered with a die; that is, he is using a loaded die. In order to test this claim, you roll the die 200 times and obtain the following frequencies:

Category	1	2	3	4	5	6
Frequency	40	35	33	30	33	29

- Specify the null and alternative hypotheses in order to test your claim.
  - Approximate the  $p$ -value.
  - At a 10% significance level, can you conclude that the die is loaded?
6. A study conducted in September and October of 2010 found that fewer than half of employers who hired new college graduates last academic year plan to definitely do so again (*The Wall Street Journal*, November 29, 2010). Suppose the hiring intentions of the respondents were as follows:

Definitely Hire	Likely to Hire	Hire Uncertain	Will not Hire
37%	17%	28%	18%



Six months later, a sample of 500 employers were asked their hiring intentions and gave the following responses:

Definitely Hire	Likely to Hire	Hire Uncertain	Will not Hire
170	100	120	110

- Specify the competing hypotheses to test whether the proportions from the initial study have changed.
  - Find the critical value at the 5% significance level.
  - Calculate the value of the test statistic.
  - What is the conclusion to the hypothesis test?  
Interpret your results.
- A rent-to-own (RTO) agreement appeals to low-income and financially distressed consumers. It allows immediate access to merchandise, and by making all payments, the consumer acquires the merchandise. At the same time, goods can be returned at any point without penalty. Suppose a recent study documents that 65% of RTO contracts are returned, 30% are purchased, and the remaining 5% default. In order to test the validity of this claim, an RTO researcher looks at the transaction data of 420 RTO contracts, of which 283 are returned, 109 are purchased, and the rest defaulted.
    - Set up the competing hypothesis to test whether the return, purchase, and default probabilities of RTO contracts differ from 0.65, 0.30, and 0.05, respectively.
    - Compute the value of the test statistic.
    - Conduct the test at the 5% level of significance and interpret the test results.
  - Despite Zimbabwe's shattered economy, with endemic poverty and widespread political strife and repression, thousands of people from overseas still head there every year (*BBC News*, August 27, 2008). Main attractions include the magnificent Victoria Falls, the ruins of Great Zimbabwe, and herds of roaming wildlife. A tourism director claims that Zimbabwe visitors are equally represented by Europe, North America, and the rest of the world. Records show that of the 380 tourists who recently visited Zimbabwe, 135 were from Europe, 126 were from North America, and 119 were from the rest of the world.
    - A recent visitor to Zimbabwe believes that the tourism director's claim is wrong. Set up the competing hypotheses such that rejection of the null hypothesis supports the visitor's belief.
    - Use the critical value approach to conduct the test at a 5% level. Do the sample data support the visitor's belief?
    - Repeat the analysis with the  $p$ -value approach.
  - In 2003, *The World Wealth Report* first started publishing market shares of global millionaires (*The Wall Street Journal*, June 25, 2008). At this time, the distribution of the world's people worth \$1 million or more was

Region	Percentage of Millionaires
Europe	35.7%
North America	31.4%
Asia Pacific	22.9%
Latin America	4.3%
Middle East	4.3%
Africa	1.4%

SOURCE: *The Wealth Report*, 2003.

A recent sample of 500 global millionaires produces the following results:

Region	Number of Millionaires
Europe	153
North America	163
Asia Pacific	139
Latin America	20
Middle East	20
Africa	5

- Test whether the distribution of millionaires today is different from the distribution in 2003 at  $\alpha = 0.05$ .
  - Would the conclusion change if we tested it at  $\alpha = 0.10$ ?
- An Associated Press/GfK Poll shows that 38% of American drivers favor U.S. cars, while 33% prefer Asian brands, with the remaining 29% going for other foreign cars ([www.msnbc.com](http://www.msnbc.com), April 21, 2010). This highlights a significant improvement for U.S. automakers, especially when just a few years ago General Motors Co. and Chrysler LLC needed government help just to survive. Perhaps Americans are giving U.S. automakers a closer look due to their buffed-up offerings. A researcher believes that the "buy American" sentiment may also be the result of watching an iconic American industry beaten down amid the Great Recession. He wonders whether the preferences for cars have changed since the Associated Press/GfK Poll. He surveys 200 Americans and finds that the number of respondents in the survey who prefer American, Asian, and other foreign cars are 66, 70, and 64, respectively. At the 5% significance level, can the researcher conclude that preferences have changed since the Associated Press/GfK Poll?
  - (Use Excel) The quality department at an electronics company has noted that, historically, 92% of the units of a specific product pass a test operation, 6% fail the test but are able to be repaired, and 2% fail the test and need to be scrapped. Due to recent process improvements, the quality department would like to confirm whether these rates are still valid. A recent sample of 500 parts revealed that 475 parts passed the test, 18 parts failed the test but were repairable, and 7 parts failed the test and were scrapped.

- a. State the appropriate null and alternative hypotheses to test if the current proportions are different than the historical proportions.
  - b. Calculate the value of the test statistic.
  - c. Use Excel's CHISQ.DIST.RT function to calculate the  $p$ -value.
  - d. At the 5% significance level, what is your conclusion? Would your conclusion change at the 1% significance level?
12. (Use Excel) An agricultural grain company processes and packages various grains purchased from farmers. A high-volume conveyor line contains four chutes at the end, each of which is designed to receive and dispense equal proportions of grain into bags. Each bag is then stamped with a date code and the number of the chute from which it came. If the chute output proportions are not relatively equal, then

a bottleneck effect is created upstream and the conveyor cannot function at peak output. Recently, a series of repairs and modifications have led management to question whether the grains still are being equally distributed among the chutes. Packaging records from 800 bags yesterday indicate that 220 bags came from Chute 1, 188 bags from Chute 2, 218 bags from Chute 3, and 174 bags from Chute 4.

- a. State the appropriate null and alternative hypotheses to test if the proportion of bags filled by any of the chutes is different from 0.25.
- b. Calculate the value of the test statistic.
- c. Use Excel's CHISQ.DIST.RT function to calculate the  $p$ -value.
- d. What is your conclusion at the 10% significance level? Would your conclusion change at the 5% significance level?

## LO 12.2

## 12.2 CHI-SQUARE TEST FOR INDEPENDENCE

Conduct a test for independence.

Recall from Chapter 4 that a contingency table is a useful tool when we want to examine or compare two qualitative variables defined on the same population.

### CONTINGENCY TABLE

A **contingency table** generally shows frequencies for two qualitative (categorical) variables,  $x$  and  $y$ , where each cell represents a mutually exclusive combination of the pair of  $x$  and  $y$  values.

In this section, we use the data in a contingency table to conduct a hypothesis test that determines whether the two qualitative variables depend upon one another. Whereas a goodness-of-fit test examines a single qualitative variable, a **test for independence**—also called a **chi-square test of a contingency table**—assesses the relationship between two qualitative variables. Many examples of the use of this test arise, especially in marketing, biomedical research, and courts of law. For instance, a retailer may be trying to determine whether there is a relationship between the age of its clientele and where it chooses to advertise. Doctors might want to investigate whether or not losing weight through stomach surgery can extend the lives of severely obese patients. Or one party in a discrimination lawsuit may be trying to show that gender and promotion are related. All of these examples lend themselves to applications of the hypothesis test discussed in this section.

In the introductory case study, we are presented with a contingency table cross-classified by the variables Age Group and Brand Name. Specifically, we want to determine whether or not the age of a consumer influences his/her decision to buy a garment from Under Armour, Nike, or Adidas. We will conduct this test at the 5% significance level.

In general, the competing hypotheses for a statistical test for independence are formulated such that rejecting the null hypothesis leads to the conclusion that the two qualitative variables are dependent. Formally,

$H_0$ : The two qualitative variables are independent.

$H_A$ : The two qualitative variables are dependent.

Since the criteria upon which we classify the data are Age Group and Brand Name, we write the competing hypotheses as

$H_0$ : Age Group and Brand Name are independent.

$H_A$ : Age Group and Brand Name are dependent.

Table 12.5 reproduces Table 12.1 of the introductory case. The variable Age Group has two possible categories: (1) Under 35 years and (2) 35 years or older. The variable Brand Name has three possible categories: (1) Under Armour, (2) Nike, and (3) Adidas. Each cell in this table represents an observed frequency  $o_{ij}$  where the subscript  $ij$  refers to the  $i$ th row and the  $j$ th column. Thus,  $o_{13}$  refers to the cell in the first row and the third column. Here,  $o_{13} = 90$ , or, equivalently, 90 customers under 35 years of age purchased an Adidas product.

**TABLE 12.5** Purchases of Compression Garments Based on Age and Brand Name

Age Group	Brand Name		
	Under Armour	Nike	Adidas
Under 35 years	174	132	90
35 years or older	54	72	78

We will use the independence assumption postulated under the null hypothesis to derive an expected frequency for each cell from the sample data. In other words, we first estimate values as if no relationship exists between the age of a consumer and the brand name of the clothing purchased. Then we will compare these expected frequencies with the observed values to compute the value of the test statistic.

## Calculating Expected Frequencies

For ease of exposition, we first denote each event using algebraic notation. We let events  $A_1$  and  $A_2$  represent “Under 35 years” and “35 years or older,” respectively; events  $B_1$ ,  $B_2$ , and  $B_3$  stand for Under Armour, Nike, and Adidas, respectively. We then sum the frequencies for each column and row. For instance, the sum of the frequencies for Event  $A_1$  is 396; this is obtained by summing the values in row  $A_1$ : 174, 132, and 90. Totals for the other rows and columns are shown in Table 12.6.

**TABLE 12.6** Row and Column Totals

Age Group	Brand Name			Row Total
	$B_1$	$B_2$	$B_3$	
$A_1$	$e_{11}$	$e_{12}$	$e_{13}$	396
$A_2$	$e_{21}$	$e_{22}$	$e_{23}$	204
Column Total	228	204	168	600

Our goal is to calculate the expected frequency  $e_{ij}$  for each cell, where again the subscript  $ij$  refers to the  $i$ th row and the  $j$ th column. Thus,  $e_{13}$  refers to the cell in the first row and the third column, or the expected number of customers who are under 35 years of age and purchase an Adidas product.

Before we can arrive at the expected frequencies, we first calculate marginal row probabilities (the proportion of people under 35 years of age and those 35 years old or older) and marginal column probabilities (the proportion of people purchasing from each brand name). We calculate marginal row (column) probabilities by dividing the row (column) sum by the total sample size:

Marginal Row Probabilities:

$$P(A_1) = \frac{396}{600} \quad \text{and} \quad P(A_2) = \frac{204}{600}$$

Marginal Column Probabilities:

$$P(B_1) = \frac{228}{600}, P(B_2) = \frac{204}{600}, \text{ and } P(B_3) = \frac{168}{600}$$

We can now calculate each cell probability by applying the multiplication rule for independent events from Chapter 4. That is, if two events are independent, say events  $A_1$  and  $B_1$  (our assumption under the null hypothesis), then their joint probability is

$$P(A_1 \cap B_1) = P(A_1)P(B_1) = \left(\frac{396}{600}\right)\left(\frac{228}{600}\right) = 0.2508.$$

Multiplying this joint probability by the sample size yields the expected frequency for  $e_{11}$ —that is, the expected number of customers who are under 35 years of age and purchase an Under Armour product:

$$e_{11} = 600(0.2508) = 150.48.$$

#### CALCULATING EXPECTED FREQUENCIES FOR A TEST FOR INDEPENDENCE

We use the following general formula to calculate the expected frequencies for each cell in a contingency table:

$$e_{ij} = \frac{(\text{Row } i \text{ total})(\text{Column } j \text{ total})}{\text{Sample Size}},$$

where  $e_{ij}$  is the expected frequency for each cell in a contingency table, and the subscript  $ij$  refers to the  $i$ th row and the  $j$ th column.

Applying the formula, we calculate all expected frequencies as

$$\begin{aligned} e_{11} &= \frac{(396)(228)}{600} = 150.48 & e_{12} &= \frac{(396)(204)}{600} = 134.64 & e_{13} &= \frac{(396)(168)}{600} = 110.88 \\ e_{21} &= \frac{(204)(228)}{600} = 77.52 & e_{22} &= \frac{(204)(204)}{600} = 69.36 & e_{23} &= \frac{(204)(168)}{600} = 57.12 \end{aligned}$$

Table 12.7 shows the expected frequency  $e_{ij}$  for each cell. In order to satisfy subsequent assumptions, each expected frequency  $e_{ij}$  *must equal five or more*. This condition is satisfied here. As we saw in Example 12.1, it may be necessary to combine two or more rows or columns to achieve this result in other applications.

**TABLE 12.7** Expected Frequencies for Contingency Table

Age Group	Brand Name			Row Total
	$B_1$	$B_2$	$B_3$	
$A_1$	150.48	134.64	110.88	396
$A_2$	77.52	69.36	57.12	204
Column Total	228	204	168	600

When conducting a test for independence, we calculate the value of the chi-square test statistic  $\chi^2_{df}$ . Analogous to the discussion in Section 12.1,  $\chi^2_{df}$  measures how much the observed frequencies differ from the expected frequencies. The smallest value that  $\chi^2_{df}$  can assume is zero—this occurs when each observed frequency equals its expected frequency. Thus, a test for independence is also implemented as a *right-tailed test*.

#### TEST STATISTIC FOR A TEST FOR INDEPENDENCE

For a test for independence applied to a contingency table with  $r$  rows and  $c$  columns, the value of the test statistic is calculated as

$$\chi^2_{df} = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

where  $df = (r - 1)(c - 1)$ , and  $o_{ij}$  and  $e_{ij}$  are the observed frequency and the expected frequency, respectively, for each cell in a contingency table.

**Note:** This test is valid when the expected frequencies for each cell are five or more.

With two rows and three columns in the contingency table, degrees of freedom are calculated as  $df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$ . We apply the formula to compute the value of the test statistic as

$$\begin{aligned}\chi^2_2 &= \frac{(174 - 150.48)^2}{150.48} + \frac{(132 - 134.64)^2}{134.64} + \frac{(90 - 110.88)^2}{110.88} \\ &\quad + \frac{(54 - 77.52)^2}{77.52} + \frac{(72 - 69.36)^2}{69.36} + \frac{(78 - 57.12)^2}{57.12} \\ &= 3.68 + 0.05 + 3.93 + 7.14 + 0.10 + 7.63 = 22.53.\end{aligned}$$

Given a significance level of 5% and  $df = 2$ , we find the critical value as  $\chi^2_{\alpha, df} = \chi^2_{0.05, 2} = 5.991$ . Hence, the decision rule is to reject  $H_0$  if  $\chi^2_2 > 5.991$ . Since  $\chi^2_2 = 22.53 > 5.991$ , we reject  $H_0$ . At the 5% significance level, we conclude that the two qualitative variables are dependent; that is, there is a relationship between the age of a consumer and the brand name of the apparel purchased.

As usual, we can conduct the above hypothesis test by using the  $p$ -value approach to hypothesis testing rather than the critical value approach, where the  $p$ -value is derived as  $P(\chi^2_2 \geq 22.53)$ . As discussed earlier, we can calculate an exact  $p$ -value using Excel by invoking “=CHISQ.DIST.RT(22.53, 2).” Excel returns an almost zero  $p$ -value of  $1.28 \times 10^{-5}$ . Again, given a significance level of 5%, we reject the null hypothesis and conclude that Age Group and Brand Name are not independent of one another.

### EXAMPLE 12.2

A recent study of gender preferences among car shoppers found that men and women equally favor economy cars (www.cargurus.com, February 14, 2011). A marketing analyst doubts these results. He believes that gender differences exist with respect to the purchase of an economy car. He collects data on 400 recent car purchases cross-classified by Gender and Car Type (economy car versus noneconomy car). The results are shown in Table 12.8. At the 10% significance level, determine whether the sample data support the marketing analyst’s claim.

**TABLE 12.8** Car Preferences by Gender

Gender	Car Type		Row Total
	Economy Car	Noneconomy Car	
Female	50	60	110
Male	120	170	290
Column Total	170	230	400

**SOLUTION:** In order to determine whether an economy car purchase depends on gender, we specify the competing hypotheses as

$H_0$ : Gender and Car Type are independent.

$H_A$ : Gender and Car Type are dependent.

With two rows ( $r = 2$ ) and two columns ( $c = 2$ ) in the contingency table, we compute degrees of freedom as  $df = (r - 1)(c - 1) = 1$ . Given  $\alpha = 0.10$ , we find the critical value as  $\chi^2_{0.10, 1} = 2.706$ .

The value of the test statistic is calculated as  $\chi^2_{df} = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ . Table 12.8 provides each cell’s observed frequency  $o_{ij}$ , so before applying the formula, we first calculate each cell’s expected frequency  $e_{ij}$ :

$$\begin{aligned}e_{11} &= \frac{(110)(170)}{400} = 46.75 & e_{12} &= \frac{(110)(230)}{400} = 63.25 \\ e_{21} &= \frac{(290)(170)}{400} = 123.25 & e_{22} &= \frac{(290)(230)}{400} = 166.75\end{aligned}$$

We then calculate

$$\begin{aligned}\chi_1^2 &= \frac{(50 - 46.75)^2}{46.75} + \frac{(60 - 63.25)^2}{63.25} + \frac{(120 - 123.25)^2}{123.25} + \frac{(170 - 166.75)^2}{166.75} \\ &= 0.23 + 0.17 + 0.09 + 0.06 = 0.55.\end{aligned}$$

Since  $\chi_1^2 = 0.55 < 2.706$ , we do not reject the null hypothesis. The sample data do not support the marketing analyst's claim that gender differences exist with respect to the purchase of an economy car.

## SYNOPSIS OF INTRODUCTORY CASE



Under Armour pioneered clothing in the compression-gear market. Compression garments are meant to keep moisture away from a wearer's body during athletic activities in warm and cool weather. Wicking moisture is a secondary characteristic of compression gear. Wicking materials were in widespread use before Under Armour and are used in most noncompression athletic wear. The characteristic that defines *compression* gear is that it is tight to help compress muscles, which supposedly helps them work better, avoid injury, and recover faster. Under Armour has experienced exponential growth since the firm went public in November 2005 (*USA Today*, June 16, 2010); however, Nike and Adidas have aggressively entered the compression-gear market as well. An analysis is conducted to examine whether the age of the customer matters when making a purchase in the compression-gear market. This information is relevant not only for Under Armour and how the firm may focus its advertising efforts, but also to competitors and retailers in this market. Data were collected on 600 recent purchases in the

compression-gear market; the data were then cross-classified by age group and brand name. A test for independence was conducted at the 5% significance level. The results suggest that a customer's age and the brand name purchased are related to one another. Given that age influences the brand name purchased, it is not surprising that Under Armour signed NFL quarterback Tom Brady (<http://cnbc.com>, October 6, 2010) to endorse its products, a move likely to attract a younger consumer. Brady had spent most of his career with Nike before breaking away to go with Under Armour.

## EXERCISES 12.2

### Mechanics

13. Suppose you are conducting a test for independence. Specify the critical value under the following scenarios:
  - a.  $r = 3$ ,  $c = 3$ , and  $\alpha = 0.10$ .
  - b.  $r = 4$ ,  $c = 5$ , and  $\alpha = 0.05$ .
14. Suppose you are conducting a test for independence. Specify the critical value under the following scenarios:
  - a.  $r = 5$ ,  $c = 2$ , and  $\alpha = 0.025$ .
  - b.  $r = 3$ ,  $c = 5$ , and  $\alpha = 0.01$ .
15. Given the following contingency table, conduct a test for independence at the 5% significance level using (a) the critical value approach and (b) the  $p$ -value approach.

Variable B	Variable A	
	1	2
1	23	47
2	32	53

16. Given the following contingency table, conduct a test for independence at the 1% significance level using (a) the  $p$ -value approach and (b) the critical value approach.

Variable B	Variable A			
	1	2	3	4
1	120	112	100	110
2	127	115	120	124
3	118	115	110	124



## Applications

17. According to an online survey by Harris Interactive for job site CareerBuilder.com (InformationWeek.com, September 27, 2007), more than half of IT workers say they have fallen asleep at work. Sixty-four percent of government workers admitted to falling asleep on the job. Assume that the following contingency table is representative of the survey results.

Slept on the Job?	Job Category	
	IT Professional	Government Professional
Yes	155	256
No	145	144

- Specify the competing hypotheses to determine whether sleeping on the job is associated with job category.
  - Calculate the value of the test statistic.
  - Approximate the  $p$ -value.
  - At the 5% significance level, can you conclude that sleeping on the job depends on job category?
18. A market researcher for an automobile company suspects differences in preferred color between male and female buyers. Advertisements targeted to different groups should take such differences into account, if they exist. The researcher examines the most recent sales information of a particular car that comes in three colors.

Color	Gender of Automobile Buyer	
	Male	Female
Silver	470	280
Black	535	285
Red	495	350

- Specify the competing hypotheses to determine whether color preference depends on gender.
  - Find the critical value at the 1% significance level.
  - Calculate the value of the test statistic.
  - Does your conclusion suggest that the company should target advertisements differently for males versus females? Explain.
19. The following sample data reflect shipments received by a large firm from three different vendors and the quality of those shipments.

Vendor	Defective	Acceptable
1	14	112
2	10	70
3	22	150

- Specify the competing hypotheses to determine whether quality is associated with the source of the shipments.
  - Conduct the test at a 1% significance level using the critical value approach.
  - Should the firm be concerned about the source of the shipments? Explain.
20. A marketing agency would like to determine if there is a relationship between union membership and type of vehicle owned (domestic or foreign brand). The goal is to develop targeted advertising campaigns for particular vehicle brands likely to appeal to specific groups of customers. A survey of 500 potential customers revealed the following results.

	Union Member	Not Union Member
Domestic brand	133	147
Foreign brand	67	153

- Specify the competing hypotheses to determine whether vehicle brand (domestic, foreign) is associated with union membership.
  - Use the critical value approach to conduct the test at the 10% significance level. What is your conclusion?
  - Is the conclusion reached in part (b) sensitive to the choice of significance level?
21. (Use Excel) The quality manager believes there may be a relationship between the experience level of an inspector and whether a product passes or fails inspection. Inspection records were reviewed for 630 units of a particular product, and the number of units which passed and failed inspection was determined based on three inspector experience levels. The results are shown in the following table.

Decision	Experience Level		
	Low (< 2 years)	Medium (2-8 years)	High (> 8 years)
Pass	152	287	103
Fail	16	46	26

- Specify the competing hypotheses to determine whether the inspector pass/fail decision depends on experience level.
- Calculate the value of the test statistic.
- Use Excel's CHISQ.DIST.RT function to calculate the  $p$ -value.
- At the 5% significance level, what is your conclusion? Does your conclusion change at the 1% significance level?

22. (Use Excel) According to a 2008 survey by the Pew Research Center, people in China are highly satisfied with their roaring economy and the direction of their nation (*USA Today*, July 22, 2008). Eighty-six percent of those who were surveyed expressed positive views of the way China is progressing and described the economic situation as good. A political analyst wants to know if this optimism among the Chinese depends on age. In an independent survey of 280 Chinese residents, the respondents are asked how happy they are with the direction that their country is taking. Their responses are tabulated in the following table.

Age	Very Happy	Somewhat Happy	Not Happy
20 up to 40	23	50	18
40 up to 60	51	38	16
60 and above	19	45	20

- Set up the appropriate hypotheses to test the claim that optimism regarding China's direction depends on the age of the respondent.
  - Calculate the value of the test statistic.
  - Use Excel's CHISQ.DIST.RT function to calculate the  $p$ -value.
  - At a 1% level of significance, can we infer that optimism among the Chinese is dependent on age?
23. A study by the Massachusetts Community & Banking Council found that blacks, and, to a lesser extent, Latinos, remain largely unable to borrow money at the same interest rate as whites (*The Boston Globe*, February 28, 2008). The following contingency table shows representative data for the city of Boston, cross-classified by race and type of interest rate received:

Race	Type of Interest Rate on Loan	
	High Interest Rate	Low Interest Rate
Black	553	480
Latino	265	324
White	491	3701

At the 5% significance level, do the data indicate that the interest rate received on a loan is dependent on race? Provide the details.

24. Founded in February 2004, Facebook is a social utility that helps people communicate with their friends and family. In just six years, Facebook had acquired more than 500 million active users, of which 50% logged on to Facebook in any given day. In a survey of 3,000 Facebook users, the designers looked at why Facebook users break up in a relationship (*The Wall Street Journal*, November 27–28, 2010).

Reasons for Breakup	Gender	
	Men	Women
Nonapproval	3%	4%
Distance	21%	16%
Cheating	18%	22%
Lost Interest	28%	26%
Other	30%	32%

SOURCE: Internal survey of 3,000 Facebook users.

Suppose the survey consisted of 1,800 men and 1,200 women. Use the data to determine whether the reasons for breakup depend on gender at the 1% significance level. Provide the details.

## 12.3 CHI-SQUARE TEST FOR NORMALITY

The goodness-of-fit test for a multinomial experiment can also be used to test a hypothesis that a population has a particular probability distribution. For instance, we can use this test to determine whether the sample data fit the binomial or the Poisson distributions. However, due to its wide applicability, we focus on the normal distribution. We describe two chi-square tests for normality: the goodness-of-fit test and the Jarque-Bera test.

### LO 12.3

Conduct a goodness-of-fit test for normality.

### The Goodness-of-Fit Test for Normality

Suppose an economist claims that annual household income in a small Midwestern city is not normally distributed. We will use the representative data (in \$1,000s) in Table 12.9 to test this claim at the 5% significance level.

**TABLE 12.9** Household Income (in \$1,000s)

90	15	85	54	62	38	38	55	62	210
19	38	57	78	98	42	19	62	66	90
25	38	14	65	77	110	22	18	180	52
44	17	45	99	250	78	58	35	57	45
37	58	62	44	35	78	35	82	94	58

**FILE**  
Household\_Income

We first use the data to compute the sample mean and the sample standard deviation as

$$\bar{x} = 63.80 \quad \text{and} \quad s = 45.78.$$

Since we want to determine whether or not the given data represents a random sample from a population having a normal distribution, we specify this in the null hypothesis, along with the above sample estimates of the population mean and the population standard deviation.

$H_0$ : Income (in \$1,000s) in a small Midwestern city follows a normal distribution with mean \$63.80 and standard deviation \$45.78.

$H_A$ : Income (in \$1,000s) in a small Midwestern city does not follow a normal distribution with mean \$63.80 and standard deviation \$45.78.

The null hypothesis implies that the underlying distribution is normal and that the population mean and the population standard deviation equal their estimates, or, equivalently,  $\mu = 63.80$  and  $\sigma = 45.78$ . As discussed in Section 12.1, the goodness-of-fit test for a multinomial experiment deals with a single population of qualitative data. Since observations that follow the normal distribution are quantitative, we essentially need to convert the data into a qualitative format. After computing the sample mean and the sample standard deviation, we subdivide the data into non-overlapping intervals (categories); in other words, we construct a frequency distribution. The first two columns of Table 12.10 show the frequency distribution for the raw data from Table 12.9.

**TABLE 12.10** Calculations for the Normality Test Example

Income (in \$1,000s)	Observed Frequency, $o_i$	$p_i$ if $H_0$ is True	Expected Frequency, $e_i = n \times p_i$	Standardized Squared Deviation, $\frac{(o_i - e_i)^2}{e_i}$
Income < 20	6	0.1685	$50 \times 0.1685 = 8.43$	$\frac{(6 - 8.43)^2}{8.43} = 0.70$
$20 \leq \text{Income} < 40$	10	0.1330	$50 \times 0.1330 = 6.65$	1.69
$40 \leq \text{Income} < 60$	13	0.1666	$50 \times 0.1666 = 8.33$	2.62
$60 \leq \text{Income} < 80$	10	0.1687	$50 \times 0.1687 = 8.44$	0.29
Income $\geq 80$	11	0.3632	$50 \times 0.3632 = 18.16$	2.82
	$n = \sum o_i = 50$	$\sum p_i = 1$	$n = \sum e_i = 50$	$\chi^2_2 = \sum \frac{(o_i - e_i)^2}{e_i} = 8.12$

Note that we have six observations (15, 19, 19, 14, 18, and 17) that are less than 20. Other frequencies are found similarly. Earlier, we were able to calculate expected frequencies by multiplying the sample size  $n$  by the hypothesized probabilities (proportions)  $p_i$  under the null hypothesis. Here, we first calculate the probabilities under the assumption that the null hypothesis is true and then use them to calculate expected frequencies. For example, under the null hypothesis that income is normally distributed with  $\mu = 63.80$  and  $\sigma = 45.78$ , we reference the  $z$  table to find the probability that an individual's income is less than 20, or

$$P(X < 20) = P\left(\frac{X - \mu}{\sigma} < \frac{20 - 63.80}{45.78}\right) = P(Z < -0.96) = 0.1685.$$

We proceed with the other intervals in a like manner.

$$\begin{aligned}
 P(20 \leq X < 40) &= P\left(\frac{20 - 63.80}{45.78} \leq \frac{X - \mu}{\sigma} < \frac{40 - 63.80}{45.78}\right) \\
 &= P(-0.96 \leq Z < -0.52) = 0.1330 \\
 P(40 \leq X < 60) &= P\left(\frac{40 - 63.80}{45.78} \leq \frac{X - \mu}{\sigma} < \frac{60 - 63.80}{45.78}\right) \\
 &= P(-0.52 \leq Z < -0.08) = 0.1666 \\
 P(60 \leq X < 80) &= P\left(\frac{60 - 63.80}{45.78} \leq \frac{X - \mu}{\sigma} < \frac{80 - 63.80}{45.78}\right) \\
 &= P(-0.08 \leq Z < 0.35) = 0.1687 \\
 P(X \geq 80) &= P\left(\frac{X - \mu}{\sigma} \geq \frac{80 - 63.80}{45.78}\right) = P(Z \geq 0.35) = 0.3632
 \end{aligned}$$

The third column of Table 12.10 shows these probabilities. We are then able to compute the expected frequencies for each interval as  $n \times p_i$ . The fourth column of Table 12.10 shows the values for the expected frequencies. As in Section 12.1, the appropriate test statistic follows the  $\chi^2_{df}$  distribution and its value is calculated as  $\chi^2_{df} = \sum \frac{(o_i - e_i)^2}{e_i}$ . The only difference is that the degrees of freedom are equal to the number of intervals minus one, minus the number of parameters estimated. Since we estimate two parameters—the mean and the standard deviation—from the sample data, the degrees of freedom for the chi-square test for normality are always  $k - 1 - 2 = k - 3$ .

#### TEST STATISTIC FOR THE GOODNESS-OF-FIT TEST FOR NORMALITY

For a goodness-of-fit test for normality, the value of the test statistic is calculated as

$$\chi^2_{df} = \sum \frac{(o_i - e_i)^2}{e_i},$$

where  $df = k - 3$ ,  $k$  is the number of intervals in the frequency distribution, and  $o_i$  and  $e_i$  are the observed frequency and the expected frequency in the  $i$ th interval, respectively.

**Note:** The test is valid when the expected frequencies for each interval are five or more.

Like before, the goodness-of-fit test for normality is a right-tailed test. When constructing the frequency distribution, we must ensure that the expected frequencies for each interval equal five or more. If necessary, the number of intervals can be adjusted by combining adjacent intervals until this condition is achieved.

Since in this example we formed five intervals ( $k = 5$ ) we calculate  $df = 5 - 3 = 2$ . Then we sum the standardized squared deviations as shown in the last column of Table 12.10 to obtain the value of the chi-square test statistic as  $\chi^2_{df} = \chi^2_2 = 8.12$ .

With a significance level of 5% and  $df$  of 2, we find the critical value as  $\chi^2_{\alpha, df} = \chi^2_{0.05, 2} = 5.991$ . Hence, our decision rule is to reject  $H_0$  if  $\chi^2_2 > 5.991$ . Since  $\chi^2_2 = 8.12 > 5.991$ , we reject  $H_0$ . At the 5% significance level, we conclude that income in this Midwestern city does not follow a normal distribution with a mean of \$63,800 and a standard deviation of \$45,780. Alternatively, we can use Excel's CHISQ.DIST.RT function by inputting “=CHISQ.DIST.RT(8.12, 2)” into an empty cell. Excel returns the exact  $p$ -value of 0.0172. At the 5% significance level, we again reject  $H_0$ .

A criticism of this test of normality is that we first have to convert raw data into a frequency distribution by grouping them into a set of arbitrary intervals or categories. The resulting value of the chi-square test statistic depends on how the data are grouped.

## The Jarque-Bera Test

### LO 12.4

Conduct the Jarque-Bera test.

An alternative to the goodness-of-fit test for normality is the **Jarque-Bera test**. In this test, it is not necessary to convert the quantitative data into a qualitative form. Instead, using the raw data, we calculate the **skewness coefficient**  $S$  and the (excess) **kurtosis coefficient**  $K$  of the sample data. A skewness coefficient of zero indicates that the data are symmetric about its mean. The kurtosis coefficient measures whether a distribution is more or less peaked than a normal distribution. The skewness coefficient and the kurtosis coefficient for the normal distribution are both equal to zero. We use Excel to calculate the skewness and the kurtosis coefficients.

When testing whether sample data are derived from the normal distribution, the null hypothesis consists of the joint hypothesis that both the skewness coefficient and the kurtosis coefficient are zero. It can be shown that the Jarque-Bera test statistic follows the  $\chi^2_{df}$  distribution with two degrees of freedom.

### THE TEST STATISTIC FOR THE JARQUE-BERA TEST

When testing whether data are derived from a normal distribution using the Jarque-Bera ( $JB$ ) test, the value of the test statistic is calculated as

$$JB = \chi^2_2 = (n/6)[S^2 + K^2/4],$$

where  $df = 2$ ,  $n$  is the sample size,  $S$  is the skewness coefficient, and  $K$  is the kurtosis coefficient.

### EXAMPLE 12.3

Using the data from Table 12.9 and the Jarque-Bera test, determine whether or not annual household income is normally distributed at the 1% significance level.

#### SOLUTION:

The competing hypotheses take the following form:

$$H_0: S = 0 \text{ and } K = 0$$

$$H_A: S \neq 0 \text{ or } K \neq 0$$

In order to compute the value of the test statistic, we first need to compute the skewness and kurtosis coefficients,  $S$  and  $K$ . We can use Excel's SKEW and KURT functions to compute values for  $S$  and  $K$ . Alternatively, we can use Excel's Data Analysis Toolpak option (from Excel's menu, choose Data > Data Analysis > Descriptive Statistics), which will report values for  $S$  and  $K$  as well as other descriptive measures. Using either method, we find that  $S = 2.32$  and  $K = 6.73$ .

The value of the test statistic is calculated as

$$JB = \chi^2_2 = (n/6)[S^2 + K^2/4] = (50/6)[2.32^2 + 6.73^2/4] = 139.21.$$

With a significance level of 1% and  $df = 2$ , we find the critical value  $\chi^2_{\alpha, df} = \chi^2_{0.01, 2} = 9.210$ . Since  $\chi^2_2 > \chi^2_{0.01, 2}$  ( $139.21 > 9.210$ ), we reject  $H_0$  and conclude that income in this Midwestern city does not follow a normal distribution.

In the above examples, the conclusion with the Jarque-Bera test and the goodness-of-fit test for normality is the same. This result is not surprising, as it is fairly well documented that income distribution, in general, is skewed to the right (not normally distributed), with a few households accounting for most of the total income. For this reason, we prefer to use the median rather than the mean to get a more accurate reflection of typical income.

## EXERCISES 12.3

### Mechanics

25. Consider the following sample data with mean and standard deviation of 20.5 and 5.4, respectively.

Class	Frequency
Less than 10	25
10 up to 20	95
20 up to 30	65
30 or more	15
	$n = 200$

- Using the goodness-of-fit test for normality, specify the competing hypotheses in order to determine whether or not the data are normally distributed.
  - At the 5% significance level, what is the critical value? What is the decision rule?
  - Calculate the value of the test statistic.
  - What is the conclusion?
26. The following frequency distribution has a sample mean of  $-3.5$  and a sample standard deviation of  $9.7$ .

Class	Frequency
Less than $-10$	70
$-10$ up to $0$	40
$0$ up to $10$	80
$10$ or more	10

At the 1% significance level, use the goodness-of-fit test for normality to determine whether or not the data are normally distributed.

27. You are given the following summary statistics from a sample of 50 observations:

Mean	77.25
Standard Deviation	11.36
Skewness	1.12
Kurtosis	1.63

- Using the Jarque-Bera test, specify the null and alternative hypotheses to determine whether or not the data are normally distributed.
- At the 5% significance level, what is the critical value?
- Calculate the value of the test statistic.
- What is the conclusion? Can you conclude that the data do not follow the normal distribution? Explain.

### Applications

28. An economics professor states on her syllabus that final grades will be distributed using the normal distribution. The final averages of 300 students are calculated, and she groups the data into a frequency distribution as shown in the accompanying table. The mean and the standard deviation of the final are  $\bar{x} = 72$  and  $s = 10$ .

Final Averages	Frequency
F: Less than 50	5
D: 50 up to 70	135
C: 70 up to 80	105
B: 80 up to 90	45
A: 90 or above	10
	Total = 300

- Using the goodness-of-fit test for normality, state the competing hypotheses in order to determine if we can reject the professor's normality claim.
  - At a 5% significance level, what is the critical value?
  - Calculate the value of the test statistic.
  - What is the conclusion to the test?
29. Fifty cities provided information on vacancy rates (in percent) in local apartments in the following frequency distribution. The sample mean and the sample standard deviation are 9% and 3.6%, respectively.

Vacancy Rate (in percent)	Frequency
Less than 6	10
6 up to 9	10
9 up to 12	20
12 or more	10

Apply the goodness-of-fit test for normality at the 5% significance level. Do the sample data suggest that vacancy rates do not follow the normal distribution?

30. The quality department at an electronics component manufacturer must ensure that their components will operate at prespecified levels. The accompanying table shows a frequency distribution with measured resistance values (in ohms) for a sample of 520 resistors. The sample mean and the sample standard deviation are 4790 ohms and 40 ohms, respectively.

Resistance (ohms)	Frequency
Under 4740	44
4740 up to 4780	145
4780 up to 4820	197
4820 up to 4860	107
4860 or more	27
	Total = 520



- a. Using the goodness-of-fit test for normality, state the competing hypotheses to test if the sample data suggest that resistance does not follow the normal distribution.
- b. At  $\alpha = 0.05$ , what is the critical value?
- c. Calculate the value of the test statistic.
- d. What is the conclusion to the test? Would your conclusion change at the 10% significance level?

31. **FILE Shaft Diameter.** Fabco, a precision machining shop, uses statistical process control (SPC) techniques to ensure quality and consistency of their steel shafts. The control limits used in their SPC charts are based on the assumption that shaft diameters are normally distributed. To verify this assumption, a quality engineer has measured the diameters for a sample of 50 of its popular 1/2-inch shafts.

- a. Using the Jarque-Bera test, state the competing hypotheses in order to determine whether or not the data follow the normal distribution.
- b. Calculate the value of the Jarque-Bera test statistic. Use Excel to calculate the  $p$ -value.
- c. At  $\alpha = 0.10$ , can you conclude that the shaft diameters are not normally distributed?
- d. Would your conclusion change at the 5% significance level?

32. Total 2005 CEO compensation for the largest U.S. companies by revenue is reported in the following frequency distribution, along with some summary statistics. Total compensation includes salary, bonuses, stock and incentives, the potential value of stock options, and gains from stock options exercised.

Total Compensation (in millions of \$)	Frequency
Less than 5	43
5 up to 10	65
10 up to 15	32
15 up to 20	38
20 or more	60
	$n = 238$

Other summary statistics for CEO compensation (in millions of \$) are as follows:

Mean	Median	Standard Deviation	Skewness	Kurtosis
19.03	11.02	27.61	5.26	35.53

- a. Conduct a goodness-of-fit test for normality of CEO compensation at the 1% significance level.

- b. Conduct the Jarque-Bera test at the 1% significance level.
- c. Does total compensation of CEOs for the largest U.S. companies not follow the normal distribution?

33. The following frequency distribution shows the distribution of monthly returns for Starbucks Corp. for the years 2003 through 2007.

Class (in percent)	Frequency
Less than -5	14
-5 up to 0	9
0 up to 5	18
5 up to 10	11
10 or more	8
	$n = 60$

SOURCE: [www.yahoo.finance.com](http://www.yahoo.finance.com).

Over this time period, the following summary statistics are provided:

Mean	Median	Standard Deviation	Skewness	Kurtosis
1.16%	1.79%	7.38%	-0.31	-0.65

- a. Conduct a goodness-of-fit test for normality at the 5% significance level. Can you conclude that monthly returns do not follow the normal distribution?
- b. Conduct the Jarque-Bera test at the 5% significance level. Can you conclude that monthly returns do not follow the normal distribution?

34. **FILE Home Depot.** The data that accompanies this exercise show weekly stock prices for Home Depot.

- a. Using the Jarque-Bera test, state the competing hypotheses in order to determine whether or not Home Depot's weekly stock prices follow the normal distribution.
- b. Calculate the value of the Jarque-Bera test statistic. Use Excel to calculate the  $p$ -value.
- c. At  $\alpha = 0.05$ , can you conclude that Home Depot's stock prices are not normally distributed?

35. **FILE MPG.** The data that accompanies this exercise show miles per gallon (MPG) for a sample of 25 cars.

- a. Using the Jarque-Bera test, state the competing hypotheses in order to determine whether or not MPG follow the normal distribution.
- b. Calculate the value of the Jarque-Bera test statistic. Use Excel to calculate the  $p$ -value.
- c. At  $\alpha = 0.05$ , can you conclude that MPG are not normally distributed?

# WRITING WITH STATISTICS



FILE

50\_Largest\_Funds

Javier Gonzalez is in the process of writing a comprehensive analysis on the three-year returns for the 50 largest mutual funds. Before he makes any inferences concerning the return data, he would first like to determine whether or not the data follow a normal distribution. Table 12.11 shows a portion of the three-year return data for the 50 largest mutual funds.

TABLE 12.11 Three-Year Returns for the 50 Largest Mutual Funds

Mutual Fund	Return (%)
American Growth	5.7
Pimco Total Return	4.7
⋮	⋮
Loomis Sayles Bond	5.4

Source: *The Boston Sunday Globe*, August 17, 2008.

Javier wants to use the sample information to:

1. Conduct a goodness-of-fit test for normality that determines, at the 5% significance level, whether or not three-year returns follow a normal distribution.
2. Perform the Jarque-Bera test that determines, at the 5% significance level, whether or not three-year returns follow a normal distribution.

## Sample Report— Assessing Whether Data Follow the Normal Distribution

As part of a broader report concerning the mutual fund industry in general, three-year return data for the 50 largest mutual funds were collected with the objective of determining whether or not the data follow a normal distribution. Information of this sort is particularly useful because much statistical inference is based on the assumption of normality. If the assumption of normality is not supported by the data, it may be more appropriate to use nonparametric techniques to make valid inferences. Table 12.A shows relevant summary statistics for three-year returns for the 50 largest mutual funds.

TABLE 12.A Three-Year Return Summary Measures for the 50 Largest Mutual Funds, August 2008

Mean	Median	Standard Deviation	Skewness	Kurtosis
5.96%	4.65%	3.39%	1.37	2.59

The average three-year return for the 50 largest mutual funds is 5.96%, with a median of 4.65%. When the mean is significantly greater than the median, it is often an indication of a positively skewed distribution. The skewness coefficient of 1.37 seems to support this claim. Moreover, the kurtosis coefficient of 2.59 suggests a distribution that is more peaked than the normal distribution. A formal test will determine whether the conclusion from the sample can be deemed real or due to chance.

The goodness-of-fit test is first applied to check for normality. The raw data is converted into a frequency distribution with five intervals ( $k = 5$ ). Expected frequencies are

calculated by multiplying the sample size  $n = 50$  by the hypothesized proportions  $p_i$  under the null hypothesis that the data follow the normal distribution with mean 5.96% and standard deviation 3.39%. Finally, the value of the chi-square test statistic is computed by summing the standardized squared deviations. All of these calculations are shown in Table 12.B.

**TABLE 12.B** Calculations for the Goodness-of-Fit Test for Normality

Return (in %)	Observed Frequency, $o_i$	$p_i$ if Return Is Normally Distributed	Expected Frequency, $e_i = n \times p_i$	Standardized Squared Deviation, $\frac{(o_i - e_i)^2}{e_i}$
Return < 2.5	7	0.1539	$50 \times 0.1539 = 7.70$	$\frac{(7 - 7.70)^2}{7.70} = 0.06$
$2.5 \leq \text{Return} < 5.0$	20	0.2359	$50 \times 0.2359 = 11.80$	5.70
$5.0 \leq \text{Return} < 7.5$	6	0.2839	$50 \times 0.2839 = 14.20$	4.74
$7.5 \leq \text{Return} < 10$	11	0.2093	$50 \times 0.2093 = 10.47$	0.03
Return $\geq 10$	6	0.1170	$50 \times 0.1170 = 5.85$	0.00
	$n = \sum o_i = 50$	$\sum p_i = 1$	$n = \sum e_i = 50$	$\chi^2_2 = \sum \frac{(o_i - e_i)^2}{e_i} = 10.53$

Table 12.C shows the competing hypotheses, the value of the test statistic, and the  $p$ -value that result from applying the goodness-of-fit test for normality and the Jarque-Bera test.

**TABLE 12.C** Test Statistics and  $p$ -values for Hypothesis Tests

Hypotheses	Test Statistic	$p$ -value
Goodness-of-Fit Test: $H_0$ : Returns are normally distributed. $H_A$ : Returns are not normally distributed.	$\chi^2_2 = 10.53$	$P(\chi^2_2 \geq 10.53) = 0.0052$
Jarque-Bera Test: $H_0$ : $S = 0$ and $K = 0$ $H_A$ : $S \neq 0$ or $K \neq 0$	$\chi^2_2 = 29.62$	$P(\chi^2_2 \geq 29.62) = 0.0000$

At the 5% significance level, the  $p$ -value of 0.0052 from the goodness-of-fit test allows us to reject the null hypothesis. The three-year returns do not follow the normal distribution at the 5% significance level.

Under the Jarque-Bera ( $JB$ ) test, the null hypothesis states that the skewness coefficient and the kurtosis coefficient are both zero. The value for the  $JB$  test statistic is 29.62 and its associated  $p$ -value is 0.0000; thus, at the 5% significance level, the null hypothesis that skewness and kurtosis are both zero is rejected. This result is consistent with the conclusion drawn from the goodness-of-fit test for normality. Both statistical tests reject the null hypothesis of normality—three-year returns do not follow the normal distribution. Statistical inference would best be conducted using nonparametric techniques that do not require the assumption of normality.

## CONCEPTUAL REVIEW

### LO 12.1 Conduct a goodness-of-fit test for a multinomial experiment.

A **multinomial experiment** consists of a series of  $n$  independent and identical trials such that on each trial there are  $k$  possible outcomes, called categories; the probability  $p_i$  associated with the  $i$ th category remains the same; and the sum of the probabilities is one.

A **goodness-of-fit test** is conducted to determine if the population proportions equal some predetermined (hypothesized) values. The value of the **test statistic** is calculated as  $\chi^2_{df} = \sum \frac{(o_i - e_i)^2}{e_i}$ , where  $df = k - 1$ , and  $o_i$  and  $e_i = np_i$  are the observed frequency and expected frequency in the  $i$ th category, respectively. The test is valid when the expected frequencies for each category are five or more. This test is always implemented as a right-tailed test.

### LO 12.2 Conduct a test for independence.

A goodness-of-fit test examines a single qualitative variable, whereas a **test for independence**, also called a **chi-square test of a contingency table**, analyzes the relationship between two qualitative variables defined on the same population. A contingency table shows frequencies for two qualitative variables,  $x$  and  $y$ , where each cell of the table represents a mutually exclusive combination of the pair of  $x$  and  $y$  values.

In order to determine whether or not the two variables are related, we again compare observed frequencies with expected frequencies. The expected frequency for each cell,  $e_{ij}$ , is calculated as  $e_{ij} = \frac{(\text{Row } i \text{ total})(\text{Column } j \text{ total})}{\text{Sample Size}}$ , where the subscript  $ij$  refers to the  $i$ th row and the  $j$ th column of the contingency table.

The value of the chi-square **test statistic** is calculated as  $\chi^2_{df} = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ , where  $o_{ij}$  is the observed frequency. Degrees of freedom are calculated as  $(r - 1)(c - 1)$  where  $r$  and  $c$  refer to the number of rows and columns, respectively, in the contingency table. The test for independence is also implemented as a right-tailed test and is valid when the expected frequencies for each cell are five or more.

### LO 12.3 Conduct a goodness-of-fit test for normality.

We can use the **goodness-of-fit test** to test the hypothesis that a population follows the **normal distribution**. Since observations that follow the normal distribution are quantitative in nature and the goodness-of-fit test is applied to qualitative data, we must first convert the data into a qualitative format.

We construct a frequency distribution with  $k$  intervals. We then calculate the probability of observing the  $i$ th interval  $p_i$  under the assumption of a normal distribution, and then use this probability to calculate the expected frequency as  $e_i = n \times p_i$ . The value of the **test statistic** is calculated as  $\chi^2_{df} = \sum \frac{(o_i - e_i)^2}{e_i}$ , with  $df = k - 3$ . Since it is a goodness-of-fit test, it is implemented as a right-tailed test and is valid when the expected frequencies in each cell are five or more.

### LO 12.4 Conduct the Jarque-Bera test.

In the goodness-of-fit test for normality, we have to first convert raw data into a frequency distribution by grouping them into a set of arbitrary intervals. The resulting value of the chi-square test statistic depends on how the data are grouped. For the **Jarque-Bera test**, it is not necessary to convert the quantitative data into a qualitative form.

Using the raw data, we use the skewness coefficient  $S$  and the (excess) kurtosis coefficient  $K$  of the sample data to conduct the test. The value of the **Jarque-Bera JB test statistic** is calculated as  $JB = \chi^2_2 = (n/6)[S^2 + K^2/4]$  where  $n$  is the sample size.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

36. The following table lists the market shares of the four firms in a particular industry in 2010 and total sales for each firm in 2011.

Firm	Market Share in 2010	Total Sales in 2011 (in billions of \$)
1	0.40	200
2	0.30	180
3	0.20	100
4	0.10	70

- Specify the competing hypotheses to test whether the market shares in 2010 are not valid in 2011.
  - At the 1% significance level, what is the critical value?
  - Calculate the value of the test statistic.
  - Do the sample data suggest that the market shares changed from 2010 to 2011?
37. A study suggests that airlines have increased restrictions on cheap fares by raising overnight requirements (*The Wall Street Journal*, August 19, 2008). This would force business travelers to pay more for their flights, since they tend to need the most flexibility and want to be home on weekends. Eight months ago, the overnight stay requirements were as follows:

One night	Two nights	Three nights	Saturday night
37%	17%	28%	18%

A recent sample of 644 flights found the following restrictions:

One night	Two nights	Three nights	Saturday night
117	137	298	92

- Specify the competing hypotheses to test whether the proportions cited by the study have changed.
  - At a 5% significance level, what is the critical value?
  - Calculate the value of the test statistic.
  - What is the conclusion to the hypothesis test? Interpret your results.
38. A local TV station claims that 60% of people support Candidate A, 30% support Candidate B,

and 10% support Candidate C. A survey of 500 registered voters is taken. The accompanying table indicates how they are likely to vote.

Candidate A	Candidate B	Candidate C
350	125	25

- Specify the competing hypotheses to test whether the TV station's claim can be rejected by the data.
  - Use the  $p$ -value approach to test the hypothesis at a 1% significance level.
39. Although founded only in 2004, Facebook has more than 500 million active users, of which 50% log on to Facebook on any given day. In a recent survey by Facebook, young users (those born after 1984) were asked about their preference for delivering the news about breaking up a relationship (*The Wall Street Journal*, November 27–28, 2010). One of the shocking results was that only 47% of users preferred to break the news in person. A researcher decides to verify the survey results of Facebook by taking her own sample of 200 young Facebook users. The preference percentages from Facebook and the researcher's survey are presented in the following table.

Delivery Method	Facebook Results	Researcher's Results
In Person	47%	55%
Phone	30%	28%
Email	4%	8%
Facebook	5%	3%
Instant Message	14%	6%

At the 5% level of significance, test if the researcher's results are inconsistent with the survey results conducted by Facebook. Provide the details, using the  $p$ -value approach.

40. A recent study in the *Journal of the American Medical Association* (February 20, 2008) found that patients who go into cardiac arrest while in the hospital are more likely to die if it happens after 11 pm. The study investigated 58,593 cardiac arrests during the day or evening. Of those, 11,604 survived to leave the hospital. There were 28,155 cardiac arrests during the shift that began at 11 pm, commonly referred to as the graveyard shift. Of those, 4,139 survived for discharge. The following contingency table summarizes the results of the study:

Shift	Survived for Discharge	Did Not Survive for Discharge	Row Totals
Day or Evening Shift	11,604	46,989	58,593
Graveyard Shift	4,139	24,016	28,155
Column Totals	15,743	71,005	86,748

- Specify the competing hypotheses to determine whether a patient's survival depends on the time at which he/she experiences cardiac arrest.
  - At a 1% significance level, what is the critical value?
  - Calculate the value of the test statistic.
  - What is the conclusion to the test? Is the timing of when a cardiac arrest occurs independent of whether or not the patient survives for discharge? Given your answer, what type of recommendations might you give to hospitals?
41. An analyst is trying to determine whether the prices of certain stocks on the NASDAQ are independent of the industry to which they belong. She examines four industries and within each industry, categorizes each stock according to its price (high-priced, average-priced, low-priced).

Stock Price	Industry			
	I	II	III	IV
High	16	8	10	14
Average	18	16	10	12
Low	7	8	4	9

- Specify the competing hypotheses to determine whether stock price depends on the industry.
  - Calculate the value of the test statistic. Approximate the  $p$ -value with the table or calculate its exact value with Excel.
  - At a 1% significance level, what can the analyst conclude?
42. Many parents have turned to St. John's wort, an herbal remedy, to treat their children with attention deficit hyperactivity disorder (ADHD). *The Journal of the American Medical Association* (June 11, 2008) recently published an article that explored the herb's effectiveness. Children with ADHD were randomly assigned to take either St. John's wort capsules or placebos. The contingency table below broadly reflects the results found in the study.

Treatment	Effect on ADHD	
	No Change in ADHD	Improvement in ADHD
St. John's wort	12	15
Placebo	14	13

At the 5% significance level, do the data indicate that St. John's wort affects children with ADHD?

43. A recent poll asked 3,228 Americans aged 16 to 21 whether they are likely to serve in the U.S. military. The following table, cross-classified by gender and race, reports that those who responded are likely or very likely to serve in the active-duty military.

Gender	Race		
	Hispanic	Black	White
Male	1098	678	549
Female	484	355	64

SOURCE: Defense Human Resources Activity telephone poll of 3,228 Americans conducted October through December 2005.

- State the competing hypotheses to test whether race and gender are dependent when making a choice to serve in the military.
  - Conduct the test using the critical value approach at the 5% significance level.
44. Given a shaky economy and high heating costs, more and more households are struggling to pay utility bills (*The Wall Street Journal*, February, 14, 2008). Particularly hard hit are households with homes heated with propane or heating oil. Many of these households are spending twice as much to stay warm this winter compared to those who heat with natural gas or electricity. A representative sample of 500 households was taken to investigate if the type of heating influences whether or not a household is delinquent in paying its utility bill. The following table reports the results.

Delinquent in Payment?	Type of Heating			
	Natural Gas	Electricity	Heating Oil	Propane
Yes	50	20	15	10
No	240	130	20	15

At the 5% significance level, test whether the type of heating influences a household's delinquency in payment. Interpret your results.

45. The following frequency distribution shows the monthly stock returns for Home Depot for the years 2003 through 2007.

Class (in percent)	Observed Frequency
Less than -5	13
-5 up to 0	16
0 up to 5	20
5 or more	11
	$n = 60$

SOURCE: www.yahoo.finance.com.



Over this time period, the following summary statistics are provided:

Mean	Median	Standard Deviation	Skewness	Kurtosis
0.31%	0.43%	6.49%	0.15	0.38

- Conduct a goodness-of-fit test for normality at the 5% significance level. Can you conclude that monthly stock returns do not follow the normal distribution?
  - Conduct the Jarque-Bera test at the 5% significance level. Are your results consistent with your answer in part (a)?
46. **FILE** *Arlington\_Homes*. The data that accompany this exercise show various variables, including price and square footage, for 36 single-family homes in Arlington, Massachusetts, sold in the first quarter of 2009.
- Use the Jarque-Bera test to test if house prices are not normally distributed at  $\alpha = 0.05$ .
  - Use the Jarque-Bera test to test if square footage is not normally distributed at  $\alpha = 0.05$ .
47. An automotive parts company has been besieged with poor publicity over the past few years due to several highly publicized product recalls which have tarnished its public image. This has prompted a series of quality improvement initiatives. Currently, the marketing manager would like to determine if these initiatives have been successful in changing public perception about the company. Below are results of two surveys, each of 600 random adults. Survey 1 was conducted prior to the quality initiatives. Survey 2 was conducted after the quality initiatives were implemented and publicized.

	Public Perception		
	Negative	Neutral	Positive
Survey 1 (previous)	324	180	96
Survey 2 (current)	246	146	208

- State the appropriate null and alternative hypotheses to test if the public perception has changed since the quality initiatives have been implemented.
  - Use the critical value approach to reach a conclusion at the 1% significance level.
  - Is your conclusion sensitive to the choice of significance level?
48. (Use Excel) Color coding is often used in manufacturing operations to display production status or to identify/prioritize materials. For

example, suppose “green” status indicates that an assembly line is operating normally, “yellow” indicates it is down waiting on personnel for set up or repair, “blue” indicates it is down waiting on materials to be delivered, and “red” indicates an emergency condition. Management has set realistic goals whereby the assembly line should be operating normally 80% of the time, waiting on personnel 9% of the time, waiting on materials 9% of the time, and in an emergency condition 2% of the time. Based on 250 recent status records, the status was green 185 times, yellow 24 times, blue 32 times, and red 9 times.

- State the appropriate null and alternative hypotheses to test if the proportions of assembly line statuses differ from the goals set by management.
  - Calculate the value of the test statistic.
  - Use Excel’s CHISQ.DIST.RT function to calculate the  $p$ -value.
  - Are management’s goals being met at  $\alpha = 0.05$ ? Will your conclusion change at  $\alpha = 0.01$ ?
49. (Use Excel) The operations manager at ElectroTech, an electronics manufacturing company, believes that workers on particular shifts may be more likely to phone in “sick” than those on other shifts. To test this belief, she has compiled the following table containing frequencies based on work shift and days absent over the past year.

	First Shift	Second Shift	Third Shift
0-2 days absent	44	20	10
3-6 days absent	38	25	12
7-10 days absent	14	9	13
11 or more days absent	4	6	5

- Specify the competing hypotheses to determine whether days absent depends on work shift.
  - Calculate the value of the test statistic.
  - Use Excel’s CHISQ.DIST.RT function to calculate the  $p$ -value.
  - What is your conclusion at the 5% significance level? What about the 1% significance level? Is the conclusion sensitive to the choice of significance level?
50. (Use Excel) The human resources department would like to consolidate the current set of retirement plan options offered to specific employee pay groups into a single plan for all pay groups (salaried, hourly, or piecework). A sample of 585 employees of various pay groups

were asked which of three potential plans they preferred (A, B, or C). The results are shown in the accompanying table. The human resources department is hoping to conclude that the retirement plan preferred by the majority of employees is independent of pay group in order to avoid the impression that the preferred plan may favor a particular group.

Preferred Plan	Employee Pay Group		
	Salaried	Hourly	Piecework
A	78	98	37
B	121	95	30
C	51	57	18

- Specify the competing hypotheses to determine whether the preferred retirement plan depends on employee pay group.
  - Calculate the value of the test statistic.
  - Use Excel's CHISQ.DIST.RT function to calculate the  $p$ -value.
  - What is your conclusion at the 10% significance level? What about the 5% significance level?
  - For any level of significance at which the preferred plan is deemed independent of employee pay group, which plan should be selected if the criterion is to choose the single plan with the highest overall employee support?
51. A software company develops and markets a popular business simulation/modeling program. A random number generator contained in the program provides random values from various probability distributions. The software design group would like to validate that the program is properly generating random numbers. Accordingly, they generated 5,000 random numbers from a normal distribution and grouped the results into the frequency distribution shown below. The sample mean and sample standard deviation are 100 and 10, respectively.

Value	Frequency
Under 70	12
70 up to 80	99
80 up to 90	658
90 up to 100	1734
100 up to 110	1681
110 up to 120	697
120 up to 130	112
130 or more	7
Total = 5,000	

- Using the goodness-of-fit test for normality, state the competing hypotheses to test if the random numbers generated do not follow the normal distribution.
  - What is the critical value at the 1% significance level?
  - Calculate the value of the test statistic.
  - What is the conclusion to the test? Is your conclusion sensitive within the range of typical significance levels?
52. **FILE Reorder\_Point.** Reorder point decisions for a particular part at an automotive parts distributor are based on the assumption that weekly demand is normally distributed. (Note: "Reorder point" is the inventory level at which a replenishment order is placed; it should be high enough to cover demand during the order fulfillment period, but low enough to avoid excessive inventory holding costs.) To examine the validity of this assumption, the logistics department has compiled weekly demand values for the past year.
- Using the Jarque-Bera test, state the competing hypotheses in order to determine whether or not weekly demand values follow the normal distribution.
  - Calculate the value of the Jarque-Bera test statistic. Use Excel to calculate the  $p$ -value.
  - At  $\alpha = 0.05$ , can you conclude that the weekly demand values are not normally distributed? Is the conclusion sensitive to the choice of significance level?

## CASE STUDIES

**CASE STUDY 12.1** A detailed study of Americans' religious beliefs and practices by the Pew Forum on Religion & Public Life revealed that religion is quite important in an individual's life (*The Boston Globe*, June 24, 2008). The second column of the accompanying table reports the proportion of Americans who feel a certain way about religion. The study also concludes that Massachusetts residents are the least likely to say that they are religious. In order to test this claim, assume 400 randomly selected Massachusetts residents are asked about the importance of religion in his/her life. The results of this survey are shown in the last column of the accompanying table.

**Data for Case Study 12.1** Importance of Religion, U.S. versus Massachusetts

Importance of Religion	U.S. Results	Responses of Massachusetts Residents
Very important	0.58	160
Somewhat important	0.25	140
Not too important	0.15	96
Don't know	0.02	4

In a report, use the sample information to:

1. Determine whether Massachusetts residents' religious beliefs differ from those based on the United States at a 5% significance level.
2. Discuss whether you would expect to find the same conclusions if you conducted a similar test for the state of Utah or states in the Southern Belt of the United States.

**CASE STUDY 12.2** A University of Utah study examined 7,925 severely obese adults who had gastric bypass surgery and an identical number of people who did not have the surgery (*The Boston Globe*, August 23, 2007). The study wanted to investigate whether losing weight through stomach surgery prolonged the lives of severely obese patients, thereby reducing their deaths from heart disease, cancer, and diabetes.

Over the course of the study, 534 of the participants died. Of those who died, the cause of death was classified as either a disease death (disease deaths include heart disease, cancer, and diabetes) or a nondisease death (nondisease deaths include suicide or accident). The following contingency table summarizes the study's findings:

**Data for Case Study 12.2** Deaths Cross-Classified by Cause and Method of Losing Weight

Cause of Death	Method of Losing Weight	
	No Surgery	Surgery
Death from disease	285	150
Death from nondisease	36	63

In a report, use the sample information to:

1. Determine at the 5% significance level whether the cause of death depends on the method of losing weight.
2. Discuss how the findings of the statistical test used in question 1 might be used by those in the health industry.

**CASE STUDY 12.3** Matthew Jordon is a research analyst for a large investment firm. He is preparing a report on the stock performance of Nike, Inc. One aspect of his report will contain inferences concerning monthly stock returns. Before making valid inferences, Matthew first wants to determine whether the return data follow the normal distribution. To this end, he constructs the following frequency distribution on monthly stock returns for the years 2006 through 2010.

Class (in percent)	Observed Frequency
Less than -5	8
-5 up to 0	20
0 up to 5	14
5 or more	18

SOURCE: [www.yahoo.finance.com](http://www.yahoo.finance.com).

He also calculates the following summary statistics over this time period:

Mean	Median	Standard Deviation	Skewness	Kurtosis
1.50%	1.31%	6.98%	0.11	-0.33

In a report, use the sample information to:

1. Conduct a goodness-of-fit test in order to determine whether the monthly stock returns are not normally distributed at the 5% significance level.
2. Perform the Jarque-Bera test in order to determine whether the monthly stock returns are not normally distributed at the 5% significance level.

## APPENDIX 12.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Where a data file is specified, copy and paste it into the relevant software spreadsheet prior to following the commands.

### Minitab

#### Goodness-of-Fit Test

- (Replicating Example 12.1) Copy the data from Table 12.4 into a Minitab spreadsheet. Remember to combine the data for firms 4 and 5 since their expected frequencies are less than 5.
- From the menu choose **Stat > Tables > Chi-Square Goodness-of-Fit Test (One Variable)**.
- Choose **Observed counts** and then select Number of Recent Customers. Under **Test**, select **Proportions specified by historical counts**, and then select Market Share. Choose **Results** and select **Display test results**.

#### Test for Independence

- (Replicating Example 12.2) Copy the data from Table 12.8 into a Minitab spreadsheet. Do not include the row totals or the column totals.
- From the menu choose **Stat > Tables > Cross Tabulation and Chi-Square**.
- Select "Summarized data in a two-way table." Under **Columns containing the table**, select Economy Car and Noneconomy Car. Choose Chi-Square and select **Chi-square test**.

#### Test of Normality

- (Confirming Normality test in Section 12.3) From the menu choose **Stat > Basic Statistics > Normality Test**.
- After **Variable**, select Household Income. Under **Tests for Normality**, select **Anderson-Darling**.

### SPSS

#### Goodness-of-Fit Test

- (Replicating Example 12.1) Copy the data from Table 12.4 into an SPSS spreadsheet. (Remember to combine the data for firms 4 and 5.)
- From the menu choose **Data > Weight Cases**. Select **Weight cases by**, and under **Frequency Variable**, select Number.

**FILE**

Household\_Income

- C. Select **Analyze > Nonparametric Tests > Legacy Dialogs > Chi-square**.
- D. Under **Test Variable List**, select **Firm**. Under **Expected Values**, select **Values**, and **Add** 0.40, 0.32, 0.24, and 0.04.

### Test for Independence

- A. (Replicating Example 12.2) In an SPSS spreadsheet, label Columns 1, 2, and 3 as “Gender,” “Type,” and “Frequency,” respectively. In the first row, enter Female, Economy, and 50; in the second row, input Female, Noneconomy, 60; in the third row, enter Male, Economy, 120; and in the fourth row, enter Male, Noneconomy, 170.
- B. From the menu choose **Data > Weight Cases**. Select **Weight cases by**, and under **Frequency Variable** select **Frequency**.
- C. From the menu select **Analyze > Descriptive Statistics > Crosstabs**.
- D. Under **Rows**, select **Gender**, and under **Columns**, select **Type**. Choose **Statistics**, check **Chi-square**.

### Test of Normality

- A. (Confirming Normality test in Section 12.3) The easiest way to test for normality in SPSS is to construct a P-P Plot (probability-probability plot or percent-percent plot). From the menu select **Analyze > Descriptive Statistics > P-P Plots**.
- B. Under **Variables**, select **Income**.

**FILE**

*Household\_Income*

## JMP

### Test for Independence

- A. (Replicating Example 12.2) In order to format the data in a JMP spreadsheet, follow step A under the SPSS instructions for Test for Independence.
- B. From the menu, select **Analyze > Fit Y by X**.
- C. Under **Select Columns**, select **Gender**, and then under **Cast Selected Columns into Roles**, select **Y, Response**. Under **Select Columns**, select **Type**, and then under **Cast Selected Columns into Roles**, select **X, Factor**. Under **Select Columns**, select **Frequency**, and then under **Cast Selected Columns into Roles**, select **Freq**.

### Test of Normality

- A. (Replicating Example 12.3) From the menu, select **Analyze > Distribution**.
- B. Under **Select Columns**, select **Household Income**, and then under **Cast Selected Columns into Roles**, select **Y, Columns**.
- C. In the red triangle next to **Household Income**, select **Continuous Fit > Normal**. In the red triangle next to **Fitted Normal**, select **Diagnostic Plot** and **Goodness of Fit**.

**FILE**

*Household\_Income*

# 13

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 13.1 Conduct and evaluate a one-way ANOVA test.
- LO 13.2 Use Fisher's LSD method and Tukey's HSD method to determine which means differ.
- LO 13.3 Conduct and evaluate a two-way ANOVA test with no interaction.
- LO 13.4 Conduct and evaluate a two-way ANOVA test with interaction.

# Analysis of Variance

In this chapter, we study analysis of variance, which is more commonly referred to as ANOVA. ANOVA is a statistical technique used to determine if differences exist between the means of three or more populations under independent sampling. The ANOVA test is actually a generalization of the two-sample  $t$  test with equal but unknown variances discussed in Chapter 10. For instance, we may want to determine whether all brands of small hybrid cars have the same average miles per gallon. Or we may wish to compare the effectiveness of different fertilizers on the average yield per acre. These are examples of one-way ANOVA, where we examine the effect of one factor on the mean. We then move on to two-way ANOVA, where the mean may be influenced by two factors. For instance, we may want to determine if average miles per gallon are affected by the brand of a hybrid and the octane level of gasoline. Or we may wish to determine if the average yield per acre is influenced by the fertilizer and the acidity level of the soil. Tests based on two-way ANOVA can be conducted *with* or *without* the interaction of the factors.



## INTRODUCTORY CASE

### Public Transportation

Sean Cox, a research analyst at an environmental organization, believes that an upswing in the use of public transportation has taken place due to environmental concerns, the volatility of gas prices, and the general economic climate. He is especially pleased with a recent study, which highlights the average annual cost savings when commuters use public transportation (*The Boston Globe*, May 8, 2009). Commuters who use public transportation save on buying, maintaining, and operating their cars, which comprise the largest household expenditure after housing. The study finds that Boston leads 20 other American cities in the amount that commuters can save if they take public transportation. Sean wonders whether or not cost savings vary dramatically by city. He collects a representative sample of public transit riders in the top four cost-savings cities: Boston, New York, San Francisco, and Chicago. Table 13.1 shows each public transit rider's annual cost savings by city.

**TABLE 13.1** Annual Cost Savings from Using Public Transportation

**FILE**  
*Public\_Transportation*

Boston	New York	San Francisco	Chicago
\$12,500	\$12,450	\$11,800	\$10,595
12,640	12,500	11,745	10,740
12,600	12,595	11,700	10,850
12,625	12,605	11,800	10,725
12,745	12,650	11,700	10,740
	12,620	11,575	
	12,560		
	12,700		
$\bar{x}_1 = \$12,622$	$\bar{x}_2 = \$12,585$	$\bar{x}_3 = \$11,720$	$\bar{x}_4 = \$10,730$
$s_1 = \$87.79$	$s_2 = \$80.40$	$s_3 = \$83.96$	$s_4 = \$90.62$
$n_1 = 5$	$n_2 = 8$	$n_3 = 6$	$n_4 = 5$

Sean wants to use the above sample information to:

1. Determine whether there are differences in mean cost savings among these four cities at the 5% significance level.
2. Establish where mean cost savings significantly differ.

A synopsis of this case is provided at the end of Section 13.2.

Conduct and evaluate a one-way ANOVA test.

We use **analysis of variance (ANOVA)** tests to determine if differences exist between the means of three or more populations under independent sampling. The ANOVA test is actually a generalization of the two-sample  $t$  test with equal but unknown variances discussed in Chapter 10. These tests are based on the  $F_{(df_1, df_2)}$  distribution that was introduced in Chapter 11. A **one-way ANOVA test** compares population means based on one categorical variable or factor. For instance, in the public transportation example from the introductory case, we want to compare cost savings of using public transportation depending on where an individual resides. We thus delineate cost savings of using public transportation by city (the categorical variable). In general, the one-way ANOVA test is used for testing  $c$  population means under the following assumptions:

1. The populations are normally distributed.
2. The population standard deviations are unknown but assumed equal.
3. The samples are selected independently.

Since we wish to test whether or not the mean annual cost savings from using public transportation is the same in Boston, New York, San Francisco, and Chicago, we formulate the following competing hypotheses:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_A: \text{Not all population means are equal.}$$

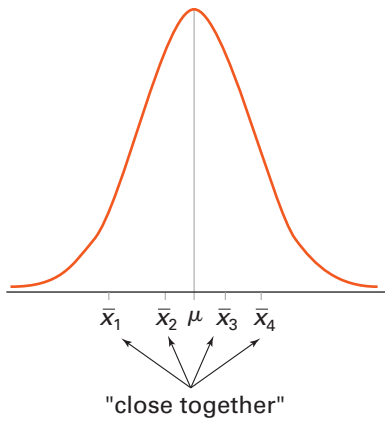
Note that  $H_A$  does not require that all means must differ from one another. In principle, the sample data may support the rejection of  $H_0$  in favor of  $H_A$  even if only two means differ.

When conducting the equality of means test, you might be tempted to set up a series of hypothesis tests, comparing  $\mu_1$  and  $\mu_2$ , then  $\mu_1$  and  $\mu_3$ , and so on, and then use the two-sample  $t$  test with equal variances discussed in Section 10.1. However, such an approach is not only cumbersome, but also flawed. In this example, where we evaluate the equality of four means, we would have to compare six combinations of two means at a time. Also, by conducting numerous pairwise comparisons, we inflate the risk of the Type I error  $\alpha$ ; that is, we increase the risk of incorrectly rejecting the null hypothesis. In other words, if we conduct all six pairwise tests at a 5% level of significance, the resulting significance level for the overall test will be greater than 5%.

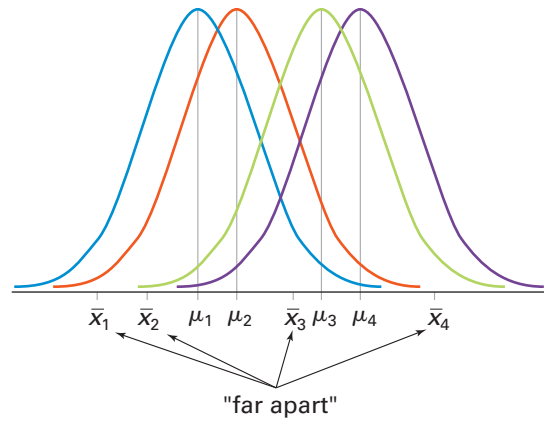
Fortunately, the ANOVA technique avoids this problem by providing one test that simultaneously evaluates the equality of several means. In the public transportation example, if the four population means are equal, we would expect the resulting sample means,  $\bar{x}_1$ ,  $\bar{x}_2$ ,  $\bar{x}_3$ , and  $\bar{x}_4$ , to be relatively close to one another. Figure 13.1a illustrates the distribution of the sample means if  $H_0$  is true. Here, the relatively small variability in the sample means can be explained by chance. What if the population means differ? Figure 13.1b shows the distributions of the sample means if the sample data support  $H_A$ . In this scenario, the sample means are relatively far apart since each sample mean is calculated from a population with a different mean. The resulting variability in the sample means cannot be explained by chance alone.

The term *treatments* is often used to identify the  $c$  populations being examined. The practice of referring to different populations as different treatments is due to the fact that many ANOVA applications were originally developed in connection with agricultural experiments where different fertilizers were regarded as different treatments applied to soil. In order to determine if significant differences exist between some of the population means, we develop two independent estimates of the common population variance  $\sigma^2$ . One estimate can be attributed to inherent differences between the  $c$  populations, while the other estimate can be attributed to chance.

a. Distribution of sample means if  $H_0$  is true



b. Distributions of sample means if  $H_0$  is false



**FIGURE 13.1**  
The logic of ANOVA

- 1. Between-Treatments Estimate of  $\sigma^2$ .** One estimate of  $\sigma^2$  is based on the variability *between* the sample means. This is referred to as **between-treatments variance**, and is denoted by *MSTR*.
- 2. Within-Treatments Estimate of  $\sigma^2$ .** The other estimate of  $\sigma^2$  is based on the variability of the data *within* each sample; that is, the variability due to chance. This estimate is generally called **within-treatments variance**, and is denoted by *MSE*.

If we find that between-treatments variance is significantly greater than within-treatments variance, then we are able to reject the null hypothesis of equal means; this is equivalent to concluding that the ratio of between-treatments variance to within-treatments variance is significantly greater than one.

### Between-Treatments Estimate of $\sigma^2$

Between-treatments variance is based on a weighted sum of squared differences between the sample means and the overall mean of the data set, referred to as the **grand mean** and denoted as  $\bar{\bar{x}}$ . We compute the grand mean by summing all observations in the data set and dividing by the total number of observations.

Each squared difference of a sample mean from the grand mean  $(\bar{x}_i - \bar{\bar{x}})^2$  is multiplied by the respective sample size for each treatment  $n_i$ . After summing the weighted squared differences, we arrive at a value called the **sum of squares due to treatments** or *SSTR*. When we average a sum of squares over its respective degrees of freedom  $c - 1$ , we obtain the **mean square for treatments** or *MSTR*.

#### CALCULATIONS FOR THE BETWEEN-TREATMENTS ESTIMATE FOR $\sigma^2$

The **grand mean**,  $\bar{\bar{x}}$ , is calculated as  $\bar{\bar{x}} = \frac{\sum_{i=1}^c \sum_{j=1}^{n_i} x_{ij}}{n_T}$ , where  $n_T = \sum_{i=1}^c n_i$  is the total sample size.

The **sum of squares due to treatments**, *SSTR*, is calculated as  $SSTR = \sum_{i=1}^c n_i (\bar{x}_i - \bar{\bar{x}})^2$ .

The **mean square for treatments**, *MSTR*, is calculated as  $MSTR = \frac{SSTR}{c - 1}$ .



The calculations for  $\bar{\bar{x}}$ ,  $SSTR$ , and  $MSTR$  for the public transportation example are as follows:

$$\bar{\bar{x}} = \frac{\sum_{j=1}^c \sum_{i=1}^{n_j} x_{ij}}{n_T} = \frac{12500 + 12640 + \cdots + 10740}{24} = 11990.$$

$$SSTR = \sum_{i=1}^c n_i (\bar{x}_i - \bar{\bar{x}})^2 = 5(12622 - 11990)^2 + 8(12585 - 11990)^2 + 6(11720 - 11990)^2 + 5(10730 - 11990)^2 = 13204720.$$

$$MSTR = \frac{SSTR}{c - 1} = \frac{13204720}{4 - 1} = 4401573.$$

### Within-Treatments Estimate of $\sigma^2$

We just calculated a value of  $MSTR$  equal to 4,401,573. Is this value of  $MSTR$  large enough to indicate that the population means differ? To answer this question, we compare  $MSTR$  to the variability that we expect due to chance. We first calculate the **sum of squares due to error**, or equivalently, the **error sum of squares**, denoted as  $SSE$ .  $SSE$  provides a measure of the degree of variability that exists even if all population means are the same. We calculate  $SSE$  as a weighted sum of the sample variances of each treatment, and the **mean square error**  $MSE$  by dividing  $SSE$  by its respective degrees of freedom,  $df = n_T - c$ .

#### CALCULATIONS FOR WITHIN-TREATMENTS ESTIMATE FOR $\sigma^2$

The **sum of squares due to error**  $SSE$  is calculated as  $SSE = \sum_{i=1}^c (n_i - 1)s_i^2$ .

The **mean square error**  $MSE$  is calculated as  $MSE = \frac{SSE}{n_T - c}$ .

The values of  $SSE$  and  $MSE$  for the public transportation example are calculated as follows:

$$\begin{aligned} SSE &= \sum_{i=1}^c (n_i - 1)s_i^2 \\ &= (5 - 1)(87.79)^2 + (8 - 1)(80.40)^2 + (6 - 1)(83.96)^2 + (5 - 1)(90.62)^2 \\ &= 144172. \end{aligned}$$

$$MSE = \frac{SSE}{n_T - c} = \frac{144172}{24 - 4} = 7209.$$

As mentioned earlier, if the ratio of the between-treatments variance to the within-treatments variance is significantly greater than one, then this finding provides evidence for rejecting the null hypothesis of equal population means. Equivalently, if this ratio is not significantly greater than one, then we are not able to reject this null hypothesis in favor of the alternative hypothesis. We use this ratio to develop the ANOVA test.

#### TEST STATISTIC FOR A ONE-WAY ANOVA TEST

The value of the **test statistic** for the hypothesis test of the equality of the population means using one-way ANOVA is computed as

$$F_{(df_1, df_2)} = \frac{MSTR}{MSE},$$

where  $df_1 = c - 1$ ,  $df_2 = n_T - c$ , and  $n_T$  is the total sample size;  $MSTR$  is the between-treatments variance and  $MSE$  is the within-treatments variance where these values are based on independent samples drawn from  $c$  normally distributed populations with a common variance.

ANOVA tests are always implemented as right-tailed tests.

We are now in a position to conduct a four-step hypothesis test for the public transportation example. Given  $MSTR = 4,401,573$ ,  $MSE = 7,209$ ,  $df_1 = c - 1 = 4 - 1 = 3$ , and  $df_2 = n_T - c = 24 - 4 = 20$ , we compute the value of the test statistic as

$$F_{(3,20)} = \frac{MSTR}{MSE} = \frac{4401573}{7209} = 610.57.$$

Since the ANOVA test is a right-tailed test, the critical value with  $\alpha = 0.05$ ,  $df_1 = 3$ , and  $df_2 = 20$  is found from the  $F$  table as  $F_{\alpha, (df_1, df_2)} = F_{0.05, (3, 20)} = 3.10$ ; we show a portion of the  $F$  table in Table 13.2. Hence, the decision rule is to reject  $H_0$  if  $F_{(3,20)} > 3.10$ .

**TABLE 13.2** Portion of the  $F$  table

$df_2$	Area in Upper Tail	$df_1$		
		1	2	3
20	0.10	2.97	2.59	2.38
	0.05	4.35	3.49	<b>3.10</b>
	0.025	5.87	4.46	3.86
	0.01	8.10	5.85	4.94

We reject the null hypothesis because the value of the test statistic falls in the rejection region (610.57 is greater than 3.10). Therefore, we conclude that the mean cost savings from using public transportation are not the same for each city.

## The One-Way ANOVA Table

Most software packages summarize the ANOVA calculations in a table. The general format of the ANOVA table is presented in Table 13.3.

**TABLE 13.3** General Format of a One-Way ANOVA Table

Source of Variation	$SS$	$df$	$MS$	$F$
Between Groups	$SSTR$	$c - 1$	$MSTR$	$F_{(df_1, df_2)} = \frac{MSTR}{MSE}$
Within Groups	$SSE$	$n_T - c$	$MSE$	
Total	$SST$	$n_T - 1$		

We should also note that **total sum of squares  $SST$**  is equal to the sum of the squared differences of each observation from the grand mean. This is equivalent to summing  $SSTR$  and  $SSE$ ; that is,  $SST = SSTR + SSE$ .

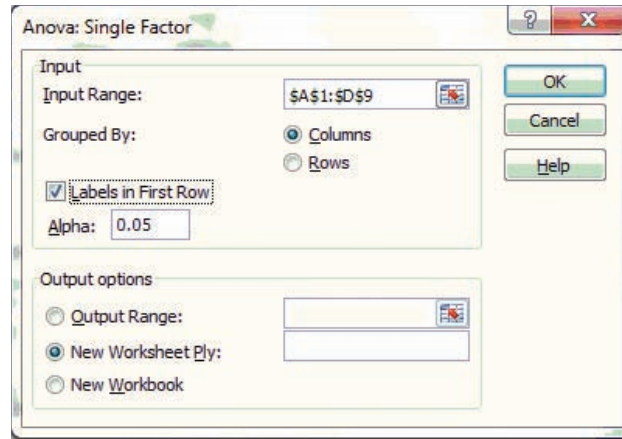
## Using Excel for a One-Way ANOVA Test

**FILE**  
*Public\_Transportation*

Fortunately, Excel provides the value of the  $F_{(df_1, df_2)}$  test statistic, the critical value, as well as the precise  $p$ -value for a one-way ANOVA test. In order to solve the public transportation example using Excel, we follow these steps.

- Open the **Public\_Transportation** data file.
- From the menu choose **Data > Data Analysis > ANOVA: Single Factor**.
- In the **ANOVA: Single Factor** dialog box shown in Figure 13.2, choose the box next to **Input range** and then select all the data, including the city names. Check the **Labels** box. If testing at a significance level other than 5%, insert the relevant  $\alpha$  value in the box next to **Alpha**. Choose an output range and click **OK**.

**FIGURE 13.2**  
Excel's ANOVA: Single  
Factor dialog box



In addition to the ANOVA table, Excel provides descriptive statistics for the sample data. Table 13.4 shows the results. You should verify that all of the hand calculations match the values produced by Excel. Any differences between the hand calculations and the computer-generated results are due to rounding.

**TABLE 13.4** Excel-Produced ANOVA Results for Public Transportation Example

SUMMARY						
Groups	Count	Sum	Average	Variance		
Boston	5	63110	12622	7707.5		
New York	8	100680	12585	6464.3		
San Francisco	6	70320	11720	7050		
Chicago	5	53650	10730	8212.5		
ANOVA						
Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	13204720	3	4401573	<b>610.57</b>	<b>7.96E-20</b>	3.098
Within Groups	144180	20	7209			
Total	13348900	23				

### EXAMPLE 13.1

Using the information in Table 13.4, repeat the ANOVA test for the public transportation example using the  $p$ -value approach.

**SOLUTION:** In order to determine whether cost savings in public transportation differ between the four cities, we again specify the competing hypotheses as

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$H_A$ : Not all population means are equal.

From Table 13.4, we find that the value of the test statistic is  $F_{3,20} = 610.57$ . Its corresponding  $p$ -value is  $7.96 \times 10^{-20}$ , or, equivalently,  $P(F_{3,20} \geq 610.57) \approx 0$ . (See the value for the test statistic and the  $p$ -value in boldface in Table 13.4.) Since the  $p$ -value is less than  $\alpha = 0.05$ , we reject the null hypothesis and again conclude that the mean cost savings from using public transportation is not the same for each city.



It is important to note that if we reject the null hypothesis, we can only conclude that not all population means are equal. The one-way ANOVA test does not allow us to infer which individual means differ. Therefore, even though the sample mean is the highest for Boston, we cannot conclude that Boston leads other cities in the amount that commuters save by taking public transportation. Further analysis of the difference between paired population means is addressed in the next section.

## EXERCISES 13.1

### Mechanics

1. A random sample of five observations from three normally distributed populations produced the following data:

Treatments		
A	B	C
22	20	19
25	25	22
27	21	24
24	26	21
22	23	19
$\bar{x}_A = 24$	$\bar{x}_B = 23$	$\bar{x}_C = 21$
$s_A^2 = 4.5$	$s_B^2 = 6.5$	$s_C^2 = 4.5$

- a. Calculate the grand mean.
  - b. Calculate  $SSTR$  and  $MSTR$ .
  - c. Calculate  $SSE$  and  $MSE$ .
  - d. Specify the competing hypotheses in order to determine whether some differences exist between the population means.
  - e. Calculate the value of the  $F_{(df_1, df_2)}$  test statistic.
  - f. Using the critical value approach at the 5% significance level, what is the conclusion to the test?
2. Random sampling from four normally distributed populations produced the following data:

Treatments			
A	B	C	D
-11	-8	-8	-12
-13	-13	-13	-13
-10	-15	-8	-15
	-12	-13	
		-10	
$\bar{x}_A = -11.3$	$\bar{x}_B = -12$	$\bar{x}_C = -10.4$	$\bar{x}_D = -13.3$
$s_A^2 = 2.33$	$s_B^2 = 8.7$	$s_C^2 = 6.3$	$s_D^2 = 2.3$

- a. Calculate the grand mean.
- b. Calculate  $SSTR$  and  $MSTR$ .
- c. Calculate  $SSE$  and  $MSE$ .
- d. Specify the competing hypotheses in order to determine whether some differences exist between the population means.
- e. Calculate the value of the  $F_{(df_1, df_2)}$  test statistic.
- f. Approximate the  $p$ -value.
- g. At the 10% significance level, what is the conclusion to the test?

3. Given the following information obtained from three normally distributed populations, construct an ANOVA table and perform an ANOVA test of mean differences at the 1% significance level.  
 $SSTR = 220.7$ ;  $SSE = 2,252.2$ ;  $c = 3$ ;  $n_1 = n_2 = n_3 = 8$
4. Given the following information obtained from four normally distributed populations, construct an ANOVA table and perform an ANOVA test of mean differences at the 5% significance level.  
 $SST = 70.47$ ;  $SSTR = 11.34$ ;  $c = 4$ ;  $n_1 = n_2 = n_3 = n_4 = 15$
5. An analysis of variance experiment produced a portion of the accompanying ANOVA table.

Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	25.08	3	?	?	0.0004	2.725
Within Groups	92.64	76	?			
Total	117.72	79				

- a. Fill in the missing statistics in the ANOVA table.
  - b. Specify the competing hypotheses in order to determine whether some differences exist between the population means.
  - c. At the 5% significance level, what is the conclusion to the test?
6. An analysis of variance experiment produced a portion of the following ANOVA table.

Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups		5	?	?	?	?
Within Groups	4321.11	54	?			
Total	4869.48	59				

- a. Fill in the missing statistics in the ANOVA table. (You may want to use Excel's F.DIST.RT and F.INV.RT functions to find the  $p$ -value and the critical value.)

- Specify the competing hypotheses in order to determine whether some differences exist between the population means.
- At the 10% significance level, what is the conclusion to the test?

## Applications

- Asian residents in Boston have the highest average life expectancy of any racial or ethnic group—a decade longer than black residents (*The Boston Globe*, August 16, 2010). Suppose sample results indicative of the overall results are as follows.

Asian	Black	Latino	White
$\bar{x}_1 = 83.7$ years	$\bar{x}_2 = 73.5$ years	$\bar{x}_3 = 80.6$ years	$\bar{x}_4 = 79.0$ years
$s_1^2 = 26.3$	$s_2^2 = 27.5$	$s_3^2 = 28.2$	$s_4^2 = 24.8$
$n_1 = 20$	$n_2 = 20$	$n_3 = 20$	$n_4 = 20$

- Construct an ANOVA table. Assume life expectancies are normally distributed.
  - Specify the competing hypotheses to test whether there are some differences in average life expectancies between the four ethnic groups.
  - At the 5% significance level, what is the conclusion to the test?
- A well-known conglomerate claims that its detergent “whitens and brightens better than all the rest.” In order to compare the cleansing action of the top three brands of detergents, 24 swatches of white cloth were soiled with red wine and grass stains and then washed in front-loading machines with the respective detergents. The following whiteness readings were obtained:

Detergent		
1	2	3
84	78	87
79	74	80
87	81	91
85	86	77
94	86	78
89	89	79
89	69	77
83	79	78
$\bar{x}_1 = 86.3$	$\bar{x}_2 = 80.3$	$\bar{x}_3 = 80.9$
$s_1^2 = 20.8$	$s_2^2 = 45.1$	$s_3^2 = 27.3$
$\bar{\bar{x}} = 82.5$		

- Construct an ANOVA table. Assume whiteness readings are normally distributed.
- Specify the competing hypotheses to test whether there are some differences in the average whitening effectiveness of the three detergents.
- At the 5% significance level, what is the conclusion to the test?

- A recent survey by Genworth Financial Inc., a financial-services company, concludes that the cost of long-term care in the United States varies significantly, depending on where an individual lives (*The Wall Street Journal*, May 16, 2009). An economist collects data from the five states with the highest annual costs (Alaska, Massachusetts, New Jersey, Rhode Island, and Connecticut) in order to determine if his sample data are consistent with the survey’s conclusions. The economist provides the following portion of an ANOVA table:

Source of Variation	SS	df	MS	F	p-value
Between Groups	635.0542	4	?	?	?
Within Groups	253.2192	20	?		
Total	888.2734	24			

- Complete the ANOVA table. Assume that long-term care costs are normally distributed. (You may want to use Excel’s F.DIST.RT function to find the  $p$ -value.)
  - Specify the competing hypotheses to test whether some differences exist in the mean long-term care costs in these five states.
  - At the 5% significance level, do mean costs differ?
- An online survey by the Sporting Goods Manufacturers Association, a trade group of sports retailers and marketers, claimed that household income of recreational athletes varies by sport (*The Wall Street Journal*, August 10, 2009). In order to verify this claim, an economist samples five sports enthusiasts participating in each of six different recreational sports and obtains each enthusiast’s income (in \$1,000s), as shown in the accompanying table.

Snorkeling	Sailing	Boardsailing/ Windsurfing	Bowling	On-Road Triathlon	Off-Road Triathlon
90.9	87.6	75.9	79.3	64.5	47.7
86.0	95.0	75.6	75.8	67.2	59.6
93.6	94.6	83.1	79.6	62.8	68.0
98.8	87.2	74.4	78.5	59.2	60.9
98.4	82.5	80.5	73.2	66.5	50.9
$\bar{x}_1 = 93.5$	$\bar{x}_2 = 89.4$	$\bar{x}_3 = 77.9$	$\bar{x}_4 = 77.3$	$\bar{x}_5 = 64.0$	$\bar{x}_6 = 57.4$
$s_1^2 = 28.8$	$s_2^2 = 28.5$	$s_3^2 = 13.8$	$s_4^2 = 7.4$	$s_5^2 = 10.3$	$s_6^2 = 66.4$
$\bar{\bar{x}} = 76.6$					

- Specify the competing hypotheses in order to test the association’s claim.
- Construct an ANOVA table. Assume incomes are normally distributed.
- At the 5% significance level, what is the critical value?
- Do some average incomes differ depending on the recreational sport? Explain.

11. The following Excel output summarizes the results of an analysis of variance experiment in which the treatments were three different hybrid cars and the variable measured was the miles per gallon (mpg) obtained while driving the same route. Assume mpg is normally distributed.

Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	1034.51	2	517.26	19.86	4.49E-07	3.182
Within Groups	1302.41	50	26.05			
Total	2336.92	52				

At the 5% significance level, can we conclude that average mpg differs between the hybrids? Explain.

12. Do energy bills vary dramatically depending on where you live in the United States? Suppose 25 households from four regions in the United States are sampled. The values for the average annual energy bill are shown in the accompanying table and are consistent with those found by The Department of Energy (*Money*, June 2009). A portion of the ANOVA table is also shown.

Region	West	Northeast	Midwest	South
Average Annual Energy Bill	\$1,491	\$2,319	\$1,768	\$1,758

Source of Variation	SS	df	MS	F	p-value
Between Groups	7531769	3	?	?	7.13E-24
Within Groups	3492385	96	?		
Total	11024154	99			

- Complete the ANOVA table. Assume energy costs are normally distributed.
  - At the 1% significance level, can we conclude that average annual energy bills vary by region?
13. Wenton Powersports produces dune buggies. They have three assembly lines, "Razor," "Blazer," and "Tracer," named after the particular dune buggy models produced on those lines. Each assembly line was originally designed using the same target production rate. However, over the years, various changes have been made to the lines. Accordingly, management wishes to determine whether the assembly lines are still operating at the same average hourly production rate. Production data (in dune buggies/hour) for the last eight hours are as follows.

Razor	Blazer	Tracer
11	10	9
10	8	9
8	11	10
10	9	9
9	11	8
9	10	7
13	11	8
11	8	9
$\bar{x}_1 = 10.1$	$\bar{x}_2 = 9.8$	$\bar{x}_3 = 8.6$
$s_1^2 = 2.4$	$s_2^2 = 1.6$	$s_3^2 = 0.8$
$\bar{x} = 9.5$		

- Construct an ANOVA table. Assume production rates are normally distributed.
  - Specify the competing hypotheses to test whether there are some differences in the mean production rates across the three assembly lines.
  - At the 5% significance level, what is the conclusion of the test? What about the 10% significance level?
14. **FILE Fill Volumes.** In the carbonated beverage industry, dispensing pressure can be an important factor in achieving accurate fill volumes. Too little pressure can slow down the dispensing process. Too much pressure can create excess "fizz" and, thus, inaccurate fill volumes. Accordingly, a leading beverage manufacturer wants to conduct an experiment at three different pressure settings to determine if differences exist in the mean fill volumes. Forty bottles with a target fill volume of 12 ounces were filled at each pressure setting, and the resulting fill volumes (in ounces) were recorded. A portion of the data is shown in the accompanying table.

Low Pressure (60 psi)	Medium Pressure (80 psi)	High Pressure (100 psi)
12.00	12.00	11.56
11.97	11.87	11.55
⋮	⋮	⋮
12.00	12.14	11.80

- Construct an ANOVA table.
  - Specify the competing hypotheses to test whether there are differences in the mean fill volumes across the three pressure settings.
  - At the 5% significance level, what is the conclusion of the test? What about the 1% significance level?
15. **FILE Exam Scores.** A statistics instructor wonders whether significant differences exist in her students' final exam scores in her three different sections. She randomly selects the scores from 10 students in each section. A portion of the data is shown in the accompanying table. Assume exam scores are normally distributed.

Section 1	Section 2	Section 3
85	91	74
68	84	69
⋮	⋮	⋮
74	75	73

Do these data provide enough evidence at the 5% significance level to indicate that there are some differences in final exam scores among these three sections?

16. **FILE Nike Revenues.** The accompanying table shows a portion of quarterly data on Nike's revenue for the fiscal years 2001 through 2010. Data for Nike's fiscal year refer to the time period from June 1 through May 31. Assume revenue is normally distributed.

Year	Quarters Ended			
	August 31	November 30	February 28	May 31
2001	2,637	2,199	2,170	2,483
2002	2,614	2,337	2,260	2,682
⋮	⋮	⋮	⋮	⋮
2010	4,799	4,406	4,733	5,077

SOURCE: Annual Reports for Nike, Inc.

Use one-way ANOVA to determine if the data provide enough evidence at the 5% significance level to indicate that there are quarterly differences in Nike's revenue.

17. **FILE Patronage.** The accompanying table shows a portion of the number of customers that frequent a restaurant on weekend days over the past 52 weeks.

Fridays	Saturdays	Sundays
391	450	389
362	456	343
⋮	⋮	⋮
443	441	376

At the 5% significance level, can we conclude that the average number of customers that frequent the restaurant differs by weekend day?

18. **FILE Field Score.** A human resource specialist wants to determine whether the average job satisfaction score (on a scale of 0 to 100) differs depending on a person's field of employment. She collects scores from 30 employees in three different fields. A portion of the data is shown in the accompanying table.

Field 1	Field 2	Field 3
80	76	81
76	73	77
⋮	⋮	⋮
79	67	80

At the 10% significance level, can we conclude that the average job satisfaction differs by field?

## LO 13.2

# 13.2 MULTIPLE COMPARISON METHODS

Use Fisher's LSD method and Tukey's HSD method to determine which means differ.

In the preceding section, we used a one-way ANOVA test to determine whether differences exist between population means. Suppose that for a given sample we reject the null hypothesis of equal means. While the ANOVA test determines that not all population means are equal, it does not indicate which ones differ. To find out which population means differ requires further analysis of the direction and the statistical significance of the difference between paired population means ( $\mu_i - \mu_j$ ). By constructing confidence intervals for all pairwise differences for the population means, we can identify which means significantly differ from one another. The first method we discuss is often referred to as **Fisher's Least Significant Difference (LSD) Method**.

We also introduce an improved method, developed by the renowned 20th-century statistician John Tukey (1915–2000), that identifies “honestly significant differences” between population means; thus, this is often referred to as **Tukey's HSD method**. Note that there are no significant differences to find if the ANOVA test does not reject the null hypothesis of equal means.

## Fisher's Least Significant Difference (LSD) Method

In Chapter 10, we stated that, when the population variances are unknown but assumed equal, the  $100(1 - \alpha)\%$  confidence interval for the difference between two population means  $\mu_i - \mu_j$  is

$$(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2, n_i + n_j - 2} \sqrt{s_p^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}.$$

Here  $s_p^2$  is a pooled estimate of the common population variance, and is computed as

$$s_p^2 = \frac{(n_i - 1)s_i^2 + (n_j - 1)s_j^2}{n_i + n_j - 2}.$$

We substitute the mean square error  $MSE$  from the one-way ANOVA test for  $s_p^2$  since  $MSE$  uses all samples whereas  $s_p^2$  uses only two for the pairwise comparison. Recall that when we conduct a one-way ANOVA test we assume that we are sampling from populations that have the same population variance  $\sigma^2$ . We still apply the  $t_{df}$  distribution, but we use the degrees of freedom corresponding to  $MSE$ , or  $df = n_T - c$ .

#### FISHER'S CONFIDENCE INTERVAL FOR $\mu_i - \mu_j$

Fisher's  $100(1 - \alpha)\%$  confidence interval for the difference between two population means  $\mu_i - \mu_j$  is given by

$$(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2, n_T - c} \sqrt{MSE \left( \frac{1}{n_i} + \frac{1}{n_j} \right)},$$

where the mean square error  $MSE$  is estimated from the one-way ANOVA test.

### EXAMPLE 13.2

Using the sample means and the ANOVA test results from the public transportation example of Section 13.1, calculate 95% confidence intervals for the difference between all possible pairings of the four population means. Comment on the direction and the significance of the differences at the 5% level.

**SOLUTION:** We are given the following sample means and sample sizes:

$$\begin{array}{ll} \bar{x}_{\text{Boston}} = 12,622 & n_{\text{Boston}} = 5 \\ \bar{x}_{\text{New York}} = 12,585 & n_{\text{New York}} = 8 \\ \bar{x}_{\text{San Francisco}} = 11,720 & n_{\text{San Francisco}} = 6 \\ \bar{x}_{\text{Chicago}} = 10,730 & n_{\text{Chicago}} = 5 \end{array}$$

In addition,  $MSE = 7,209$  (derived from the ANOVA test in Section 13.1),  $n_T - c = 24 - 4 = 20$ , and  $t_{\alpha/2, n_T - c} = t_{0.025, 20} = 2.086$ . Table 13.5 shows the 95% confidence intervals.

**TABLE 13.5** Fisher's 95% Confidence Intervals for Example 13.2

Population Mean Differences	Confidence Interval
$\mu_{\text{Boston}} - \mu_{\text{New York}}$	$(12622 - 12585) \pm 2.086 \sqrt{7209 \left( \frac{1}{5} + \frac{1}{8} \right)} = 37 \pm 100.97$ or $[-63.97, 137.97]$
$\mu_{\text{Boston}} - \mu_{\text{San Francisco}}$	$(12622 - 11720) \pm 2.086 \sqrt{7209 \left( \frac{1}{5} + \frac{1}{6} \right)} = 902 \pm 107.25$ or $[794.75, 1009.25]$
$\mu_{\text{Boston}} - \mu_{\text{Chicago}}$	$(12622 - 10730) \pm 2.086 \sqrt{7209 \left( \frac{1}{5} + \frac{1}{5} \right)} = 1892 \pm 112.02$ or $[1779.98, 2004.02]$
$\mu_{\text{New York}} - \mu_{\text{San Francisco}}$	$(12585 - 11720) \pm 2.086 \sqrt{7209 \left( \frac{1}{8} + \frac{1}{6} \right)} = 865 \pm 95.65$ or $[769.35, 960.65]$
$\mu_{\text{New York}} - \mu_{\text{Chicago}}$	$(12585 - 10730) \pm 2.086 \sqrt{7209 \left( \frac{1}{8} + \frac{1}{5} \right)} = 1855 \pm 100.97$ or $[1754.03, 1955.97]$
$\mu_{\text{San Francisco}} - \mu_{\text{Chicago}}$	$(11720 - 10730) \pm 2.086 \sqrt{7209 \left( \frac{1}{6} + \frac{1}{5} \right)} = 990 \pm 107.25$ or $[882.75, 1097.25]$

The 95% confidence interval for  $\mu_{\text{Boston}} - \mu_{\text{New York}}$  is given by  $37 \pm 100.97$ , which is  $[-63.97, 137.97]$ . Since this interval contains the value zero, we cannot reject the null hypothesis, given by  $H_0: \mu_{\text{Boston}} - \mu_{\text{New York}} = 0$ , at the 5% significance level. In other words, the average cost savings from using public transportation in Boston and in New York are not significantly different.

For  $\mu_{\text{Boston}} - \mu_{\text{San Francisco}}$ , the entire interval, ranging from 794.75 to 1,009.25, is above the value zero. Thus, we conclude at the 5% significance level that the average cost savings from using public transportation in Boston are different from the average cost savings in San Francisco. In fact, the remaining intervals are all above zero, suggesting that average cost savings are different between the corresponding cities. In other words, we conclude at the 5% significance level that the average cost savings from using public transportation are different between Boston and Chicago, New York and San Francisco, New York and Chicago, and San Francisco and Chicago.

These pairwise comparisons of the means use Fisher's least significant difference (LSD) method, although the LSD tests are motivated within a hypothesis-testing framework. (Recall that we can conduct two-tailed tests with confidence intervals.) We can apply this method only if the ANOVA test has rejected the null hypothesis of equal means. However, as mentioned earlier, some issues arise when inferring the equality of means by conducting paired tests. In Example 13.2, we have six paired tests. Therefore, if we use the 5% significance level for each test, the probability that we would make a Type I error (incorrectly rejecting a null hypothesis of equal means) on *at least* one of these individual tests will be greater than 5%. The more means we compare, the more the Type I error becomes inflated.

One way to avoid this problem is to perform each individual paired test at a reduced significance level, which ensures that the overall significance level for the equality of all means does not exceed  $\alpha$ . The resulting confidence intervals are wider and hence reduce the probability of incorrectly rejecting the null hypothesis of equal means. However, this technique reduces the power of the test and thus results in an increased risk of a Type II error (incorrectly failing to reject a null hypothesis of equal means).

## Tukey's Honestly Significant Differences (HSD) Method

An improved multiple comparison technique is Tukey's honestly significant differences (HSD) method. The original Tukey's HSD method was introduced with **balanced** data, but it was subsequently modified for **unbalanced** data. If there are an equal number of observations in each sample—that is, when  $n_1 = n_2 = \dots = n_c$ —then the data are balanced. In situations where different numbers of observations occur in each sample—that is, when  $n_i \neq n_j$ —the data are unbalanced. Tukey's method uses the studentized range distribution, which has broader, flatter, and thicker tails than the  $t_{df}$  distribution. In other words, for a given probability under the right tail of the distribution, the studentized range value will be larger than the corresponding  $t_{df}$  value. Therefore, Tukey's HSD method protects against an inflated risk of a Type I error.

### TUKEY'S CONFIDENCE INTERVAL FOR $\mu_i - \mu_j$

Tukey's  $100(1 - \alpha)\%$  confidence interval for the difference between two population means  $\mu_i - \mu_j$  is given by

$$(\bar{x}_i - \bar{x}_j) \pm q_{\alpha, (c, n_T - c)} \sqrt{\frac{MSE}{n}} \quad \text{for balanced data } (n = n_i = n_j), \text{ and}$$

$$(\bar{x}_i - \bar{x}_j) \pm q_{\alpha, (c, n_T - c)} \sqrt{\frac{MSE}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} \quad \text{for unbalanced data } (n_i \neq n_j),$$

where  $q_{\alpha, (c, n_T - c)}$  is the studentized range value.



The studentized range value  $q_{\alpha,(c,n_T-c)}$  varies with the significance level  $\alpha$ , the number of populations  $c$ , and  $n_T - c$ . Table 13.6 shows a portion of the studentized range table; Table 5 in Appendix A provides a more comprehensive table. For example, with  $\alpha = 0.05$ ,  $c = 6$ , and  $n_T - c = 19$ , we find  $q_{0.05,(6,19)} = 4.47$ . With  $\alpha = 0.01$ ,  $c = 3$ , and  $n_T - c = 20$ , we find  $q_{0.01,(3,20)} = 4.64$ . These values are in boldface in Table 13.6.

**TABLE 13.6** Portion of Values for  $q_{\alpha,(c,n_T-c)}$  in Tukey's HSD Method

$n_T - c$	$\alpha$	$c = \text{number of means}$							
		2	3	4	5	6	7	8	9
19	0.05	2.96	3.59	3.98	4.25	<b>4.47</b>	4.65	4.79	4.92
	0.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02
20	0.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90
	0.01	4.02	<b>4.64</b>	5.02	5.29	5.51	5.69	5.84	5.97

#### FISHER'S LSD METHOD VERSUS TUKEY'S HSD METHOD

When using Fisher's LSD method at some stated significance level  $\alpha$ , the probability of committing a Type I error increases as the number of pairwise comparisons increases; that is, the likelihood of incorrectly rejecting the null hypothesis of equal means in at least one of the pairwise comparisons is greater than  $\alpha$ . By using the studentized range distribution over the  $t_{df}$  distribution, Tukey's HSD method ensures that the probability of a Type I error equals  $\alpha$ , irrespective of the number of pairwise comparisons.

#### EXAMPLE 13.3

A consumer advocate in California is concerned with the price of a common generic drug. Specifically, he feels that one region of the state has significantly different prices for the drug than two other regions. He divides the state into three regions and collects the generic drug's price from 10 pharmacies in each region. He produces the summary statistics and ANOVA results shown in Table 13.7.

**TABLE 13.7** Summary Statistics and ANOVA Results for Example 13.3

SUMMARY					
Groups	Count	Sum	Average	Variance	
Region 1	10	350	35.0	3.78	
Region 2	10	342	34.2	5.07	
Region 3	10	395	39.5	2.72	

ANOVA						
Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	163.27	2	81.63	21.17	2.95E-06	3.35
Within Groups	104.10	27	3.86			
Total	267.37	29				

- a. At the 5% significance level, do differences exist between the mean drug prices in the three regions?
- b. If significant differences exist, use Tukey's HSD method to determine which regions' means differ at the 5% significance level.

**SOLUTION:**

- a. In order to test differences between the mean drug prices in the three regions, we specify the competing hypotheses as

$$H_0: \mu_{\text{Region1}} = \mu_{\text{Region2}} = \mu_{\text{Region3}}$$

$$H_A: \text{Not all mean drug prices are equal.}$$

The ANOVA table shows the value of the test statistic as  $F_{(2,27)} = 21.17$  with a  $p$ -value of  $2.95 \times 10^{-6}$ , or  $P(F_{2,27} \geq 21.17) \approx 0$ . Since the  $p$ -value is less than the significance level of 0.05, we reject  $H_0$  and conclude that not all mean drug prices are equal.

- b. We use Tukey's confidence interval for balanced data since each sample size is the same ( $n_1 = n_2 = n_3 = 10$ ). Given  $\alpha = 0.05$ ,  $c = 3$ , and  $n_T - c = 30 - 3 = 27$ , we find  $q_{\alpha, (c, n_T - c)} = q_{0.05, (3, 27)} = 3.51$ . We use  $(\bar{x}_i - \bar{x}_j) \pm q_{\alpha, (c, n_T - c)} \sqrt{\frac{MSE}{n}}$  to compute 95% confidence intervals for all pairwise differences of the means. The results are shown in Table 13.8.

**TABLE 13.8** Tukey's 95% Confidence Intervals for Example 13.3

Population Mean Differences	Confidence Interval
$\mu_1 - \mu_2$	$(35.0 - 34.2) \pm 3.51 \sqrt{\frac{3.86}{10}}$ or $[-1.38, 2.98]$
$\mu_1 - \mu_3$	$(35.0 - 39.5) \pm 3.51 \sqrt{\frac{3.86}{10}}$ or $[-6.68, -2.32]^*$
$\mu_2 - \mu_3$	$(34.2 - 39.5) \pm 3.51 \sqrt{\frac{3.86}{10}}$ or $[-7.48, -3.12]^*$

The asterisk \* shows that the confidence interval does not include the value zero, thus indicating that the corresponding means are different at the 5% significance level. The consumer advocate's claim is supported by the data. At the 5% significance level, the average price of generic drugs in region 3 is different from the average prices in regions 1 and 2. At the 5% significance level, the consumer advocate cannot conclude that average prices in regions 1 and 2 differ.

We also employed Tukey's method for unbalanced data using the ANOVA results from the public transportation example. While the resulting intervals (not reported) became wider than those reported in Table 13.5, the inference regarding the population means remains the same.

## SYNOPSIS OF INTRODUCTORY CASE

A recent report by the American Public Transportation Association suggests that commuters who use public transportation can save a substantial amount of money annually. Sean Cox, a research analyst at an environmental firm, conducts a survey to determine whether average cost savings differ depending on where the commuters reside. He collects data on public transit riders in the top four cost-savings cities: Boston, New York, San Francisco, and Chicago. Table 13.9 shows summary statistics and relevant ANOVA results.



**TABLE 13.9** Summary Statistics and Relevant ANOVA Results

Boston	New York	San Francisco	Chicago
$\bar{x}_1 = \$12,622$	$\bar{x}_2 = \$12,585$	$\bar{x}_3 = \$11,720$	$\bar{x}_4 = \$10,730$
$s_1 = \$87.79$	$s_2 = \$80.40$	$s_3 = \$83.96$	$s_4 = \$90.62$
$n_1 = 5$	$n_2 = 8$	$n_3 = 6$	$n_4 = 5$
Mean Square Error (MSE): 7,209			
Calculated $F$ -statistic and its $p$ -value: 610.6 and 0.0000, respectively.			

Sean reports that in all four major cities the sample average cost savings are above \$10,000. Since the  $p$ -value of the ANOVA test is close to zero, he concludes that there are differences in cost savings between the cities at the 5% significance level. Sean also constructs confidence intervals for all pairwise differences to identify pairings of cities that have statistically different cost savings. Commuters in Boston and New York have the highest cost savings; however, at the 5% significance level, their average cost savings do not differ from one another. For every other pairing, average cost savings from using public transportation are statistically different between cities. In other words, he concludes at the 5% significance level that average cost savings from using public transportation are different between Boston and San Francisco, Boston and Chicago, New York and San Francisco, New York and Chicago, and San Francisco and Chicago.

## EXERCISES 13.2

### Mechanics

19. The following statistics are computed by sampling from three normal populations whose variances are equal:  
 $\bar{x}_1 = 25.3$ ,  $n_1 = 8$ ;  $\bar{x}_2 = 31.5$ ,  $n_2 = 10$ ;  $\bar{x}_3 = 32.3$ ,  $n_3 = 6$ ;  
 $MSE = 27.2$ 
  - a. Calculate 95% confidence intervals for  $\mu_1 - \mu_2$ ,  $\mu_1 - \mu_3$ , and  $\mu_2 - \mu_3$  using Fisher's LSD approach.
  - b. Repeat the analysis with Tukey's HSD approach.
  - c. Which of these two approaches would you use to determine whether differences exist between the population means? Explain.
20. The following statistics are calculated by sampling from four normal populations whose variances are equal:  
 $\bar{x}_1 = 149$ ,  $n_1 = 10$ ;  $\bar{x}_2 = 154$ ,  $n_2 = 10$ ;  $\bar{x}_3 = 143$ ,  $n_3 = 10$ ;  
 $\bar{x}_4 = 139$ ,  $n_4 = 10$ ;  $MSE = 51.3$ 
  - a. Use Fisher's LSD method to determine which population means differ at  $\alpha = 0.01$ .

- b. Use Tukey's HSD method to determine which population means differ at  $\alpha = 0.01$ .
  - c. Do all population means differ? Explain.
21. A one-way analysis of variance experiment produced the following ANOVA table.

SUMMARY						
Groups	Count		Average			
Column 1	6		0.57			
Column 2	6		1.38			
Column 3	6		2.33			
Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	9.12	2	4.56	12.84	0.0006	3.68
Within Groups	5.33	15	0.36			
Total	14.46	17				

- Conduct an ANOVA test at the 5% significance level to determine if some population means differ.
  - Calculate 95% confidence interval estimates for  $\mu_1 - \mu_2$ ,  $\mu_1 - \mu_3$ , and  $\mu_2 - \mu_3$  with Tukey's HSD approach.
  - Given your response to part (b), which means significantly differ?
22. A one-way analysis of variance experiment produced the following ANOVA table.

SUMMARY						
Groups	Count			Average		
Column 1	10			349		
Column 2	10			348		
Column 3	10			366		
Column 4	10			365		
Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	2997.11	3	999.04	15.54	1.2E-06	2.866
Within Groups	2314.71	36	64.30			
Total	5311.82	39				

- Use Fisher's LSD method to determine which means differ at the 5% level of significance.
  - Use Tukey's HSD method to determine which means differ at the 5% level of significance.
  - Given your responses to parts (a) and (b), do the population means differ at the 5% significance level?
23. The following Excel output summarizes the results for a one-way analysis of variance experiment in which the treatments were three different hybrid cars and the variable measured was the miles per gallon (mpg) obtained while driving the same route.

Hybrid 1: $\bar{x}_1 = 38$ , $n_1 = 20$
Hybrid 2: $\bar{x}_2 = 48$ , $n_2 = 15$
Hybrid 3: $\bar{x}_3 = 39$ , $n_3 = 18$

Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	1034.51	2	517.26	19.86	4.49E-07	3.182
Within Groups	1302.41	50	26.05			
Total	2336.92	52				

- At the 5% significance level, can we conclude that average mpg differs between the hybrids?
- If significant differences exist, use Tukey's HSD method at the 5% significance level to determine which hybrids' means differ.

## Applications

24. In an attempt to improve efficiency, Starbucks has implemented "lean" Japanese techniques at many of its 11,000 U.S. stores (*The Wall Street Journal*, August 4, 2009). By reducing the time baristas (employees) spend on bending, reaching, and walking, they will have more time to interact with customers and improve the Starbucks experience. Suppose Starbucks adopts the lean technique at Store 1, but makes no changes at Stores 2 and 3. On a recent Monday morning between the hours of 7:00 AM and 8:00 AM, the following statistics were obtained relating to average time per order (in seconds):

Store 1: $\bar{x}_1 = 56$ , $n_1 = 18$
Store 2: $\bar{x}_2 = 66$ , $n_2 = 12$
Store 3: $\bar{x}_3 = 63$ , $n_3 = 14$

Excel produced the following ANOVA table:

Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	811.70	2	405.85	52.11	5.5E-12	3.226
Within Groups	319.30	41	7.79			
Total	1131.00	43				

- Compute 95% confidence interval estimates for all paired differences for the means using Fisher's LSD approach.
  - Repeat the analysis with Tukey's HSD approach.
  - Which of these two approaches is more reliable? Explain.
  - Do the data suggest that the lean technique is improving efficiency? Explain.
25. Do energy bills vary dramatically depending on where you live in the United States? Suppose 25 households from four regions in the United States are sampled. The values for the average annual energy bill are shown below and are consistent with those found by The Department of Energy (*Money*, June 2009).

Region	West	Northeast	Midwest	South
Average Annual Energy Bill	\$1,491	\$2,319	\$1,768	\$1,758

A portion of the ANOVA calculations are below:

Source of Variation	SS	df	MS	F	p-value
Between Groups	7531769	3	?	?	71 3E-24
Within Groups	3492385	96	?		
Total	11024154	99			

- Complete the ANOVA table.
- At the 1% significance level, can we conclude that average annual energy bills vary by region?
- If significant differences exist, use Tukey's HSD method at the 1% significance level to determine which regions' means differ.

26. Elastotech, a plastics company, is trying to determine whether four successive daily batches of their Lexan polycarbonate have the same mean hardness value. Four daily samples of polycarbonate were taken, and the resulting hardness values were measured (using the Rockwell *R* scale). The following Excel output was obtained.

Day 1	Day 2	Day 3	Day 4
$\bar{x}_1 = 115.75$	$\bar{x}_2 = 108.00$	$\bar{x}_3 = 121.39$	$\bar{x}_4 = 119.53$
$n_1 = 20$	$n_2 = 16$	$n_3 = 18$	$n_4 = 17$

Source of Variation	SS	df	MS	F	p-value	F crit at 1%
Between Groups	1751.62	3	583.87	19.29	0.000	4.09
Within Groups	2028.26	67	30.27			
Total	3779.89	70				

- At the 1% significance level, can we conclude that the mean hardness differs among the four daily batches?
  - If significant differences exist, use Tukey's HSD method at the 1% significance level to determine which batches have different mean hardness values.
27. Producers of a new grass seed called Pearl's Premium claim that grass grown using its seed blend requires less maintenance as compared to other brands (*The Boston Globe*, July 4, 2009). For instance, grass grown using Pearl's Premium needs mowing only once a month. Suppose an independent tester wants to test whether the average height of grass after one month's growth is the same between Pearl's Premium and the other two top-selling brands. The independent tester measures 25 grass blades using each of the three seeds (grass blades are measured in inches). Using Excel, he constructs the following ANOVA table with supporting descriptive statistics.

SUMMARY		
Groups	Count	Average
Pearl's Premium	25	4.83
Top Brand 1	25	6.50
Top Brand 2	25	6.99

Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	64.43	2	32.21	121.67	8.09E-24	3.123
Within Groups	19.06	72	0.26			
Total	83.49	74				

- At the 5% significance level, can we conclude that the average heights of grass blades differ by brand?
  - If significant differences exist, use Tukey's HSD method at the 5% significance level to determine which brands differ.
28. **FILE Employee\_Absences.** A production manager is examining whether work shift is related to employee absenteeism. The number of days absent over the past year was tallied for random samples of 25 workers on each shift. A portion of the data is shown in the accompanying table.

First Shift	Second Shift	Third Shift
14	5	19
7	10	9
⋮	⋮	⋮
2	9	10

- Construct an ANOVA table.
  - At the 5% significance level, can we conclude that the mean days of absenteeism differ among the three shifts? Does the conclusion change at the 10% significance level?
  - If significant differences exist, use Fisher's LSD method at the 10% significance level to determine which shifts have different mean days of absenteeism.
29. **FILE Patronage.** The accompanying table shows a portion of the number of customers that ate at a restaurant on weekend days over the past 52 weeks.

Fridays	Saturdays	Sundays
391	450	389
362	456	343
⋮	⋮	⋮
443	441	376

- Verify that the average number of customers that frequent the restaurant differs by weekend day at the 5% significance level.
- Use Tukey's HSD method at the 5% significance level to determine which weekend days differ.

Conduct and evaluate a two-way ANOVA test with no interaction.

One-way ANOVA tests are used to compare population means based on one categorical variable or factor. For instance, we can use a one-way ANOVA test to determine whether differences exist in average miles per gallon depending on the brand name of hybrid cars. Two-way ANOVA tests extend the analysis by measuring the effects of two factors simultaneously. Suppose we want to determine if the brand of a hybrid car and the octane level of gasoline influence average miles per gallon. Whereas one-way ANOVA tests are able to assess either the brand effect or the octane-level effect in isolation, two-way ANOVA tests are able to assess the effect of a factor while controlling for the other factor. The additional factor explains some of the unexplained variation in miles per gallon, or equivalently, reduces the sum of squares due to error for a more discriminating  $F_{(df_1, df_2)}$  test statistic.

Another feature of two-way ANOVA is that it can be extended to capture the interaction between the factors. In the above example, if we believe that some brands of a hybrid car react more positively to the octane levels than others, then we can include the interaction of these factors in examining miles per gallon. We use tests that determine whether the factors do indeed interact.

**Two-way ANOVA** tests are used to simultaneously examine the effect of two factors on the mean. These tests can be conducted with or without the interaction of the factors.

In the following example, we initially conduct a one-way ANOVA test and quickly recognize its limitations. We then introduce a two-way ANOVA test without interaction. In Section 13.4, we discuss a two-way ANOVA test with interaction.

#### EXAMPLE 13.4

Julia Hayes is an undergraduate who is completely undecided as to what career she should pursue. To help in her decision process, she wants to determine whether or not there are significant differences in annual incomes depending on the field of employment. Initially, she confines her analysis to the following three fields: educational services, financial services, and medical services. As a preliminary experiment, she surveys four workers from each of these three fields and asks how much he/she earns annually. Table 13.10 shows the results (in \$1,000s) from the experiment.

**TABLE 13.10** Data for Example 13.4

Educational Services	Financial Services	Medical Services
35	58	110
18	90	62
75	25	26
46	45	43

Conduct a one-way ANOVA test to determine if differences exist among the fields' average incomes at the 5% significance level.

**SOLUTION:** Table 13.11 shows the relevant results from implementing a one-way ANOVA test.

#### FILE

*One-Factor\_Income*



**TABLE 13.11** ANOVA Results for Example 13.4

ANOVA						
Source of Variation	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i> -value	<i>F</i> crit
Between Groups	579.5	2	289.75	0.32998	0.727281	4.2565
Within Groups	7902.75	9	878.0833			
Total	8482.25	11				

In order to determine whether mean incomes differ by field of employment, we specify the following hypotheses:

$$H_0: \mu_{\text{Education}} = \mu_{\text{Financial}} = \mu_{\text{Medical}}$$

$H_A$ : Not all population means are equal.

The value of the test statistic is  $F_{(2,9)} = 0.33$ . Using the critical value approach, the decision rule is to reject  $H_0$  if  $F_{(2,9)} > 4.26$ . We do not reject the null hypothesis since the value of the test statistic does not fall in the rejection region (0.33 is not greater than 4.26). At the 5% significance level, we cannot conclude that average incomes differ by field.

Julia is surprised by these results, since she feels that those in the educational services industry probably earn less than those in the other two fields. Julia is advised that she must interpret these results with caution because many other factors influence annual income—one of which is an individual's educational attainment. We can capture the true influence of field of employment on income only when educational attainment is held fixed.

As mentioned earlier, a two-way ANOVA test helps us find a more discriminating  $F_{(df_1, df_2)}$  test statistic, since the additional factor reduces the resulting sum of squares due to error. An added requirement for a two-way ANOVA test is that all groups must have the same sample size.

To show how a two-way ANOVA test works, we redo the analysis from Example 13.4, but this time we allow the variation in income to be affected by field (factor *A*) and educational attainment (factor *B*). We match a worker from each field according to his or her highest educational attainment. For example, we randomly select a worker from the educational services industry whose highest educational attainment is a high school degree. We then randomly select three more workers from this field whose highest educational attainment is a bachelor's degree, a master's degree, and a Ph.D. (or its equivalent), respectively. We repeat this process for the other two fields. The outcomes in this experiment are matched or blocked in the sense that one worker is randomly selected from each field of employment depending on his/her educational attainment. In general, blocks are the levels at which we hold an extraneous factor fixed, so that we can measure its contribution to the total variation of the variable. This experimental design is called a **randomized block design**. The experiment in this example is designed to eliminate the variability in income attributable to differences in educational attainment.

Table 13.12 shows the incomes (in \$1,000s) for 12 workers according to their field of employment and highest educational attainment. Also included in the table are the factor means.

**TABLE 13.12** Data for Two-Factor Income Example (No Interaction)

Education Level (Factor <i>B</i> )	Field of Employment (Factor <i>A</i> )			Factor B Means
	Educational Services	Financial Services	Medical Services	
High School	18	25	26	$\bar{x}_{\text{high school}} = 23.00$
Bachelor's	35	45	43	$\bar{x}_{\text{bachelor's}} = 41.00$
Master's	46	58	62	$\bar{x}_{\text{master's}} = 55.33$
Ph.D.	75	90	110	$\bar{x}_{\text{Ph.D.}} = 91.67$
<b>Factor A Means</b>	$\bar{x}_{\text{education}} = 43.50$	$\bar{x}_{\text{financial}} = 54.50$	$\bar{x}_{\text{medical}} = 60.25$	$\bar{\bar{x}} = 52.75$

**FILE***Two-Factor\_Income*

The goal of the analysis is to answer the following two questions:

- A. At the 5% significance level, do average annual incomes differ by field of employment?
- B. At the 5% significance level, do average annual incomes differ by level of educational attainment?

A one-way ANOVA test is based on one factor for which we used the notation “sum of squares due to treatments  $SSTR$ ” to capture the variability *between* the levels of this factor. Since we are now examining two factors, we use the notation  $SSA$  to capture the variability *between* the levels of factor  $A$  and  $SSB$  to capture the variability *between* the levels of factor  $B$ .

#### A TWO-WAY ANOVA TEST WITHOUT INTERACTION

In a **two-way ANOVA test without interaction**, the total sum of squares,  $SST$ , of the variable is partitioned into three distinct components: the sum of squares for factor  $A$ ,  $SSA$ ; the sum of squares for factor  $B$ ,  $SSB$ ; and the sum of squares due to error,  $SSE$ . That is,  $SST = SSA + SSB + SSE$ .

### The Sum of Squares for Factor $A$ , $SSA$

We calculate the sum of squares for factor  $A$ ,  $SSA$ , as we did before; that is, we first calculate the sum of the squared differences between the mean for each level of factor  $A$  and the grand mean. We then multiply this sum by the number of rows in the randomized block design  $r$ . For this example,  $r$  equals four. We calculate  $SSA$  as

$$\begin{aligned} SSA &= r \sum_{i=1}^c (\bar{x}_i - \bar{\bar{x}})^2 \\ &= 4[(43.50 - 52.75)^2 + (54.50 - 52.75)^2 + (60.25 - 52.75)^2] \\ &= 579.50. \end{aligned}$$

Dividing  $SSA$  by its degrees of freedom,  $c - 1$ , (where  $c$  is the number of columns in the randomized block design) yields the **mean square for factor  $A$** ,  $MSA$ . For this example,  $c$  equals three, so  $MSA$  is

$$MSA = \frac{SSA}{c - 1} = \frac{579.50}{3 - 1} = 289.75.$$

### The Sum of Squares for Factor $B$ , $SSB$

In order to obtain the sum of squares for factor  $B$ ,  $SSB$ , we calculate the sum of the squared differences between the mean for each level of factor  $B$  and the grand mean. We multiply this sum by  $c$ ; thus, we calculate  $SSB$  as

$$\begin{aligned} SSB &= c \sum_{j=1}^r (x_j - \bar{\bar{x}})^2 \\ &= 3[(23.00 - 52.75)^2 + (41.00 - 52.75)^2 + (55.33 - 52.75)^2 + (91.67 - 52.75)^2] \\ &= 7633.64. \end{aligned}$$

Dividing  $SSB$  by its degrees of freedom,  $r - 1$ , yields the **mean square for factor  $B$** ,  $MSB$ , or

$$MSB = \frac{SSB}{r - 1} = \frac{7633.64}{4 - 1} = 2544.55.$$

### The Sum of Squares Due to Error, $SSE$

As mentioned earlier, in a two-way ANOVA test without interaction,  $SST = SSA + SSB + SSE$ . We can calculate  $SSE$  by rewriting this expression as  $SSE = SST - (SSA + SSB)$ .

We calculate  $SST$  as the sum of squared differences between each data point and the grand mean, or equivalently,  $SST = \sum_{i=1}^c \sum_{j=1}^r (x_{ij} - \bar{\bar{x}})^2$ . For this example, we calculate  $SST$  as

$$\begin{aligned} SST &= \sum_{i=1}^c \sum_{j=1}^r (x_{ij} - \bar{\bar{x}})^2 \\ &= (18 - 52.75)^2 + (35 - 52.75)^2 + \cdots + (110 - 52.75)^2 \\ &= 8482.25. \end{aligned}$$

We then compute  $SSE$  as

$$SSE = SST - (SSA + SSB) = 8482.25 - (579.50 + 7633.64) = 269.11.$$

We can make some generalizations about the difference in the magnitudes of the  $SSE$  values for the one-way ANOVA versus the two-way ANOVA examples. When we used one factor (field of employment) to explain annual incomes, the value of  $SSE$  was 7,902.75 (see Table 13.11). By ignoring the second factor (level of educational attainment), we could not establish that annual incomes were different by field of employment. However, once we include this second factor, the value of  $SSE$  declines dramatically to 269.11. We will show shortly that by accounting for the effect of educational attainment on income, the  $F_{(df_1, df_2)}$  test allows Julia to conclude that significant differences do exist among annual incomes by field of employment.

Dividing  $SSE$  by its degrees of freedom ( $n_T - c - r + 1$ ) yields the **mean square error**,  $MSE$ , or

$$MSE = \frac{SSE}{n_T - c - r + 1} = \frac{269.11}{12 - 3 - 4 + 1} = 44.85.$$

Most software packages easily provide these statistics. Table 13.13 shows the general format of an ANOVA table when conducting a two-way ANOVA test without interaction.

**TABLE 13.13** General Format of ANOVA Table for Randomized Block Design

Source of Variation	$SS$	$df$	$MS$	$F$
Rows	$SSB$	$r - 1$	$MSB = \frac{SSB}{r - 1}$	$F_{(df_1, df_2)} = \frac{MSB}{MSE}$
Columns	$SSA$	$c - 1$	$MSA = \frac{SSA}{c - 1}$	$F_{(df_1, df_2)} = \frac{MSA}{MSE}$
Error	$SSE$	$n_T - c - r + 1$	$MSE = \frac{SSE}{n_T - c - r + 1}$	
Total	$SST$	$n_T - 1$		

Table 13.13 shows values for two  $F_{(df_1, df_2)}$  test statistics. We use the first statistic ( $F_{(df_1, df_2)} = \frac{MSB}{MSE}$ , where  $df_1 = r - 1$ ,  $df_2 = n_T - c - r + 1$ ) to test whether significant differences exist between the factor  $B$  means. We use the second statistic ( $F_{(df_1, df_2)} = \frac{MSA}{MSE}$ , where  $df_1 = c - 1$ ,  $df_2 = n_T - c - r + 1$ ) to test whether significant differences exist between the factor  $A$  means.

## Using Excel to Solve a Two-Way ANOVA Test without Interaction

**FILE**  
*Two-Factor\_Income*

In order to reproduce our manual calculations for the two-factor income example using Excel, we follow these steps.

- A. Open the *Two-factor\_Income* data file.
- B. From the menu choose **Data > Data Analysis > ANOVA: Two Factor Without Replication**.

- C. In the *ANOVA: Two Factor Without Replication* dialog box shown in Figure 13.3, choose the box next to *Input range* and then select all the data, including the labels. Check the *Labels* box. If testing at a significance level other than 5%, insert the relevant significance level in the box next to *Alpha*. Choose an output range and click **OK**.

**FIGURE 13.3**  
Excel's ANOVA:  
Two-Factor Without  
Replication dialog box.

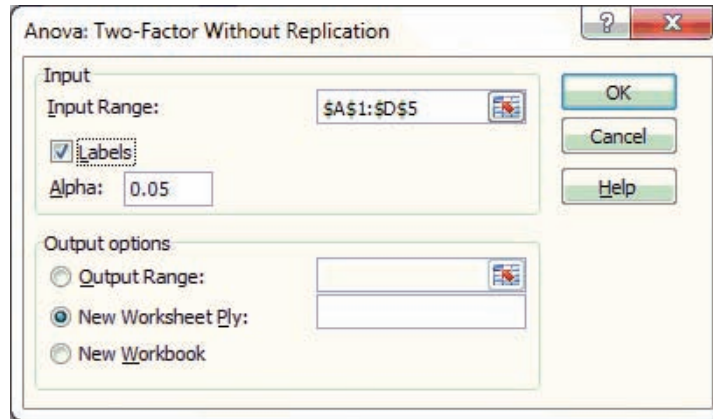


Table 13.14 shows a portion of the results. Any differences between our manual calculations and the values that appear in the table are due to rounding. Note that Excel provides precise  $p$ -values for the calculated test statistics and critical values at the 5% significance level. (This is the value of  $\alpha$  specified in Excel's dialog box.)

**TABLE 13.14** Excel's ANOVA Output for the Two-Factor Income Example

Source of Variation	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i> -value	<i>F</i> crit
Rows	7632.92	3	2544.31	56.58	8.6E-05	4.76
Columns	579.50	2	289.75	6.44	0.03207	5.14
Error	269.83	6	44.97			
Total	8482.25	11				

### EXAMPLE 13.5

Use the Excel output from Table 13.14 to conduct the following hypothesis tests.

- At the 5% significance level, do average annual incomes differ by field of employment?
- At the 5% significance level, do average annual incomes differ by level of educational attainment?

#### SOLUTION:

- Using the critical value approach, we determine whether average annual incomes differ by field of employment. The competing hypotheses are

$$H_0: \mu_{\text{Education}} = \mu_{\text{Financial}} = \mu_{\text{Medical}}$$

$$H_A: \text{Not all population means are equal.}$$

When testing whether the factor  $A$  (column) means differ, the value of the test statistic is  $F_{(df_1, df_2)} = \frac{MSA}{MSE} = \frac{289.75}{44.97} = 6.44$  where  $df_1 = c - 1 = 3 - 1 = 2$  and  $df_2 = n_T - c - r + 1 = 12 - 3 - 4 + 1 = 6$ ; that is,  $F_{(2,6)} = 6.44$ . At the 5%

significance level, the critical value is  $F_{0.05,(2,6)} = 5.14$ . The decision rule is to reject  $H_0$  if  $F_{(2,6)} > 5.14$ . We reject  $H_0$  because the value of the test statistic is greater than the critical value. Therefore, contrary to the results derived earlier with a one-way ANOVA test, average annual salaries do differ by field of employment at the 5% significance level.

- b. Using the  $p$ -value approach, we determine whether average annual incomes differ by level of educational attainment. The competing hypotheses are

$$H_0: \mu_{\text{High School}} = \mu_{\text{Bachelor's}} = \mu_{\text{Master's}} = \mu_{\text{Ph.D.}}$$

$$H_A: \text{Not all population means are equal.}$$

When testing whether the factor  $B$  (row) means differ, the value of the test statistic is  $F_{(df_1, df_2)} = \frac{MSB}{MSE} = \frac{2544.31}{44.97} = 56.58$  where  $df_1 = r - 1 = 4 - 1 = 3$  and  $df_2 = n_T - c - r + 1 = 12 - 3 - 4 + 1 = 6$ ; that is,  $F_{(3,6)} = 56.58$ . From Table 13.14, the  $p$ -value associated with this test statistic is  $8.6 \times 10^{-5}$ , or  $P(F_{(3,6)} \geq 56.58) \approx 0.0000$ . We reject  $H_0$  because the  $p$ -value is less than the significance level of 5%. We conclude at the 5% significance level that average annual salaries also differ by level of educational attainment. Since level of educational attainment exerts a significant influence on salaries, it must be incorporated in ANOVA testing.

We would like to point out that, analogous to the last section, we can apply the  $MSE$  estimate from the two-way ANOVA test to construct useful confidence intervals for the paired differences in population means using Fisher's LSD method. The only significant modification to these confidence intervals is with respect to degrees of freedom, which are now given by  $df = n_T - c - r + 1$ . Similarly, we can also use Tukey's  $HSD$  method to determine which column means or row means are significantly different from one another. The value for the margin of error in the confidence interval will depend on whether we are assessing differences between the column means or the row means. When constructing the confidence interval for the difference between two column means, we calculate the margin of error as  $q_{\alpha,(c,n_T-c)} \sqrt{\frac{MSE}{n}}$ , where  $n$  is the number of observations in each column. When constructing the confidence interval for the difference between two row means, we calculate the margin of error as  $q_{\alpha,(r,n_T-r)} \sqrt{\frac{MSE}{n}}$ , where  $n$  is the number of observations in each row.

## EXERCISES 13.3

### Mechanics

30. The following observations were obtained when conducting a two-way ANOVA experiment with no interaction.

Factor B	Factor A			$\bar{x}_j$ for Factor B
	1	2	3	
1	5	15	12	10.6
2	2	10	8	6.7
3	0	-9	-2	-3.7
4	-3	-14	-8	-8.3
$\bar{x}_i$ for Factor A	1.0	0.5	2.5	$\bar{x} = 1.33$

- Calculate  $SST$ ,  $SSA$ ,  $SSB$ , and  $SSE$ .
- Calculate  $MSA$ ,  $MSB$ , and  $MSE$ .
- Construct an ANOVA table.

- At a 1% significance level, can you conclude that the column means differ?
- At a 1% significance level, can you conclude that the row means differ? In other words, is blocking necessary?

31. The following observations were obtained when conducting a two-way ANOVA experiment with no interaction.

Factor B	Factor A				$\bar{x}_j$ for Factor B
	1	2	3	4	
1	2	3	2	4	2.8
2	6	5	7	6	6.0
3	8	10	9	10	9.3
$\bar{x}_i$ for Factor A	5.3	6.0	6.0	6.7	$\bar{x} = 6$

- Calculate  $SST$ ,  $SSA$ ,  $SSB$ , and  $SSE$ .
- Calculate  $MSA$ ,  $MSB$ , and  $MSE$ .
- Construct an ANOVA table.
- At the 5% significance level, do the levels of Factor  $B$  differ?
- At the 5% significance level, do the levels of Factor  $A$  differ?

32. A two-way analysis of variance experiment with no interaction is conducted. Factor  $A$  has four levels (columns) and Factor  $B$  has three levels (rows). The results include the following sum of squares terms:

$$SST = 1,630.7 \quad SSB = 532.3 \quad SSE = 374.5$$

- Construct an ANOVA table.
  - At the 1% significance level, can you conclude that the factor  $A$  means differ?
  - At the 1% significance level, can you conclude that the factor  $B$  means differ?
33. A two-way analysis of variance experiment with no interaction is conducted. Factor  $A$  has three levels (columns) and Factor  $B$  has five levels (rows). The results include the following sum of squares terms:

$$SST = 311.7 \quad SSA = 201.6 \quad SSE = 69.3$$

- Construct an ANOVA table.
  - At the 5% significance level, can you conclude that the row means differ?
  - At the 5% significance level, can you conclude that the column means differ?
34. The following table summarizes a portion of the results for a two-way analysis of variance experiment with no interaction.

Source of Variation	$SS$	$df$	$MS$	$F$	$p$ -value	$F$ crit
Rows	1057	5	$MSB = ?$	$F_{\text{Factor B}} = ?$	0.0064	3.326
Columns	7	2	$MSA = ?$	$F_{\text{Factor A}} = ?$	0.9004	4.103
Error	330	10	$MSE = ?$			
Total	1394	17				

- Find the missing values in the ANOVA table.
  - At the 5% significance level, can you conclude that the column means differ?
  - At the 5% significance level, can you conclude that the row means differ?
35. The following table summarizes a portion of the results for a two-way analysis of variance experiment with no interaction.

Source of Variation	$SS$	$df$	$MS$	$F$	$p$ -value	$F$ crit
Rows	25.17	2	$MSB = ?$	$F_{\text{Factor B}} = ?$	0.0832	5.143
Columns	142.25	3	$MSA = ?$	$F_{\text{Factor A}} = ?$	0.0037	4.757
Error	19.50	6	$MSE = ?$			
Total	186.92	11				

- Find the missing values in the ANOVA table.
- At the 5% significance level, can you conclude that the column means differ?
- At the 5% significance level, can you conclude that the row means differ?

## Applications

36. During a typical Professional Golf Association (PGA) tournament, the competing golfers play four rounds of golf, where the hole locations are changed for each round. Here are the scores for the top five finishers at the 2009 U.S. Open.

Golfer	Round			
	1	2	3	4
Lucas Glover	69	64	70	73
Phil Mickelson	69	70	69	70
David Duval	67	70	70	71
Ricky Barnes	67	65	70	76
Ross Fisher	70	78	79	72
Grand mean: $\bar{x} = 70.45$				

The following statistics were computed:

$$SST = 272.95 \quad SSB = 93.2 \quad SSE = 127.6$$

- Construct the ANOVA table.
  - At the 5% significance level, can you conclude that the average scores produced by the four different rounds differ?
  - At the 5% significance level, can you conclude that the average scores produced by the five different players differ?
37. The following output summarizes a portion of the results for a two-way analysis of variance experiment with no interaction. Factor  $A$  consists of four different kinds of organic fertilizers, factor  $B$  consists of three different kinds of soil acidity levels, and the variable measured is the height (in inches) of a plant at the end of four weeks.

Source of Variation	$SS$	$df$	$MS$	$F$	$p$ -value	$F$ crit
Rows	0.13	2	$MSB = ?$	$F_{\text{Factor B}} = ?$	0.8182	5.143
Columns	44.25	3	$MSA = ?$	$F_{\text{Factor A}} = ?$	0.0001	4.757
Error	1.88	6	$MSE = ?$			
Total	46.25	11				

- Find the missing values in the ANOVA table.
- At the 5% significance level, can you conclude that the average growth of the plant differs by organic fertilizer?
- At the 5% significance level, can you conclude that the average growth of the plant differs by acidity level?



38. **FILE Shift Output.** Metalworks, a supplier of fabricated industrial parts, wants to determine if the average output rate for a particular component is the same across the three work shifts. However, since any of four machines can be used, the machine effect must be controlled for within the sample. The accompanying table shows output rates (in units) for the previous day.

Machine (Factor B)	Shift (Factor A)		
	1	2	3
A	1392	1264	1334
B	1228	1237	1107
C	1173	1108	1186
D	1331	1342	1387

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
  - At the 5% significance level, can you conclude that the average output rate differs across the work shifts?
  - At the 5% significance level, can you conclude that the average output rate differs across the machines?
  - If significant differences exist across the machines, use Tukey's HSD method at the 5% significance level to determine which machines have different average output rates.
39. **FILE Restaurants.** Given a recent outbreak of illness caused by *E. coli* bacteria, the mayor in a large city is concerned that some of his restaurant inspectors are not consistent with their evaluations of a restaurant's cleanliness. In order to investigate this possibility, the mayor has five restaurant inspectors grade (scale of 0 to 100) the cleanliness of three restaurants. The results are shown in the accompanying table.

Inspector	Restaurant		
	1	2	3
1	72	54	84
2	68	55	85
3	73	59	80
4	69	60	82
5	75	56	84

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
  - At the 5% significance level, can you conclude that the average grades differ by restaurant?
  - If the average grades differ by restaurant, use Tukey's HSD method at the 5% significance level to determine which averages differ.
  - At the 5% significance level, can you conclude that the average grades differ by inspector? Does the mayor have cause for concern?
40. **FILE YumYum.** The marketing manager at YumYum, a large deli chain, is testing the effectiveness of four

potential advertising strategies. After a two-week trial period for each advertising strategy, sales were determined. However, since some store locations have higher customer traffic than other locations, the effect of location on sales must be controlled for within the sample. The accompanying table shows sales (in thousands of dollars) achieved over the 2-week trial period using each advertising strategy at three different store locations.

Store Location (Factor B)	Advertising Strategy (Factor A)			
	Newspaper only	Internet only	TV only	Internet & TV
City	511	644	585	712
Suburban	458	548	503	614
Rural	388	298	347	421

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
  - At the 5% significance level, can you conclude that the mean sales differ among the advertising strategies? What about the 10% significance level?
  - At the 5% significance level, can you conclude that the mean sales differ across the store locations?
  - If significant differences exist across advertising strategies, use Fisher's LSD method at the 10% significance level to find which strategies have different mean sales.
41. **FILE Houses.** First National Bank employs three real estate appraisers whose job is to establish a property's market value before the bank offers a mortgage to a prospective buyer. It is imperative that each appraiser values a property with no bias. Suppose First National Bank wishes to check the consistency of the recent values that its appraisers have established. The bank asked the three appraisers to value (in \$1,000s) three different types of homes: a cape, a colonial, and a ranch. The results are shown in the accompanying table.

House Type	Appraiser		
	1	2	3
Cape	425	415	430
Colonial	530	550	540
Ranch	390	400	380

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
- At the 5% significance level, can you conclude that the average values differ by appraiser? Should the bank be concerned with appraiser inconsistencies?
- At the 5% significance level, can you conclude that the average values differ by house type?
- If average values differ by house type, use Tukey's HSD method at the 5% significance level to determine which averages differ.

Conduct and evaluate a two-way ANOVA test with interaction.

We use a two-way ANOVA test with interaction to capture the possible relationship between factors  $A$  and  $B$ . Such tests allow the influence of factor  $A$  to change over levels of factor  $B$  and the influence of factor  $B$ , to change over levels of factor  $A$ . In the annual income example from Section 13.3, field of employment may interact with educational attainment. In other words, the influence of field of employment may vary between levels of educational attainment. Similarly, the differences between educational attainment may not be the same for all fields of employment.

#### TWO-WAY ANOVA WITH INTERACTION

In a **two-way ANOVA test with interaction**, the total sum of squares  $SST$  of the variable is partitioned into four distinct components: the sum of squares for factor  $A$ ,  $SSA$ ; the sum of squares for factor  $B$ ,  $SSB$ ; the sum of squares for the interaction between the two factors,  $SSAB$ ; and the sum of squares due to error,  $SSE$ . That is,

$$SST = SSA + SSB + SSAB + SSE.$$

While we still use a randomized block design, we need at least two observations for each combination of the  $i$ th level of factor  $A$  and the  $j$ th level of factor  $B$ . In other words, we need more than one observation per cell. In a two-way ANOVA test with interaction, we let  $w$  equal the number of observations for each combination of the  $i$ th level of factor  $A$  and the  $j$ th level of factor  $B$ .

To illustrate two-way ANOVA with interaction, we reanalyze the income example, using new data with three incomes for each combination; thus,  $w = 3$ . Given the data in Table 13.15, we ultimately want to determine whether interaction is present between educational attainment and field of employment.

**TABLE 13.15** Data for Two-Factor Income Example with Interaction

Education Level (Factor $B$ )	Field of Employment (Factor $A$ )		
	Educational Services	Financial Services	Medical Services
High School	20	27	26
	25	25	24
	22	25	25
Bachelor's	30	44	42
	35	46	43
	34	48	45
Master's	46	50	62
	47	58	56
	50	56	60
Ph.D.	79	90	90
	78	92	100
	74	95	105

We are specifically interested in whether field of employment and education level interact with respect to average annual income. In order to find the relevant sum of squares for the test, we first compute the cell means and the factor means. For example, the cell mean for workers in the field of educational services with a high school education is computed as  $(20 + 25 + 22)/3 = 22.33$ . Factor means are based on one row or one column of the data. Table 13.16 shows the cell means  $\bar{x}_{ij}$ , factor means  $\bar{x}_i$  and  $\bar{x}_j$ , and the grand mean  $\bar{\bar{x}}$  for the data.

**TABLE 13.16** Cell and Factor Means for Two-Factor Income Example with Interaction

Education Level (Factor <i>B</i> )	Field of Employment (Factor <i>A</i> )			Factor <i>B</i> Means
	Educational Services	Financial Services	Medical Services	
High School	22.33	25.67	25.00	24.33
Bachelor's	33.00	46.00	43.33	40.78
Master's	47.67	54.67	59.33	53.89
Ph.D.	77.00	92.33	98.33	89.22
Factor <i>A</i> Means	45.00	54.67	56.50	$\bar{\bar{X}} = 52.06$

### The Total Sum of Squares, *SST*

*SST* is computed as  $SST = \sum_{i=1}^c \sum_{j=1}^r \sum_{k=1}^w (x_{ijk} - \bar{\bar{X}})^2$ . Using all observations from Table 13.15 and the grand mean from Table 13.16, we calculate

$$SST = (20 - 52.06)^2 + (25 - 52.06)^2 + \cdots + (105 - 52.06)^2 = 22008.$$

### The Sum of Squares for Factor *A*, *SSA*, and the Sum of Squares for Factor *B*, *SSB*

The calculations for *SSA* and *SSB* are analogous to the earlier two-way ANOVA discussion, with one minor modification. For two-way ANOVA without interaction, *SSA* and *SSB* were calculated as  $r \sum_{i=1}^c (\bar{x}_i - \bar{\bar{X}})^2$  and  $c \sum_{j=1}^r (\bar{x}_j - \bar{\bar{X}})^2$ , respectively. Now each formula is multiplied by the number of observations per cell *w*. So,  $SSA = wr \sum_{i=1}^c (\bar{x}_i - \bar{\bar{X}})^2$  and  $SSB = wc \sum_{j=1}^r (\bar{x}_j - \bar{\bar{X}})^2$ . Given the means in Table 13.16 with *w* = 3, *c* = 3, and *r* = 4, we calculate

$$SSA = (3 \times 4)[(45 - 52.06)^2 + (54.67 - 52.06)^2 + (56.5 - 52.06)^2] = 916,$$

and

$$SSB = (3 \times 3)[(24.33 - 52.06)^2 + (40.78 - 52.06)^2 + (53.89 - 52.06)^2 + (89.22 - 52.06)^2] = 20524.$$

We divide by the respective degrees of freedom to obtain the **mean square for factor A**, *MSA*, and the **mean square for factor B**, *MSB*, as

$$MSA = \frac{SSA}{c - 1} = \frac{916}{3 - 1} = 458 \quad \text{and} \\ MSB = \frac{SSB}{r - 1} = \frac{20524}{4 - 1} = 6841.$$

### The Sum of Squares for the Interaction of Factor *A* and Factor *B*, *SSAB*

When two factors interact, the effect of one factor on the mean depends upon the specific value or level present for the other factor. Interaction exists between these factors when two mathematical expressions, denoted Expression 1 and Expression 2, are significantly different from one another.

Expression 1 is defined as the difference of a cell mean from the grand mean, or equivalently,  $(\bar{x}_{ij} - \bar{\bar{X}})$ . Using the data from Table 13.16, one such difference would be  $(\bar{x}_{11} - \bar{\bar{X}}) = (22.33 - 52.06)$ .

Expression 2 is defined as the *combined* differences of the corresponding factor *A* mean from the grand mean and the corresponding factor *B* mean from the grand mean, or equivalently,  $(\bar{x}_i - \bar{\bar{X}}) + (\bar{x}_j - \bar{\bar{X}})$ . Again using the first cell in Table 13.16 as a reference, the corresponding difference would be  $(45 - 52.06) + (24.33 - 52.06)$ .

If the difference between Expression 1 and Expression 2 is nonzero, then there is evidence of interaction. If we let  $I$  denote interaction, then we can measure  $I$  as

$$I = (\bar{x}_{ij} - \bar{\bar{x}}) - [(\bar{x}_i - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}})].$$

This expression can be simplified to

$$I = \bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}.$$

The sum of squares for the interaction between factor  $A$  and factor  $B$ ,  $SSAB$ , is then based on a weighted sum of the squared interactions ( $I^2$ ) where the weight equals the number of observations per cell  $w$ :

$$SSAB = w \sum_{i=1}^c \sum_{j=1}^r (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2.$$

Using the means in Table 13.16, we calculate

$$\begin{aligned} SSAB &= 3[(22.33 - 45 - 24.33 + 52.06)^2 + (33 - 45 - 40.78 + 52.06)^2 \\ &\quad + \dots + (98.33 - 56.5 - 89.22 + 52.06)^2] \\ &= 318. \end{aligned}$$

We obtain the **mean square for interaction**,  $MSAB$ , by dividing  $SSAB$  by its degrees of freedom  $(c - 1)(r - 1)$ , or

$$MSAB = \frac{SSAB}{(c - 1)(r - 1)} = \frac{318}{(3 - 1)(4 - 1)} = 53.$$

### The Sum of Squares due to Error, $SSE$

We solve for  $SSE$  by rearranging  $SST = SSA + SSB + SSAB + SSE$ ; that is,

$$SSE = SST - (SSA + SSB + SSAB) = 22008 - (916 + 20524 + 318) = 250.$$

Finally, we divide  $SSE$  by its degrees of freedom  $rc(w - 1)$  and obtain the **mean square error**,  $MSE$ , as

$$MSE = \frac{SSE}{rc(w - 1)} = \frac{250}{(4 \times 3)(3 - 1)} = 10.4.$$

#### FILE

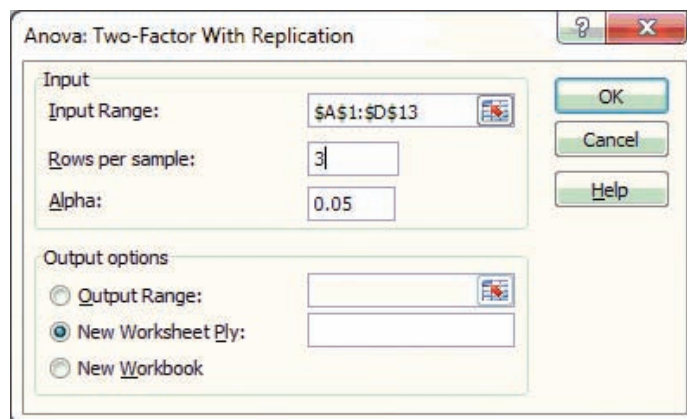
*Income\_Interaction*

## Using Excel to Solve a Two-Way ANOVA Test with Interaction

Fortunately, Excel easily calculates all of these statistics. We follow these steps.

- A. Open the *Income\_Interaction* data file.
- B. From the menu choose **Data > Data Analysis > ANOVA: Two Factor With Replication**.
- C. In the *ANOVA: Two Factor With Replication* dialog box shown in Figure 13.4, choose the box next to *Input range* and then select all the data, including the labels.

**FIGURE 13.4**  
Excel's ANOVA: Two-Factor With Replication dialog box



Enter 3 for *Rows Per Sample*. If testing at a significance level other than 5%, insert the relevant significance level in the box next to *Alpha*. Choose an output range and click **OK**.

Table 13.17 shows the relevant portion of the output. Any differences in our manual calculations and the values found in the table are due to rounding.

**TABLE 13.17** Excel's ANOVA Output for the Two-Factor Income With Interaction Example

Source of Variation	SS	df	MS	F	p-value	F crit
Sample (Rows)	20524	3	6841	658.5	3.58E-23	3.009
Columns	916.2	2	458.1	44.1	9.18E-09	3.403
Interaction	318.4	6	53.07	5.109	0.001659	2.508
Within (Error)	249.3	24	10.39			
Total	22008	35				

Table 13.17 shows values for three  $F_{(df_1, df_2)}$  test statistics. The first two test statistics are used to examine the **main effects**—potential differences in factor *B* or the row means ( $F_{(df_1, df_2)} = \frac{MSB}{MSE}$ , where  $df_1 = r - 1$ ,  $df_2 = rc(w - 1)$ ) and potential differences in factor *A* or the column means ( $F_{(df_1, df_2)} = \frac{MSA}{MSE}$ , where  $df_1 = c - 1$ ,  $df_2 = rc(w - 1)$ ). The third test statistic ( $F_{(df_1, df_2)} = \frac{MSAB}{MSE}$ , where  $df_1 = (r - 1)(c - 1)$ ,  $df_2 = rc(w - 1)$ ) is used to test whether there is interaction between factor *A* and factor *B*.

### EXAMPLE 13.6

Use the Excel output from Table 13.17 to determine whether field of employment and education level interact with respect to the average annual income at the 5% significance level.

**SOLUTION:** We set up the following competing hypotheses:

$H_0$ : There is no interaction between factors *A* and *B*.

$H_A$ : There is interaction between factors *A* and *B*.

The value of the test statistic is  $F_{(df_1, df_2)} = \frac{MSAB}{MSE} = \frac{53.07}{10.39} = 5.11$  where  $df_1 = (r - 1)(c - 1) = (4 - 1)(3 - 1) = 6$  and  $df_2 = rc(w - 1) = (4 \times 3)(3 - 1) = 24$ , or  $F_{(6, 24)} = 5.11$  with a corresponding *p*-value of 0.0017. At the 5% significance level, we reject  $H_0$  and conclude that sufficient evidence exists of an interaction effect between educational attainment and field of employment. This result implies that the average annual income for attaining an advanced degree is higher in some fields than in others.

Note that due to the interaction, the differences between educational attainment are not the same for all fields of employment. Such an outcome serves to complicate the interpretation of the main effects, since differences in one factor are not consistent across the other factor. This is why we should perform the interaction test before making any conclusions using the other two  $F_{(df_1, df_2)}$  statistics. If the interaction effect is not significant, then we can proceed by focusing on the main effects: testing whether or not the row means or the column means differ. If the interaction effect is significant, as it is here, one option is to use another technique called regression analysis. Regression analysis is discussed in the next four chapters.

## EXERCISES 13.4

### Mechanics

42. A two-way analysis of variance experiment with interaction was conducted. Factor *A* had four levels (columns), factor *B* had three levels (rows), and five observations were obtained for each combination. The results include the following sum of squares terms:

$$SST = 2500 \quad SSA = 1200 \quad SSB = 1000 \quad SSE = 280$$

- Construct an ANOVA table. (You may want to use Excel's F.DIST.RT and F.INV.RT functions to find the *p*-values and the critical values.)
- At the 5% significance level, can you conclude that there is interaction between factor *A* and factor *B*?
- At the 5% significance level, can you conclude that the factor *A* means differ?
- At the 5% significance level, can you conclude that the factor *B* means differ?

43. A two-way analysis of variance experiment with interaction was conducted. Factor *A* had three levels (columns), factor *B* had five levels (rows), and six observations were obtained for each combination. The results include the following sum of squares terms:

$$SST = 1558 \quad SSA = 1008 \quad SSB = 400 \quad SSAB = 30$$

- Construct an ANOVA table. (You may want to use Excel's F.DIST.RT and F.INV.RT functions to find the *p*-values and the critical values.)
- At the 1% significance level, can you conclude that there is interaction between factor *A* and factor *B*?
- At the 1% significance level, can you conclude that the factor *A* means differ?
- At the 1% significance level, can you conclude that the factor *B* means differ?

44. A researcher conducts a two-way ANOVA test with interaction and provides the following ANOVA table.

Source of Variation	SS	df	MS	F	p-value
Sample	30.827	1	30.827	11.690	0.0031
Columns	169.861	2	84.930	32.208	1.13E-06
Interaction	4.241	2	2.120	0.804	0.4629
Within	47.465	18	2.637		
Total	252.393	23			

- At the 1% significance level, can you conclude that there is interaction between the two factors?
  - Are you able to conduct tests based on the main effects? If yes, conduct these tests at the 1% significance level. If no, explain.
45. A researcher conducts a two-way ANOVA test with interaction and provides the following ANOVA table.

Source of Variation	SS	df	MS	F	p-value	F crit
Sample	752.78	2	$MSB = ?$	$F_{\text{Factor B}} = ?$	0.0116	3.885
Columns	12012.50	1	$MSA = ?$	$F_{\text{Factor A}} = ?$	5.62E-09	4.747
Interaction	58.33	2	$MSAB = ?$	$F_{\text{Interaction}} = ?$	0.6117	3.885
Within	683.33	12	$MSE = ?$			
Total	13506.94	17				

- Find the missing values in the ANOVA table.
- At the 5% significance level, can you conclude that there is an interaction effect?
- At the 5% significance level, can you conclude that the column means differ?
- At the 5% significance level, can you conclude that the row (sample) means differ?

### Applications

46. The engineering department at a steel mill is studying the tensile strength of a particular grade of steel when fabricated at various pressures (Factor *A*) and temperatures (Factor *B*). The accompanying ANOVA table shows a portion of the results from conducting a two-way ANOVA test with interaction.

Source of Variation	SS	df	MS	F	p-value	F crit @5%
Sample ( <i>B</i> , temperature)	150.22	1	150.22			
Columns ( <i>A</i> , pressure)			62.06			
Interaction	24.11	2	12.06			
Within			26.11			
Total	611.78	17				

- How many levels did pressure have?
  - How many observations were run for each combination of pressure-temperature settings?
  - At the 5% significance level, can you conclude that there is interaction between pressure and temperature?
  - At the 5% significance level, can you conclude that the main effect of pressure is significant?
  - At the 5% significance level, can you conclude that the main effect of temperature is significant?
47. The effects of detergent brand name (factor *A*) and the temperature of the water (factor *B*) on the brightness of washed fabrics are being studied. Four brand names and two temperature levels are used, and six observations for each combination are examined. The following ANOVA table is produced.



Source of Variation	SS	df	MS	F	p-value	F crit
Sample	75	1	75	63.38	8.92E-10	4.084
Columns	130.25	3	43.42	36.69	1.45E-11	2.839
Interaction	8.67	3	2.89	2.44	0.0783	2.839
Within	47.33	40	1.18			
Total	261.25	47				

- Can you conclude that there is interaction between the detergent brand name and the temperature of the water at the 5% significance level?
  - Are you able to conduct tests based on the main effects? If yes, conduct these tests at the 5% significance level. If no, explain.
48. **FILE Buy4Less.** The marketing group at Buy4Less, a local retail chain, is examining the effect of advertising at various times of day and on various local television channels. Based on 12-week cycles (4 time periods x 3 local channels), three observations (cycles) of weekly sales data have been obtained for each time period-channel combination. The results (in thousands of dollars) are shown in the accompanying table.

Local Channel (Factor B)	Time of Day (Factor A)			
	Morning (6am-11am)	Mid-Day (11am-5pm)	Evening (5pm-10pm)	Late Night (10pm-2am)
WTZX	69.6	84.4	83.9	60.0
	74.6	79.9	85.9	63.0
	72.8	73.8	78.0	58.3
WABC	70.1	84.7	81.2	64.4
	68.4	80.0	89.8	69.7
	70.0	82.8	80.2	67.9
WXAQ	76.1	76.0	76.3	56.7
	75.3	83.5	81.4	68.2
	68.7	71.7	81.2	64.4

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
  - At the 5% significance level, can you conclude that there is interaction between the time of day and the local channel used for advertising?
  - Are you able to conduct tests based on the main effects? If yes, conduct them at the 5% significance level. If no, explain why.
49. **FILE Brand\_Garage.** A consumer advocate examines whether the longevity of car batteries (measured in years) is affected by the brand name (factor A) and whether or not the car is kept in a garage (factor B). Interaction is suspected. The results are shown in the accompanying table.

Kept in Garage?	Brand Name of Battery		
	A	B	C
Yes	7, 8, 8	6, 7, 7	8, 9, 9
No	5, 6, 6	4, 5, 4	6, 7, 7

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
  - At the 5% significance level, is there interaction between the brand name and whether a car is garaged?
  - At the 5% significance level, can you conclude that the average battery lives differ by brand name?
  - At the 5% significance level, can you conclude that the average battery lives differ depending on whether a car is garaged?
50. **FILE Job\_Satisfaction.** A human resource specialist wants to determine whether the average job satisfaction score (on a scale of 0 to 100) is the same for three different industries and three types of work experience. A randomized block experiment with interaction is performed. The results are shown in the accompanying table.

Work Experience	Industry		
	A	B	C
Less than 5 years	77	66	81
	67	58	59
	82	54	64
Five up to 10 years	93	65	57
	92	60	49
	97	68	72
10 years or more	58	75	60
	78	57	45
	91	47	59

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
  - At the 5% significance level, is there interaction between industry and work experience?
  - At the 5% significance level, can you conclude that job satisfaction differs by industry?
  - At the 5% significance level, can you conclude that job satisfaction differs by work experience?
51. **FILE Salaries.** It is generally believed that a practical major such as business or engineering can really pay off for college graduates (CNNMoney.com, July 22, 2010). Other studies have shown that it is not just the major but also how students perform, as measured by their GPA, that influences their salaries. Henry Chen, an employee of PayScale.com, wants to measure the effect of major and GPA on starting salaries of graduates of the University of California at Irvine. He samples starting salaries of five graduates for a given GPA range from the schools of business, engineering, and social sciences. The sample data are shown in the following table.

GPA	Business	Engineering	Social Sciences
3.5–4.0	68	70	44
	54	66	52
	78	62	66
	80	56	42
	58	72	72
3.0–3.5	48	66	48
	76	54	48
	60	70	58
	48	52	56
	64	64	38
<3.0	54	60	44
	42	42	38
	66	48	42
	58	59	52
	66	65	50

- At the 5% significance level, is there interaction between major and GPA?
- At the 5% significance level, can you conclude that starting salary differs between majors?
- At the 5% significance level, can you conclude that starting salary depends on GPA?

## WRITING WITH STATISTICS



The Texas Transportation Institute, one of the finest higher-education-affiliated transportation research agencies in the nation, recently published its highly anticipated *2009 Annual Urban Mobility Report* (July 8, 2009). The study finds that the average U.S. driver languished in rush-hour traffic for 36.1 hours, as compared to 12 hours in 1982 when the records begin. This congestion also wasted approximately 2.81 billion gallons in fuel, or roughly three weeks' worth of gas per traveler. John Farnham, a research analyst at an environmental firm, is stunned by some of the report's conclusions. John is asked to conduct an independent study in order to see if differences exist in congestion depending on the city where the traveler drives. He selects 25 travelers from each of the five cities that suffered from the worst

congestion. He asks each traveler to approximate the time spent in traffic (in hours) over the last calendar year. Table 13.18 shows a portion of his sample results.

**TABLE 13.18** Annual Hours of Delay per Traveler in Five Cities

Los Angeles	Washington, DC	Atlanta	Houston	San Francisco
71	64	60	58	57
60	64	58	56	56
⋮	⋮	⋮	⋮	⋮
68	57	57	59	56

**FILE**  
Congestion

John wants to use the sample information to:

- Determine whether significant differences exist in congestion, depending on the city where the traveler drives.
- Use Tukey's method to establish in which of the five cities travelers experience the least and the worst delays.

Does traffic congestion vary by city? *The 2009 Annual Urban Mobility Report* found that traffic congestion, measured by annual hours of delay per traveler, was the worst in Los Angeles, followed by Washington, DC, Atlanta, Houston, and then San Francisco. An independent survey was conducted to verify some of the findings. Twenty-five travelers in each of these cities were asked how many hours they wasted in traffic over the past calendar year. Table 13.A reports the summary statistics. The sample data indicate that Los Angeles residents waste the most time sitting in traffic with an average of 69.2 hours per year. Washington, DC, residents rank a close second, spending an average of 62 hours per year in traffic. Residents in Atlanta, Houston, and San Francisco spend on average, 57.0, 56.5, and 55.6 hours per year in traffic, respectively. Houston had the highest variability of congestion and San Francisco had the lowest variability as measured by their respective standard deviations.

**TABLE 13.A** Summary Statistics and Relevant ANOVA Results

Los Angeles	Washington, DC	Atlanta	Houston	San Francisco
$\bar{x}_1 = 69.2$	$\bar{x}_2 = 62.0$	$\bar{x}_3 = 57.0$	$\bar{x}_4 = 56.5$	$\bar{x}_5 = 55.6$
$s_1 = 4.6$	$s_2 = 4.7$	$s_3 = 4.8$	$s_4 = 5.4$	$s_5 = 3.7$
$n_1 = 25$	$n_2 = 25$	$n_3 = 25$	$n_4 = 25$	$n_5 = 25$

A one-way ANOVA test was conducted to determine if significant differences exist in the average number of hours spent in traffic in these five worst-congested cities. The value of the test statistic is  $F_{4,120} = 37.3$  with a  $p$ -value of approximately zero. Therefore, at the 5% level of significance, we reject the null hypothesis of equal means and conclude that traffic congestion does vary by city.

In order to determine which cities had significantly different average delays per traveler, Tukey's *HSD* method was used. The 95% confidence interval for the difference between two population means  $\mu_i - \mu_j$  was computed as  $(\bar{x}_i - \bar{x}_j) \pm q_{\alpha, (c, n_T - c)} \sqrt{\frac{MSE}{n}}$ . Referencing the studentized range table, the approximate value of  $q_{0.05, (5, 115)}$  is 3.92. The one-way ANOVA test produced an *MSE* of 21.82; thus, the margin of error for the confidence interval was  $3.92 \sqrt{\frac{21.82}{25}}$ , which equals 3.66. Therefore, we can conclude with 95% confidence that travelers in Los Angeles suffered the most hours of congestion followed by travelers in Washington, DC. Congestion was not significantly different in the cities of Atlanta, Houston, and San Francisco.

## CONCEPTUAL REVIEW

### LO 13.1 Conduct and evaluate a one-way ANOVA test.

A **one-way analysis of variance (ANOVA)** test is used to determine if differences exist between the means of three or more populations. This test examines the amount of variability *between* the samples relative to the amount of variability *within* the samples.

The value of the **test statistic** for the hypothesis test of the equality of the  $c$  population means is calculated as  $F_{(df_1, df_2)} = \frac{MSTR}{MSE}$ , where *MSTR* is the mean square for treatments, *MSE* is the mean square error,  $df_1 = c - 1$ ,  $df_2 = n_T - c$ , and  $n_T$  is the total sample size. *MSTR* and *MSE* are based on independent samples drawn from  $c$  normally distributed populations with a common variance. An ANOVA test is always specified as a right-tailed test.

### LO 13.2 Use Fisher's LSD method and Tukey's HSD method to determine which means differ.

The ANOVA test can determine whether significant differences exist between the population means. However, it cannot indicate which population means differ. By constructing confidence

intervals for all pairwise differences for the population means, we can identify which means differ. Note that if the ANOVA test does not reject  $H_0$ , then there are no differences to find.

**Fisher's LSD method** is implemented by computing  $100(1 - \alpha)\%$  confidence intervals for all mean differences  $\mu_i - \mu_j$  as  $(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2, n_T - c} \sqrt{MSE \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$  where  $MSE$  is estimated from the ANOVA test. If the computed interval does not include the value zero, then we reject the null hypothesis  $H_0: \mu_i - \mu_j = 0$ .

When pairwise comparisons are made with Fisher's LSD method, we inflate the risk of the Type I error  $\alpha$ ; that is, we increase the risk of incorrectly rejecting the null hypothesis. In other words, if we conduct all pairwise tests at  $\alpha = 0.05$ , the resulting  $\alpha$  for the overall test will be greater than 0.05.

An improved multiple comparison technique is **Tukey's HSD method**, which seeks out "honestly significant differences" between paired means. Tukey's method uses the **studentized range distribution**, which has broader, flatter, and thicker tails than the  $t_{df}$  distribution, and therefore protects against an inflated risk of a Type I error. Tukey's  $100(1 - \alpha)\%$  confidence interval for  $\mu_i - \mu_j$  is computed as  $(\bar{x}_i - \bar{x}_j) \pm q_{\alpha, (c, n_T - c)} \sqrt{\frac{MSE}{n}}$  for **balanced data** ( $n = n_i = n_j$ ) and  $(\bar{x}_i - \bar{x}_j) \pm q_{\alpha, (c, n_T - c)} \sqrt{\frac{MSE}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$  for **unbalanced data** ( $n_i \neq n_j$ ), where  $q_{\alpha, (c, n_T - c)}$  is the studentized range value.

### LO 13.3 Conduct and evaluate a two-way ANOVA test with no interaction.

Whereas a one-way ANOVA test is used to compare population means based on one factor, a two-way ANOVA test extends the analysis to measure the effects of two factors simultaneously. The additional factor explains some of the unexplained variation, or equivalently, reduces the sum of squares due to error for a more discriminating  $F_{(df_1, df_2)}$  statistic. A two-way ANOVA test can be conducted with or without interaction between the factors.

In a **two-way ANOVA test without interaction**, we partition the total variation  $SST$  into the sum of squares for factor  $A$  ( $SSA$ ), the sum of squares for factor  $B$  ( $SSB$ ), and the sum of squares due to error ( $SSE$ ); that is,  $SST = SSA + SSB + SSE$ . We find the value of two  $F_{(df_1, df_2)}$  test statistics. The value of the first test statistic  $F_{(df_1, df_2)} = \frac{MSB}{MSE}$ , where  $df_1 = r - 1$  and  $df_2 = n_T - c - r + 1$  is used to test whether significant differences exist between the factor  $B$  means (the row means). The value of the second test statistic  $F_{(df_1, df_2)} = \frac{MSA}{MSE}$ , where  $df_1 = c - 1$  and  $df_2 = n_T - c - r + 1$  is used to test whether significant differences exist between the factor  $A$  means (the column means).

### LO 13.4 Conduct and evaluate a two-way ANOVA test with interaction.

In a **two-way ANOVA test with interaction**, we partition the total variation  $SST$  into four components: the sum of squares for factor  $A$  ( $SSA$ ), the sum of squares for factor  $B$  ( $SSB$ ), the sum of squares for the interaction of the two factors ( $SSAB$ ), and the sum of squares due to error ( $SSE$ ); that is,  $SST = SSA + SSB + SSAB + SSE$ . Here we find the values of three  $F_{(df_1, df_2)}$  test statistics. The first two test statistics are used to examine the **main effects**—differences in the levels of factor  $B$  ( $F_{(df_1, df_2)} = \frac{MSB}{MSE}$ , where  $df_1 = r - 1$  and  $df_2 = rc(w - 1)$ ) and differences in the levels of factor  $A$  ( $F_{(df_1, df_2)} = \frac{MSA}{MSE}$ , where  $df_1 = c - 1$  and  $df_2 = rc(w - 1)$ ). The value of the third test statistic  $F_{(df_1, df_2)} = \frac{MSAB}{MSE}$ , where  $df_1 = (r - 1)(c - 1)$  and  $df_2 = rc(w - 1)$ , is used to test whether there is interaction between factor  $A$  and factor  $B$ .

Interaction between the factors complicates the interpretation of the main effects. This is why we should perform the interaction test before testing the main effects. If the interaction effect is not significant, then we can proceed by focusing on the main effects. If the interaction effect is significant, one option is to use another technique called regression analysis. Regression analysis is discussed in the next four chapters.

## ADDITIONAL EXERCISES AND CASE STUDIES

52. A government agency wants to determine whether the average salaries of four various kinds of transportation operators differ. A random sample of five employees in each of the four categories yields the salary data given in the accompanying table.

Average Salaries of Transportation Operators (\$1,000s)			
Locomotive Engineer	Truck Driver	Bus Driver	Taxi and Limousine Driver
54.7	40.5	32.4	26.8
53.2	42.7	31.2	27.1
55.1	41.6	30.9	28.3
54.3	40.9	31.8	27.9
51.5	39.2	29.8	29.9
$\bar{x}_1 = 53.76$	$\bar{x}_2 = 40.98$	$\bar{x}_3 = 31.22$	$\bar{x}_4 = 28.00$
$s_1^2 = 2.10$	$s_2^2 = 1.69$	$s_3^2 = 0.96$	$s_4^2 = 1.49$
Grand mean: $\bar{\bar{X}} = 38.49$			

- Construct an ANOVA table and estimate the  $p$ -value.
  - Specify the competing hypotheses in order to determine whether the average salaries of the transportation operators differ.
  - At the 5% significance level, can we conclude that the average salaries of the four transportation operators differ?
53. The Marketing Manager at Foodco, a large grocery store, wants to determine if store display location influences sales of a particular grocery item. He instructs employees to rotate the display location of that item every week and then tallies the weekly sales at each location over a 24-week period (8 weeks per location). The following sales results were obtained.

Front of store (units sold)	Center of store (units sold)	Side aisle (units sold)
947	858	1096
1106	780	1047
1143	786	910
1162	816	823
967	800	919
956	770	924
1057	876	1091
996	802	1027
$\bar{x}_1 = 1041.8$	$\bar{x}_2 = 811.0$	$\bar{x}_3 = 979.6$
$s_1^2 = 7583.4$	$s_2^2 = 1418.3$	$s_3^2 = 9840.0$
$\bar{\bar{X}} = 944.1$		

- Construct an ANOVA table. Assume sales are normally distributed.
- Specify the competing hypotheses to test whether there are some differences in the mean weekly sales across the three store display locations.
- At the 5% significance level, what is the conclusion of the test?

54. **FILE Concrete Mixing.** Compressive strength of concrete is affected by several factors, including composition (sand, cement, etc.), mixer type (batch vs. continuous), and curing procedure. Accordingly, a concrete company is conducting an experiment to determine how mixing technique affects the resulting compressive strength. Four potential mixing techniques have been identified. Subsequently, samples of 20 specimens have been subjected to each mixing technique, and the resulting compressive strengths (in pounds per square inch, psi) were measured. A portion of the data is shown in the accompanying table.

Mix Tech. 1 (psi)	Mix Tech. 2 (psi)	Mix Tech. 3 (psi)	Mix Tech. 4 (psi)
2972	2794	2732	2977
2818	3162	2905	2986
:	:	:	:
2665	2837	3073	3081

- Construct an ANOVA table. Assume compressive strengths are normally distributed.
  - Specify the competing hypotheses to test whether there are some differences in the mean compressive strengths across the four mixing techniques.
  - At the 5% significance level, what is the conclusion of the test? What about the 1% significance level?
55. **FILE SAT Ethnicity.** The manager of an SAT review program wonders whether average SAT scores differ depending on the ethnicity of the test taker. Thirty test scores for four ethnicities are collected. A portion of the data is shown in the accompanying table.

White	Black	Asian-American	Mexican-American
1587	1300	1660	1366
1562	1255	1576	1531
:	:	:	:
1500	1284	1584	1358



At the 5% significance level, can we conclude that the average SAT scores differ by ethnicity?

56. **FILE Plywood.** An engineer wants to determine whether the average strength of plywood boards (in pounds per square inch, psi) differs depending on the type of glue used. For three types of glue, she measures the strength of 20 plywood boards. A portion of the data is shown in the accompanying table.

Glue 1	Glue 2	Glue 3
38	41	42
34	38	38
⋮	⋮	⋮
38	49	50

At the 5% significance level, can she conclude that the average strength of the plywood boards differs by the type of glue used? Assume that the strength of plywood boards is normally distributed.

57. An employee of a small software company in Minneapolis bikes to work during the summer months. He can travel to work using one of three routes and wonders whether the average commute times (in minutes) differ between the three routes. He obtains the following data after traveling each route for one week.

Route 1	29	30	33	30	32
Route 2	27	32	28	30	29
Route 3	25	27	24	29	26

The following one-way ANOVA results were obtained for  $\alpha = 0.01$ :

Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	54.53	2	27.27	8.099	0.0059	6.93
Within Groups	40.40	12	3.37			
Total	94.93	14				

- Determine at the 1% significance level whether the average commute times differ between the three routes.
  - If differences exist, use Tukey's *HSD* method at the 1% significance level to determine which routes' average times differ.
58. An economist wants to determine whether average Price/Earnings (P/E) ratios differ for firms in three industries. Independent samples of five firms in each industry show the following results:

Industry A	12.19	12.44	7.28	9.96	10.51	$\bar{x}_A = 10.48, s_A^2 = 4.32$
Industry B	14.34	17.80	9.32	14.90	9.41	$\bar{x}_B = 13.15, s_B^2 = 13.69$
Industry C	26.38	24.75	16.88	16.87	16.70	$\bar{x}_C = 20.32, s_C^2 = 23.30$
Grand Mean: $\bar{\bar{x}} = 14.65$						

- Construct an ANOVA table.
  - At the 5% significance level, determine whether average P/E ratios differ in the three industries.
  - If differences exist, use Tukey's method at the 5% significance level to determine which industries' mean P/E ratios differ.
59. Before the Great Recession, job-creating cities in the Sunbelt, like Las Vegas, Phoenix, and Orlando saw their populations, income levels, and housing prices surge. Las Vegas, however, offered something that often eluded these other cities: upward mobility for the working class. For example, hard-working hotel maids were able to prosper during the boom times. According to the Bureau of Labor Statistics, the average hourly rate for hotel maids was \$14.25 in Las Vegas, versus \$9.25 in Phoenix and \$8.84 in Orlando (*The Wall Street Journal*, July 20, 2009). Suppose the following summary statistics and ANOVA table were produced for  $\alpha = 0.05$  from a sample of hourly wages of 25 hotel maids in each city.

SUMMARY						
Groups	Count	Average				
Las Vegas	25	13.91				
Phoenix	25	8.82				
Orlando	25	8.83				
Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	430.87	2	215.44	202.90	2.58E-30	3.124
Within Groups	76.44	72	1.06			
Total	507.31	74				

- At the 5% significance level, do mean hourly rates for hotel maids differ between the three cities?
  - If differences exist, use Tukey's method to determine which cities' mean hourly rates differ at the 5% significance level.
60. The marketing department for an upscale retail catalog company wants to determine if there are differences in the mean customer purchase amounts across the available purchase sources (Internet, phone, or mail-in). Accordingly, samples were taken for 20 random orders for each purchase source. The following Excel output was compiled.



Internet Purchase	Phone Purchase	Mail-in Purchase
$\bar{x}_1 = \$214.05$ $n_1 = 20$	$\bar{x}_2 = 212.45$ $n_2 = 20$	$\bar{x}_3 = 182.40$ $n_3 = 20$

Source of Variation	SS	df	MS	F	p-value	F crit at 5%
Between Groups	12715.23	2	6357.62	0.433	0.651	3.16
Within Groups	836704.70	57	14679.03			
Total	849419.93	59				

- At the 5% significance level, can we conclude that the mean purchase amount is different across the three purchase sources?
  - If significant differences exist, use Fisher's LSD method at the 5% significance level to determine which purchase sources have different mean purchase amounts.
61. An accounting professor wants to know if students perform the same on the departmental final exam irrespective of the accounting section they attend. She randomly selects the exam scores of 20 students from three sections. A portion of the output from conducting a one-way ANOVA test is shown in the accompanying table. Assume exam scores are normally distributed.

Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	57.39	2	$MSTR = ?$	$F_{2,57} = ?$	0.3461	3.159
Within Groups	$SSE = ?$	57	$MSE = ?$			
Total	1570.19	59				

- Find the missing values in the ANOVA table.
  - At the 5% significance level, can you conclude that average grades differ in the accounting sections?
62. **FILE Job\_Satisfaction.** The accompanying table shows a portion of job satisfaction scores (on a scale of 0 to 100) categorized by a person's field of expertise.

Field 1	Field 2	Field 3
80	76	81
76	73	77
:	:	:
79	67	80

- At the 5% significance level, can we conclude that average job satisfaction differs by field?
- If significant differences exist, use Fisher's LSD method at the 5% significance level to determine which fields differ.

63. **FILE Battery\_Times.** Electrobat, a battery manufacturer, is investigating how storage temperature affects the performance of one of its popular deep-cell battery models used in recreational vehicles. Samples of 30 fully charged batteries were subjected to a light load under each of four different storage temperature levels. The hours until deep discharge (meaning  $\leq 20\%$  of charge remaining) was measured. A portion of the data is shown in the accompanying table.

0 degrees F	30 degrees F	60 degrees F	90 degrees F
3	6	12	12
5	8	13	15
:	:	:	:
4	9	9	15

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
  - At the 5% significance level, what is the conclusion of the test? What about the 1% significance level?
  - If significant differences exist, use Tukey's HSD method at the 5% significance level to determine which temperature levels have different mean discharge times.
64. The accompanying table shows a portion of the results from conducting a two-way ANOVA test with no interaction in which five different production methods (factor A, columns) were evaluated in terms of labor cost per unit (dependent variable). Operator experience was used as a blocking factor (factor B, rows) and was considered at four levels.

Source of Variation	SS	df	MS	F	p-Value	F crit @5%
Rows (Experience level)	$SSB = ?$	3	$MSB = ?$	$F_B = ?$	$p\text{-val}_B = ?$	$F_{critB} = ?$
Columns (Prod. method)	$SSA = ?$	$df_A = ?$	$MSA = 0.720$	$F_A = ?$	$p\text{-val}_A = ?$	$F_{critA} = ?$
Error	3.072	12	$MSE = ?$			
Total	10.998	19				

- Find the missing values in the ANOVA table. Use Excel's F.DIST.RT to find the p-values and a 5% significance level to find the critical values.
- At the 5% significance level, can you conclude that labor cost per unit differs by production method? What about the 10% significance level?

- c. At the 5% significance level, can you conclude that labor cost per unit differs by operator experience level?

65. The following Excel output for  $\alpha = 0.05$  summarizes a portion of the results for a two-way ANOVA test without interaction where factor A (column) represents three income categories (low, medium, high), factor B (rows) consists of three different kinds of political parties (Democrat, Republican, Independent), and the variable measured was the amount (in \$) contributed to the political party during the 2008 presidential election.

Source of Variation	SS	df	MS	F	p-value	F crit
Rows	25416.67	2	MSB = ?	$F_{\text{Factor B}} = ?$	0.0990	6.944
Columns	42916.67	2	MSA = ?	$F_{\text{Factor A}} = ?$	0.0457	6.944
Error	11666.67	4	MSE = ?			
Total	80000	8				

- a. Find the missing values in the ANOVA table.  
b. At the 5% significance level, can you conclude that average contributions differ by political party?  
c. At the 5% significance level, can you conclude that average contributions differ by income level?

66. **FILE Headlight Design.** An automotive parts manufacturer is testing three potential halogen headlight designs, one of which ultimately will be promoted as providing best-in-class nighttime vision. The distance at which a traffic sign can be read in otherwise total darkness is the variable of interest. Since older drivers often have lower visual acuity, driver age must be controlled in this experiment. The following results (in feet) were obtained from sampling 12 drivers (four age groups for each headlight design).

Driver Age (Factor B)	Headlight Design (Factor A)		
	Design 1	Design 2	Design 3
Below 30	293	268	270
30-45	254	243	254
46-59	224	249	231
60-up	238	214	205

- a. Use Excel to generate the appropriate ANOVA table at the 5% significance level.  
b. At the 5% significance level, can you conclude that the mean nighttime viewing distance is different among the headlight designs? Practically speaking, what does your conclusion imply?

- c. At the 5% significance level, was including the blocking variable *Driver Age* beneficial to this experiment?

67. At a gymnastics meet, three judges evaluate the balance beam performances of five gymnasts. The judges use a scale of 1 to 10, where 10 is a perfect score.

Gymnast	Judge			Means
	1	2	3	
1	8.0	8.5	8.2	$\bar{X}_{\text{Gymnast 1}} = 8.2$
2	9.5	9.2	9.7	$\bar{X}_{\text{Gymnast 2}} = 9.5$
3	7.3	7.5	7.7	$\bar{X}_{\text{Gymnast 3}} = 7.5$
4	8.3	8.7	8.5	$\bar{X}_{\text{Gymnast 4}} = 8.5$
5	8.8	9.2	9.0	$\bar{X}_{\text{Gymnast 5}} = 9.0$
Means	$\bar{X}_{\text{Judge 1}} = 8.4$	$\bar{X}_{\text{Judge 2}} = 8.6$	$\bar{X}_{\text{Judge 3}} = 8.6$	

A statistician wants to examine the objectivity and consistency of the judges. She performs a two-way ANOVA analysis without interaction using  $\alpha = 0.01$  and obtains the following results:

Source of Variation	SS	df	MS	F	p-value	F crit
Rows	6.742	4	1.686	44.75	1.62E-05	7.006
Columns	0.192	2	0.096	2.55	0.1392	8.649
Error	0.301	8	0.038			
Total	7.235	14				

- a. At the 1% significance level, can you conclude that average scores differ by judge?  
b. At the 1% significance level, can you conclude that average scores differ by gymnast?  
c. If average scores differ by gymnast, use Tukey's HSD method at the 1% significance level to determine which gymnasts' performances differ.

68. **FILE Fuel Hybrid.** An environmentalist wants to examine whether average fuel consumption (measured in miles per gallon) is affected by fuel type (factor A) and type of hybrid (factor B). A two-way ANOVA experiment with interaction is performed. The results are shown in the accompanying table.

Car Type	Fuel A	Fuel B	Fuel C
Hybrid I	36	36	36
	43	43	43
	48	48	48
Hybrid II	36	36	36
	43	43	43
	48	48	48
Hybrid III	36	36	36
	43	43	43
	48	48	48

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
- At the 5% significance level, is there interaction between fuel type and hybrid type?
- At the 5% significance level, can you conclude that average fuel consumption differs by fuel type?
- At the 5% significance level, can you conclude that average fuel consumption differs by type of hybrid?

69. A management consultant wants to determine whether the age and gender of a restaurant's wait staff influence the size of the tip the customer leaves. Three age brackets (factor *A* in columns: young, middle-age, older) and gender (factor *B* in rows: male, female) are used to construct a two-way ANOVA experiment with interaction. For each combination, the percentage of the total bill left as a tip for 10 wait staff is examined. The following ANOVA table with  $\alpha = 0.01$  is produced.

Source of Variation	SS	df	MS	F	p-value	F crit
Sample	0.04278	1	0.04278	16.5951	0.00015	7.129
Columns	0.01793	2	0.00897	3.47884	0.03792	5.021
Interaction	0.00561	2	0.00281	1.08872	0.34392	5.021
Within	0.1392	54	0.00258			
Total	0.20552	59				

- Can you conclude that there is interaction between age and gender at the 1% significance level?
- Are you able to conduct tests based on the main effects? If yes, conduct these tests at the 1% significance level. If no, explain.

70. **FILE Training Experience.** A production manager is investigating whether the operator training method (factor *A*) will affect the resulting output for a particular product. Three training methods were studied: a full-day workshop (most intensive), in-line training, and as-needed training (least intensive). Since the value of training likely depends on the operator experience level, experience level (factor *B*) was also studied. Five operators for each training method-experience level category were randomly chosen. The accompanying table shows the total output rates (in units produced) for the previous week.

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
- At the 5% significance level, can you conclude that there is interaction between the training method and operator experience level? Explain why this is reasonable from a practical standpoint.

- Are you able to conduct tests based on the main effects? If yes, conduct them at the 5% significance level. If no, explain why. Explain why your conclusion is reasonable from a practical standpoint.

Experience Level (Factor B)	Training Method (Factor A)		
	Full-Day Workshop	In-Line Training	As Needed Training
Under 3 months	492	405	363
	460	408	313
	433	384	325
	483	447	378
	439	370	338
3–6 months	495	482	425
	457	478	381
	523	435	414
	539	489	387
	497	501	422
Over 6 months	562	511	526
	518	527	526
	567	511	525
	533	599	577
	588	534	547

71. **FILE BestCuts.** The cutting department at BestCuts, a furniture manufacturer, is examining the effect of depth of cut and feed rate on the surface roughness of table legs used in a popular dining room table model. The accompanying table shows the surface roughness results for six replicates involving three different depth-of-cut settings and two different feed rate settings.

Feed Rate (Factor B)	Depth of Cut (Factor A)		
	0.8 mm	1.0 mm	1.2 mm
6 cm/minute	10	10	12
	10	11	10
	8	10	11
	9	9	12
	9	10	12
	9	11	11
8 cm/minute	12	13	15
	12	10	12
	9	14	15
	13	13	13
	11	13	15
	9	12	14

- Use Excel to generate the appropriate ANOVA table at the 5% significance level.
- At the 5% significance level, can you conclude that there is interaction between depth of cut and feed rate?
- Are you able to conduct tests based on the main effects? If yes, conduct them at the 5% significance level. If no, explain why.
- Explain why your conclusions in part (c) are reasonable from a practical standpoint.

## CASE STUDIES

**CASE STUDY 13.1** Lisa Grattan, a financial analyst for a small investment firm, collects annual stock return data for 10 firms in the energy industry, 13 firms in the retail industry, and 16 firms in the utilities industry. A portion of the data is shown in the accompanying table.

**FILE**  
Industry\_Returns

**Data for Case Study 13.1** Annual Stock Returns (in %)

Energy	Retail	Utilities
12.5	6.6	3.5
8.2	7.4	6.4
⋮	⋮	⋮
6.9	7.9	4.3

In a report, use the sample information to:

- Determine whether significant differences exist in the annual returns for the three industries at the 5% significance level.
- Construct 95% confidence intervals for the difference between annual returns for each pairing using Tukey's HSD method.
- Evaluate which means (if any) significantly differ from one another using the results from part 2.

**CASE STUDY 13.2** In 2007, the United States experienced the biggest jump in food prices in 17 years (*The Wall Street Journal*, April 1, 2008). A variety of reasons led to this result, including rising demand for meat and dairy products in emerging overseas markets, increased use of grains for alternative fuels, and bad weather in some parts of the world. A recent survey compared prices of selected products at grocery stores in the Boston area. The accompanying table shows the results.

**FILE**  
Grocery\_Prices

**Data for Case Study 13.2** Prices of Select Groceries at Three Stores

Item	Crosby's	Shaw's	Market Basket
Two-liter Coke	\$1.79	\$1.59	\$1.50
Doritos chips	4.29	4.99	3.50
Cheerios cereal	3.69	2.99	3.00
Prince spaghetti	1.59	1.69	1.99
Skippy peanut butter	5.49	4.49	3.99
Cracker Barrel cheese	4.99	4.99	3.49
Pepperidge Farm white bread	3.99	3.99	3.99
Oreo cookies	4.69	3.39	3.00
One dozen eggs*	2.49	2.69	1.59
Coffee*	4.49	4.79	3.99
Gallon of milk*	3.69	3.19	1.59

\*Store brand items; data collected October 5–6, 2011.

In a report, use the sample information to:

1. Determine whether differences exist in the average prices of products sold at the three stores at the 5% significance level.
2. Determine whether differences exist in the average prices of the 11 products at the 5% significance level.
3. If differences exist between the average prices of products sold at the three stores, use Tukey's HSD method to determine which stores' prices differ.

**CASE STUDY 13.3** The manager of an SAT review program wonders whether the ethnic background of a student and the program's instructor affect the student's performance on the SAT. Four ethnicities and three instructors are examined. Ten student scores for each combination are sampled. A portion of the data is shown in the following table.

**Data for Case Study 13.3** Ethnic Background and SAT Scores

	White	Black	Asian-American	Mexican-American
Instructor A	1587	1300	1660	1366
	1562	1255	1576	1531
	⋮	⋮	⋮	⋮
Instructor B	1598	1296	1535	1345
	1539	1286	1643	1357
	⋮	⋮	⋮	⋮
Instructor C	1483	1289	1641	1400
	1525	1272	1633	1421
	⋮	⋮	⋮	⋮

**FILE**  
ANOVA\_SAT

In a report, use the sample information and  $\alpha = 0.05$  to:

1. Determine if there is any interaction between instructor and ethnicity.
2. Establish whether average SAT scores differ by instructor.
3. Establish whether average SAT scores differ by ethnicity.

## APPENDIX 13.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### One-Way ANOVA; Fisher and Tukey Confidence Intervals

- A. (Replicating Examples 13.1 and 13.2) From the menu choose **Stat > ANOVA > One-Way**.
- B. Select **"Response data are in a separate column for each factor level."**
- C. Under **Responses**, select Boston, NY, SF, and Chicago. Choose **Comparisons**. Enter the value 5 after **Error rate for comparisons**. After **Comparisons procedures assuming equal variances**, select **Tukey** and **Fisher**.

**FILE**  
Public\_Transportation

#### Two-Way ANOVA (No Interaction)

- A. (Replicating Example 13.5) Stack all salary values into Column 1 and label Salary. In Column 2 (label Education), denote all salary values associated with a high school education with H, all salary values associated with a bachelor's degree with B, all salary values associated with a master's degree with M, and all salary values

**FILE**  
Two-factor\_Income

associated with a PhD with P. In Column 3 (label Field), denote all income values associated with Education with E, all salary values associated with Financial with F, and all salary values associated with Medical with Med.

- B. From the menu choose **Stat > ANOVA > Balanced ANOVA**.
- C. For **Response** select Salary, for **Model** select Education and Field.

### Two-Way ANOVA (with Interaction)

- A. (Replicating Example 13.6) In order to arrange the data, follow the Minitab instructions for Two-Way ANOVA (No Interaction), step A.
- B. From the menu choose **Stat > ANOVA > General Linear Model > Fit General Linear Model**.
- C. For **Response** select Salary, for **Factors** select Education and Field.
- D. Choose **Model** and under **Factors and covariates**, select Education and Field, then **Add**.

## SPSS

### One-Way ANOVA; Fisher and Tukey Confidence Intervals

- A. (Replicating “Examples” 13.1 and 13.2) Stack all cost values in one column and label Cost. In adjacent column (label City), denote all Boston costs with value 1, all New York costs with value 2, etc.
- B. From the menu choose **Analyze > Compare Means > One-Way ANOVA**.
- C. Under **Dependent, List**, select Cost, and under **Factor** select City. Choose **Post Hoc** and select **LSD (Fisher)** and **Tukey**.

### Two-Way ANOVA (No Interaction)

- A. (Solving Example 13.5) In order to arrange the data, follow the Minitab instructions for Two-Way ANOVA (No Interaction), step A.
- B. From the menu select **Analyze > General Linear Model > Univariate**.
- C. Under **Dependent Variable**, select Salary, and under **Fixed Factor(s)**, select Education and Field. Choose **Model**. In the Model dialog box, select **Custom**. Under **Model** select Education and Field. Under **Type**, select **All 2-way**. Deselect **Include Intercept in Model**.

### Two-Way ANOVA (with Interaction)

- A. (Replicating Example 13.6) In order to arrange the data, follow the Minitab instructions for Two-Way ANOVA (No Interaction), step A.
- B. From the menu select **Analyze > General Linear Model > Univariate**.
- C. In the Univariate dialog box, under **Dependent Variable** select Salary, and under **Fixed Factor(s)** select Education and Field. Choose **Model**. In the Model dialog box, select **Full factorial**. Deselect **Include Intercept in Model**. Click **Continue** and then click **OK**.

## JMP

### One-Way ANOVA; Fisher and Tukey Confidence Intervals

- A. (Replicating Examples 13.1 and 13.2) In order to arrange the data, follow the SPSS instructions for One-Way ANOVA, step A.
- B. From the menu select **Analyze > Fit Y by X**.
- C. Under **Select Columns**, select Pooled, then under **Cast Selected Columns into Roles**, select **Y, Columns**. Under **Select Columns**, select Group, then under **Cast Selected Columns into Roles**, select **X, Factor**.

#### FILE

Income\_Interaction

#### FILE

Public\_Transportation

#### FILE

Two-factor\_Income

#### FILE

Income\_Interaction

#### FILE

Public\_Transportation



- D. Click on the red triangle next to **Oneway Analysis of Pooled by Group** and select **Means/Anova**.
- E. For Fisher confidence intervals, click on the red triangle next to **Oneway Analysis of Pooled by Group** and select **Compare Means > Each Pair, Student's t**.
- F. For Tukey confidence intervals, click on the red triangle next to **Oneway Analysis of Pooled by Group** and select **Compare Means > All Pairs, Tukey HSD**.

### Two-Way ANOVA (No Interaction)

- A. (Replicating Example 13.5) In order to arrange the data, follow the Minitab instructions for Two-Way ANOVA (No Interaction), step A.
- B. From the menu select **Analyze > Fit Model**.
- C. Under **Select Columns**, select Salary, and then under **Pick Role Variables**, select Y. Under **Select Columns**, simultaneously select Education and Field, and then select **Macros > Full Factorial**. Double-click on Education\*Field in order to deselect this variable.

**FILE**

*Two-factor\_Income*

### Two-Way ANOVA (with Interaction)

- A. (Replicating Example 13.6) In order to arrange the data, follow the Minitab instructions for Two-Way ANOVA (No Interaction), step A.
- B. From the menu select **Analyze > Fit Model**.
- C. Under **Select Columns**, select Salary, and then under **Pick Role Variables**, select Y. Under **Select Columns**, simultaneously select Education and Field, and then select **Macros > Full Factorial**.

**FILE**

*Income\_Interaction*

# 14

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 14.1 Conduct a hypothesis test for the population correlation coefficient.
- LO 14.2 Discuss the limitations of correlation analysis.
- LO 14.3 Estimate the simple linear regression model and interpret the coefficients.
- LO 14.4 Estimate the multiple linear regression model and interpret the coefficients.
- LO 14.5 Calculate and interpret the standard error of the estimate.
- LO 14.6 Calculate and interpret the coefficient of determination,  $R^2$ .
- LO 14.7 Differentiate between  $R^2$  and adjusted  $R^2$ .

# Regression Analysis

As researchers or analysts, we often need to examine the relationship between two or more variables. We begin this chapter with a review of the correlation coefficient, first discussed in Chapter 3, and then conduct a hypothesis test to determine if two variables are significantly correlated. Although the correlation analysis may establish a linear relationship between two variables, it does not demonstrate that one variable causes change in the other variable. In this chapter, we introduce a method called regression analysis, which captures the causal relationship between the variables. In particular, it captures the effect of one or more variables, called the explanatory variables, on the variable of interest, called the response variable. We first explore the procedures for estimating a linear relationship between two variables, commonly referred to as the simple linear regression model. We then extend the simple linear regression model to the case involving several variables, called the multiple regression model. Finally, we examine goodness-of-fit measures to assess how well the estimated model fits the data and use them to select the best-fitting regression model.



## INTRODUCTORY CASE

### Consumer Debt Payments

A recent study found that American consumers are making average monthly debt payments of \$983 (Experian.com, November 11, 2010). However, the study of 26 metropolitan areas reveals quite a bit of variation in debt payments, depending on where the consumer lives. For instance, in Washington, DC, residents pay the most (\$1,285 per month), while Pittsburgh residents pay the least (\$763 per month). Madelyn Davis, an economist at a large bank, believes that income differences between cities are the primary reason for the disparate debt payments. For example, the Washington, DC, area's high incomes have likely contributed to its placement on the list. She is unsure about the likely effect of the unemployment rate on consumer debt payments. On the one hand, higher unemployment rates may reduce consumer debt payments, as consumers forgo making major purchases such as large appliances and cars. On the other hand, higher unemployment rates may raise consumer debt payments as consumers struggle to pay their bills. In order to analyze the relationship between income, the unemployment rate, and consumer debt payments, Madelyn gathers data from the same 26 metropolitan areas used in the debt payment study. Specifically, she collects each area's 2010–2011 median household income as well as the monthly unemployment rate and average consumer debt for August 2010. Table 14.1 shows a portion of the data.

**TABLE 14.1** Income, the Unemployment Rate, and Consumer Debt Payments, 2010–2011

Metropolitan Area	Income (in \$1,000s)	Unemployment	Debt
Washington, D.C.	\$103.50	6.3%	\$1,285
Seattle	81.70	8.5	1,135
⋮	⋮	⋮	⋮
Pittsburgh	63.00	8.3	763

SOURCE: eFannieMae.com reports 2010–2011 Area Median Household Incomes; bls.com gives monthly unemployment rates for August 2010; Experian.com collected average monthly consumer debt payments in August 2010 and published the data in November 2010.

Madelyn would like to use the sample information in Table 14.1 to:

1. Determine if debt payments and income are significantly correlated.
2. Use regression analysis to make predictions for debt payments for given values of income and the unemployment rate.
3. Use various goodness-of-fit measures to determine the regression model that best fits the data.

A synopsis of this case is provided at the end of Section 14.4.

## 14.1 THE COVARIANCE AND THE CORRELATION COEFFICIENT

### LO 14.1

Conduct a hypothesis test for the population correlation coefficient.

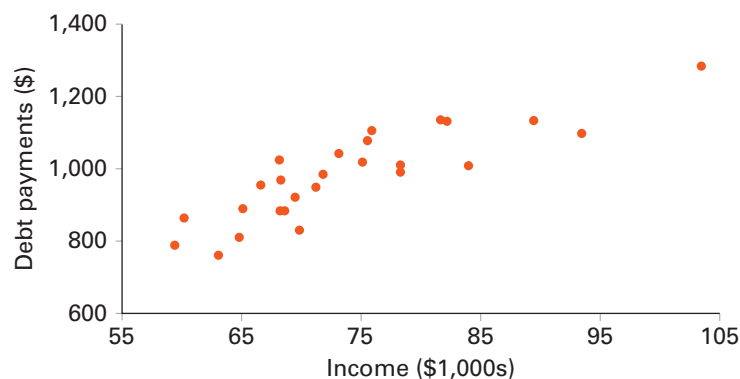
It will be useful in this section to review a scatterplot, as well as the calculation of the sample covariance and the sample correlation coefficient—these concepts were discussed in Chapters 2 and 3. A scatterplot graphically shows the relationship between two variables, while the covariance and the correlation coefficient quantify the direction and the strength of the linear relationship between two variables.

#### A SCATTERPLOT

A **scatterplot** is a graphical tool that helps in determining whether or not two variables are related in some systematic way. Each point in the diagram represents a pair of observed values of the two variables.

Using the data from the introductory case, Figure 14.1 shows a scatterplot depicting the relationship between income and debt payments. We may infer that the two variables have a positive relationship; as one increases, the other one tends to increase.

**FIGURE 14.1**  
Scatterplot of debt payments against income



A numerical measure that reveals the direction of the linear relationship between two variables  $x$  and  $y$  is called the **covariance**. It assesses whether a positive or a negative linear relationship exists between  $x$  and  $y$ .

#### THE SAMPLE COVARIANCE

The **sample covariance** is a measure of the linear relationship between two variables  $x$  and  $y$ . We compute the sample covariance  $s_{xy}$  as

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1},$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means of  $x$  and  $y$ , respectively, and  $n$  represents the number of observations.

A positive value of the sample covariance implies that, on average, when  $x$  is above its mean,  $y$  is also above its mean;  $x$  and  $y$  have a positive linear relationship. Similarly, a negative value suggests that, on average, when  $x$  is above its mean,  $y$  is below its mean; this is indicative of a negative linear relationship between the two variables. If the covariance is zero, then the two variables have no linear relationship. Further interpretation of the covariance is difficult because it is sensitive to the units of measurement. For instance, the covariance between two variables might be 100 and the covariance between two other variables might be 1,000, yet all we can conclude is that both sets of variables are positively related. In other words, we cannot comment on the strength of the relationships.

An easier measure to interpret is the **correlation coefficient**, which describes both the direction and strength of the relationship between  $x$  and  $y$ .

### THE SAMPLE CORRELATION COEFFICIENT

The **sample correlation coefficient** gauges the strength of the linear relationship between two variables  $x$  and  $y$ . We calculate the sample correlation coefficient  $r_{xy}$  as

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

where  $s_x$  and  $s_y$  are the sample standard deviations of  $x$  and  $y$ , respectively, and  $-1 \leq r_{xy} \leq 1$ .

In short, the sample correlation coefficient  $r_{xy}$  is unit-free and its value falls between  $-1$  and  $1$ . If  $r_{xy}$  equals  $1$ , then a perfect positive linear relationship exists between  $x$  and  $y$ . Similarly, a perfect negative linear relationship exists if  $r_{xy}$  equals  $-1$ . If  $r_{xy}$  equals zero, then no linear relationship exists between  $x$  and  $y$ . Other values for  $r_{xy}$  must be interpreted with reference to  $-1$ ,  $0$ , and  $1$ . As the absolute value of  $r_{xy}$  approaches  $1$ , the stronger the linear relationship. For instance,  $r_{xy} = -0.80$  indicates a strong negative relationship, whereas  $r_{xy} = 0.12$  indicates a weak positive relationship. However, we should comment on the direction of the relationship only if the correlation coefficient is found to be statistically significant—a topic which we address shortly.

### EXAMPLE 14.1

Calculate the sample covariance and the sample correlation coefficient between debt payments and income from the data in Table 14.1. Interpret these values.

**SOLUTION:** Let  $x$  denote income (in \$1,000s) and  $y$  denote average monthly consumer debt payments (in \$). We first compute the sample mean and the sample standard deviation of these variables as  $\bar{x} = 74.05$ ,  $\bar{y} = 983.46$ ,  $s_x = 10.35$ , and  $s_y = 124.61$ . We then calculate deviations from the mean for each variable. The first two columns in Table 14.2 show a portion of these calculations. Then we find the product of each pairing and sum these products. These calculations are shown in the third column.

**TABLE 14.2** Calculations for Example 14.1

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
$103.50 - 74.05 = 29.45$	$1285 - 983.46 = 301.54$	$(29.45)(301.54) = 8880.31$
$81.70 - 74.05 = 7.65$	$1135 - 983.46 = 151.54$	$(7.65)(151.54) = 1159.27$
$\vdots$	$\vdots$	$\vdots$
$63.00 - 74.05 = -11.05$	$763 - 983.46 = -220.46$	$(-11.05)(-220.46) = 2436.10$
		$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 27979.50$

Using the overall sum from the third column in Table 14.2 and  $n = 26$ , we calculate the covariance as

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{27979.50}{26 - 1} = 1119.18.$$

Given that  $s_x = 10.35$  and  $s_y = 124.61$ , the correlation coefficient is calculated as

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1119.18}{(10.35)(124.61)} = 0.87.$$

Thus, the covariance of 1,119.18 indicates that income and debt payments have a positive linear relationship. In addition, the correlation coefficient of 0.87 indicates that the strength of the positive linear relationship is strong. We will soon see that the correlation coefficient is statistically significant.

## Using Excel to Calculate the Covariance and the Correlation Coefficient

As discussed in Chapter 3, Excel easily produces the covariance and the correlation coefficients. For the next example, refer to Table 3.15, which outlines Excel's functions for the covariance and the correlation coefficient.

### FILE

*Debt\_Payments*

### EXAMPLE 14.2

Use Excel to recalculate the sample covariance and the sample correlation coefficient between debt payments and income from the data in Table 14.1.

**SOLUTION:** To calculate the sample covariance, we open *Debt\_Payments*, find an empty cell, and insert “=COVARIANCE.S(D2:D27,B2:B27)”. Note that the data for Debt are stored in cells D2 through D27 (array1) and the data for Income are stored in cells B2 through B27 (array2). After we click **<Enter>**, Excel returns a value of 1119.18, which matches the value that we calculated by hand. Similarly, when we insert “=CORREL(D2:D27,B2:B27),” Excel returns a value of 0.87.

## Testing the Correlation Coefficient

We conduct a hypothesis test to determine whether the apparent relationship between the two variables, implied by the sample correlation coefficient, is real or due to chance. Let  $\rho_{xy}$  denote the population correlation coefficient. A two-tailed test of whether the population correlation coefficient differs from zero takes the following form:

$$H_0: \rho_{xy} = 0$$

$$H_A: \rho_{xy} \neq 0$$

Note that we can easily modify the test to a one-tailed test. As in all hypothesis tests, the next step is to specify and calculate the value of the test statistic.

### TEST STATISTIC FOR $\rho_{xy}$

The value of the **test statistic** for the hypothesis test concerning the significance of the **population correlation coefficient**  $\rho_{xy}$  is calculated as

$$t_{df} = \frac{r_{xy}}{s_r},$$

where  $df = n - 2$  and  $s_r = \sqrt{(1 - r_{xy}^2)/(n - 2)}$  is the standard error of  $r_{xy}$ . Or,

$$\text{equivalently, } t_{df} = \frac{r_{xy} \sqrt{n - 2}}{\sqrt{1 - r_{xy}^2}}.$$

### EXAMPLE 14.3

Using the critical value approach to hypothesis testing, determine whether the correlation coefficient between income and debt payments is significant at the 5% level.

**SOLUTION:** When testing whether the correlation coefficient between income  $x$  and debt payments  $y$  is significant, we set up the following competing hypotheses:



$$H_0: \rho_{xy} = 0$$

$$H_A: \rho_{xy} \neq 0$$

With  $\alpha = 0.05$   $df = n - 2 = 24$ ,  $t_{\alpha/2, df} = t_{0.025, 24} = 2.064$ . Thus, the decision rule is to reject  $H_0$  if  $t_{24} > 2.064$  or  $t_{24} < -2.064$ . Using  $r_{xy} = 0.87$  and  $n = 26$  from Example 14.1, we calculate the value of the test statistic as

$$t_{24} = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0.87 \sqrt{26-2}}{\sqrt{1-(0.87)^2}} = 8.64.$$

Since  $8.64 > 2.064$ , we reject  $H_0$ . At the 5% significance level, the correlation coefficient between income and debt payments is significantly different from zero.

If we conduct the test with the  $p$ -value approach, we can use the  $t$  table to approximate  $2P(T_{24} \geq 8.64)$ . Alternatively, we can obtain the exact  $p$ -value as  $7.87 \times 10^{-9} \approx 0$  using the T.DIST.2T function in Excel (we input ‘= T.DIST.2T(8.64, 24)’). Consistent with the critical value approach, we reject  $H_0$  because the  $p$ -value  $< \alpha$ .

## Limitations of Correlation Analysis

Several limitations apply to correlation analysis.

- A.** The correlation coefficient captures only a linear relationship. Two variables can have a very low correlation coefficient yet have a strong *nonlinear* relation. Consider the following sample data:

$x$	-20	-15	-10	-5	0	5	10	15	20
$y$	380	210	90	20	0	30	110	240	420

The sample correlation coefficient for these data is  $r_{xy} = 0.09$ , implying an extremely weak positive linear relationship. However, further analysis of the data would reveal a perfect nonlinear relationship given by  $y_i = x_i + x_i^2$ .

- B.** The correlation coefficient may not be a reliable measure when *outliers* are present in one or both of the variables. Recall that outliers are a small number of extreme high or low values in the data set. As a general rule, we must determine whether the sample correlation coefficient varies dramatically by removing a few outliers. However, we must use judgment to determine whether those outliers contain important information about the relationship between the two variables (and should be included in the correlation analysis) or do not contain important information (and should be excluded).
- C.** Correlation does not imply causation. Even if two variables are highly correlated, one does not necessarily cause the other. *Spurious correlation* can make two variables appear closely related when no causal relation exists. Spurious correlation between two variables is *not* based on any theoretical relationship, but rather on a relation that arises in the data solely because each of the two variables is related to some third variable. For example, Robert Matthews in his article “Storks Bring Babies” (*Teaching Statistics*, Summer 2000) finds that the correlation coefficient between stork breeding pairs and the human birth rate for 17 European countries is 0.62. Further, he finds that the correlation is significantly different from zero at the 5% significance level. He stresses that the most plausible explanation for this observed correlation—and absurd conclusion—is the existence of a confounding variable, namely land area. That is, we are likely to see higher human birth rates in more densely populated areas. More densely populated areas also provide more chimneys, where stork breeding pairs prefer to nest.

### LO 14.2

Discuss the limitations of correlation analysis.

## EXERCISES 14.1

### Mechanics

1. Consider the following sample data:

x	8	5	3	10	2
y	380	210	90	20	2

- Construct and interpret a scatterplot.
  - Calculate and interpret the sample covariance.
  - Calculate and interpret the sample correlation coefficient.
2. Consider the following sample data:

x	-30	10	0	23	16
y	44	-15	-10	-2	5

- Construct and interpret a scatterplot.
  - Calculate and interpret  $s_{xy}$ .
  - Calculate and interpret  $r_{xy}$ .
3. Consider the following competing hypotheses:

$$H_0: \rho_{xy} = 0$$

$$H_A: \rho_{xy} \neq 0$$

The sample consists of 25 observations and the sample correlation coefficient is 0.15.

- At the 5% significance level, specify the critical value(s) and the decision rule.
  - Calculate the value of the test statistic.
  - What is the conclusion to the test? Explain.
4. Consider the following competing hypotheses:

$$H_0: \rho_{xy} \geq 0$$

$$H_A: \rho_{xy} < 0$$

The sample consists of 30 observations and the sample correlation coefficient is -0.60.

- Calculate the value of the test statistic.
  - Approximate the  $p$ -value.
  - At the 5% significance level, what is the conclusion to the test? Explain.
5. A sample of 10 observations provides the following statistics:
- $$s_x = 13, \quad s_y = 18, \quad \text{and} \quad s_{xy} = 117.22$$
- Calculate and interpret the sample correlation coefficient  $r_{xy}$ .
  - Specify the hypotheses to determine whether the population correlation coefficient is positive.
  - Calculate the value of the test statistic. Approximate the  $p$ -value.
  - At the 5% significance level, what is the conclusion to the test? Explain.
6. A sample of 25 observations provides the following statistics:

$$s_x = 2, \quad s_y = 5, \quad \text{and} \quad s_{xy} = -1.75$$

- Calculate and interpret the sample correlation coefficient  $r_{xy}$ .

- Specify the competing hypotheses in order to determine whether the population correlation coefficient differs from zero.
- Make a conclusion at the 5% significance level.

### Applications

7. In June 2009, an onslaught of miserable weather in New England played havoc with people's plans and psyches. However, the dreary weather brought a quiet benefit to many city neighborhoods. Police reported that the weather was a key factor in reducing fatal and nondeadly shootings (*The Boston Globe*, July 3, 2009). For instance, it rained in Boston on 22 days in June, when 15 shootings occurred. In 2008, the city saw rain on only eight days and 38 shootings occurred. The accompanying table shows the number of rainy days and the number of shootings that occurred in June from 2005 to 2009.

	Number of Rainy Days	Number of Shootings
June 2005	7	31
June 2006	15	46
June 2007	10	29
June 2008	8	38
June 2009	22	15

SOURCE: *The Boston Globe*, July 3, 2009.

- Calculate and interpret the covariance and the correlation coefficient.
  - Specify the competing hypotheses in order to determine whether there is a negative population correlation between the number of rainy days and crime.
  - Calculate the value of the test statistic and approximate its  $p$ -value.
  - At the 5% significance level, what is the conclusion to the test? Does it appear that dreary weather and crime are negatively correlated?
8. **FILE 2010\_StockReturns.** Diversification is considered important in finance because it allows investors to reduce risk by investing in a variety of assets. It is especially effective when the correlation between the assets is low. Consider the accompanying table, which shows a portion of monthly data on closing stock prices of four companies in 2010.

Month	Microsoft	Coca Cola	Bank of America	General Electric
Jan	27.61	49.52	15.13	15.64
Feb	28.22	54.88	16.61	15.72
⋮	⋮	⋮	⋮	⋮
Dec	27.91	55.58	13.34	18.29

SOURCE: finance.yahoo.com.

- Compute the correlation coefficients between all pairs of stock prices.
  - Suppose an investor already has a stake in Microsoft and would like to add another asset to her portfolio. Which of the remaining three assets will give her the maximum benefit of diversification? (*Hint: Find the asset with the lowest correlation with Microsoft.*)
  - Suppose an investor does not own any of the above four stocks. Pick two stocks so that she gets the maximum benefit of diversification.
9. A realtor studies the relationship between the size of a house (in square feet) and the property taxes owed by the owner. He collects the following data on six homes in an affluent suburb 60 miles outside of New York City.

	Square Feet	Property Taxes (\$)
Home 1	4,182	12,540
Home 2	2,844	9,363
Home 3	5,293	22,717
Home 4	2,284	6,508
Home 5	1,586	5,355
Home 6	3,394	7,901

- Construct and interpret a scatterplot.
  - Calculate and interpret  $s_{xy}$  and  $r_{xy}$ .
  - Specify the competing hypotheses in order to determine whether the population correlation between the size of a house and property taxes differs from zero.
  - Calculate the value of the test statistic and approximate its  $p$ -value.
  - At the 5% significance level, what is the conclusion to the test?
10. **FILE Happiness\_Age.** Many attempts have been made to relate happiness with various factors. One such study relates happiness with age and finds that holding everything else constant, people are least happy when they are in their mid-40s (*The Economist*, December 16, 2010). The accompanying table shows a portion of data on

a respondent's age and his/her perception of well-being on a scale from 0 to 100.

Age	Happiness
49	62
51	66
⋮	⋮
69	72

- Calculate and interpret the sample correlation coefficient between age and happiness.
  - Is the population correlation coefficient statistically significant at the 1% level?
  - Construct a scatterplot to point out a flaw with the above correlation analysis.
11. **FILE Points.** The following table lists the National Basketball Association's leading scorers, their average points per game (PPG), and their average minutes per game (MPG) for 2008.

	PPG	MPG
D. Wade	30.2	38.6
L. James	28.4	37.7
K. Bryant	26.8	36.1
D. Nowitzki	25.9	37.3
D. Granger	25.8	36.2
K. Durant	25.3	39.0
C. Paul	22.8	38.5
C. Anthony	22.8	34.5
C. Bosh	22.7	38.0
B. Roy	22.6	37.2

SOURCE: [www.espn.com](http://www.espn.com).

- Calculate and interpret the sample correlation coefficient between PPG and MPG.
- Specify the competing hypotheses in order to determine whether the population correlation between PPG and MPG is positive.
- Calculate the value of the test statistic and the corresponding  $p$ -value.
- At the 5% significance level, what is the conclusion to the test? Is this result surprising? Explain.

## 14.2 THE SIMPLE LINEAR REGRESSION MODEL

**LO 14.3**

As mentioned earlier, the covariance and the correlation coefficient may establish a linear relationship between two variables, but the measures do not suggest that one variable causes change in the other variable. With **regression analysis**, we change the emphasis from correlation to causation. Here, we explicitly assume that one variable, called the **response variable**, is influenced or caused by other variables, called the **explanatory variables**. Consequently, we use information on the explanatory variables to predict and/or describe changes in the response variable. Alternative names for the explanatory

Estimate the simple linear regression model and interpret the coefficients.

variables are independent variables, predictor variables, control variables, or regressors, while the response variable is often referred to as the dependent variable, the explained variable, the predicted variable, or the regressand.

Regression analysis is one of the most widely used statistical methodologies in business, engineering, and the social sciences. In the introductory case, Madelyn is interested in examining how income and the unemployment rate might influence debt payments. In another scenario, we may want to predict a firm's sales based on its advertising; estimate an individual's salary based on education and years of experience; predict the selling price of a house on the basis of its size and location; or describe auto sales with respect to consumer income, interest rates, and price discounts. In all of these examples, we can use regression analysis to describe the causal relationship between the variables of interest.

No matter the response variable that we choose to examine, we cannot expect to predict its exact value because some omitted explanatory variables may also influence it. If the value of the response variable is uniquely determined by the values of the explanatory variables, we say that the relationship between the variables is **deterministic**. This is often the case in the physical sciences. For example, momentum  $p$  is the product of the mass  $m$  and velocity  $v$  of an object; that is,  $p = mv$ . In most fields of research, however, we tend to find that the relationship between the explanatory variables and the response variable is **inexact**, due to the omission of relevant factors (sometimes not measurable) that influence the response variable. For instance, debt payments are likely to be influenced by the household size costs—a variable that is not included in the introductory case. Similarly, when trying to predict an individual's salary, the individual's natural ability is often omitted since it is extremely difficult, if not impossible, to quantify.

#### DETERMINISTIC VERSUS INEXACT RELATIONSHIPS

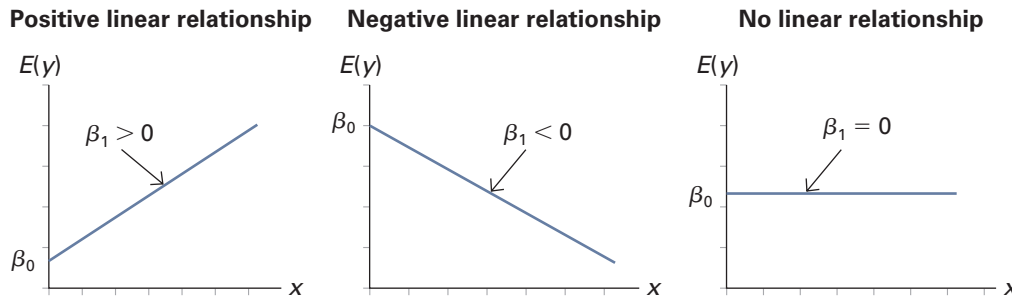
The relationship between the response variable and the explanatory variables is **deterministic** if the value of the response variable is uniquely determined by the explanatory variables; otherwise, the relationship is **inexact**.

Our objective is to develop a mathematical model that captures the relationship between the response variable  $y$  and the  $k$  explanatory variables  $x_1, x_2, \dots, x_k$ . The model must also account for the randomness that is a part of real life. In order to develop a linear regression model, we start with a deterministic component that approximates the relationship we want to model, and then add a random term to it, making the relationship inexact.

In this section, we focus on the **simple linear regression model**, which uses one explanatory variable, denoted  $x_1$ , to explain the variation in the response variable, denoted  $y$ . For ease of exposition when discussing the simple linear regression model, we often drop the subscript on the explanatory variable and refer to it solely as  $x$ . In the next section, we extend the simple linear regression model to the **multiple linear regression model**, where more than one explanatory variable is presumed to have a linear relationship with the response variable.

A fundamental assumption underlying the simple linear regression model is that the expected value of  $y$  lies on a straight line, denoted by  $\beta_0 + \beta_1 x$ , where  $\beta_0$  and  $\beta_1$  (the Greek letters read as betas) are the unknown intercept and slope parameters, respectively. (You have actually seen this relationship before, but you just used different notation. Recall the equation for a line:  $y = mx + b$ , where  $b$  and  $m$  are the intercept and the slope, respectively, of the line.)

The expression  $\beta_0 + \beta_1 x$  is the deterministic component of the simple linear regression model, which can be thought of as the expected value of  $y$  for a given value of  $x$ . In other words, conditional on  $x$ ,  $E(y) = \beta_0 + \beta_1 x$ . The slope parameter  $\beta_1$  determines whether the linear relationship between  $x$  and  $E(y)$  is positive ( $\beta_1 > 0$ ) or negative ( $\beta_1 < 0$ );  $\beta_1 = 0$  indicates that there is no linear relationship. Figure 14.2 shows the expected value of  $y$  for various values of the intercept  $\beta_0$  and the slope  $\beta_1$  parameters.



**FIGURE 14.2** Various examples of a simple linear regression model

As noted earlier, the actual value  $y$  may differ from the expected value  $E(y)$ . Therefore, we add a random error term  $\varepsilon$  (the Greek letter read as epsilon) to develop a simple linear regression model.

#### THE SIMPLE LINEAR REGRESSION MODEL

The simple linear regression model is defined as

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where  $y$  and  $x$  are the response variable and the explanatory variable, respectively, and  $\varepsilon$  is the random error term. The coefficients  $\beta_0$  and  $\beta_1$  are the unknown parameters to be estimated.

### Determining the Sample Regression Equation

The population parameters  $\beta_0$  and  $\beta_1$  used in the simple linear regression model are unknown, and, therefore, must be estimated. As always, we use sample data to estimate the population parameters of interest. Here sample data consist of  $n$  pairs of observations on  $y$  and  $x$ .

Let  $b_0$  and  $b_1$  represent the estimates of  $\beta_0$  and  $\beta_1$ , respectively. We form the **sample regression equation** as  $\hat{y} = b_0 + b_1 x$ , where  $\hat{y}$  (read as y-hat) is the predicted value of the response variable given a specified value of the explanatory variable  $x$ . For a given value of  $x$ , the observed and the predicted values of the response variable are likely to be different since many factors besides  $x$  influence  $y$ . We refer to the difference between the observed and the predicted values of  $y$ , that is  $y - \hat{y}$ , as the **residual**  $e$ .

The **sample regression equation** for the simple linear regression model is denoted as

$$\hat{y} = b_0 + b_1 x,$$

where  $b_0$  and  $b_1$  are the estimates of  $\beta_0$  and  $\beta_1$ , respectively.

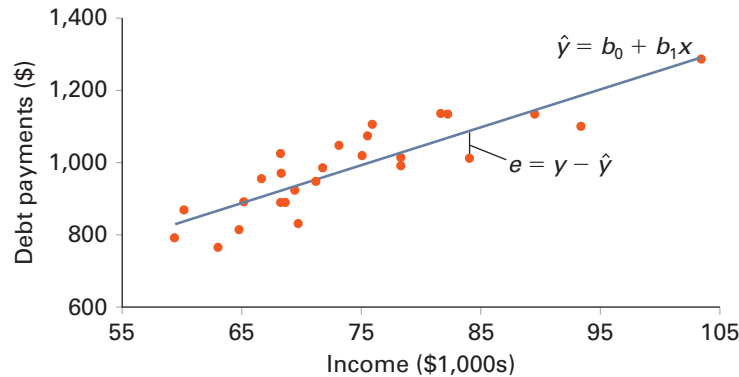
The difference between the observed and the predicted values of  $y$  represents the **residual**  $e$ ; that is,  $e = y - \hat{y}$ .

Before estimating a simple linear regression model, it is useful to visualize the relationship between  $y$  and  $x$  by constructing a scatterplot. Here, we explicitly place  $y$  on the vertical axis and  $x$  on the horizontal axis, implying that  $x$  influences the variation in  $y$ . In Figure 14.3, we use the data from the introductory case to show a scatterplot of debt payments against income. We then superimpose a linear trendline through the points on the scatterplot.

The superimposed line in Figure 14.3 is the sample regression equation,  $\hat{y} = b_0 + b_1 x$ , where  $y$  and  $x$  represent debt payments and income, respectively. The upward slope of the

line suggests that as income increases, the predicted debt payments also increase. Also, the vertical distance between any data point on the scatterplot and the corresponding point on the line,  $y$  and  $\hat{y}$ , represents the residual,  $e = y - \hat{y}$ .

**FIGURE 14.3**  
Scatterplot with  
a superimposed  
trendline



**FILE**

*Debt\_Payments*

## Using Excel to Construct a Scatterplot and a Trendline

In order to replicate Figure 14.3 using Excel, we follow these steps.

- A.** Open the *Debt\_Payments* data file. For the purpose of creating a scatterplot of debt payments against income, disregard the column with the unemployment data.
- B.** Simultaneously select the data for Income and Debt and choose **Insert > Scatter**. Select the graph on the top left.
- C.** Right-click on the scatter points, choose **Add Trendline**, and then choose **Linear**.
- D.** Further formatting regarding colors, axes, etc. can be done by selecting **Layout** from the menu.

A common approach to fitting a line to the scatterplot is the **method of least squares**, also referred to as **ordinary least squares (OLS)**. In other words, we use OLS to estimate the parameters  $\beta_0$  and  $\beta_1$ . OLS estimators have many desirable properties if certain assumptions hold (these assumptions are discussed in the next chapter). The OLS method chooses the line whereby the **sum of squares due to error, SSE**, also referred to as the **error sum of squares**, is minimized, where  $SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$ .  $SSE$  is the sum of the squared differences between the observed values  $y$  and their predicted values  $\hat{y}$ , or equivalently, the sum of the squared distances from the regression equation. Thus, using this distance measure, we say that the OLS method produces the straight line that is “closest” to the data. In the context of Figure 14.3, the superimposed line has been estimated by OLS.

Using calculus, equations have been developed for  $b_0$  and  $b_1$  that satisfy the OLS criterion.

### CALCULATING THE COEFFICIENTS $b_1$ AND $b_0$

The slope  $b_1$  and the intercept  $b_0$  of the sample regression equation are calculated as

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \text{and}$$

$$b_0 = \bar{y} - b_1\bar{x}.$$

It is important to be able to interpret the estimated regression coefficients. As we will see in the following example, it is not always possible to provide an economic interpretation of the intercept estimate  $b_0$ ; mathematically, however, it represents the predicted



value of  $\hat{y}$  when  $x$  has a value of zero. The slope estimate  $b_1$  represents the change in  $\hat{y}$  when  $x$  increases by one unit.

### EXAMPLE 14.4

Using the data from Table 14.1, let debt payments represent the response variable and income represent the explanatory variable in a simple linear regression model.

- Calculate and interpret  $b_1$ .
- Calculate and interpret  $b_0$ .
- What is the sample regression equation?
- Predict debt payments if income is \$80,000.

**SOLUTION:** The simple linear regression model,  $\text{Debt} = \beta_0 + \beta_1 \text{Income} + \varepsilon$ , or simply,  $y = \beta_0 + \beta_1 x + \varepsilon$ , where  $y$  and  $x$  represent debt payments and income, respectively.

- We first find the sample mean of income and debt payments and obtain  $\bar{x} = 74.05$  and  $\bar{y} = 983.46$  (calculations not shown). In order to obtain  $b_1$ , we calculate deviations from the mean for both  $x$  and  $y$ , as shown in the first two columns of Table 14.3. We then calculate the product of deviations from the mean, as shown in the third column. The sum of the products of the deviations from the mean is the numerator in the formula for  $b_1$ ; this value is found in the last cell of the third column. The fourth column shows the squared deviations from the mean for  $x$ . The sum of these squared deviations, found in the last cell of the fourth column, is the denominator in the formula for  $b_1$ .

**TABLE 14.3** Calculations for Example 14.4

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
$103.50 - 74.05 = 29.45$	$1285 - 983.46 = 301.54$	$(29.45)(301.54) = 8880.31$	$(29.45)^2 = 867.30$
$81.70 - 74.05 = 7.65$	$1135 - 983.46 = 151.54$	$(7.65)(151.54) = 1159.27$	$(7.65)^2 = 58.52$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$63.00 - 74.05 = -11.05$	$763 - 983.46 = -220.46$	$(-11.05)(-220.46) = 2436.10$	$(-11.05)^2 = 122.10$
		$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 27979.50$	$\Sigma(x_i - \bar{x})^2 = 2679.75$

Using the summations from the last cells of the last two columns, we compute

$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{27979.50}{2679.75} = 10.44$ . As anticipated, the slope is positive, suggesting a positive relationship between income and debt payments. Since income is measured in \$1,000s, our interpretation is that if median household income increases by \$1,000, then on average, we predict consumer debt payments to increase by  $b_1$ ; that is, by \$10.44.

- Using  $b_1 = 10.44$  and the sample means,  $\bar{x} = 74.05$  and  $\bar{y} = 983.46$ , we obtain an estimate for  $b_0$  as  $b_0 = \bar{y} - b_1 \bar{x} = 983.46 - 10.44(74.05) = 210.38$ ; the exact value computed without rounding is 210.30. This estimated intercept coefficient of 210.30 suggests that if income equals zero, then predicted debt payments are \$210.30. In this particular application, this conclusion makes some sense, since a household with no income still needs to make debt payments for any credit card use, for example, automobile loans, etc. However, we should be careful about predicting  $y$  when we use a value for  $x$  that is not included in the sample range of  $x$ . In the *Debt\_Payments* data set, the lowest and highest values for income (in \$1,000s) are \$59.40 and \$103.50, respectively; plus the scatterplot suggests that a line fits the data well within this range of the explanatory variable. Unless we assume that income and debt payments

will maintain the same linear relationship at income values less than \$59.40 and more than \$103.50, we should refrain from making predictions based on values of the explanatory variable outside the sample range.

- c. With  $b_0 = 210.30$  and  $b_1 = 10.44$ , we write the sample regression equation as  $\hat{y} = 210.30 + 10.44x$ ; that is,  $\widehat{\text{Debt}} = 210.30 + 10.44\text{Income}$ .
- d. Note that income is measured in \$1,000s; therefore, if income equals \$80,000, we input  $\text{Income} = 80$  in the sample regression equation and find predicted debt payments as  $\text{Debt} = 210.30 + 10.44(80) = \$1,045.50$ .

## Using Excel to Find the Sample Regression Equation

Fortunately, we rarely have to calculate a sample regression equation by hand. Virtually every statistical software package computes the necessary output to construct a sample regression equation. In addition, values of all relevant statistics for assessing the model, discussed shortly, are also included.

### FILE

*Debt\_Payments*

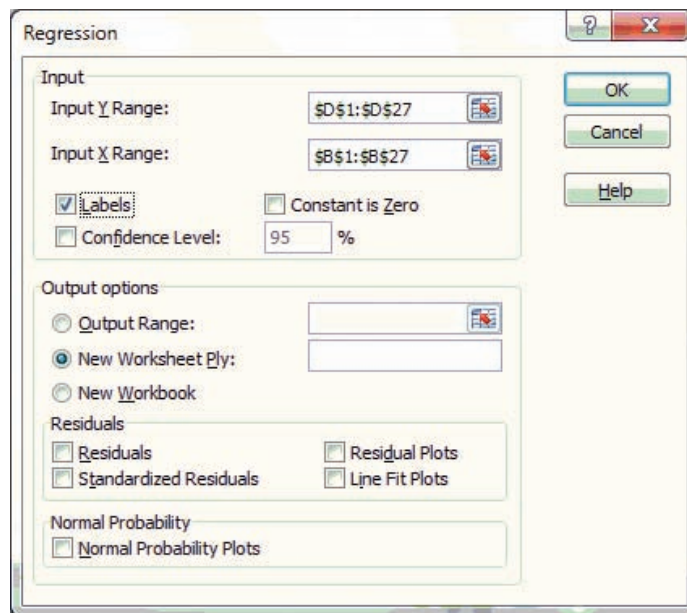
### EXAMPLE 14.5

Given the data from Table 14.1, use Excel to reestimate the sample regression equation with debt payments as the response variable and income as the explanatory variable.

#### SOLUTION:

- A. Open the *Debt\_Payments* data file.
- B. Choose **Data > Data Analysis > Regression** from the menu.
- C. See Figure 14.4. In the *Regression* dialog box, click on the box next to *Input Y Range*, then select the Debt data, including its heading. For *Input X Range*, select the Income data, including its heading. Check *Labels*, since we are using Debt and Income as headings.
- D. Click **OK**.

**FIGURE 14.4** Regression dialog box for Example 14.5



The Excel output is presented in Table 14.4.

**TABLE 14.4** Regression Results for Example 14.5

Regression Statistics						
Multiple R	0.8675					
R Square	0.7526					
Adjusted R Square	0.7423					
Standard Error	63.2606					
Observations	26					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	292136.9086	292136.9086	72.9996	9.6603E-09	
Residual	24	96045.5529	4001.8980			
Total	25	388182.4615				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p-Value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	<b>210.2977</b>	91.3387	2.3024	0.0303	21.7838	398.8116
Income	<b>10.4411</b>	1.2220	8.5440	9.6603E-09	7.9189	12.9633

As Table 14.4 shows, Excel produces quite a bit of statistical information. In order to answer the questions in Example 14.5, we only need the estimated coefficients, which we have put in boldface. We will address the remaining information at the end of this chapter as well as in Chapter 15. The estimates for  $\beta_0$  and  $\beta_1$  are  $b_0 = 210.2977$  and  $b_1 = 10.4411$ . The sample regression equation,  $\widehat{\text{Debt}} = 210.30 + 10.44\text{Income}$ , is virtually the same as the one we calculated with the formulas.

## EXERCISES 14.2

### Mechanics

12. In a simple linear regression, the following information is given:

$$\bar{x} = 34; \bar{y} = 44; \Sigma(x_i - \bar{x})(y_i - \bar{y}) = 1250;$$

$$\Sigma(x_i - \bar{x})^2 = 925$$

- Calculate  $b_1$ .
- Calculate  $b_0$ .
- What is the sample regression equation? Predict  $y$  if  $x$  equals 40.

13. In a simple linear regression, the following information is given:

$$\bar{x} = -25; \bar{y} = 56; \Sigma(x_i - \bar{x})(y_i - \bar{y}) = -866;$$

$$\Sigma(x_i - \bar{x})^2 = 711$$

- Calculate  $b_1$ .
- Calculate  $b_0$ .
- What is the sample regression equation? Predict  $y$  if  $x$  equals  $-20$ .

14. In a simple linear regression, the following sample regression equation is obtained:

$$\hat{y} = 15 + 2.5x.$$

- Predict  $y$  if  $x$  equals 10.
- What happens to this prediction if  $x$  doubles in value?

15. In a simple linear regression, the following sample regression equation is obtained:

$$\hat{y} = 436 - 17x.$$

- Interpret the slope coefficient.
- Predict  $y$  if  $x$  equals  $-15$ .

16. Consider the following sample data:

<i>x</i>	12	23	11	23	14	21	18	16
<i>y</i>	28	43	21	40	33	41	37	32

- Construct a scatterplot and verify that estimating a simple linear regression model is appropriate in this problem.

- Calculate  $b_1$  and  $b_0$ . What is the sample regression equation?
- Find the predicted value for  $y$  if  $x$  equals 10, 15, and 20.

17. Consider the following sample data:

x	22	24	27	21	23	14	14	15
y	101	139	250	88	87	14	16	20

- Construct a scatterplot and verify that estimating a simple linear regression model is appropriate in this problem.
- Calculate  $b_1$  and  $b_0$ . What is the sample regression equation?
- Find the predicted value for  $y$  if  $x$  equals 15, 20, and 25.

18. Thirty observations were used to estimate  $y = \beta_0 + \beta_1 x + \varepsilon$ . A portion of the Excel results is as follows:

	Coefficients	Standard Error	t Stat	p-value
Intercept	41.82	8.58	4.87	3.93E-05
x	0.49	0.10	4.81	4.65E-05

- What is the estimate for  $\beta_1$ ? Interpret this value.
- What is the sample regression equation?
- If  $x = 30$ , what is  $\hat{y}$ ?

19. Twenty-four observations were used to estimate  $y = \beta_0 + \beta_1 x + \varepsilon$ . A portion of the Excel results is as follows:

	Coefficients	Standard Error	t Stat	p-value
Intercept	2.25	2.36	0.95	0.3515
x	-0.16	0.30	-0.53	0.6017

- What is the estimate for  $\beta_1$ ? Interpret this value.
- What is the sample regression equation?
- What is the predicted value for  $y$  if  $x = 2$ ? If  $x = -2$ ?

## Applications

20. The director of graduate admissions at a large university is analyzing the relationship between scores on the math portion of the Graduate Record Examination (GRE) and subsequent performance in graduate school, as measured by a student's grade point average (GPA). She uses a sample of 8 students who graduated within the past five years. The data are as follows:

GRE	700	720	650	750	680	730	740	780
GPA	3.0	3.5	3.2	3.7	3.1	3.9	3.3	3.5

- Construct a scatterplot placing GRE on the horizontal axis.
- Find the sample regression equation for the model:  $GPA = \beta_0 + \beta_1 GRE + \varepsilon$ .
- What is a student's predicted GPA if he/she scored 710 on the math portion of the GRE?

21. A social scientist would like to analyze the relationship between educational attainment and salary. He collects the following sample data, where Education refers to years of higher education and Salary is the individual's annual salary (in \$1,000s):

Education	3	4	6	2	5	4	8	0
Salary	40	53	80	42	70	50	110	38

- Find the sample regression equation for the model:  $Salary = \beta_0 + \beta_1 Education + \varepsilon$ .
  - Interpret the coefficient for Education.
  - What is the predicted salary for an individual who completed 7 years of higher education?
22. The owner of several used-car dealerships believes that the selling price of a used car can best be predicted using the car's age. He uses data on the recent selling price and age of 20 used sedans to estimate  $Price = \beta_0 + \beta_1 Age + \varepsilon$ . A portion of the Excel results is as follows:

	Coefficients	Standard Error	t Stat	p-value
Intercept	21187.94	733.42	28.89	1.56E-16
Age	-1208.25	128.95	-9.37	2.41E-08

- What is the estimate for  $\beta_1$ ? Interpret this value.
  - What is the sample regression equation?
  - Predict the selling price of a 5-year-old sedan.
23. If a firm spends more on advertising, is it likely to increase sales? Data on annual sales (in \$100,000s) and advertising expenditures (in \$10,000s) were collected for 20 firms in order to estimate the model  $Sales = \beta_0 + \beta_1 Advertising + \varepsilon$ . A portion of the Excel results is as follows:

	Coefficients	Standard Error	t Stat	p-value
Intercept	-7.42	1.46	-5.09	7.66E-05
Advertising	0.42	0.05	8.70	7.26E-08

- Is the sign on the slope as expected? Explain.
  - What is the sample regression equation?
  - Predict the sales for a firm that spends \$500,000 annually on advertising.
24. **FILE Consumption Function.** The consumption function captures one of the key relationships in economics that was first developed by John Maynard Keynes. It expresses consumption as a function of disposable income, where disposable income is income after taxes. The accompanying table shows a portion of average U.S. annual consumption and disposable income for the years 1985–2006.

	Consumption	Disposable Income
1985	23490	22887
1986	23866	23172
⋮	⋮	⋮
2006	48398	58101

SOURCE: *The Statistical Abstract of the United States.*

- Use Excel to estimate the model:  $\text{Consumption} = \beta_0 + \beta_1 \text{Disposable Income} + \varepsilon$ .
  - What is the sample regression equation?
  - In this model, the slope coefficient is called the marginal propensity to consume. Interpret its meaning.
  - What is predicted consumption if disposable income is \$57,000?
25. **FILE MLB Pitchers.** The following table lists Major League Baseball's (MLB's) leading pitchers, their earned run average (ERA), and their salary (in \$1,000,000s) for 2008.

	ERA	Salary (in \$1,000,000s)
J. Santana	2.53	17.0
C. Lee	2.54	4.0
T. Lincecum	2.62	0.4
C. Sabathia	2.70	11.0
R. Halladay	2.78	10.0
J. Peavy	2.85	6.5
D. Matsuzaka	2.90	8.3
R. Dempster	2.96	7.3
B. Sheets	3.09	12.1
C. Hamels	3.09	0.5

SOURCE: [www.ESPN.com](http://www.ESPN.com).

- Use Excel to estimate the model:  $\text{Salary} = \beta_0 + \beta_1 \text{ERA} + \varepsilon$  and interpret the coefficient of ERA.
  - Use the estimated model to predict the salary for each player, given his ERA. For example, use the sample regression equation to predict the salary for J. Santana with ERA = 2.53.
  - Derive the corresponding residuals and explain why the residuals might be so high.
26. **FILE Happiness Age.** Refer to the accompanying data file on happiness and age to answer the following questions.
- Use Excel to estimate a simple linear regression model with Happiness as the response variable and Age as the explanatory variable.
  - Use the sample regression equation to predict Happiness when Age equals 25, 50, and 75.

- Construct a scatterplot of Happiness against Age. Discuss why your predictions might not be accurate.

27. **FILE Property Taxes.** The accompanying table shows a portion of data that refers to the size of a home (in square feet) and its property taxes owed by the owner (in \$) in an affluent suburb 30 miles outside New York City.

Size (in square feet)	Property Taxes (in \$)
2449	21928
2479	17339
⋮	⋮
2864	29235

- Determine the sample regression equation that enables us to predict property taxes on the basis of the size of the home.
  - Interpret the slope coefficient.
  - Predict the property taxes for a 1500-square-foot home.
28. **FILE Test Scores.** The accompanying table shows a portion of the scores that 32 students obtained on the midterm and the final in a course in statistics.

Midterm	Final
78	86
97	94
⋮	⋮
47	91

- Determine the sample regression equation that enables us to predict a student's final score on the basis of his/her midterm score.
  - Predict the final score of a student who received an 80 on the midterm.
29. **FILE Fertilizer.** A horticulturist is studying the relationship between tomato plant height and fertilizer amount. Thirty tomato plants grown in similar conditions were subjected to various amounts of fertilizer over a four-month period, and then their heights were measured. A portion of the data is shown in the accompanying table.

Tomato Plant Height (inches)	Fertilizer Amount (ounces)
20.4	1.9
49.2	5.0
⋮	⋮
46.4	3.1

- Use Excel to estimate the regression model:  $\text{Height} = \beta_0 + \beta_1 \text{Fertilizer} + \varepsilon$ .
- Interpret the coefficient of Fertilizer. Does the y-intercept make practical sense?

- c. Use the estimated model to predict, after four months, the height of a tomato plant which received 3.0 ounces of fertilizer.

30. **FILE Dexterity.** Finger dexterity, the ability to make precisely coordinated finger movements to grasp or assemble very small objects, is important in jewelry making. Subsequently, the manufacturing manager at Gemco, a manufacturer of high-quality watches, wants to develop a regression model to predict the productivity (in watches per shift) of new employees based on dexterity. He has subjected a sample of 20 current employees to the O'Connor dexterity test in which the time required to place 3 pins in each of 100 small holes using tweezers is measured. A portion of the data is shown in the accompanying table.

Time (seconds)	Watches per shift
513	23
608	19
⋮	⋮
437	20

- Use Excel to estimate the model:  $\text{Watches} = \beta_0 + \beta_1 \text{Time} + \varepsilon$ .
- Interpret the coefficient of Time.
- Explain why the  $y$ -intercept makes no practical sense in this particular problem.
- Suppose a new employee takes 550 seconds on the dexterity test. How many watches per shift is she expected to produce?

#### LO 14.4

Estimate the multiple linear regression model and interpret the coefficients.

## 14.3 THE MULTIPLE LINEAR REGRESSION MODEL

The simple linear regression model allows us to analyze the linear relationship between one explanatory variable and the response variable. However, by restricting the number of explanatory variables to one, we sometimes reduce the potential usefulness of the model. In Chapter 15, we will discuss how the OLS estimates can be quite misleading when important explanatory variables are excluded. A **multiple linear regression model** allows us to study how the response variable is influenced by two or more explanatory variables. The choices of the explanatory variables are based on economic theory, intuition, and/or prior research. The multiple linear regression model is a straightforward extension of the simple linear regression model.

### THE MULTIPLE LINEAR REGRESSION MODEL

The multiple linear regression model is defined as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

where  $y$  is the response variable,  $x_1, x_2, \dots, x_k$  are the  $k$  explanatory variables, and  $\varepsilon$  is the random error term. The coefficients  $\beta_0, \beta_1, \dots, \beta_k$  are the unknown parameters to be estimated.

The difference between the observed and the predicted values of  $y$  represents the residual  $e$ ; that is,  $e = y - \hat{y}$ .

### Determining the Sample Regression Equation

As in the case of the simple linear regression model, we apply the OLS method that minimizes  $SSE$ , where  $SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$ .

The **sample regression equation** for the multiple linear regression model is denoted as

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k,$$

where  $b_0, b_1, \dots, b_k$  are the estimates of  $\beta_0, \beta_1, \dots, \beta_k$ .

For each explanatory variable  $x_j$  ( $j = 1, \dots, k$ ), the corresponding slope coefficient  $b_j$  is the estimate of  $\beta_j$ . We slightly modify the interpretation of the slope coefficients in the context



of a multiple linear regression model. Here  $b_j$  measures the change in the predicted value of the response variable  $\hat{y}$  given a unit increase in the associated explanatory variable  $x_j$ , *holding all other explanatory variables constant*. In other words, it represents the partial influence of  $x_j$  on  $\hat{y}$ .

When we used formulas to estimate the simple linear regression model, we found that the calculations were quite cumbersome. As you might imagine, if we were to estimate the multiple linear regression model by hand, the calculations would become even more tedious. Thus, we rely solely on using statistical packages to estimate a multiple linear regression model.

### EXAMPLE 14.6

In the previous section, we analyzed how debt payments are influenced by income, ignoring the possible effect of the unemployment rate.

- Given the data from Table 14.1, estimate the multiple linear regression model with debt payments as the response variable, and income and the unemployment rate as the explanatory variables.
- Interpret the regression coefficients.
- Predict debt payments if income is \$80,000 and the unemployment rate is 7.5%.

#### SOLUTION:

- We will use Excel to estimate the multiple linear regression model,  $\text{Debt} = \beta_0 + \beta_1 \text{Income} + \beta_2 \text{Unemployment} + \varepsilon$ . We follow similar steps as we did when we estimated the simple linear regression model.
  - Open the *Debt\_Payments* data file. *Debt\_Payments*.
  - Choose **Data > Data Analysis > Regression** from the menu.
  - In the *Regression* dialog box, click on the box next to *Input Y Range*, then select the data for Debt. For *Input X Range*, *simultaneously* select the data for Income and Unemployment. Select *Labels*, since we are using Debt, Income, and Unemployment as headings.
  - Click **OK**.

We show the Excel output in Table 14.5.

**TABLE 14.5** Regression Results for Example 14.6

Regression Statistics						
Multiple R		0.8676				
R Square		0.7527				
Adjusted R Square		0.7312				
Standard Error		64.6098				
Observations		26				
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	292170.8	146085.4	34.99536	1.05E-07	
Residual	23	96011.7	4174.4			
Total	25	388182.5				
	Coefficients	Standard Error	t Stat	p-Value	Lower 95%	Upper 95%
Intercept	198.9956	156.3619	1.2727	0.2159	-124.464	522.455
Income	10.5122	1.4765	7.1120	2.98E-07	7.458	13.567
Unemployment	0.6186	6.8679	0.0901	0.9290	-13.589	14.826

**FILE**  
*Debt\_Payments*

Using the boldface estimates from Table 14.5,  $b_0 = 198.9956$ ,  $b_1 = 10.5122$ , and  $b_2 = 0.6186$ , we derive the sample regression equation as

$$\widehat{\text{Debt}} = 199.00 + 10.51\text{Income} + 0.62\text{Unemployment}.$$

- b. The regression coefficient of Income is 10.51. Since income is measured in \$1,000s, the model suggests that if income increases by \$1,000, then debt payments are predicted to increase by \$10.51, holding the unemployment rate constant. Similarly, the regression coefficient of Unemployment is 0.62, implying that a one percentage point increase in the unemployment rate leads to a predicted increase in debt payments of \$0.62, holding income constant. It seems that the predicted impact of Unemployment, with Income held constant, is rather small. In fact, the influence of the unemployment rate is not even statistically significant at any reasonable level; we will discuss such tests of significance in the next chapter.
- c. If income is \$80,000 and the unemployment rate is 7.5%, predicted debt payments are

$$\widehat{\text{Debt}} = 199.00 + 10.51(80) + 0.62(7.5) = \$1,044.45.$$

## EXERCISES 14.3

### Mechanics

31. In a multiple regression, the following sample regression equation is obtained:

$$\hat{y} = 152 + 12.9x_1 + 2.7x_2.$$

- a. Predict  $y$  if  $x_1$  equals 20 and  $x_2$  equals 35.  
b. Interpret the slope coefficient of  $x_1$ .
32. In a multiple regression, the following sample regression equation is obtained:

$$\hat{y} = -8 + 2.6x_1 - 47.2x_2.$$

- a. Predict  $y$  if  $x_1$  equals 40 and  $x_2$  equals -10.  
b. Interpret the slope coefficient of  $x_2$ .
33. Thirty observations were used to estimate  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon$ . A portion of the Excel results is as follows:

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p-value</i>
Intercept	21.97	2.98	7.37	6.31E-08
$x_1$	30.00	2.23	13.44	1.75E-13
$x_2$	-1.88	0.27	-6.96	1.75E-07

- a. What is the estimate for  $\beta_1$ ? Interpret this value.  
b. What is the sample regression equation?  
c. If  $x_1 = 30$  and  $x_2 = 20$ , what is  $\hat{y}$ ?
34. Forty observations were used to estimate  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon$ . A portion of the regression results is shown in the accompanying table.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p-value</i>
Intercept	13.83	2.42	5.71	1.56E-06
$x_1$	-2.53	0.15	-16.87	5.84E-19
$x_2$	0.29	0.06	4.83	2.38E-05

- a. What is the estimate for  $\beta_1$ ? Interpret this value.  
b. What is the sample regression equation?  
c. What is the predicted value for  $y$  if  $x_1 = -9$  and  $x_2 = 25$ .

35. Consider the following sample data:

$y$	46	51	28	55	29	53	47	36
$x_1$	40	48	29	44	30	58	60	29
$x_2$	13	28	24	11	28	28	29	14

- a. Estimate a multiple linear regression model and interpret its coefficients.  
b. Find the predicted value for  $y$  if  $x_1$  equals 50 and  $x_2$  equals 20.

36. Consider the following sample data:

$y$	52	49	45	54	45	52	40	34
$x_1$	11	10	9	13	9	13	6	7
$x_2$	25	39	25	24	31	22	28	21

- a. Estimate a multiple linear regression model and interpret the coefficient for  $x_2$ .  
b. Find the predicted value for  $y$  if  $x_1$  equals 12 and  $x_2$  equals 30.

## Applications

37. Using data from 50 workers, a researcher estimates  $\text{Wage} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Experience} + \beta_3 \text{Age} + \varepsilon$ , where Wage is the hourly wage rate and Education, Experience, and Age are the years of higher education, the years of experience, and the age of the worker, respectively. A portion of the regression results is shown in the following table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	7.87	4.09	1.93	0.0603
Education	1.44	0.34	4.24	0.0001
Experience	0.45	0.14	3.16	0.0028
Age	-0.01	0.08	-0.14	0.8920

- Interpret the estimates for  $\beta_1$  and  $\beta_2$ .
  - What is the sample regression equation?
  - Predict the hourly wage rate for a 30-year-old worker with 4 years of higher education and 3 years of experience.
38. On the first day of class, an economics professor administers a test to gauge the math preparedness of her students. She believes that the performance on this math test and the number of hours studied per week on the course are the primary factors that predict a student's score on the final exam. Using data from her class of 60 students, she estimates  $\text{Final} = \beta_0 + \beta_1 \text{Math} + \beta_2 \text{Hours} + \varepsilon$ . A portion of the regression results is shown in the following table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	40.55	3.37	12.03	2.83E-17
Math	0.25	0.04	6.06	1.14E-07
Hours	4.85	0.57	8.53	9.06E-12

- What is the slope coefficient of Hours?
  - What is the sample regression equation?
  - What is the predicted final exam score for a student who has a math score of 70 and studies 4 hours per week?
39. Osteoporosis is a degenerative disease that primarily affects women over the age of 60. A research analyst wants to forecast sales of StrongBones, a prescription drug for treating this debilitating disease. She uses the model  $\text{Sales} = \beta_0 + \beta_1 \text{Population} + \beta_2 \text{Income} + \varepsilon$ , where Sales refers to the sales of StrongBones (in \$1,000,000s), Population is the number of women over the age of 60 (in millions), and Income is the average income of women over the age of 60 (in \$1,000s). She collects data on 38 cities across the United States and obtains the following relevant regression results:

	Coefficients	Standard Error	t Stat	p-value
Intercept	10.35	4.02	2.57	0.0199
Population	8.47	2.71	3.12	0.0062
Income	7.62	6.63	1.15	0.2661

- What is the sample regression equation?
  - Interpret the slope coefficients.
  - Predict sales if a city has 1.5 million women over the age of 60 and their average income is \$44,000.
40. A sociologist believes that the crime rate in an area is significantly influenced by the area's poverty rate and median income. Specifically, she hypothesizes crime will increase with poverty and decrease with income. She collects data on the crime rate (crimes per 100,000 residents), the poverty rate (in %), and the median income (in \$1,000s) from 41 New England cities. A portion of the regression results is shown in the following table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	-301.62	549.71	-0.55	0.5864
Poverty	53.16	14.22	3.74	0.0006
Income	4.95	8.26	0.60	0.5526

- Are the signs as expected on the slope coefficients?
  - Interpret the slope coefficient for Poverty.
  - Predict the crime rate in an area with a poverty rate of 20% and a median income of \$50,000.
41. **FILE Arlington\_Homes.** A realtor in Arlington, Massachusetts, is analyzing the relationship between the sale price of a home (Price), its square footage (Sqft), the number of bedrooms (Beds), and the number of bathrooms (Baths). She collects data on 36 recent sales in Arlington in the first quarter of 2009 for the analysis. A portion of the data is shown in the accompanying table.

Price	Sqft	Beds	Baths
840000	2768	4	3.5
822000	2500	4	2.5
⋮	⋮	⋮	⋮
307500	850	1	1

SOURCE: <http://Newenglandmoves.com>.

- Estimate the model  $\text{Price} = \beta_0 + \beta_1 \text{Sqft} + \beta_2 \text{Beds} + \beta_3 \text{Baths} + \varepsilon$ .
  - Interpret the slope coefficients.
  - Predict the price of a 2,500-square-foot home with three bedrooms and two bathrooms.
42. **FILE Engine\_Overhaul.** The maintenance manager at a trucking company wants to build a regression model to forecast the time until the first engine overhaul based on four

independent variables: (1) annual miles driven, (2) average load weight, (3) average driving speed, (4) oil change interval. Based on driver logs and onboard computers, data have been obtained for a sample of 25 trucks. A portion of the data is shown in the accompanying table.

Time until First Engine Overhaul (Years)	Annual Miles Driven (1,000 miles)	Average Load Weight (tons)	Average Driving Speed (mph)	Oil Change Interval (1,000 miles)
79	42.8	19	46	15
0.9	98.5	25	46	29
⋮	⋮	⋮	⋮	⋮
6.1	61.2	24	58	19

- For each explanatory variable, discuss whether it is likely to have a positive or negative causal effect on time until the first engine overhaul.
  - Use Excel to estimate the regression model (use all four explanatory variables).
  - Based on part (a), are the signs of the regression coefficients logical?
  - Predict the time before the first engine overhaul for a particular truck driven 60(000) miles per year with an average load of 22 tons, an average driving speed of 57 mph, and 18(000) miles between oil changes.
43. **FILE MCAS.** Education reform is one of the most hotly debated subjects on both state and national policy makers' list of socioeconomic topics. Consider a linear regression model that relates school expenditures and family background to student performance in Massachusetts using 224 school districts. The response variable is the mean score on the MCAS (Massachusetts Comprehensive Assessment System) exam given in May 1998 to 10th-graders. Four explanatory variables are used: (1) STR is the student-to-teacher ratio, (2) TSAL is the average teacher's salary, (3) INC is the median household income, and (4) SGL is the percentage of single family households. A portion of the data is shown in the accompanying table.

Score	STR (%)	TSAL (in \$1,000)	INC (in \$1,000)	SGL (%)
227.00	19.00	44.01	48.89	4.70
230.67	17.90	40.17	43.91	4.60
⋮	⋮	⋮	⋮	⋮
230.67	19.20	44.79	47.64	5.10

SOURCE: Massachusetts Department of Education and the Census of Population and Housing.

- For each explanatory variable, discuss whether it is likely to have a positive or negative causal effect on Score.

- Find the sample regression equation. Are the signs of the slope coefficients as expected?
  - What is the predicted score if STR = 18, TSAL = 50, INC = 60, SGL = 5?
  - What is the predicted score if everything else is the same as in part (c) except INC = 80?
44. **FILE Electricity\_Cost.** The facility manager at a pharmaceutical company wants to build a regression model to forecast monthly electricity cost. Three main variables are thought to dictate electricity cost: (1) average outdoor temperature, (2) working days per month, (3) tons of product produced. A portion of the past year's monthly data is shown in the accompanying table.
- | Cost (\$) | Ave Temp (°F) | Work Days | Tons Produced |
|-----------|---------------|-----------|---------------|
| 24100     | 26            | 24        | 80            |
| 23700     | 32            | 21        | 73            |
| ⋮         | ⋮             | ⋮         | ⋮             |
| 26000     | 39            | 22        | 69            |
- For each explanatory variable, discuss whether it is likely to have a positive or negative causal effect on monthly electricity cost.
  - Use Excel to estimate the regression model.
  - Are the signs of the regression coefficients as expected? If not, speculate as to why this could be the case.
  - What is the predicted electricity cost in a month during which the average outdoor temperature is 65°, there are 23 working days, and 76 tons are produced?
45. **FILE Quarterback\_Salaries.** American football is the highest paying sport on a per-game basis. The quarterback, considered the most important player on the team, is appropriately compensated. A sports statistician wants to use 2009 data to estimate a multiple linear regression model that links the quarterback's salary with his pass completion percentage (PCT), total touchdowns scored (TD), and his age. A portion of the data is shown in the accompanying table.

Name	Salary (in \$ millions)	PCT	TD	Age
Philip Rivers	25.5566	65.2	28	27
Jay Cutler	22.0441	60.5	27	26
⋮	⋮	⋮	⋮	⋮
Tony Romo	0.6260	63.1	26	29

SOURCE: *USA Today* database for salaries; <http://NFL.com> for other data.

- Estimate the model defined as  $\text{Salary} = \beta_0 + \beta_1 \text{PCT} + \beta_2 \text{TD} + \beta_3 \text{Age} + \varepsilon$ .
- Are you surprised by the estimated coefficients?
- Drew Brees earned 12.9895 million dollars in 2009. According to the model, what is his predicted salary if  $\text{PCT} = 70.6$ ,  $\text{TD} = 34$ , and  $\text{Age} = 30$ ?
- Tom Brady earned 8.0073 million dollars in 2009. According to the model, what is his predicted salary if  $\text{PCT} = 65.7$ ,  $\text{TD} = 28$ , and  $\text{Age} = 32$ ?
- Compute and interpret the residual salary for Drew Brees and Tom Brady.

46. **FILE AnnArbor\_Rental.** The accompanying table shows a portion of data consisting of the rent, the number of bedrooms, the number of bathrooms, and the square footage for 40 apartments in the college town of Ann Arbor, Michigan.

Rent	Bed	Bath	Sqft
645	1	1	500
675	1	1	648
⋮	⋮	⋮	⋮
2400	3	2.5	2700

- Determine the sample regression equation that enables us to predict the rent of an Ann Arbor apartment on the basis of the number of bedrooms, the number of bathrooms, and the square footage.
  - Interpret the slope coefficient of Bath.
  - Predict the rent for a 1500-square-foot apartment with 2 bedrooms and 1 bathroom.
47. **FILE Car\_Prices.** The accompanying table shows a portion of data consisting of the selling price, the age, and the mileage for 20 used sedans.

Selling Price	Age	Mileage
13590	6	61485
13775	6	54344
⋮	⋮	⋮
11988	8	42408

- Determine the sample regression equation that enables us to predict the price of a sedan on the basis of its age and mileage.
- Interpret the slope coefficient of Age.
- Predict the selling price of a five-year-old sedan with 65,000 miles.

## 14.4 GOODNESS-OF-FIT MEASURES

By simply observing the sample regression equation, we cannot assess how well the explanatory variables explain the variation in the response variable. However, several objective “goodness-of-fit” measures do exist that summarize how well the sample regression equation fits the data. If all the observations lie on the sample regression equation, then we have a perfect fit. Since that almost never happens, we evaluate the models on a relative basis.

We will study three goodness-of-fit measures: the standard error of the estimate, the coefficient of determination, and the adjusted coefficient of determination. The relevant formulas used to derive these measures are applicable for both simple and multiple linear regression models.

In the introductory case study, we were interested in analyzing consumer debt payments. We estimated two models. Let Model 1 represent the simple linear regression model,  $\text{Debt} = \beta_0 + \beta_1 \text{Income} + \varepsilon$ , and Model 2 represent the multiple linear regression model,  $\text{Debt} = \beta_0 + \beta_1 \text{Income} + \beta_2 \text{Unemployment} + \varepsilon$ . (For ease of exposition, we use the same notation to refer to the coefficients in Models 1 and 2. We note, however, that these coefficients and their estimates may have a different meaning depending on which model we are referencing.)

If you had to choose one of these models to predict debt payments, which model would you choose? It may be that by using more explanatory variables, you can better describe the response variable. However, for a given sample, more is not always better. In order to select the preferred model, we need to examine goodness-of-fit measures.

### The Standard Error of the Estimate

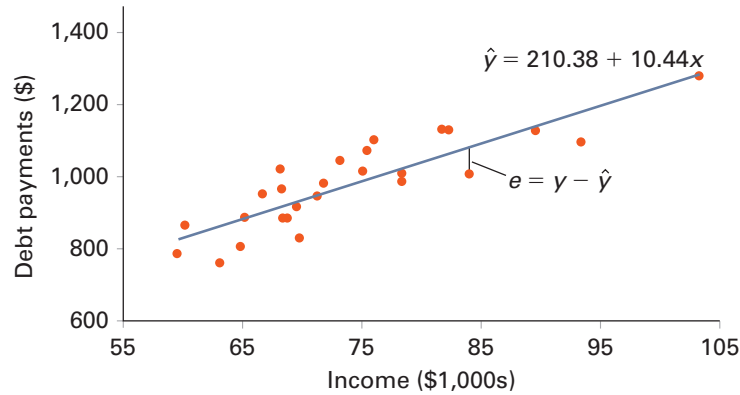
We first describe goodness-of-fit measures in the context of a simple linear regression model, or Model 1. Figure 14.5 reproduces the scatterplot of debt payments against income, as well as the sample regression line. Recall that the residual  $e$  represents the

#### LO 14.5

Calculate and interpret the standard error of the estimate.

difference between an observed value and the predicted value of the response variable; that is,  $e = y - \hat{y}$ . If all the data points had fallen on the line, then each residual would be zero; in other words, there would be no dispersion between the observed and the predicted values. Since in practice we rarely, if ever, obtain this result, we evaluate models on the basis of the relative magnitude of the residuals. The sample regression equation provides a good fit when the dispersion of the residuals is relatively small.

**FIGURE 14.5**  
Scatterplot of debt  
payments  $y$   
against income  $x$



A numerical measure that gauges dispersion from the sample regression equation is the sample variance of the residual, denoted  $s_e^2$ . We generally report the standard deviation of the residual, denoted  $s_e$ , more commonly referred to as the **standard error of the estimate**. The variance  $s_e^2$  is defined as the average squared difference between  $y_i$  and  $\hat{y}_i$ . The numerator of the formula is the sum of squares due to error,  $SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$ . Dividing  $SSE$  by its respective degrees of freedom  $n - k - 1$  yields  $s_e^2$ . Recall that  $k$  denotes the number of explanatory variables in the linear regression model; thus, for a simple linear regression model,  $k$  equals one. The standard error of the estimate  $s_e$  is the positive square root of  $s_e^2$ . The less the dispersion, the smaller the  $s_e$ , which implies a better fit to the model.

#### THE STANDARD ERROR OF THE ESTIMATE

The **standard error of the estimate**  $s_e$  is calculated as

$$s_e = \sqrt{s_e^2} = \sqrt{\frac{SSE}{n - k - 1}} = \sqrt{\frac{\sum e_i^2}{n - k - 1}}.$$

Theoretically,  $s_e$  can assume any value between zero and infinity,  $0 \leq s_e < \infty$ . The closer  $s_e$  is to zero, the better the model fits.

For a given sample size  $n$ , increasing the number  $k$  of the explanatory variables reduces both the numerator ( $SSE$ ) and the denominator ( $n - k - 1$ ) in the formula for  $s_e$ . The net effect, shown by the value of  $s_e$ , allows us to determine if the added explanatory variables improve the fit of the model.

#### EXAMPLE 14.7

Consider the sample data in Table 14.1 and the regression output for Model 1 in Table 14.4. Use the sample regression equation,  $\text{Debt} = 210.30 + 10.44 \text{ Income}$ , to calculate the standard error of the estimate  $s_e$ .



**SOLUTION:** First, we calculate the variance of the residual  $s_e^2$  of the residual. Let  $y$  and  $x$  denote Debt and Income, respectively. The first two columns of Table 14.6 show the values of these variables. The third column shows the predicted values  $\hat{y}$  and the fourth column shows the squared residuals,  $e^2 = (y - \hat{y})^2$ . The sum of the squared residuals, shown in the last cell of the last column, is the familiar  $SSE$  and is the numerator in the formula for  $s_e^2$ .

**TABLE 14.6** Calculations for Example 14.7

$y$	$x$	$\hat{y} = 210.30 + 10.44x$	$e^2 = (y - \hat{y})^2$
1285	103.50	$210.30 + 10.44 \times 103.50 = 1290.95$	$(1285 - 1290.84)^2 = 35.43$
1135	81.70	$210.30 + 10.44 \times 81.70 = 1063.34$	$(1135 - 1063.25)^2 = 5135.73$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
763	63.00	$210.30 + 10.44 \times 63.00 = 868.09$	$(763 - 868.02)^2 = 11043.35$
			$\Sigma e_i^2 = \Sigma (y_i - \hat{y}_i)^2 = 96045.55$

Using  $SSE = \Sigma e_i^2 = 96045.55$ , we determine  $s_e^2$  as

$$s_e^2 = \frac{SSE}{n - k - 1} = \frac{96045.55}{26 - 1 - 1} = 4001.90.$$

Taking the square root of the variance, we obtain

$$s_e = \sqrt{s_e^2} = \sqrt{4001.90} = 63.26.$$

The standard error of the estimate  $s_e$  is measured in the same units of measurement as the response variable. Since debt payments are in dollars, we report  $s_e$  as \$63.26.

Most of the time we rely on statistical software packages to report  $s_e$ . (If  $s_e$  is not explicitly given, other statistics like  $SSE$  are generally provided, which then greatly facilitate the calculation of  $s_e$ .) Excel reports the value for  $s_e$  in the regression output section entitled *Regression Statistics*. It is simply referred to as Standard Error. In column 2 of Table 14.7, we report the Excel regression statistics for Model 1. Note that  $s_e = 63.26$  is the same as the one calculated above.

**TABLE 14.7** Regression Statistics for Model 1 and Model 2

	Model 1	Model 2
Multiple R	0.8675	0.8676
R Square	0.7526	0.7527
Adjusted R Square	0.7423	0.7312
Standard Error	63.26	64.61
Observations	26	26
Regression Equation	$\hat{y} = 210.30 + 10.44x$	$\hat{y} = 199 + 10.51x_1 + 0.62x_2$

Our objective in adding another explanatory variable to the linear regression model is to increase the model's usefulness. In Model 2, we use income  $x_1$  and the unemployment rate  $x_2$  to explain debt payments  $y$ . If Model 2 is an improvement over Model 1, then we would expect it to have a smaller standard error of the estimate. Table 14.7 also shows the relevant regression statistics for Model 2. We note that the standard error of the estimate for Model 2,  $s_e = \$64.61$ , is actually greater than that for Model 1 ( $64.61 > 63.26$ ). In other words, there is less dispersion between the observed values of debt payments and

the predicted values of debt payments when we include only one explanatory variable in the model. So far, this suggests that Model 1 provides a better fit for the sample data. In general, we use the standard error of the estimate in conjunction with other measures to judge the overall usefulness of a model.

#### LO 14.6

Calculate and interpret the coefficient of determination,  $R^2$ .

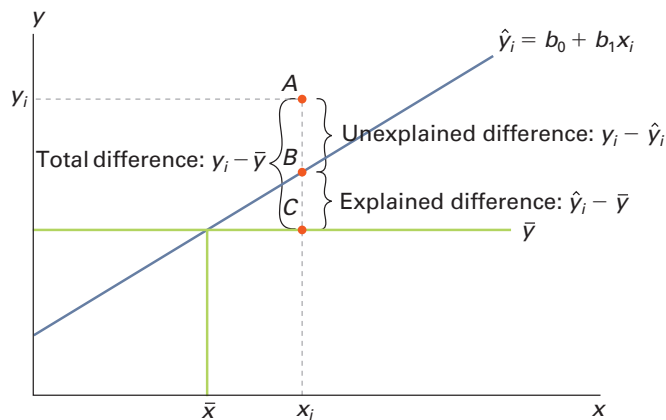
### The Coefficient of Determination, $R^2$

In Example 14.7, we calculated  $s_e = \$63.26$ . It is difficult to interpret this value in isolation. The fact that a value closer to 0 implies a better fit does not allow us to conclude whether or not a value of \$63.26 is close to zero. The standard error of the estimate is a useful goodness-of-fit measure when we are comparing various models—the model with the smaller  $s_e$  provides the better relative fit. However, it proves less useful when we are assessing a single model. One reason for this is due to the fact that  $s_e$  has no predefined upper limit; that is,  $0 \leq s_e < \infty$ . The **coefficient of determination**, commonly referred to as  $R^2$ , is another goodness-of-fit measure that is easier to interpret than the standard error of the estimate. As we will see shortly,  $R^2$  has both lower and upper bounds that make its interpretation quite a bit more intuitive.

Like the standard error of the estimate, the coefficient of determination evaluates how well the sample regression equation fits the data. In particular,  $R^2$  quantifies the sample variation in the response variable  $y$  that is explained by sample regression equation. It is computed as the ratio of the explained variation of the response variable to its total variation. We generally convert this ratio into a percent by multiplying it by 100. For example, if  $R^2 = 0.72$ , we say that 72% of the sample variation in the response variable is explained by the sample regression equation. Other factors, which have not been included in the model, account for the remaining 28% of the sample variation.

We use analysis of variance (ANOVA) in the context of the linear regression model, to derive  $R^2$ . We denote the **total variation** in  $y$  as  $\sum(y_i - \bar{y})^2$ , which is the numerator in the formula for the variance of  $y$ . This value, called the **total sum of squares, SST**, can be broken down into two components: **explained variation** and **unexplained variation**. Figure 14.6 illustrates the decomposition of the total variation in  $y$  into its two components.

**FIGURE 14.6**  
Total, explained,  
and unexplained  
differences



For ease of exposition, we show a scatterplot with all the points removed except one (point A). Point A refers to the observation  $(x_i, y_i)$ . The blue line represents the estimated regression equation based on the entire sample data; the horizontal and vertical green lines represent the sample means  $\bar{y}$  and  $\bar{x}$ , respectively. The vertical distance between the data point A and  $\bar{y}$  (point C) is the total difference  $y_i - \bar{y}$  (distance AC). For each data point, we square these differences and then find their sum—this amounts to  $SST = \sum(y_i - \bar{y})^2$ .  $SST$  is a measure of the total variation in  $y$ .

Now, we focus on the distance between the predicted value  $\hat{y}_i$  (point  $B$ ) and  $\bar{y}$ ; that is, the explained difference (distance  $BC$ ). It is called explained because the difference between  $\hat{y}_i$  and  $\bar{y}$  can be explained by changes in  $x$ . Squaring all such differences and summing them yields the **sum of squares due to regression**,  $SSR$ , where  $SSR = \sum(\hat{y}_i - \bar{y})^2$ .  $SSR$  is a measure of the explained variation in  $y$ .

The distance between the particular observation and its predicted value (distance  $AB$ ) is the unexplained difference. This is the portion that remains unexplained; it is due to random error or chance. Squaring all such differences and summing them yields the familiar sum of squares due to error,  $SSE = \sum(y_i - \hat{y}_i)^2$ .  $SSE$  is a measure of the unexplained variation in  $y$ .

Thus, the total variation in  $y$  can be decomposed into explained and unexplained variation as follows:

$$SST = SSR + SSE.$$

Dividing both sides by  $SST$  and rearranging yields:

$$\frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Each side of the above equation shows two equivalent formulas for the coefficient of determination  $R^2$ ; that is,  $R^2 = SSR/SST$  or  $R^2 = 1 - SSE/SST$ . The value of  $R^2$  falls between zero and one,  $0 \leq R^2 \leq 1$ . The closer  $R^2$  is to one, the stronger the fit; the closer it is to zero, the weaker the fit.

#### THE COEFFICIENT OF DETERMINATION, $R^2$

The coefficient of determination,  $R^2$ , is the proportion of the sample variation in the response variable that is explained by the sample regression equation. We compute  $R^2$  as

$$R^2 = \frac{SSR}{SST}, \text{ or equivalently, } R^2 = 1 - \frac{SSE}{SST},$$

where  $SSR = \sum(\hat{y} - \bar{y})^2$ ,  $SSE = \sum(y_i - \hat{y}_i)^2$ , and  $SST = \sum(y_i - \bar{y})^2$ . The coefficient of determination  $R^2$  can also be computed as  $R^2 = r_{yy}^2$ , where  $r_{yy}$  is the sample correlation coefficient between  $y$  and  $\hat{y}$ .

The value of  $R^2$  falls between zero and one; the closer the value is to one, the better the fit.

Most statistical packages, including Excel, provide the calculations for  $SSR$ ,  $SSE$ , and  $SST$  in an ANOVA table for regression. The general format of the ANOVA table that accompanies regression output is shown in Table 14.8. We will discuss the values for  $F_{(df_1, df_2)}$  and the  $p$ -value for  $F_{(df_1, df_2)}$  in the next chapter.

**TABLE 14.8** General Format of ANOVA Table That Accompanies Regression Results

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	<i>k</i>	<i>SSR</i>	<i>MSR</i>	$F_{(df_1, df_2)} = \frac{MSR}{MSE}$	<i>p</i> -value for $F_{(df_1, df_2)}$
Residual	$n - k - 1$	<i>SSE</i>	<i>MSE</i>		
Total	$n - 1$	<i>SST</i>			

#### EXAMPLE 14.8

Calculate and interpret the coefficient of determination  $R^2$  given the sample data in Table 14.1 and the sample regression equation for Model 1:  $\widehat{\text{Debt}} = 210.30 + 10.44\text{Income}$ .

**SOLUTION:** In Example 14.7, we calculated  $SSE$  for Model 1 as 96045.55. Using  $R^2 = 1 - SSE/SST$ , the only missing part is  $SST$ . Given  $\bar{y} = 983.46$ , we calculate  $SST$  as

$$\begin{aligned}\Sigma(y_i - \bar{y})^2 &= (1285 - 983.46)^2 + (1135 - 983.46)^2 + \cdots + (763 - 983.46)^2 \\ &= 388182.46.\end{aligned}$$

The  $SSE$  and  $SST$  values are identical to the Excel values reported in Table 14.4. Therefore,

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{96045.55}{388182.46} = 0.7526.$$

Note that this value matches the Excel estimate shown in Tables 14.4 and 14.7. The coefficient of determination  $R^2$  shows that 75.26% of the sample variation in debt payments is explained by changes in income.

Another interesting statistic in Table 14.7 is **Multiple  $R$** . This measure is simply the sample correlation coefficient between the response variable  $y$  and its predicted value  $\hat{y}$  which, using our earlier notation, implies that Multiple  $R = r_{y\hat{y}}$ . In a simple linear regression model, Multiple  $R$  also represents the absolute value of the correlation between  $y$  and  $x$ , or  $|r_{yx}|$ . Moreover,  $R^2$  is the square of Multiple  $R$ ; that is,  $R^2 = r_{y\hat{y}}^2$ .

In Chapter 16, we will use several examples to compute  $r_{y\hat{y}}$  from scratch. Here, we simply note that Multiple  $R$  in Table 14.7 implies that the sample correlation coefficient between  $y$  and  $\hat{y}$  for Model 2, for example, is  $r_{y\hat{y}}^2 = 0.8676$ . Thus,  $R^2 = r_{y\hat{y}}^2 = 0.8676^2 = 0.7527$ , which is the same value for  $R^2$  that is shown in Table 14.7.

Recall that the standard error of the estimate for Model 1 ( $s_e = 63.26$ ) was smaller than that for Model 2 ( $s_e = 64.61$ ), suggesting that Model 1 provides a better fit. Since the coefficient of determination for Model 2 ( $R^2 = 0.7527$ ) is slightly higher than that of Model 1 ( $R^2 = 0.7526$ ), one may think that Model 2 explains more of the variation in debt payments. How do we resolve these apparent conflicting results? It turns out that we cannot use  $R^2$  for model comparison when the competing models do not include the same number of explanatory variables. This occurs because  $R^2$  never decreases as we add more explanatory variables to the model. A popular model selection method in such situations is to choose a model that has the highest adjusted  $R^2$  value.

#### LO 14.7

Differentiate between  $R^2$  and adjusted  $R^2$ .

### The Adjusted $R^2$

Since  $R^2$  never decreases as we add more explanatory variables to the linear regression model, it is possible to increase its value unintentionally by including a group of explanatory variables that may have no economic or intuitive foundation in the linear regression model. This is true especially when the number of explanatory variables  $k$  is large relative to the sample size  $n$ . In order to avoid the possibility of  $R^2$  creating a false impression, virtually all software packages include **adjusted  $R^2$** . Unlike  $R^2$ , adjusted  $R^2$  explicitly accounts for the sample size  $n$  and the number of explanatory variables  $k$ . It is common to use adjusted  $R^2$  for model selection because it imposes a penalty for any additional explanatory variable that is included in the analysis.

#### ADJUSTED $R^2$

The adjusted coefficient of determination is calculated as

$$\text{Adjusted } R^2 = 1 - (1 - R^2) \left( \frac{n-1}{n-k-1} \right).$$

Adjusted  $R^2$  is used to compare competing linear regression models with different numbers of explanatory variables; the higher the adjusted  $R^2$ , the better the model.

If  $SSE$  is substantially greater than zero and  $k$  is large compared to  $n$ , then adjusted  $R^2$  will differ substantially from  $R^2$ . Adjusted  $R^2$  may be negative, if the correlation between the response variable and the explanatory variables is sufficiently low.

We would also like to point out that both the standard error of the estimate and the adjusted  $R^2$  are useful for comparing the linear regression models with different numbers of explanatory variables. Adjusted  $R^2$ , however, is the more commonly used criterion for model selection.

### EXAMPLE 14.9

Given the regression statistics from Table 14.7, use the value of the adjusted  $R^2$  for model comparison.

**SOLUTION:** We note from Table 14.7 that Model 1 has an adjusted  $R^2$  of 0.7423, whereas its value for Model 2 is 0.7312. Therefore, given its higher adjusted  $R^2$ , we choose Model 1 to predict debt payments.

## SYNOPSIS OF INTRODUCTORY CASE

A recent study shows substantial variation in consumer debt payments depending on where the consumer resides (Experian.com, November 11, 2010). A possible explanation is that a linear relationship exists between consumer debt payments and an area's median household income. In order to substantiate this claim, relevant data on 26 metropolitan areas are collected. The correlation coefficient between debt payments and income is computed as 0.87, suggesting a strong positive linear relationship between the two variables. A simple test confirms that the correlation coefficient is statistically significant at the 5% level.

Two regression models are also estimated for the analysis. A simple linear regression model (Model 1), using consumer debt payments as the response variable and median household income as the explanatory variable, is estimated as  $\widehat{\text{Debt}} = 210.30 + 10.44\text{Income}$ . For every \$1,000 increase in median household income, consumer debt payments are predicted to increase by \$10.44. In an attempt to improve upon the prediction, a multiple regression model (Model 2) is proposed, where median household income and the unemployment rate are used as explanatory variables. The sample regression line for Model 2 is  $\widehat{\text{Debt}} = 199.00 + 10.51\text{Income} + 0.62\text{Unemployment}$ . Given its slope coefficient of only 0.62, the economic impact of the unemployment rate on consumer debt payments, with median household income held fixed, seems extremely weak. Goodness-of-fit measures confirm that Model 1 provides a better fit than Model 2. The standard error of the estimate is smaller for Model 1, suggesting less dispersion of the data from the sample regression equation. In addition, the adjusted  $R^2$  is higher for Model 1, implying that more of the sample variation in consumer debt payments is explained by the simple linear regression model. Using Model 1 and an area's median household income of \$80,000, consumer debt payments are predicted to be \$1,045.50.



## EXERCISES 14.4

### Mechanics

48. In a simple linear regression based on 25 observations, the following intermediate data are given:  $\Sigma(y_i - \hat{y})^2 = 1250$  and  $\Sigma(y_i - \bar{y})^2 = 1500$ .
- Calculate  $s_e^2$  and  $s_e$ .
  - Calculate  $R^2$ .
49. In a simple linear regression based on 30 observations, it is found that  $SSE = 2540$  and  $SST = 13,870$ .
- Calculate  $s_e^2$  and  $s_e$ .
  - Calculate  $R^2$ .
50. In a multiple regression with two explanatory variables, the following intermediate data are given:  $n = 50$ ,  $\Sigma(y_i - \hat{y})^2 = 35$ , and  $\Sigma(y_i - \bar{y})^2 = 90$ .
- Calculate the standard error of the estimate.
  - Calculate the coefficient of determination  $R^2$ .
51. In a multiple regression with four explanatory variables and 100 observations, it is found that  $SSR = 4.75$  and  $SST = 7.62$ .
- Calculate the standard error of the estimate.
  - Calculate the coefficient of determination  $R^2$ .
  - Calculate adjusted  $R^2$ .
52. The following ANOVA table was obtained when estimating a simple linear regression model.

ANOVA	df	SS	MS	F	Significance F
Regression	1	81.58	81.58	366.50	1.27E-17
Residual	28	6.23	0.22		
Total	29	87.81			

- Calculate the standard error of the estimate.
  - Calculate the coefficient of determination. Interpret this value.
53. The following ANOVA table was obtained when estimating a simple linear regression model.

ANOVA	df	SS	MS	F	Significance F
Regression	1	75.92	75.92	178.76	1.12E-13
Residual	28	11.89	0.42		
Total	29	87.81			

- Calculate  $s_e$ .
  - Calculate and interpret  $R^2$ .
54. The following ANOVA table was obtained when estimating a multiple linear regression model.

ANOVA	df	SS	MS	F	Significance F
Regression	2	161478.4	80739.19	11.5854	0.0002
Residual	27	188163.9	6969.03		
Total	29	349642.2			

- Calculate the standard error of the estimate.
- Calculate and interpret the coefficient of determination.
- Calculate the adjusted  $R^2$ .

55. The following ANOVA table was obtained when estimating a multiple regression model.

ANOVA	df	SS	MS	F	Significance F
Regression	2	188246.8	94123.4	35.2	9.04E-07
Residual	17	45457.32	2673.96		
Total	19	233704.1			

- Calculate the standard error of the estimate.
- Calculate and interpret the coefficient of determination.
- Calculate the adjusted  $R^2$ .

### Applications

56. When estimating the selling price of a used sedan as a function of its age using the model  $\text{Price} = \beta_0 + \beta_1 \text{Age} + \varepsilon$ , a researcher gets the following ANOVA results:

ANOVA	df	SS	MS	F	Significance F
Regression	1	199.93	199.93	87.80	2.41E-08
Residual	18	40.99	2.28		
Total	19	240.92			

- How many observations did the researcher use?
- What is the standard error of the estimate?
- Calculate and interpret the coefficient of determination.

57. When estimating the sales of a firm as a function of its advertising expenditures using the model  $\text{Sales} = \beta_0 + \beta_1 \text{Advertising} + \varepsilon$ , an analyst obtains the following ANOVA results:

ANOVA	df	SS	MS	F	Significance F
Regression	1	199.93	199.93	87.80	2.41E-08
Residual	18	40.99	2.28		
Total	19	240.92			

- What proportion of the sample variation in sales is explained by advertising expenditures?
- What proportion of the sample variation in sales is unexplained by advertising expenditures?

58. The director of college admissions at a local university is trying to determine whether a student's high school GPA or SAT score is a better predictor of the student's subsequent college GPA. She formulates two models:

$$\text{Model 1. College GPA} = \beta_0 + \beta_1 \text{High School GPA} + \varepsilon$$

$$\text{Model 2. College GPA} = \beta_0 + \beta_1 \text{SAT Score} + \varepsilon$$

She estimates these models using data from a sample of 10 recent college graduates. A portion of the results are as follows:

ANOVA Results for Model 1

	df	SS	MS	F	Significance F
Regression	1	1.4415	1.4415	11.5032	0.0095
Residual	8	1.0025	0.1253		
Total	9	2.4440			



ANOVA Results for Model 2					
	df	SS	MS	F	Significance F
Regression	1	1.0699	1.0699	6.2288	0.0372
Residual	8	1.3741	0.1718		
Total	9	2.4440			

- Calculate the standard error of the estimate for Model 1 and Model 2.
  - Calculate the coefficient of determination for Model 1 and Model 2.
  - Given these two measures, which model is a better fit? Explain.
59. For a sample of 41 New England cities, a sociologist studies the crime rate in each city (crimes per 100,000 residents) as a function of its poverty rate (in %) and its median income (in \$1,000s). A portion of the regression results is shown in the following table.

ANOVA	df	SS	MS	F	Significance F
Regression	2	3549788	1774894	16.12513	8.5E-06
Residual	38	4182663	110070.1		
Total	40	7732451			

- Calculate the standard error of the estimate.
  - What proportion of the sample variation in crime rate is explained by the variability in the explanatory variables? What proportion is unexplained?
60. A financial analyst uses the following model to estimate a firm's stock return:  $\text{Return} = \beta_0 + \beta_1 P/E + \beta_2 P/S + \varepsilon$ , where  $P/E$  is a firm's price-to-earnings ratio and  $P/S$  is a firm's price-to-sales ratio. A portion of the regression results is shown in the following table.

ANOVA	df	SS	MS	F	Significance F
Regression	2	918.746	459.373	2.817	0.0774
Residual	27	4402.786	163.066		
Total	29	5321.532			

- Calculate the standard error of the estimate.
  - Calculate and interpret the coefficient of determination.
  - Calculate the corresponding adjusted  $R^2$ .
61. **FILE Test Scores.** The accompanying data file shows the midterm and final scores for 32 students in a statistics course.
- Estimate a student's final score as a function of his/her midterm score.
  - Find the standard error of the estimate.
  - Find and interpret the coefficient of determination.
62. **FILE Property Taxes.** The accompanying data file shows the square footage and associated property taxes for 20 homes in an affluent suburb 30 miles outside New York City.
- Estimate a home's property taxes as a linear function of the size of the home (measured by its square footage).

- What proportion of the sample variation in property taxes is explained by the home's size?
- What proportion of the sample variation in property taxes is unexplained by the home's size?

63. **FILE Football.** Is it defense or offense that wins football games? Consider the following portion of data, which includes a team's winning record (Win), the average number of yards gained, and the average number of yards allowed during the 2009 NFL season.

Team	Win (%)	Yards Gained	Yards Allowed
Arizona Cardinals	62.50	344.40	346.40
Atlanta Falcons	56.30	340.40	348.90
⋮	⋮	⋮	⋮
Washington Redskins	25.00	312.50	319.70

Source: NFL website.

- Compare two simple linear regression models, where Model 1 predicts the winning percentage based on Yards Gained and Model 2 uses Yards Allowed.
  - Estimate a multiple linear regression model, Model 3, that applies both Yards Gained and Yards Allowed to forecast the winning percentage. Is this model an improvement over the other two models? Explain.
64. **FILE Executive Compensation.** Executive compensation has risen dramatically beyond the rising levels of an average worker's wage over the years. The government is even considering a cap on high-flying salaries for executives (*The New York Times*, February 9, 2009). Consider the following portion of data which links total compensation of the 455 highest-paid CEOs in 2006 with three measures: (industry-adjusted return on assets (Adj ROA), industry-adjusted stock return (Adj Stock Return) and the firm's size (Total Assets)).

Compensation (in \$ millions)	Adj ROA	Adj Stock Return	Total Assets (in \$ millions)
16.58	2.53	-0.15	20,917.5
26.92	1.27	0.57	32,659.5
⋮	⋮	⋮	⋮
2.3	0.45	0.75	44,875.0

Source: SEC website and Compustat.

- Estimate three simple linear regression models that use Compensation as the response variable with Adj ROA, Adj Stock Return, or Total Assets as the explanatory variable. Which model do you select? Explain.
- Estimate multiple linear regression models that use various combinations of two, or all three explanatory variables. Which model do you select? Explain.

# WRITING WITH STATISTICS



Matthew Farnham is an investment consultant who always recommends a well-diversified portfolio of mutual funds to his clients. He knows that a key concept in benefiting from diversification is correlation. Correlation is the extent to which assets perform in relation to one another. If all of an investor's assets move in lock-step, or are highly correlated, then the investor is either all right or all wrong. In order to reduce risk, it is considered good practice to invest in assets whose values rise and fall independently of one another. Matthew is approached by a client who has already invested in Vanguard's 500 Index Fund—a fund that mimics the Standard & Poor's 500 Index. She seeks advice for choosing her next investment from one of the following Vanguard funds:

- Inflation-Protected Securities Index
- Intermediate-Term Bond Index
- Real Estate Investment Trust Index
- Small Cap Index

Matthew collects 10 years of monthly return data for each mutual fund for the analysis. A portion of the data is shown in Table 14.9.

**TABLE 14.9** Monthly Return Data for Five Mutual Funds, January 2001–December 2010

	500 Index	Inflation-Protected Securities	Intermediate-Term Bond	Real Estate	Small Cap
January 2001	0.0342	0.0205	0.0181	0.0044	0.0527
February 2001	−0.1007	0.0161	0.0129	−0.0175	−0.0658
⋮	⋮	⋮	⋮	⋮	⋮
December 2010	0.0585	−0.0284	−0.0320	0.0355	0.0663

SOURCE: [finance.yahoo.com](http://finance.yahoo.com); data retrieved January 4, 2011.

Matthew would like to use the sample information in Table 14.9 to:

1. Calculate and interpret the sample correlation coefficient of each fund with Vanguard's 500 Index.
2. Make a recommendation for a mutual fund that is not correlated with Vanguard's 500 Index.

## Sample Report— Making Investment Decisions by Diversifying

In attempting to create a well-diversified portfolio, an analysis of the correlation between assets' returns is crucial. The correlation coefficient measures the direction and the strength of the linear relationship between assets' returns. This statistic can aid in the hunt for assets that form a portfolio. An investor has already chosen Vanguard's 500 Index mutual fund as part of her portfolio. When choosing to add to her portfolio, she considers these four mutual funds from the Vanguard family:

- Inflation-Protected Securities Index
- Intermediate-Term Bond Index
- Real Estate Investment Trust Index
- Small Cap Index

Ten years of monthly return data for each of these prospective funds, as well as the 500 Index, are collected. The first row of Table 14.A shows the sample correlation coefficients between the 500 Index and each mutual fund.

**TABLE 14.A** Analysis of Correlations between the 500 Index and Each Mutual Fund

	Inflation-Protected Securities	Intermediate-Term Bond	Real Estate	Small Cap
Correlation Coefficient	0.0796	−0.0408	0.6630	0.9030
Test Statistic	0.87	−0.44	9.62	22.83
<i>p</i> -value	0.39	0.66	0.00	0.00

The correlation coefficient always assumes a value between −1 and 1; an absolute value close to 1 implies that the two assets move in sync. In this sample, the highest sample correlation coefficient is between the 500 Index and the Small Cap Index, with a value of 0.9030. Next on the list is the correlation of 0.6630 between the 500 Index and the Real Estate Index. Investors often choose to invest across a range of asset classes that earn respectable returns but are relatively uncorrelated. This way, if one asset in a portfolio suffers, the rest may be unaffected. Given that the Inflation-Protected Securities Index and the Intermediate-Term Bond Index have correlation coefficients close to zero, these may prove to be desirable additions to the investor's portfolio.

A hypothesis test is conducted in order to determine whether the population correlation coefficients are significantly different from zero at the 5% significance level. The null hypothesis is that the returns are uncorrelated and the alternative hypothesis suggests either a positive or a negative correlation. Rows 2 and 3 of Table 14.A show the value of the test statistics and the corresponding *p*-values. For instance, given a *p*-value of 0.39, the correlation coefficient between the 500 Index and the Inflation-Protected Securities Index is not significantly different from zero at the 5% significance level. This same conclusion holds for the correlation between the 500 Index and the Intermediate-Term Bond Index. On the other hand, given *p*-values of 0.00 for both of the test statistics associated with the correlation coefficient between the 500 Index and the Real Estate Index and the 500 Index and the Small Cap Index, we can conclude that these correlation coefficients are significantly different from zero.

Assuming that the correlation between assets is likely to remain stable in the future, then it appears that the investor should add either the Inflation-Protected Securities Index or the Intermediate Bond Index to her portfolio. Compared to the other two funds, these funds would offer the maximum benefit from diversification in the sense of reducing volatility.

## CONCEPTUAL REVIEW

### LO 14.1 Conduct a hypothesis test for the population correlation coefficient.

The **covariance** is a measure of the linear relationship between two variables  $x$  and  $y$ . We compute the sample covariance as  $s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$ , where  $n$  represents the number of observations. The **correlation coefficient** is a unit-free measure that gauges the strength of the linear relationship between two variables  $x$  and  $y$ . We calculate the sample correlation coefficient as  $r_{xy} = \frac{s_{xy}}{s_x s_y}$ , where  $-1 \leq r_{xy} \leq 1$  and  $s_x$  and  $s_y$  are the sample standard deviations of  $x$  and  $y$ , respectively.

When determining whether the correlation coefficient differs from zero, the null and alternative hypotheses are formulated as  $H_0: \rho_{xy} = 0$  and  $H_A: \rho_{xy} \neq 0$ , where  $\rho_{xy}$  is the

population correlation coefficient; one-tailed tests are constructed similarly. The value of the test statistic is calculated as  $t_{df} = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$  with  $df = n - 2$ .

#### LO 14.2 Discuss the limitations of correlation analysis.

There are several **limitations to correlation analysis**. These include: (1) two variables may have a very low correlation coefficient, yet a strong *nonlinear* relation; (2) the existence of *outliers* may blur the interpretation of the covariance and the correlation coefficient; and (3) *spurious correlation* can make two variables appear closely related when no causal relationship exists.

#### LO 14.3 Estimate the simple linear regression model and interpret the coefficients.

**Regression analysis** explicitly assumes that one variable, called the **response variable**, is influenced by other variables, called the **explanatory variables**.

The **simple linear regression model** uses only one explanatory variable to predict and/or describe changes in the response variable. The model is expressed as  $y = \beta_0 + \beta_1 x + \varepsilon$ , where  $y$  and  $x$  are the response variable and the explanatory variable, respectively, and  $\varepsilon$  is the random error term. The coefficients  $\beta_0$  and  $\beta_1$  are the unknown parameters to be estimated.

We apply the **ordinary least squares (OLS)** method to find a sample regression equation  $\hat{y} = b_0 + b_1 x$ , where  $\hat{y}$  is the predicted value of the response variable and  $b_0$  and  $b_1$  are the estimates of  $\beta_0$  and  $\beta_1$ , respectively. The estimated slope coefficient  $b_1$  represents the change in  $\hat{y}$  when  $x$  changes by one unit. The units of  $b_1$  are the same as those of  $y$ .

#### LO 14.4 Estimate the multiple linear regression model and interpret the coefficients.

The **multiple regression model** allows more than one explanatory variable to be linearly related with the response variable  $y$ . It is defined as  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$ , where  $y$  is the response variable,  $x_1, x_2, \dots, x_k$  are the  $k$  explanatory variables and  $\varepsilon$  is the random error term. The coefficients  $\beta_0, \beta_1, \dots, \beta_k$  are the unknown parameters to be estimated. We again use the OLS method to arrive at the following sample regression equation:  $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$ , where  $b_0, b_1, \dots, b_k$  are the estimates of  $\beta_0, \beta_1, \dots, \beta_k$ , respectively.

For each explanatory variable  $x_j$  ( $j = 1, \dots, k$ ), the corresponding slope coefficient  $b_j$  is the estimated regression coefficient. It measures the change in the predicted value of the response variable  $\hat{y}$  given a unit increase in the associated explanatory variable  $x_j$ , *holding all other explanatory variables constant*. In other words, it represents the partial influence of  $x_j$  on  $\hat{y}$ .

#### LO 14.5 Calculate and interpret the standard error of the estimate.

The **standard error of the estimate**  $s_e$  is calculated as  $s_e = \sqrt{s_e^2} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{\sum e_i^2}{n-k-1}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-k-1}}$ . Theoretically,  $s_e$  can assume any value between zero and infinity,  $0 \leq s_e < \infty$ ; the closer  $s_e$  is to zero, the better the model fits. Since  $s_e$  has no predetermined upper limit, it is difficult to interpret it in isolation. The standard error of the estimate is a useful goodness-of-fit measure when comparing models; the model with the smaller  $s_e$  provides the better fit.

#### LO 14.6 Calculate and interpret the coefficient of determination, $R^2$ .

The **coefficient of determination**  $R^2$  is the proportion of the sample variation in the response variable that is explained by the sample regression equation. It falls between 0 and 1;

the closer the value is to 1, the better the model fits. For example, if  $R^2 = 0.72$ , we say that 72% of the sample variation in  $y$  is explained by the estimated model.

We compute the coefficient of determination as  $R^2 = SSR/SST = 1 - \frac{SSE}{SST}$  where  $SSR = \sum(\hat{y} - \bar{y})^2$ ,  $SSE = \sum(y_i - \hat{y}_i)^2$ , and  $SST = \sum(y_i - \bar{y})^2$ . Alternatively, we can compute it as  $R^2 = r_{yy}^2$ , where  $r_{yy}$  is the sample correlation coefficient between  $y$  and  $\hat{y}$ .

#### LO 14.7 Differentiate between $R^2$ and adjusted $R^2$ .

**Adjusted  $R^2$**  adjusts  $R^2$  by accounting for the sample size  $n$  and the number of explanatory variables  $k$  used in the regression. It is calculated as adjusted  $R^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$ . In comparing competing models with different numbers of explanatory variables, the preferred model will have the highest adjusted  $R^2$ .

## ADDITIONAL EXERCISES AND CASE STUDIES

65. The following table shows the annual returns for two of Vanguard's mutual funds: the Vanguard Energy Fund and the Vanguard Healthcare Fund.

Year	Annual Total Returns (in percent)	
	Energy $x$	Healthcare $y$
2004	36.65	9.51
2005	44.60	15.41
2006	19.68	10.87
2007	37.00	4.43
2008	-42.87	-18.45
	$\bar{x} = 19.01$	$\bar{y} = 4.35$
	$s_x = 35.77$	$s_y = 13.34$
	$s_{xy} = 447.68$	

SOURCE: [www.finance.yahoo.com](http://www.finance.yahoo.com).

- Calculate and interpret the sample correlation coefficient  $r_{xy}$ .
  - Specify the competing hypotheses in order to determine whether the population correlation coefficient is significantly different from zero.
  - At the 5% significance level, what is the conclusion to the test? Are the returns on the mutual funds significantly correlated?
66. **FILE Yields.** In response to the Great Recession, Federal Reserve leaders continue to keep the short-run target interest rate near zero. While the Fed controls short-term interest rates, long-term interest rates essentially depend on supply/demand dynamics, as well as longer-term interest rate expectations. Consider the following annualized rates for 3-month Treasury yields and 10-year Treasury yields.

Year	3-Month Yield (%)	10-Year Yield (%)
2001	3.47	5.02
2002	1.63	4.61
2003	1.03	4.02
2004	1.40	4.27
2005	3.21	4.29
2006	4.85	4.79
2007	4.47	4.63
2008	1.39	3.67
2009	0.15	3.26
2010	0.14	3.21

SOURCE: Federal Reserve Bank of Dallas.

- Construct and interpret a scatterplot of a 10-year treasury yield against a 3-month yield.
  - Calculate and interpret the sample correlation coefficient. Use  $\alpha = 0.05$  to test if the population correlation coefficient is significantly different from zero.
  - Estimate and interpret a sample regression equation using the 10-year yield as the response variable and the 3-month yield as the explanatory variable.
67. **FILE Home Ownership.** The homeownership rate in the U.S. was 67.4% in 2009. In order to determine if homeownership is linked with income, 2009 state level data on the homeownership rate (Ownership) and median household income (Income) were collected. A portion of the data is shown in the accompanying table.

State	Income	Ownership
Alabama	\$39,980	74.1%
Alaska	\$61,604	66.8%
⋮	⋮	⋮
Wyoming	\$52,470	73.8%

SOURCE: www.census.gov.

- Estimate and interpret the model:  $\text{Ownership} = \beta_0 + \beta_1 \text{Income} + \varepsilon$ .
  - What is the standard error of the estimate?
  - Interpret the coefficient of determination.
68. **FILE Dow\_2010.** A research analyst is trying to determine whether a firm's price-earnings (P/E) and price-sales (P/S) ratios can explain the firm's stock performance over the past year. A P/E ratio is calculated as a firm's share price compared to the income or profit earned by the firm per share. Generally, a high P/E ratio suggests that investors are expecting higher earnings growth in the future compared to companies with a lower P/E ratio. The P/S ratio is calculated by dividing a firm's share price by the firm's revenue per share for the trailing 12 months. In short, investors can use the P/S ratio to determine how much they are paying for a dollar of the firm's sales rather than a dollar of its earnings (P/E ratio). In general, the lower the P/S ratio, the more attractive the investment. The accompanying table shows the year-to-date (YTD) returns and the P/E and P/S ratios for a portion of the 30 firms included in the Dow Jones Industrial Average.

	YTD return (in %)	P/E ratio	P/S ratio
1. 3M Co.	4.4	14.37	2.41
2. Alcoa Inc.	-4.5	11.01	0.78
⋮	⋮	⋮	⋮
30. Walt Disney Company	16.3	13.94	1.94

SOURCE: The 2010 returns (January 1, 2010–December 31, 2010) were obtained from *The Wall Street Journal*, January 3, 2010; the P/E ratios and the P/S ratios were obtained from finance.yahoo.com on January 20, 2011.

- Estimate:  $\text{Return} = \beta_0 + \beta_1 \text{P/E} + \beta_2 \text{P/S} + \varepsilon$ . Are the signs on the coefficients as expected? Explain.
- Interpret the slope coefficient of the P/S ratio.
- What is the predicted return for a firm with a P/E ratio of 10 and a P/S ratio of 2?
- What is the standard error of the estimate?
- Interpret  $R^2$ .

69. **FILE SAT.** There has been a lot of discussion regarding the relationship between Scholastic Aptitude Test (SAT) scores and test-takers' family income (*The New York Times*, August 27, 2009). It is generally believed that the wealthier a student's family, the higher the SAT score. Another commonly used predictor for SAT scores is the student's grade point average (GPA). Consider the following portion of data collected on 24 students.

SAT	Income	GPA
1651	47000	2.79
1581	34000	2.97
⋮	⋮	⋮
1940	113000	3.96

- Estimate three models:
    - $\text{SAT} = \beta_0 + \beta_1 \text{Income} + \varepsilon$ ,
    - $\text{SAT} = \beta_0 + \beta_1 \text{GPA} + \varepsilon$ , and
    - $\text{SAT} = \beta_0 + \beta_1 \text{Income} + \beta_2 \text{GPA} + \varepsilon$ .
  - Use goodness-of-fit measures to select the best-fitting model.
  - Predict SAT given the mean value of the explanatory variable(s).
70. **FILE Startups.** Many of today's leading companies, including Google, Microsoft, and Facebook, are based on technologies developed within universities. Lisa Fisher is a business school professor who would like to analyze university factors that enhance innovation. She collects data on 143 universities in 2008 for a regression where the response variable is the number of startups (Startups), which is used as a measure for innovation. The explanatory variables include the university's research expenditure in millions of dollars (Research), the number of patents issued (Patents), and the age of its technology transfer office in years (Duration). A portion of the data is shown in the accompanying table.

Startups	Research (\$ millions)	Patents	Duration
1	145.52	8	23
1	237.52	16	23
⋮	⋮	⋮	⋮
1	154.38	3	9

SOURCE: Association of University Managers and National Science Foundation.

- Estimate:  $\text{Startups} = \beta_0 + \beta_1 \text{Research} + \beta_2 \text{Patents} + \beta_3 \text{Duration} + \varepsilon$ .
- Predict the number of startups for a university that spent \$120 million on research, issued 8 patents, and has had a technology transfer office for 20 years.



- c. How much more research expenditure is needed for the university to have an additional predicted startup, with everything else being the same?
71. **FILE Hourly\_Wage.** A researcher interviews 50 employees of a large manufacturer and collects data on each worker's hourly wage (Wage), years of higher education (EDUC), experience (EXPER), and age (AGE).
- Estimate:  $\text{Wage} = \beta_0 + \beta_1 \text{EDUC} + \beta_2 \text{EXPER} + \beta_3 \text{AGE} + \varepsilon$ .
  - Are the signs as expected?
  - Interpret the coefficient of EDUC.
  - Interpret the coefficient of determination.
  - Predict the hourly wage of a 40-year-old employee who has 5 years of higher education and 8 years of experience.

## CASE STUDIES

**CASE STUDY 14.1** A local university offers its employees the following Fidelity investment products for their retirement plans:

- Fidelity Total Bond Fund
- Fidelity Short-Term Bond Fund
- Fidelity Magellan Fund
- Fidelity International Small Cap Fund
- Fidelity Freedom Income Fund

After working at the university for a year, Minori Vardan is now eligible to participate in the retirement plan. She has already decided to invest a portion of her retirement funds in the Magellan fund. She would like to choose one other fund that has the smallest correlation, preferably zero, with the Magellan fund. She collects 5 years of monthly return data for each mutual fund, a portion of which is shown in the accompanying table.

**Data for Case Study 14.1** Monthly Return Data for Five Mutual Funds

	Magellan	Total Bond	Short-Term Bond	Int'l Small Cap	Freedom Income
January 2006	0.0477	0.0026	0.0013	0.0904	0.0107
February 2006	-0.0149	0.0026	0.0027	-0.0152	0.0000
⋮	⋮	⋮	⋮	⋮	⋮
December 2010	0.0586	-0.0183	-0.0035	0.0478	0.0036

SOURCE: finance.yahoo.com; data retrieved January 6, 2011.

**FILE**  
*Fidelity\_Retirement*

In a report, use the sample information to:

1. Calculate and interpret the sample correlation coefficient of each fund with Magellan.
2. Discuss the statistical significance of the correlation coefficients.
3. Make an investment recommendation for Minori.

**CASE STUDY 14.2** Akiko Hamaguchi, the manager at a small sushi restaurant in Phoenix, Arizona, is concerned that the weak economic environment has hampered foot traffic in her area, thus causing a dramatic decline in sales. Her cousin in San Francisco, Hiroshi Sato, owns a similar restaurant, but he has seemed to prosper during these rough economic times. Hiroshi agrees that higher unemployment rates have likely forced

some customers to dine out less frequently, but he maintains an aggressive marketing campaign to thwart this apparent trend. For instance, he advertises in local papers with valuable two-for-one coupons and promotes early-bird specials over the airwaves. Despite the fact that advertising increases overall costs, he believes that this campaign has positively affected sales at his restaurant. In order to support his claim, Hiroshi provides monthly sales data and advertising costs pertaining to his restaurant, as well as the monthly unemployment rate from San Francisco County. A portion of the data is shown in the accompanying table.

**Data for Case Study 14.2** Hiroshi's Sales, Advertising Costs, and Unemployment Data

Month	Year	Sales (in \$1,000s)	Advertising Costs (in \$)	Unemployment Rate (in percent)
January	2008	27.0	550	4.6
February	2008	24.2	425	4.3
⋮	⋮	⋮	⋮	⋮
May	2009	27.4	550	9.1

SOURCE FOR UNEMPLOYMENT RATE DATA: Development Department, State of California, June 2009.

In a report, use the sample information to:

1. Estimate a simple regression model,  $\text{Sales} = \beta_0 + \beta_1 \text{Advertising} + \epsilon$ , as well as a multiple regression model,  $\text{Sales} = \beta_0 + \beta_1 \text{Advertising} + \beta_2 \text{Unemployment} + \epsilon$ .
2. Show that the multiple regression model is more appropriate for making predictions.
3. Make predictions for sales with an unemployment rate of 6% and advertising costs of \$400 and \$600.

**CASE STUDY 14.3** Megan Hanson, a realtor in Brownsburg, Indiana, would like to use estimates from a multiple regression model to help prospective sellers determine a reasonable asking price for their homes. She believes that the following four factors influence the asking price (Price) of a house: (1) the square footage of the house (SQFT), (2) the number of bedrooms (Bed), (3) the number of bathrooms (Bath), and (4) the lot size (LTSZ) in acres. She randomly collects online listings for 50 single-family homes. A portion of the data is presented in the accompanying table.

**Data for Case Study 14.3** Real Estate Data for Brownsburg, Indiana

Price	SQFT	Bed	Bath	LTSZ
399900	5026	4	4.5	0.3
375000	3200	4	3	5
⋮	⋮	⋮	⋮	⋮
102900	1938	3	1	0.1

SOURCE: *Indianapolis Star*, February 27, 2008.

In a report, use the sample information to:

1. Provide summary statistics on the asking price, square footage, the number of bedrooms, the number of bathrooms, and the lot size.
2. Estimate and interpret a multiple regression model where the asking price is the response variable and the other four factors are the explanatory variables.
3. Interpret the resulting coefficient of determination.

**FILE**

*Sushi\_Restaurant*

**FILE**

*Indiana\_Real Estate*

## APPENDIX 14.1 Guidelines for Other Software Packages

The following section provides brief commands for specific software packages: Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### Constructing a Scatterplot with Trendline

(Replicating Figure 14.3) From the menu choose **Graph > Scatterplot > With Regression**. Under **Y variables**, select Debt, and under **X variables**, select Income.

**FILE**  
*Debt\_Payments*

#### Simple Linear Regression

(Replicating Example 14.5) From the menu choose **Stat > Regression > Regression > Fit Regression Model**. Select Debt for **Responses** and select Income for **Continuous predictors**.

#### Multiple Regression

(Replicating Example 14.6) From the menu choose **Stat > Regression > Regression > Fit Regression Model**. Select Debt for **Responses**, and select Income and Unemployment for **Continuous predictors**.

### SPSS

#### Simple Linear Regression

(Replicating Example 14.5) From the menu choose **Analyze > Regression-Linear**. Select Debt as **Dependent**, and Income as **Independent(s)**.

**FILE**  
*Debt\_Payments*

#### Multiple Regression

(Replicating Example 14.6) From the menu choose **Analyze > Regression-Linear**. Select Debt as **Dependent**, and Income and Unemployment as **Independent(s)**.

### JMP

#### Constructing a Scatterplot with Trendline and Simple Linear Regression

- A. (Replicating Figure 14.3 and Example 14.5) From the menu choose **Analyze > Fit Y by X**. Select Debt as **Y, Response**, and select Income as **X, Factor**.
- B. Click on the red triangle next to the header that reads **Bivariate Fit of Debt by Income** and select **Fit line**.

**FILE**  
*Debt\_Payments*

#### Multiple Regression

(Replicating Example 14.6) From the menu choose **Analyze > Fit Model**. Under **Pick Role Variables**, select Debt, and under **Construct Model Effects**, select Income and Unemployment, and choose **Add**.

# 15

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 15.1 Conduct tests of individual significance.
- LO 15.2 Conduct a test of joint significance.
- LO 15.3 Conduct a general test of linear restrictions.
- LO 15.4 Calculate and interpret confidence intervals and prediction intervals.
- LO 15.5 Explain the role of the assumptions on the OLS estimators.
- LO 15.6 Describe common violations of the assumptions and offer remedies.

# Inference with Regression Models

In Chapter 14, we employed simple and multiple regression models to capture a relationship between a response variable and one or more explanatory variables. We also studied objective goodness-of-fit measures that assess how well the sample regression equation fits the data. While the estimated regression models and goodness-of-fit measures are useful, it is not clear if the conclusions based on the estimated coefficients are real or due to chance. In this chapter, we focus on statistical inference with regression models. In particular, we develop hypothesis tests that enable us to determine the individual and joint significance of the explanatory variables. We also develop interval estimates for a prediction from the sample regression equation. Finally, we examine the importance of the assumptions on the statistical properties of the ordinary least squares (OLS) estimator, as well as the validity of the testing procedures. We address common violations to the model assumptions, the consequences when these assumptions are violated, and offer some remedial measures.





## INTRODUCTORY CASE

### Analyzing the Winning Percentage in Baseball

On a recent radio talk show, two sports analysts quarreled over which statistic was a better predictor of a Major League Baseball team's winning percentage (Win). One argued that the team's batting average (BA) was a better predictor of a team's success since the team with the higher batting average has won approximately 75% of the World Series contests. The other insisted that a team's pitching is clearly the main factor in determining wins—the lower a team's earned run average (ERA), the higher the team's winning percentage. In order to determine if these claims are backed by the data, relevant information is collected for the 14 American League (AL) and 16 National League (NL) teams during the regular season of 2010. A portion of the data is shown in Table 15.1.

**TABLE 15.1** Winning Percentage, Batting Average, and Earned Run Average in Baseball

<b>FILE</b> <i>Baseball</i>	Team	League	Win	BA	ERA
	Baltimore Orioles	AL	0.407	0.259	4.59
	Boston Red Sox	AL	0.549	0.268	4.20
	⋮	⋮	⋮	⋮	⋮
	Washington Nationals	NL	0.426	0.250	4.13

SOURCE: mlb.mlb.com.

Use the sample information to:

1. Employ goodness-of-fit measures to determine the best-fitting regression model for the winning percentage.
2. Determine the statistical significance of the batting average and earned run average variables.

A synopsis of this case is provided at the end of Section 15.1.

# 15.1 TESTS OF SIGNIFICANCE

This section continues from Chapter 14 with the assessment of linear regression models. Then we turn our attention to hypothesis testing about the unknown parameters (coefficients)  $\beta_0, \beta_1, \dots, \beta_k$ . In particular, we test for the individual and joint significance of the regression coefficients to determine whether there is evidence of a linear relationship between the response and the explanatory variables. We note that for the tests to be valid, the ordinary least squares (OLS) estimators  $b_0, b_1, \dots, b_k$  must be normally distributed. This condition is satisfied if the random error term,  $\varepsilon$ , is normally distributed. If we cannot assume the normality of  $\varepsilon$ , then the tests are valid only for large sample sizes. We discuss the underlying assumptions of the linear regression model in Section 15.4.

The objective outlined in the introductory case study is to predict a baseball team's winning percentage, denoted Win, either on the basis of its batting average BA or its earned run average ERA, or jointly by BA and ERA. For those readers who do not follow baseball, BA is a ratio of hits divided by times at bat, and ERA is the average number of earned runs given up by a pitcher per nine innings pitched. A priori, we expect that a higher BA positively influences a team's winning percentage, while a higher ERA negatively affects a team's winning percentage. We define

- Model 1 as  $\text{Win} = \beta_0 + \beta_1 \text{BA} + \varepsilon$ ,
- Model 2 as  $\text{Win} = \beta_0 + \beta_1 \text{ERA} + \varepsilon$ , and
- Model 3 as  $\text{Win} = \beta_0 + \beta_1 \text{BA} + \beta_2 \text{ERA} + \varepsilon$ .

Before we develop the tests of significance, we will use goodness-of-fit measures discussed in Chapter 14 to choose the appropriate model. Table 15.2 shows the relevant regression output for the three models; we advise you to replicate these results using the sample data in Table 15.1.

TABLE 15.2 Relevant Regression Output to Compare the Models

	Model 1	Model 2	Model 3
Multiple $R$	0.4596	0.6823	0.8459
$R$ Square	0.2112	0.4656	0.7156
Adjusted $R$ Square	0.1830	0.4465	0.6945
Standard Error	0.0614	0.0505	0.0375
Observations	30	30	30

We choose Model 3 to predict the winning percentage because it has the lowest standard error of the estimate and the highest adjusted  $R^2$ . Remember, we cannot compare on the basis of  $R^2$ , because Models 1 and 2 use only one explanatory variable, whereas Model 3 uses two.

## LO 15.1

Conduct tests of individual significance.

## Tests of Individual Significance

Tests of individual significance can be implemented in the context of the simple and the multiple regression models. Consider the following multiple regression model, which links the response variable  $y$  with  $k$  explanatory variables  $x_1, x_2, \dots, x_k$ :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon.$$

If the slope coefficient  $\beta_j$  equals zero, then the explanatory variable  $x_j$  basically drops out of the above equation, implying that  $x_j$  does not influence  $y$ . In other words, if  $\beta_j$  equals zero, there is no linear relationship between  $x_j$  and  $y$ . Conversely, if  $\beta_j$  does not equal zero, then  $x_j$  influences  $y$ .

Following the hypothesis testing methodology introduced in earlier chapters, we want to test whether the population coefficient  $\beta_j$  is different from, greater than, or less than  $\beta_{j0}$ , where  $\beta_{j0}$  is the hypothesized value of  $\beta_j$ . That is, the competing hypotheses take one of the following forms:



Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: \beta_j = \beta_{j0}$	$H_0: \beta_j \leq \beta_{j0}$	$H_0: \beta_j \geq \beta_{j0}$
$H_A: \beta_j \neq \beta_{j0}$	$H_A: \beta_j > \beta_{j0}$	$H_A: \beta_j < \beta_{j0}$

When testing whether  $x_j$  significantly influences  $y$ , we set  $\beta_{j0} = 0$  and specify a two-tailed test as  $H_0: \beta_j = 0$  and  $H_A: \beta_j \neq 0$ . We could easily specify one-tailed competing hypotheses for a positive linear relationship ( $H_0: \beta_j \leq 0$  and  $H_A: \beta_j > 0$ ) or a negative linear relationship ( $H_0: \beta_j \geq 0$  and  $H_A: \beta_j < 0$ ).

Although tests of significance are commonly based on  $\beta_{j0} = 0$ , in some situations we might wish to determine whether the slope coefficient differs from a nonzero value. For instance, if we are analyzing the relationship between students' exam scores on the basis of hours studied, we may want to determine if an extra hour of review before the exam will increase a student's score by more than 5 points. Here, we formulate the hypotheses as  $H_0: \beta_j \leq 5$  and  $H_A: \beta_j > 5$ ; in this example,  $\beta_{j0} = 5$ . Finally, although in most applications we are interested in conducting hypothesis tests on the slope coefficient(s), there are instances where we may also be interested in testing the intercept,  $\beta_0$ . The testing framework for the intercept remains the same; that is, if we want to test whether the intercept differs from zero, we specify the competing hypotheses as  $H_0: \beta_0 = 0$  and  $H_A: \beta_0 \neq 0$ .

As in all hypothesis tests, the next essential piece of information is how we define the appropriate test statistic.

#### TEST STATISTIC FOR THE TEST OF INDIVIDUAL SIGNIFICANCE

The value of the test statistic for a **test of individual significance** is calculated as

$$t_{df} = \frac{b_j - \beta_{j0}}{se(b_j)},$$

where  $df = n - k - 1$ ,  $n$  is the sample size,  $k$  is the number of explanatory variables,  $b_j$  is the estimate for  $\beta_j$ ,  $se(b_j)$  is the standard error of the OLS estimator  $b_j$ , and  $\beta_{j0}$  is the hypothesized value of  $\beta_j$ . If  $\beta_{j0} = 0$ , the value of the test statistic reduces

$$\text{to } t_{df} = \frac{b_j}{se(b_j)}.$$

Tests of significance can be implemented in the context of the simple and the multiple regression models. In Example 15.1 and Example 15.2, the tests are based on the multiple regression model; Example 15.3 uses the simple regression model for the test.

#### EXAMPLE 15.1

Let's revisit Model 3,  $\text{Win} = \beta_0 + \beta_1 \text{BA} + \beta_2 \text{ERA} + \varepsilon$ , estimated with the sample data in Table 15.1. We reproduce a portion of the regression results in Table 15.3. Conduct a hypothesis test to determine whether batting average influences winning percentage at the 5% significance level.

**TABLE 15.3** Portion of Regression Results for Model 3:  $\text{Win} = \beta_0 + \beta_1 \text{BA} + \beta_2 \text{ERA} + \varepsilon$

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	0.1269	0.1822	0.6964	0.4921	-0.2470	0.5008
BA	3.2754	0.6723	4.8719	0.0000	1.8960	4.6549
ERA	-0.1153	0.0167	-6.9197	0.0000	-0.1494	-0.0811

**SOLUTION:** We use the  $p$ -value approach for the test. We set up the following competing hypotheses in order to determine whether winning percentage and batting average have a linear relationship:

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

From Table 15.3, we find that  $b_1 = 3.2754$  and  $se(b_1) = 0.6723$ . In addition, given  $n = 30$  and  $k = 2$ , we find  $df = n - k - 1 = 30 - 2 - 1 = 27$ . So we calculate the value of the test statistic as  $t_{27} = \frac{b_1 - \beta_{10}}{se(b_1)} = \frac{3.2754 - 0}{0.6723} = 4.8719$ . Note that this calculation is not necessary since Excel automatically provides the value of the test statistic and its associated  $p$ -value.

It is important to note that the Excel-reported results are valid only in a standard case where a two-tailed test is implemented to determine whether a regression coefficient differs from zero. Here, we can use the Excel-reported results because the above example represents a standard case. Shortly, we will see an application with a nonstandard case.

As usual, the decision rule is to reject  $H_0$  if the  $p$ -value  $< \alpha$ . Since the  $p$ -value is approximately zero, we reject  $H_0$ . At the 5% significance level, there is a linear relationship between winning percentage and batting average; in other words, batting average is significant in explaining winning percentage.

## Using a Confidence Interval to Determine Individual Significance

In earlier chapters, we constructed a confidence interval to conduct a two-tailed hypothesis test. When assessing whether the regression coefficient differs from zero, we can apply the same methodology.

### CONFIDENCE INTERVAL FOR $\beta_j$

A  $100(1 - \alpha)\%$  confidence interval for the regression coefficient  $\beta_j$  is computed as

$$b_j \pm t_{\alpha/2, df} se(b_j) \quad \text{or} \quad [b_j - t_{\alpha/2, df} se(b_j), b_j + t_{\alpha/2, df} se(b_j)],$$

where  $se(b_j)$  is the standard error of  $b_j$  and  $df = n - k - 1$ .

Excel automatically provides a 95% confidence interval for the regression coefficients; it will provide other levels if prompted. In general, if the confidence interval for the slope coefficient contains the value zero, then the explanatory variable associated with the regression coefficient is not significant. Conversely, if the confidence interval does not contain the value zero, then the explanatory variable associated with the regression coefficient is statistically significant. The next example is based on the the confidence interval approach.

### EXAMPLE 15.2

Let's revisit Model 3 with a portion of the regression results in Table 15.3. Construct the 95% confidence interval to determine whether earned run average is significant in explaining winning percentage.

**SOLUTION:** For testing whether earned run average is statistically significant, we set up the following competing hypotheses:

$$H_0: \beta_2 = 0$$

$$H_A: \beta_2 \neq 0$$

For the 95% confidence interval,  $\alpha = 0.05$  and  $\alpha/2 = 0.025$ . With  $n = 30$  and  $k = 2$ , we use  $df = 30 - 2 - 1 = 27$  and reference the  $t$  table to find  $t_{\alpha/2, df} = t_{0.025, 27} = 2.052$ . Given  $b_2 = -0.1153$  and  $se(b_2) = 0.0167$  (from Table 15.3), the 95% confidence interval for the population coefficient  $\beta_2$  is

$$b_2 \pm t_{\alpha/2, df} se(b_2) = -0.1153 \pm 2.052 \times 0.0167 = -0.1153 \pm 0.0343.$$

Thus, the lower and upper limits of the confidence interval are  $-0.1496$  and  $-0.0810$ , respectively. Note that Table 15.3 also provides these values. (Slight differences are due to rounding.) Since the 95% confidence interval does not contain the value zero, we can conclude that earned run average is significant in explaining the winning percentage at  $\alpha = 0.05$ .

So far, we have only considered examples of two-tailed tests to determine if a regression coefficient differs from zero. As mentioned earlier, for such standard cases, we can use the computer-generated value of the test statistic as well as the corresponding  $p$ -value. For a one-tailed test with  $\beta_{j0} = 0$ , the value of the test statistic is valid, but the  $p$ -value is not; in most cases, the computer-generated  $p$ -value must be divided in half. For a one- or two-tailed test to determine if the regression coefficient differs from a nonzero value, both the computer-generated value of the test statistic and the  $p$ -value become invalid. These facts are summarized below.

#### COMPUTER-GENERATED TEST STATISTIC AND THE $P$ -VALUE

Excel and virtually all other statistical packages report a value of the test statistic and its associated  $p$ -value for a two-tailed test that assesses whether the regression coefficient differs from zero.

- If we specify a one-tailed test, then we need to divide the computer-generated  $p$ -value in half.
- If we test whether the coefficient differs from a nonzero value, then we cannot use the value of the computer-generated test statistic and its  $p$ -value.

We would also like to point out that for a one-tailed test with  $\beta_{j0} = 0$ , there are rare instances when the computer-generated  $p$ -value is invalid. This occurs when the sign of  $b_j$  (and the value of the accompanying test statistic) is not inconsistent with the null hypothesis. For example, for a right-tailed test,  $H_0: \beta_j \leq 0$  and  $H_A: \beta_j > 0$ , the null hypothesis cannot be rejected if the estimate  $b_j$  (and the value of the accompanying test statistic  $t_{df}$ ) is negative. Similarly, no further testing is necessary if  $b_j > 0$  (and thus  $t_{df} > 0$ ) for a left-tailed test.

### A Test for a Nonzero Slope Coefficient

In Examples 15.1 and 15.2, the null hypothesis included a zero value for the slope coefficient; that is,  $\beta_{j0} = 0$ . We now motivate a test where the hypothesized value is not zero by using a renowned financial application—the capital asset pricing model (CAPM).

Let  $R$  represent the return on a stock or portfolio of interest. Given the market return  $R_M$  and the risk-free return  $R_f$ , the CAPM expresses the risk-adjusted return of an asset,  $R - R_f$ , as a function of the risk-adjusted market return,  $R_M - R_f$ . It is common to use the return of the S&P 500 index for  $R_M$  and the return on a Treasury bill for  $R_f$ . For empirical estimation, we express the CAPM as

$$R - R_f = \alpha + \beta(R_M - R_f) + \varepsilon.$$

We can rewrite the model as  $y = \alpha + \beta x + \varepsilon$ , where  $y = R - R_f$  and  $x = R_M - R_f$ . Note that this is essentially a simple linear regression model that uses  $\alpha$  and  $\beta$ , in place of the usual  $\beta_0$  and  $\beta_1$ , to represent the intercept and the slope coefficients, respectively. The slope coefficient  $\beta$ , called the stock's **beta**, measures how sensitive the stock's return is to changes in the level of the overall market. When  $\beta$  equals 1, any change in the market return leads to an identical change in the given stock return. A stock for which  $\beta > 1$  is considered more “aggressive” or riskier than the market, whereas one for which  $\beta < 1$  is considered “conservative” or less risky. We also give importance to the intercept coefficient  $\alpha$ , called the stock's **alpha**. The CAPM theory predicts  $\alpha$  to be zero, and thus a nonzero estimate indicates abnormal returns. Abnormal returns are positive when  $\alpha > 0$  and negative when  $\alpha < 0$ .

### EXAMPLE 15.3

Johnson & Johnson (J&J) was founded more than 120 years ago on the premise that doctors and nurses should use sterile products to treat people's wounds. Since that time, J&J products have become staples in most people's homes. Consider the CAPM where the J&J risk-adjusted stock return  $R - R_f$  is used as the response variable and the risk-adjusted market return  $R_M - R_f$  is used as the explanatory variable. A portion of 60 months of data is shown in Table 15.4.

**TABLE 15.4** Risk-Adjusted Stock Return of J&J and Market Return

Date	$R - R_f$	$R_M - R_f$
1/1/2006	-4.59	2.21
2/1/2006	0.39	-0.31
⋮	⋮	⋮
12/1/2010	0.48	2.15

SOURCE: finance.yahoo.com and U.S. Treasury.

- Since consumer staples comprise many of the products sold by J&J, its stock is often considered less risky; that is, people need these products whether the economy is good or bad. At the 5% significance level, is the beta coefficient less than one?
- At the 5% significance level, are there abnormal returns? In other words, is the alpha coefficient significantly different from zero?

**SOLUTION:** We use the critical value approach for part a and the  $p$ -value approach for part b. Using the CAPM notation, we estimate the model,  $R - R_f = \alpha + \beta(R_M - R_f) + \varepsilon$ ; the relevant portion of the regression output is presented in Table 15.5.

**TABLE 15.5** Portion of CAPM Regression Results for J&J

	Coefficients	Standard Error	$t$ Stat	$p$ -value
Intercept	0.2666	0.4051	0.6580	0.5131
$R_M - R_f$	0.5844	0.0803	7.2759	0.0000

- The estimate for the beta coefficient is 0.5844 and its standard error is 0.0803. Interestingly, our estimate is identical to the beta reported in the popular press (www.dailyfinance.com, March 4, 2011). In order to determine whether the beta coefficient is significantly less than one, we formulate the hypotheses as

$$H_0: \beta \geq 1$$

$$H_A: \beta < 1$$

Given 60 data points,  $df = n - k - 1 = 60 - 1 - 1 = 58$ . With  $df = 58$  and  $\alpha = 0.05$ , the critical value for a left-tailed test is  $-t_{0.05,58} = -1.672$ . We cannot use the test statistic value reported in Table 15.5, since the hypothesized value of  $\beta$  is not zero. We calculate the value of the test statistic as  $t_{58} = \frac{b_j - \beta_0}{se(b_j)} = \frac{0.5844 - 1}{0.0803} = -5.18$ . The decision rule is to reject  $H_0$  if  $t_{58} < -t_{0.05,58}$ . Since  $-5.18 < -1.672$ , we reject  $H_0$  and conclude that  $\beta$  is significantly less than one; that is, the return on J&J stock is less risky than the return on the market.

- Abnormal returns exist when  $\alpha$  is significantly different from zero. Thus, the competing hypotheses are  $H_0: \alpha = 0$  versus  $H_A: \alpha \neq 0$ . Since it is a standard case, where the hypothesized value of the coefficient is zero, we can use the reported test statistic value of 0.6580 with an associated  $p$ -value of 0.5131. We cannot reject  $H_0$  at any reasonable level of significance. Therefore, we cannot conclude that there are abnormal returns for J&J stock.

#### FILE

Johnson\_Johnson

## Test of Joint Significance

### LO 15.2

Conduct a test of joint significance.

So far we considered tests of individual significance of explanatory variables. For instance, we used a  $t$  test to determine whether the batting average has a statistically significant influence on the winning percentage. When we assess a multiple linear regression model, it is also important to conduct a **test of joint significance**. A test of joint significance is often regarded as a test of the overall usefulness of a regression. This test determines whether the explanatory variables  $x_1, x_2, \dots, x_k$  have a joint statistical influence on  $y$ .

In the null hypothesis of the test of joint significance, *all* of the slope coefficients are assumed zero. The competing hypotheses for a test of joint significance are specified as

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_A: \text{At least one } \beta_j \neq 0.$$

You might be tempted to implement this test by performing a series of tests of individual significance with the  $t$  statistic. However, such an option is not appropriate. The test of joint significance determines if at least one of the explanatory variables is significant. Therefore, it is not clear if one or all of the explanatory variables must be significant in order to document a joint significance. In addition, recall from the discussion of ANOVA in Chapter 13 that if we conduct many individual tests at (say) a 5% level of significance, the resulting significance level for the joint test will be greater than 5%.

Testing a series of individual hypotheses is not equivalent to testing the same hypotheses jointly.

To conduct the test of joint significance, we employ a right-tailed  $F$  test. (Recall that the  $F_{(df_1, df_2)}$  distribution introduced in Chapter 11 was used for hypothesis testing in Chapters 11 and 13.) The test statistic measures how well the regression equation explains the variability in the response variable. It is defined as the ratio of the **mean square regression (MSR)** to the **mean square error (MSE)** where  $MSR = SSR/k$  and  $MSE = SSE/(n - k - 1)$ . Recall from Chapter 14 that  $SSR$  is the sum of squares due to regression and  $SSE$  is the sum of squares due to error. These values, including the  $F_{(df_1, df_2)}$  test statistic, are provided in the ANOVA portion of the regression results.

### TEST STATISTIC FOR THE TEST OF JOINT SIGNIFICANCE

The value of the test statistic for a **test of joint significance** is calculated as

$$F_{(df_1, df_2)} = \frac{SSR/k}{SSE/(n - k - 1)} = \frac{MSR}{MSE},$$

where  $df_1 = k$ ,  $df_2 = n - k - 1$ ,  $SSR$  is the sum of squares due to regression,  $SSE$  is the sum of squares due to error,  $MSR$  is the mean square regression, and  $MSE$  is the mean square error.

In general, a large value of  $F_{(df_1, df_2)}$  indicates that a large portion of the sample variation in  $y$  is explained by the regression model; thus, the model is useful. A small value of  $F_{(df_1, df_2)}$  implies that a large portion of the sample variation in  $y$  remains unexplained. In fact, the test of joint significance is sometimes informally referred to as the test of the significance of  $R^2$ . Note that while the test of joint significance is important for a multiple regression model, it is redundant for a simple regression model. In fact, in a simple regression model, the  $p$ -value of the  $F$  test is identical to that of the  $t$  test; we advise you to verify this fact.

### EXAMPLE 15.4

Let's revisit Model 3,  $\text{Win} = \beta_0 + \beta_1\text{BA} + \beta_2\text{ERA} + \varepsilon$ , estimated with the sample data in Table 15.1. We reproduce the ANOVA portion of the regression results in Table 15.6. Conduct a test to determine if batting average and earned run average are jointly significant in explaining the winning percentage at  $\alpha = 0.05$ .

**TABLE 15.6** Portion of Regression Results for Model 3:  $\text{Win} = \beta_0 + \beta_1\text{BA} + \beta_2\text{ERA} + \varepsilon$

ANOVA	df	SS	MS	F	Significance F
Regression	2	0.0958	0.0479	33.9663	0.0000
Residual	27	0.0381	0.0014		
Total	29	0.1338			

**SOLUTION:** When testing whether the explanatory variables are jointly significant in explaining winning percentage, we set up the following competing hypotheses:

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_A: \text{At least one } \beta_j \neq 0.$$

Given  $n = 30$  and  $k = 2$ , we find that  $df_1 = k = 2$  and  $df_2 = n - k - 1 = 27$ . From Table 15.6, we calculate the value of the test statistic as

$$F_{(2,27)} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE} = \frac{0.0958/2}{0.0381/(30-2-1)} = \frac{0.0479}{0.0014} = 34.21.$$

Note that Excel reports the test statistic value as 33.9663. The values are slightly different due to rounding.

Given the computer-generated regression output, the easiest way to conduct a test of joint significance is with the  $p$ -value approach, since the ANOVA table from the computer output provides both the value of the test statistic and its associated  $p$ -value. The value under the heading *Significance F* is the  $p$ -value, or  $P(F_{(2,27)} > 33.9663) \approx 0.0000$ . Since the  $p$ -value is less than  $\alpha = 0.05$ , we reject  $H_0$ . At the 5% significance level, the batting average and the earned run average variables are jointly significant in explaining winning percentage. The result is not surprising since both BA and ERA were found to be individually significant in Examples 15.1 and 15.2. It is worth noting that sometimes we may find joint significance when only one, sometimes even none, of the explanatory variables is individually significant.

## Reporting Regression Results

Regression results are often reported in a “user-friendly” table. Table 15.7 reports the regression results for the three models discussed in this section that attempt to explain a baseball team's winning percentage. The explanatory variables are batting average in Model 1, earned run average in Model 2, and both batting average and earned run average in Model 3. If we were supplied with only this table, we would be able to compare these models, construct the sample regression equation of the chosen model, and perform a respectable assessment of the model with the statistics provided. Many tables contain a Notes section at the bottom explaining some of the notation. We choose to put the  $p$ -values in parentheses under all estimated coefficients; however, some researchers place the standard errors of the coefficients or the values of the test statistics in parentheses. Whichever format is chosen, it must be made clear to the reader in the Notes section.



**TABLE 15.7** Estimates of the Alternative Regression Models

Variable	Model 1	Model 2	Model 3
Intercept	-0.2731 (0.3421)	0.9504* (0.0000)	0.1269 (0.4921)
Batting Average	3.0054* (0.0106)	NA	3.2754* (0.0000)
Earned Run Average	NA	-0.1105* (0.0000)	-0.1153* (0.0000)
$s_e$	0.0614	0.0505	0.0375
$R^2$	0.2112	0.4656	0.7156
Adjusted $R^2$	0.1830	0.4465	0.6945
F-test ( $p$ -value)			33.9663*(0.0000)

NOTES: Parameter estimates are in the top half of the table with the  $p$ -values in parentheses; \* represents significance at the 5% level. NA denotes not applicable. The lower part of the table contains goodness-of-fit measures.

## SYNOPSIS OF INTRODUCTORY CASE

Two sports analysts have conflicting views over how best to predict a Major League Baseball team's winning percentage. One argues that the team's batting average is a better predictor of a team's success, since the team with the higher batting average has won approximately 75% of the World Series contests. The other analyst insists that a team's pitching is clearly the main factor in determining wins. Three regression models are used to analyze a baseball team's winning percentage (Win). The explanatory variables are batting average (BA) in Model 1, earned run average (ERA) in Model 2, and both BA and ERA in Model 3.



After estimating the models using data from 14 American League teams and 16 National League teams during the regular season of 2010, it is found that Model 2 has a lower standard error and a higher  $R^2$  than Model 1. Therefore, if simply choosing between these two models, Model 2 appears better for prediction. However, Model 3 provides the best overall fit, as measured by its highest adjusted  $R^2$  value. The sample regression equation for Model 3 is  $\widehat{\text{Win}} = 0.13 + 3.28\text{BA} - 0.12\text{ERA}$ . Further testing of this preferred model reveals that the two explanatory variables are jointly as well as individually significant in explaining the winning percentage at the 5% significance level. It appears that neither analyst is totally right or totally wrong. Given  $R^2 = 0.7156$ , approximately 72% of the sample variability in the winning percentage is explained by the estimated Model 3. However, 28% of the sample variability in the winning percentage remains unexplained. This is not entirely surprising, since other factors, besides batting average and earned run average, influence a baseball team's winning percentage.

## EXERCISES 15.1

### Mechanics

- In a simple linear regression based on 30 observations, it is found that  $b_1 = 3.25$  and  $se(b_1) = 1.36$ . Consider the hypotheses:

$$H_0: \beta_1 = 0 \text{ and } H_A: \beta_1 \neq 0.$$

- Calculate the value of the test statistic.
- Approximate the  $p$ -value.

- At the 5% significance level, what is the conclusion? Is the explanatory variable statistically significant?
- In a simple linear regression based on 25 observations, it is found that  $b_1 = 0.5$  and  $se(b_1) = 0.3$ . Consider the hypotheses:

$$H_0: \beta_1 \leq 0 \text{ and } H_A: \beta_1 > 0.$$

- At the 5% significance level, find the critical value(s).
  - Calculate the value of the test statistic.
  - What is the conclusion to the test?
3. In a simple linear regression based on 30 observations, it is found that  $b_1 = 7.2$  and  $se(b_1) = 1.8$ . Consider the hypotheses:

$$H_0: \beta_1 \geq 10 \text{ and } H_A: \beta_1 < 10.$$

- At the 5% significance level, find the critical value(s).
  - Calculate the value of the test statistic.
  - What is the conclusion to the test?
4. Consider the following regression results based on 20 observations.

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	34.2123	4.5665	7.4920	0.0000	24.62	43.81
$x_1$	0.1223	0.1794	0.6817	0.5041	-0.25	0.50

- Specify the hypotheses to determine if the intercept differs from zero. Perform this test at the 5% significance level.
  - Construct the 95% confidence interval for the slope coefficient. At the 5% significance level, does the slope differ from zero? Explain.
5. Consider the following regression results based on 40 observations.

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	43.1802	12.6963	3.4010	0.0016	17.48	68.88
$x_1$	0.9178	0.9350	0.9816	0.3325	-0.97	2.81

- Specify the hypotheses to determine if the slope differs from minus one.
  - At the 5% significance level, find the critical value(s).
  - Calculate the value of the test statistic.
  - Does the slope differ from minus one? Explain.
6. When estimating a multiple linear regression model based on 30 observations, the following results were obtained.

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	152.27	119.70	1.27	0.2142	-93.34	397.87
$x_1$	12.91	2.68	4.81	5.06E-05	7.40	18.41
$x_2$	2.74	2.15	1.28	0.2128	-1.67	7.14

- Specify the hypotheses to determine whether  $x_1$  is linearly related to  $y$ . At the 5% significance level, use the  $p$ -value approach to complete the test. Are  $x_1$  and  $y$  linearly related?
- What is the 95% confidence interval for  $\beta_2$ ? Using this confidence interval, is  $x_2$  significant in explaining  $y$ ? Explain.
- At the 5% significance level, can you conclude that  $\beta_1$  is less than 20? Show the relevant steps of the appropriate hypothesis test.

7. The following ANOVA table was obtained when estimating a multiple linear regression model.

ANOVA	df	SS	MS	F	Significance F
Regression	2	22016.75	11008.38		0.0228
Residual	17	39286.93	2310.996		
Total	19	61303.68			

- How many explanatory variables were specified in the model? How many observations were used?
- Specify the hypotheses to determine whether the explanatory variables are jointly significant.
- Compute the value of the test statistic.
- At the 5% significance level, what is the conclusion to the test? Explain.

## Applications

8. A marketing manager analyzes the relationship between the annual sales of a firm (in \$100,000s) and its advertising expenditures (in \$10,000s). He collects data from 20 firms and estimates  $\text{Sales} = \beta_0 + \beta_1 \text{Advertising} + \epsilon$ . A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	-7.42	1.46	-5.09	7.66E-05
Advertising	0.42	0.05		1.21E-07

- Specify the competing hypotheses in order to determine whether advertising expenditures and sales have a positive linear relationship.
  - Calculate the value of the test statistic.
  - At the 5% significance level, what is the conclusion to the test? Do advertising expenditures and sales have a positive linear relationship?
9. In order to examine the relationship between the selling price of a used car and its age, an analyst uses data from 20 recent transactions and estimates  $\text{Price} = \beta_0 + \beta_1 \text{Age} + \epsilon$ . A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	21187.94	733.42	28.89	1.56E-16
Age	-1208.25	128.95		2.41E-08

- Specify the competing hypotheses in order to determine whether the selling price of a used car and its age are linearly related.
- Calculate the value of the test statistic.
- At the 5% significance level, what is the conclusion to the test? Is the age of a used car significant in explaining its selling price?
- Conduct a hypothesis test at the 5% significance level in order to determine if  $\beta_1$  differs from -1000. Show all of the relevant steps.

10. A recent study on the evolution of mankind shows that, with a few exceptions, world-record holders in the 100-meter dash have progressively gotten bigger over time (*The Wall Street Journal*, July 22, 2009). The following table shows runners who have held the record, along with their record-holding times and heights:

Record Holder/Year	Time (in seconds)	Height (in inches)
Eddie Tolan (1932)	10.30	67
Jesse Owens (1936)	10.20	70
Charles Greene (1968)	9.90	68
Eddie Hart (1972)	9.90	70
Carl Lewis (1991)	9.86	74
Asafa Powell (2007)	9.74	75
Usain Bolt (2008)	9.69	77

A portion of the regression results from estimating  $\text{Time} = \beta_0 + \beta_1 \text{Height} + \varepsilon$  is:

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	13.353	1.1714	11.3990	9.1E-05	10.34	16.36
Height	-0.0477	0.0163			-0.09	-0.01

- Formulate the estimated regression equation.
  - Specify the hypotheses to determine whether Height is linearly related to Time.
  - Calculate the value of the test statistic.
  - At the 5% significance level, is Height statistically significant? Explain.
11. An economist examines the relationship between changes in short-term interest rates and long-term interest rates. He believes that changes in short-term rates are significant in explaining long-term interest rates. He estimates the model  $\text{Dlong} = \beta_0 + \beta_1 \text{Dshort} + \varepsilon$ , where  $\text{Dlong}$  is the change in the long-term interest rate (10-year Treasury bill) and  $\text{Dshort}$  is the change in the short-term interest rate (3-month Treasury bill). Monthly data from January 2006 through December 2010 ( $n = 60$ ) were obtained from the St. Louis Federal Reserve's website. A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	-0.0038	0.0088	-0.4273	0.6708	-0.02	0.01
Dshort	0.0473	0.0168	2.8125	0.0067	0.01	0.08

Use a 5% significance level to determine whether there is a linear relationship between  $\text{Dshort}$  and  $\text{Dlong}$ .

12. For a sample of 20 New England cities, a sociologist studies the crime rate in each city (crimes per 100,000 residents) as a function of its poverty rate (in %) and its median income (in \$1,000s). A portion of the regression results is shown in the accompanying table.

ANOVA	df	SS	MS	F	Significance F
Regression	2	188246.8	94123.4		9.04E-07
Residual	17	45457.32	2673.96		
Total	19	233704.1			

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	-301.62	549.7135	-0.5487	0.5903	-1,461.52	858.28
Poverty	53.1597	14.2198	3.7384	0.0016	23.16	83.16
Income	4.9472	8.2566	0.5992	0.5569	-12.47	22.37

- Specify the sample regression equation.
  - At the 5% significance level, show whether the poverty rate and the crime rate are linearly related.
  - Construct the 95% confidence interval for the slope coefficient of income. Using the confidence interval, determine whether income influences the crime rate at the 5% significance level.
  - At the 5% significance level, are the poverty rate and income jointly significant in explaining the crime rate?
13. Akiko Hamaguchi is a manager at a small sushi restaurant in Phoenix, Arizona. Akiko is concerned that the weak economic environment has hampered foot traffic in her area, thus causing a dramatic decline in sales. In order to offset the decline in sales, she has pursued a strong advertising campaign. She believes advertising expenditures have a positive influence on sales. To support her claim, Akiko estimates the following linear regression model:  $\text{Sales} = \beta_0 + \beta_1 \text{Unemployment} + \beta_2 \text{Advertising} + \varepsilon$ . A portion of the regression results is shown in the accompanying table.

ANOVA	df	SS	MS	F	Significance F
Regression	2	72.6374	36.3187	8.760	0.0034
Residual	14	58.0438	4.1460		
Total	16	130.681			

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	17.5060	3.9817	4.3966	0.0006	8.97	26.05
Unemployment	-0.6879	0.2997	-2.2955	0.0377	-1.33	-0.05
Advertising	0.0266	0.0068	3.9322	0.0015	0.01	0.04

- At the 5% significance level, test whether the explanatory variables jointly influence sales.
  - At the 1% significance level, test whether the unemployment rate is negatively related with sales.
  - At the 1% significance level, test whether advertising expenditures are positively related with sales.
14. A researcher estimates the following model relating the return on a firm's stock as a function of its price-to-earnings ratio and its price-to-sales ratio:  $\text{Return} = \beta_0 + \beta_1 \text{P/E} + \beta_2 \text{P/S} + \varepsilon$ . A portion of the regression results is shown in the accompanying table.

ANOVA	df	SS	MS	F	Significance F
Regression	2	918.746	459.3728	2.817095	0.077415
Residual	27	4402.786	163.0661		
Total	29	5321.532			

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	-12.0243	7.886858	-1.5246	0.1390	-28.21	4.16
P/E	0.1459	0.4322	0.3376	0.7383	-0.74	1.03
P/S	5.4417	2.2926	2.3736	0.0250	0.74	10.15

- Specify the sample regression equation.
  - At the 10% significant level, are P/E and P/S jointly significant? Show the relevant steps of the appropriate hypothesis test.
  - Are both explanatory variables individually significant at the 10% significance level? Show the relevant steps of the appropriate hypothesis tests.
- FILE Test\_Scores.** The accompanying data file shows midterm and final grades for 32 students. Estimate a student's final grade as a linear function of a student's midterm grade. At the 1% significance level, is a student's midterm grade significant in explaining a student's final grade? Show the relevant steps of the test.
  - FILE Property\_Taxes.** The accompanying data file shows the square footage and associated property taxes for 20 homes in an affluent suburb 30 miles outside of New York City. Estimate a home's property taxes as a linear function of its size. At the 5% significance level, do home size and property taxes have a linear relationship? Show the relevant steps of the test.
  - FILE Fertilizer.** A horticulturist is studying the relationship between tomato plant height and fertilizer amount. Thirty tomato plants grown in similar conditions were subjected to various amounts of fertilizer over a four-month period, and then their heights were measured.
    - Estimate the regression model:  

$$\text{Height} = \beta_0 + \beta_1 \text{Fertilizer} + \varepsilon.$$
    - At the 5% significance level, determine if a unit of fertilizer increases height by more than 3 units. Show the relevant steps of the test.
  - FILE Dexterity.** Finger dexterity, the ability to make precisely coordinated finger movements to grasp or assemble very small objects, is important in jewelry making. Thus, the manufacturing manager at Gemco, a manufacturer of high-quality watches, wants to develop a regression model to predict the productivity (in watches per shift) of new employees based on dexterity. He has subjected a sample of 20 current employees to the O'Connor dexterity test in which the time required to place 3 pins in each of 100 small holes using tweezers is measured.
    - Estimate the regression model:  $\text{Watches} = \beta_0 + \beta_1 \text{Time} + \varepsilon.$

- At the 5% significance level, determine if a unit of Time decreases Watches by more than 0.02. Show the relevant steps of the test.

- FILE Engine\_Overhaul.** The maintenance manager at a trucking company wants to build a regression model to forecast the time until the first engine overhaul based on four explanatory variables: (1) annual miles driven, (2) average load weight, (3) average driving speed, and (4) oil change interval. Based on driver logs and onboard computers, data have been obtained for a sample of 25 trucks.
  - Estimate the time until the first engine overhaul as a function of all four explanatory variables.
  - At the 10% significance level, are the explanatory variables jointly significant? Show the relevant steps of the test.
  - Are the explanatory variables individually significant at the 10% significance level? Show the relevant steps of the test.
- FILE Electricity\_Cost.** The facility manager at a pharmaceutical company wants to build a regression model to forecast monthly electricity cost. Three main variables are thought to dictate electricity cost: (1) average outdoor temperature, (2) working days per month, and (3) tons of product produced.
  - Estimate the regression model.
  - At the 10% significance level, are the explanatory variables jointly significant? Show the relevant steps of the test.
  - Are the explanatory variables individually significant at the 10% significance level? Show the relevant steps of the test.
- FILE Caterpillar.** Caterpillar, Inc. manufactures and sells heavy construction equipment worldwide. The performance of Caterpillar's stock is likely to be strongly influenced by the economy. For instance, during the subprime mortgage crisis, the value of Caterpillar's stock plunged dramatically. Monthly data for Caterpillar's risk-adjusted return and the risk-adjusted market return are collected for a five-year period ( $n = 60$ ). A portion of the data is shown in the accompanying table.

Date	$R - R_f$	$R_M - R_f$
1/1/2006	17.66	2.21
2/1/2006	7.27	-0.31
⋮	⋮	⋮
11/1/2010	3.37	2.15

Source: <http://finance.yahoo.com> and U.S. Treasury.

- Estimate the CAPM model for Caterpillar, Inc. Show the regression results in a well-formatted table.
- At the 5% significance level, determine if investment in Caterpillar is riskier than the market (beta significantly greater than 1).
- At the 5% significance level, is there evidence of abnormal returns?

22. **FILE** *Arlington\_Homes*. A realtor examines the factors that influence the price of a house in Arlington, Massachusetts. He collects data on recent house sales (Price) and notes each house's square footage (Sqft) as well as its number of bedrooms (Beds) and number of bathrooms (Baths). A portion of the data is shown in the accompanying table.

Price	Sqft	Beds	Baths
840000	2768	4	3.5
822000	2500	4	2.5
⋮	⋮	⋮	⋮
307500	850	1	1

- a. Estimate:  $\text{Price} = \beta_0 + \beta_1 \text{Sqft} + \beta_2 \text{Beds} + \beta_3 \text{Baths} + \varepsilon$ . Show the regression results in a well-formatted table.
- b. At the 5% significance level, are the explanatory variables jointly significant in explaining Price?
- c. At the 5% significance level, are all explanatory variables individually significant in explaining Price?
23. **FILE** *Final\_Test*. On the first day of class, an economics professor administers a test to gauge the

math preparedness of her students. She believes that the performance on this math test and the number of hours studied per week on the course are the primary factors that predict a student's score on the final exam. She collects data from 60 students, a portion of which is shown in the accompanying table.

Final	Math	Hours
94	92	5
74	90	3
⋮	⋮	⋮
63	64	2

- a. Estimate the sample regression equation that enables us to predict a student's final exam score on the basis of his/her math score and the number of hours studied per week.
- b. At the 5% significance level, are a student's math score and the number of hours studied per week jointly significant in explaining a student's final exam score?
- c. At the 5% significance level, is each explanatory variable individually significant in explaining a student's final exam score?

## 15.2 A GENERAL TEST OF LINEAR RESTRICTIONS

LO 15.3

The significance tests discussed in the preceding section can also be labeled as tests of linear restrictions. For example, the  $t$  test is a test of one restriction that determines whether or not a slope coefficient differs from zero. Similarly, the  $F$  test is a test of  $k$  restrictions that determines whether or not at least one of the slope coefficients is nonzero. In this section, we introduce a general test of linear restrictions; the resulting  $F$  test is often referred to as the **partial  $F$**  test. We can apply this test to any subset of the regression coefficients.

Consider a multiple regression model with three explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

As mentioned earlier, we use a  $t$  test for a test of one restriction,  $\beta_j = 0$ , and an  $F$  test for a test of three restrictions,  $\beta_1 = \beta_2 = \beta_3 = 0$ . What if we wanted to test if  $x_2$  and  $x_3$  are jointly significant? This is an example of a test of two restrictions,  $\beta_2 = \beta_3 = 0$ . Similarly, we may wish to test if the influence of  $x_3$  is identical to that of  $x_2$ . This would be a test of one restriction,  $\beta_2 = \beta_3$ . When conducting a partial  $F$  test, the null hypothesis implies that the restrictions are valid. In these two examples, the null hypothesis would be specified as  $H_0: \beta_2 = \beta_3 = 0$  and  $H_0: \beta_2 = \beta_3$ , respectively. As usual, the alternative hypothesis implies that the null hypothesis is not true. We conclude that the restrictions implied by the null hypothesis are not valid if we reject the null hypothesis.

In order to conduct the partial  $F$  test, we estimate the model with and without the restrictions. The **restricted model** is a reduced model where we do not estimate the coefficients that are restricted to a specific value under the null hypothesis. The **unrestricted model** is a complete model that imposes no restrictions on the coefficients; therefore, all coefficients are estimated. If the restrictions are valid—that is, the null hypothesis is true—then the sum of squares due to error of the restricted model  $SSE_R$  will not be significantly larger than the sum of squares due to error of the unrestricted model  $SSE_U$ . With the partial  $F$  test, we basically analyze the ratio of  $(SSE_R - SSE_U)$  to  $SSE_U$ . If this ratio, suitably adjusted for the degrees of freedom, is significantly large, then we reject the null hypothesis and conclude that the restrictions implied by the null hypothesis are not valid.

Conduct a general test of linear restrictions.



### TEST STATISTIC FOR THE TEST OF LINEAR RESTRICTIONS

When testing linear restrictions, the value of the test statistic is calculated as

$$F_{(df_1, df_2)} = \frac{(SSE_R - SSE_U)/df_1}{SSE_U/df_2},$$

where  $df_1$  is equal to the number of linear restrictions,  $df_2 = n - k - 1$  where  $k$  is the number of explanatory variables in the unrestricted model;  $SSE_R$  and  $SSE_U$  are the sum of squares due to error for the restricted and the unrestricted models, respectively.

We will consider two examples of the partial  $F$  test.

### EXAMPLE 15.5

A manager at a car wash company in Missouri wants to measure the effectiveness of price discounts and various types of advertisement expenditures on sales. For the analysis, he uses varying price discounts (Discount) and advertisement expenditures on radio (Radio) and newspapers (Newspaper) in 40 counties in Missouri. A portion of the monthly data on sales (in \$1,000s), price discounts (in percent), and advertisement expenditures (in \$1,000s) on radio and newspapers are shown in Table 15.8. At the 5% level, determine if the advertisement expenditures on radio and newspapers have a statistically significant influence on sales.

**TABLE 15.8** Sales, Price Discounts, and Advertising Expenditures,  $n = 40$

County	Sales (in \$1,000s)	Discount (in %)	Radio (in \$1,000s)	Newspaper (in \$1,000s)
1	62.72	40	2.27	3.00
2	49.65	20	3.78	1.78
⋮	⋮	⋮	⋮	⋮
40	49.95	40	3.57	1.57

**FILE**  
Car\_Wash

**SOLUTION:** Consider the unrestricted model (U) that does not impose restrictions on the coefficients and is specified as

$$(U) \quad \text{Sales} = \beta_0 + \beta_1 \text{Discount} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper} + \varepsilon.$$

A test that determines whether advertisement expenditures on radio and newspapers have a significant influence on sales is equivalent to a test that determines whether the Radio and Newspaper variables are jointly significant. We formulate the competing hypotheses as

$$H_0: \beta_2 = \beta_3 = 0$$

$H_A$ : At least one of the coefficients is nonzero.

In order to implement the partial  $F$  test, we then create the restricted (R) model. Note that we do not estimate the coefficients that are restricted to zero under the null hypothesis. Therefore, we exclude Radio and Newspaper and specify the model as

$$(R) \quad \text{Sales} = \beta_0 + \beta_1 \text{Discount} + \varepsilon.$$

For ease of exposition, we use the same notation to refer to the coefficients in models U and R. We note, however, that these coefficients and their estimates have a different meaning depending on which model we are referencing. Table 15.9 shows the relevant portion of the regression results.



**TABLE 15.9** Relevant Regression Output for Example 15.5

Variable	Restricted	Unrestricted
Intercept	43.4541* (0.0000)	6.7025 (0.3559)
Discount	0.4016* (0.0001)	0.3417* (0.0000)
Radio	NA	6.0624* (0.0007)
Newspaper	NA	9.3968* (0.0001)
SSE	2182.5649	1208.1348

NOTES: Parameter estimates are in the main body of the table with the  $p$ -values in parentheses;

\* represents significance at the 5% level. NA denotes not applicable. The last row presents the sum of squares due to error.

We will use the critical value approach at the 5% significance level ( $\alpha = 0.05$ ) to conduct the test. We use  $df_1 = 2$ , since we are testing for two restrictions,  $\beta_2 = 0$  and  $\beta_3 = 0$ , and  $df_2 = n - k - 1 = 40 - 3 - 1 = 36$ . Referencing the  $F$  table, we find the approximate critical value as  $F_{0.05, (2, 36)} = 3.26$ , or equivalently, we use Excel's function  $F.INV.RT(0.05, 2, 36)$  and obtain 3.26.

Taking the appropriate  $SSE$  values from Table 15.9, we calculate the value of the relevant test statistic as

$$F_{(2, 36)} = \frac{(SSE_R - SSE_U)/df_1}{SSE_U/df_2} = \frac{(2182.5649 - 1208.1348)/2}{1208.1348/36} = 14.52.$$

We reject  $H_0$  since 14.52 is greater than the critical value of 3.26. At the 5% level, we conclude that the advertisement expenditures on radio and newspapers have a significant influence on sales.

### EXAMPLE 15.6

In Example 15.5, we showed that the advertisement expenditures on radio and newspapers have a statistically significant influence on sales. The manager believes that the influence of the advertisement expenditure on radio differs from the influence of the advertisement expenditure on newspapers. Conduct the appropriate partial  $F$  test at the 5% level to verify the manager's belief.

**SOLUTION:** We again specify the unrestricted model (U) as

$$(U) \quad \text{Sales} = \beta_0 + \beta_1 \text{Discount} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper} + \varepsilon.$$

Because  $\beta_2$  and  $\beta_3$  capture the influence of the advertisement expenditures on radio and newspapers, respectively, we formulate the competing hypotheses as

$$H_0: \beta_2 = \beta_3$$

$$H_A: \beta_2 \neq \beta_3$$

In order to implement the partial  $F$  test, we then create the restricted (R) model. Note that under the restriction that  $\beta_2 = \beta_3$ , the unrestricted model simplifies to

$$\text{Sales} = \beta_0 + \beta_1 \text{Discount} + \beta_2 \text{Radio} + \beta_2 \text{Newspaper} + \varepsilon; \text{ that is,}$$

$$(R) \quad \text{Sales} = \beta_0 + \beta_1 \text{Discount} + \beta_2 (\text{Radio} + \text{Newspaper}) + \varepsilon.$$

Thus, the restricted model uses only two explanatory variables, where the new second explanatory variable is defined as the sum of Radio and Newspaper. For this regression, we have to first create a new explanatory variable by adding up the Radio and Newspaper values. Further, given the restriction,  $\beta_2 = \beta_3$ , the estimated coefficient for this new explanatory variable applies to both Radio and Newspaper.

**FILE**  
Car\_Wash

Note that the restricted model imposes one restriction, as there is one fewer coefficient to estimate. Table 15.10 presents the relevant portion of the regression results.

**TABLE 15.10** Relevant Regression Output for Example 15.6

Variable	Restricted	Unrestricted
Intercept	7.9524 (0.2740)	6.7025 (0.3559)
Discount	0.3517*(0.0000)	0.3417* (0.0000)
Radio	NA	6.0624* (0.0007)
Newspaper	NA	9.3968* (0.0001)
Radio + Newspaper	7.1831* (0.0000)	NA
SSE	1263.6243	1208.1348

NOTES: Parameter estimates are in the main body of the table with the  $p$ -values in parentheses; \* represents significance at the 5% level. NA denotes not applicable. The last row presents the sum of squares due to error.

We will use the  $p$ -value approach to conduct the test. We use  $df_1 = 1$  since we are testing for only one restriction, and  $df_2 = n - k - 1 = 40 - 3 - 1 = 36$ . Using the appropriate  $SSE$  values from Table 15.10, we calculate the value of the test statistic as

$$F_{(1,36)} = \frac{(SSE_R - SSE_U)/df_1}{SSE_U/df_2} = \frac{(1263.6243 - 1208.1348)/1}{1208.1348/36} = 1.65.$$

With  $df_1 = 1$  and  $df_2 = 36$ , we use Excel to compute the  $p$ -value as 0.2072 (F.DIST.RT(1.65, 1, 36) = 0.2072). We do not reject  $H_0$  since the  $p$ -value = 0.2072 > 0.05 =  $\alpha$ . At the 5% significance level, we cannot conclude that the influence of the advertisement expenditures on radio is different from the influence of the advertisement expenditures on newspapers.

## EXERCISES 15.2

### Mechanics

24. Consider the multiple linear regression model,  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ . You wish to test whether the slope coefficients  $\beta_1$  and  $\beta_3$  are jointly significant. Define the restricted and unrestricted models needed to conduct the test.
25. Consider the multiple linear regression model,  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ . You wish to test whether the slope coefficients  $\beta_1$  and  $\beta_3$  are statistically different from each other. Define the restricted and unrestricted models needed to conduct the test.
26. Consider the multiple linear regression model,  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ . Define the restricted and unrestricted models if the hypotheses are
 
$$H_0: \beta_1 + \beta_2 = 1 \quad \text{and} \quad H_A: \beta_1 + \beta_2 \neq 1.$$
27. Consider a portion of simple linear regression results,
 
$$\hat{y} = 105.40 + 39.17x_1; \quad SSE = 407,308; \quad n = 30$$
 In an attempt to improve the results, two explanatory variables are added. The relevant regression results are the following:

$$\hat{y} = 4.87 + 19.47x_1 - 26.31x_2 + 7.31x_3;$$

$$SSE = 344,784; \quad n = 30$$

- a. Formulate the hypotheses to determine whether  $x_2$  and  $x_3$  are jointly significant in explaining  $y$ .
- b. Calculate the value of the test statistic.
- c. At the 5% significance level, find the critical value(s).
- d. What is the conclusion to the test?

### Applications

28. A real estate analyst estimates the following regression, relating a house price to its square footage (Sqft):

$$\widehat{\text{Price}} = 48.39 + 52.74\text{Sqft}; \quad SSE = 56,944; \quad n = 50$$

In an attempt to improve the results, he adds two more explanatory variables: the number of bedrooms (Beds) and the number of bathrooms (Baths). The estimated regression equation is

$$\widehat{\text{Price}} = 28.11 + 40.17\text{Sqft} + 10.08\text{Beds} + 16.14\text{Baths}; \\ SSE = 48,074; \quad n = 50$$

- a. Formulate the hypotheses to determine whether Beds and Baths are jointly significant in explaining Price.

- b. Calculate the value of the test statistic.
- c. At the 5% significance level, find the critical value(s).
- d. What is the conclusion to the test?

29. A financial analyst believes that the best way to predict a firm's returns is by using the firm's price-to-earnings ratio (P/E) and its price-to sales ratio (P/S) as explanatory variables. He estimates the following regression, using 30 large firms:

$$\widehat{\text{Return}} = -33.40 + 3.97P/E - 3.37P/S;$$

$$SSE = 5,021.63; \quad n = 30$$

A colleague suggests that he can improve on his prediction if he also includes the P/E-to-growth ratio (PEG) and the dividend yield (DIV). He re-estimates the model by including these explanatory variables and obtains

$$\widehat{\text{Return}} = -31.84 + 4.26P/E - 2.16P/S - 11.49PEG + 3.82DIV; \quad SSE = 4,149.21; \quad n = 30$$

At the 5% significance level, is the colleague's claim substantiated by the data? Explain.

30. Lisa Fisher is a business school professor who would like to analyze university factors that enhance innovation. She collects data on 143 universities in 2008 for a regression where the response variable is the number of startups (Startups), which is used as a measure for innovation. Lisa believes that the amount of money that a university directs toward research (Research) is the most important factor influencing Startups. She estimates Startups as a function of Research and obtains

$$\widehat{\text{Startups}} = 0.21 + 0.01\text{Research}; \quad SSE = 1,434.78; \quad n = 143$$

Two other explanatory variables are also likely to influence Startups: the number of patents issued (Patents), and the age of its technology transfer office in years (Duration). Lisa then includes these additional variables in the model and obtains

$$\widehat{\text{Startups}} = 0.42 + 0.01\text{Research} + 0.05\text{Patents} - 0.02\text{Duration}; \quad SSE = 1,368.14; \quad n = 143$$

At the 5% significance level, should Lisa include Patents and Duration in the model predicting Startups?

31. **FILE Hourly\_Wage.** A researcher interviews 50 employees of a large manufacturer and collects data on each worker's hourly wage (Wage), years of higher education (EDUC), experience (EXPER), and age (AGE). A portion of the data is shown in the accompanying table.

Wage	EDUC	EXPER	AGE
\$37.85	11	2	40
21.72	4	1	39
⋮	⋮	⋮	⋮
24.18	8	11	64

- a. Estimate:
 
$$\text{Wage} = \beta_0 + \beta_1\text{EDUC} + \beta_2\text{EXPER} + \beta_3\text{AGE} + \varepsilon.$$

- b. The researcher wonders if the influence of experience is different from that of age, or if  $\beta_2 \neq \beta_3$ . Specify the competing hypotheses for this test.
- c. What is the restricted model given that the null hypothesis is true? Estimate this model.
- d. At the 5% significance level, can you conclude that the influence of experience is different from that of age?

32. **FILE Mobile\_Phones.** The manager of a local Costo store is in the process of making hiring decisions for selling mobile phone contracts. She believes that the sale of mobile phone contracts depends crucially on the number of hours clocked by male and female employees. She collects the weekly data on last year's sales of mobile phone contracts (Sale) along with work hours of male (Hours Males) and female (Hours Females) employees. A portion of the data is shown in the accompanying table.

Sale	Hours Males	Hours Females
59	30	32
65	33	36
⋮	⋮	⋮
64	35	35

- a. Report the sample regression equation of the appropriate model.
- b. At the 5% significance level, are the explanatory variables jointly significant? Are they individually significant? Use the  $p$ -value approach for the tests.
- c. The manager would like to determine whether there is a difference in productivity of male and female employees. In other words, for the same work hours, whether the number of sales of mobile contracts varies between male and female employees. Conduct the appropriate test at the 5% level of significance. Provide the details.

33. **FILE Football.** A multiple regression model is used to predict an NFL team's winning record (Win). For the explanatory variables, the average rushing yards (Rush) and the average passing yards (Pass) are used to capture offense and the average yards allowed are used to capture defense. A portion of the data for the 2009 NFL season is shown in the accompanying table.

Team	Win (%)	Rush	Pass	Yards Allowed
Arizona Cardinals	62.50	93.40	251.00	346.40
Atlanta Falcons	56.30	117.21	223.19	348.90
⋮	⋮	⋮	⋮	⋮
Washington Redskins	25.00	94.38	218.13	319.70

SOURCE: NFL website.

- a. Estimate the model:  $\text{Win} = \beta_0 + \beta_1\text{Rush} + \beta_2\text{Pass} + \beta_3\text{Yards Allowed} + \varepsilon.$
- b. Conduct a test at the 10% significance level to determine whether the impact of Rush is different from that of Pass in explaining Win, or  $\beta_1 \neq \beta_2$ . Provide the relevant steps.

34. **FILE Union\_Pay.** An automotive workers union, in conjunction with top management, is negotiating a new hourly pay policy for union workers based on three variables: (1) job class, (2) years with the company, and (3) years as a union member at any company. The goal is to develop an equitable model that can objectively specify hourly pay, thereby reducing pay disparity grievances. Fifty union workers have been sampled and will be used as the basis for the pay model. A portion of the data is shown in the accompanying table.

Hourly Pay (\$)	Job Class	Years with Company	Years in Union
15.90	24	12	7
23.70	52	17	14
⋮	⋮	⋮	⋮
26.70	43	2	2

- Report the sample regression equation of the appropriate model.
- At the 5% significance level, are the explanatory variables jointly significant? Are they individually significant? Use the  $p$ -value approach for the tests.
- Predict hourly pay for a worker in Job Class 48 with 18 years experience at the company and 14 years with the union.
- A manager wonders if the years with the company and the years as a union member matter in negotiating hourly pay. At the 5% significance level, can you conclude that the influence of these two explanatory variables is jointly significant? Provide the details.

## 15.3 INTERVAL ESTIMATES FOR THE RESPONSE VARIABLE

### LO 15.4

Calculate and interpret confidence intervals and prediction intervals.

In the introductory case, we analyzed the winning percentage of a baseball team on the basis of its batting average (BA) and earned run average (ERA). We estimated the most appropriate model to be  $\widehat{\text{Win}} = 0.1269 + 3.2754\text{BA} - 0.1153\text{ERA}$ . Suppose we want to predict a team's winning percentage given its batting average of 0.25 and earned run average of 4. We can use the above estimated model with  $\text{BA} = 0.25$  and  $\text{ERA} = 4$  to derive the predicted winning percentage as

$$\widehat{\text{Win}} = 0.1269 + 3.2754 \times 0.25 - 0.1153 \times 4 = 0.4846$$

Predictions, such as the one above, are certainly useful, but we need to be aware that such predictions are subject to sampling variations. In other words, the prediction will change if we use a different sample to estimate the regression model. Recall from Chapter 8 that the point estimate along with the margin of error is used to construct the relevant interval estimate. In the above example, 0.4846 represents the point estimate.

In this section, we will make a distinction between the interval estimate for the mean (expected value) of the response variable  $y$  and the interval estimate for the individual value of  $y$ . It is common to refer to the former as the **confidence interval** and the latter as the **prediction interval**. For given values of the explanatory variables, we can think of the confidence interval as the range that contains the mean of  $y$  and the prediction interval as the range that contains the individual value of  $y$ . We use the same point estimate for constructing both interval estimates. In the context of the above example, 0.4846 is the point estimate for the mean winning percentage as well as the individual winning percentage given the team's batting average of 0.25 and earned run average of 4. Due to the added uncertainty in predicting the individual value of  $y$ , the prediction interval is always wider than the corresponding confidence interval.

### CONFIDENCE INTERVAL AND PREDICTION INTERVAL

We construct two types of interval estimates regarding the response variable  $y$  for given values of the explanatory variables. The interval estimate for the mean (expected value) of  $y$  is referred to as the **confidence interval**. It is common to refer to the interval estimate for an individual value of  $y$  as the **prediction interval**. The prediction interval is always wider than the corresponding confidence interval.

We will now describe the general procedure for constructing the confidence interval and the prediction interval. For the sake of simplicity, we let  $x$ 's denote all explanatory variables. In the context of the preceding example,  $x_1$  and  $x_2$  represent the team's batting average and earned run average, respectively.

Consider a multiple regression model  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$  with  $k$  explanatory variables,  $x_1, x_2, \dots, x_k$ .

Moreover, let

$$y^0 = \beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \cdots + \beta_k x_k^0 + \varepsilon^0,$$

where  $x_1^0, x_2^0, \dots, x_k^0$  denote specific values for  $x_1, x_2, \dots, x_k$  at which  $y^0$  is evaluated and  $\varepsilon^0$  is the (unobserved) random error term. In the baseball example, we used  $x_1^0 = 0.25$  and  $x_2^0 = 4$ . Alternatively, we can evaluate the expected value of the response variable at  $x_1^0, x_2^0, \dots, x_k^0$  as

$$E(y^0) = \beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \cdots + \beta_k x_k^0.$$

The expected value equation uses the fact that the expected value of the random error term is assumed to be zero; that is,  $E(\varepsilon^0) = 0$ . We discuss this assumption in the next section. Note that the prediction interval is wider than the confidence interval because it also incorporates the additional uncertainty due to  $\varepsilon^0$ . We first derive a confidence interval for  $E(y^0)$ , followed by a prediction interval for  $y^0$ .

The predicted value,  $\hat{y}^0 = b_0 + b_1 x_1^0 + b_2 x_2^0 + \cdots + b_k x_k^0$ , is the point estimate for  $E(y^0)$ . In the baseball example, 0.4846 is the point estimate of  $E(y^0)$  when  $x_1^0 = 0.25$  and  $x_2^0 = 4$ . We form the  $100(1 - \alpha)\%$  confidence interval for  $E(y^0)$  as  $\hat{y}^0 \pm t_{\alpha/2, df} se(\hat{y}^0)$ , where  $se(\hat{y}^0)$  is the estimated standard error of  $\hat{y}^0$ . While there is a simple formula to compute the standard error  $se(\hat{y}^0)$  for a simple linear regression model, it is very cumbersome to do so for a multiple linear regression model. Next we describe a relatively easy way to construct a confidence interval that works for both simple and multiple linear regression models.

#### CONFIDENCE INTERVAL FOR THE EXPECTED VALUE OF $Y$

For specific values of  $x_1, x_2, \dots, x_k$ , denoted by  $x_1^0, x_2^0, \dots, x_k^0$ , the  $100(1 - \alpha)\%$  confidence interval for the expected value of  $y$  is computed as

$$\hat{y}^0 \pm t_{\alpha/2, df} se(\hat{y}^0),$$

where  $\hat{y}^0 = b_0 + b_1 x_1^0 + b_2 x_2^0 + \cdots + b_k x_k^0$ ,  $se(\hat{y}^0)$  is the standard error of  $\hat{y}^0$ , and  $df = n - k - 1$ .

To derive  $\hat{y}^0$  together with  $se(\hat{y}^0)$ , we first estimate a modified regression model where  $y$  is the response variable and the explanatory variables are defined as  $x_1^* = x_1 - x_1^0, x_2^* = x_2 - x_2^0, \dots, x_k^* = x_k - x_k^0$ . The resulting estimate of the intercept and its standard error equal  $\hat{y}^0$  and  $se(\hat{y}^0)$ , respectively.

#### EXAMPLE 15.7

We again reference the data from Table 15.1 and the regression model  $\text{Win} = \beta_0 + \beta_1 \text{BA} + \beta_2 \text{ERA} + \varepsilon$ . Construct the 95% confidence interval for expected winning percentage if BA is 0.25 and ERA is 4.

**SOLUTION:** Let  $y, x_1$ , and  $x_2$  denote Win, BA, and ERA, respectively. In order to construct a confidence interval for  $E(y^0)$ , we follow the above-mentioned procedure to derive  $\hat{y}^0$  as well as  $se(\hat{y}^0)$ . First, given  $x_1^0 = 0.25$  and  $x_2^0 = 4$ , we define two modified explanatory variables as  $x_1^* = x_1 - 0.25$  and  $x_2^* = x_2 - 4$ . Table 15.11 shows the computed values.

**TABLE 15.11** Computing the Values of Modified Explanatory Variables (Example 15.7)

$y$	$x_1$	$x_2$	$x_1^* = x_1 - 0.25$	$x_2^* = x_2 - 4$
0.407	0.259	4.59	$0.259 - 0.25 = 0.009$	$4.59 - 4 = 0.59$
0.549	0.268	4.20	$0.268 - 0.25 = 0.018$	$4.20 - 4 = 0.20$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0.426	0.250	4.13	$0.250 - 0.25 = 0.000$	$4.13 - 4 = 0.13$

The regression output for a multiple regression model that uses  $y$  as the response variable and  $x_1^*$  and  $x_2^*$  as the explanatory variables is presented in Table 15.12.

**TABLE 15.12** Regression Results with Modified Explanatory Variables (Example 15.7)

Regression Statistics						
Multiple R		0.8459				
R Square		0.7156				
Adjusted R Square		0.6945				
Standard Error		<b>0.0375</b>				
Observations		30				
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	0.0958	0.0479	33.9663	4.25E-08	
Residual	27	0.0381	0.0014			
Total	29	0.1338				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	<b>0.4847</b>	<b>0.0085</b>	57.2582	0.0000	<b>0.4673</b>	<b>0.5021</b>
$x_1^*$	3.2754	0.6723	4.8719	0.0000	1.8960	4.6549
$x_2^*$	-0.1153	0.0167	-6.9197	0.0000	-0.1494	-0.0811

We note that the modified regression output is identical to the original regression output (refer to the summarized results for Model 3 in Table 15.7) except for the estimates of the intercept term. The boldface intercept estimate is 0.4847 and its standard error is 0.0085. Therefore, we use  $\hat{y}^0 = 0.4847$  and  $se(\hat{y}^0) = 0.0085$  in constructing the confidence interval. Note that Excel's calculation for  $\hat{y}^0$  is the same as our earlier estimate,  $\hat{y}^0 = 0.1269 + 3.2754 \times 0.25 - 0.1153 \times 4 = 0.4846$ , except for rounding.

For the 95% confidence level and  $df = n - k - 1 = 30 - 2 - 1 = 27$ , we find  $t_{\alpha/2, df} = t_{0.025, 27} = 2.052$ . The 95% confidence interval for  $E(y^0)$  is

$$\hat{y}^0 \pm t_{\alpha/2, df} se(\hat{y}^0) = 0.4847 \pm 2.052 \times 0.0085 = 0.4847 \pm 0.0174.$$

Or, with 95% confidence,

$$0.4673 \leq E(y^0) \leq 0.5021.$$

Using this 95% confidence interval, we can state that the mean winning percentage of a team with a batting average of 0.25 and earned run average of 4 falls between 0.4673 and 0.5021. Note that these limits are also provided in the boldface Lower 95% and Upper 95% values of Table 15.12.

As mentioned earlier, the prediction interval pertains to the individual value of the response variable defined for specific explanatory variables as  $y^0 = \beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \cdots + \beta_k x_k^0 + \varepsilon^0$ . The prediction interval is wider than the confidence interval because it incorporates the variability of the random error term  $\varepsilon^0$ .



### PREDICTION INTERVAL FOR AN INDIVIDUAL VALUE OF $y$

For specific values of  $x_1, x_2, \dots, x_k$ , denoted by  $x_1^0, x_2^0, \dots, x_k^0$ , the  $100(1 - \alpha)\%$  prediction interval for an individual value of  $y$  is computed as

$$\hat{y}^0 \pm t_{\alpha/2, df} \sqrt{(se(\hat{y}^0))^2 + s_e^2},$$

where  $df = n - k - 1$ ,  $se(\hat{y}^0)$  is the standard error of  $\hat{y}^0$ , and  $s_e$  is the standard error of the estimate.

Note that the standard error of the estimate,  $s_e$ , captures the variability of the random error term,  $\epsilon^0$ .

### EXAMPLE 15.8

Reconsider the estimated model,  $\hat{y}^0 = 0.1269 + 3.2754x_1 - 0.1153x_2$ , where  $y$ ,  $x_1$ , and  $x_2$  denote Win, BA, and ERA, respectively.

- Construct the 95% prediction interval for Win if BA is 0.25 and ERA is 4.
- Comment on any differences between this prediction interval and the confidence interval constructed in Example 15.7.

#### SOLUTION:

- As in the calculation for the confidence interval, we compute  $\hat{y}^0 = 0.4847$ ,  $se(\hat{y}^0) = 0.0085$ , and  $t_{\alpha/2, df} = t_{0.025, 27} = 2.052$ . The only thing missing from the prediction interval formula is the standard error of the estimate  $s_e$ . From Table 15.12, we extract the boldface value,  $s_e = 0.0375$ . The 95% prediction interval is then

$$\begin{aligned} \hat{y}^0 \pm t_{\alpha/2, df} \sqrt{(se(\hat{y}^0))^2 + s_e^2} &= 0.4847 \pm 2.052 \sqrt{0.0085^2 + 0.0375^2} \\ &= 0.4847 \pm 0.0789. \end{aligned}$$

Or, with 95% confidence,

$$0.4058 \leq y^0 \leq 0.5636.$$

- Using this 95% prediction interval, we can state that the winning percentage of a team with a batting average of 0.25 and earned run average of 4 falls between 0.4058 and 0.5636. In the previous example, we used the 95% confidence interval to state that the mean winning percentage of a team with a batting average of 0.25 and earned run average of 4 falls between 0.4673 and 0.5021. As expected, the prediction interval is wider than the corresponding confidence interval. In forming the prediction interval, we also have to account for a very important source of variability caused by the random error term. This is captured by the standard error of the estimate,  $s_e$ , in the prediction interval formula. The higher variability makes it more difficult to predict accurately, thus necessitating a wider interval.

## EXERCISES 15.3

### Mechanics

- In a simple linear regression based on 30 observations, the following information is provided:  $\hat{y} = -6.92 + 1.35x$  and  $s_e = 2.78$ . Also,  $se(\hat{y}^0)$  evaluated at  $x = 30$  is 1.02.
  - Construct the 95% confidence interval for  $E(y)$  if  $x = 30$ .
  - Construct the 95% prediction interval for  $y$  if  $x = 30$ .
  - Which interval is narrower? Explain.
- In a multiple regression with 40 observations, the following sample regression equation is obtained:  $\hat{y} = 12.8 + 2.6x_1 - 1.2x_2$  with  $s_e = 5.84$ . Also, when  $x_1$  equals 15 and  $x_2$  equals 6,  $se(\hat{y}^0) = 2.20$ .
  - Construct the 95% confidence interval for  $E(y)$  if  $x_1$  equals 15 and  $x_2$  equals 6.
  - Construct the 95% prediction interval for  $y$  if  $x_1$  equals 15 and  $x_2$  equals 6.
  - Which interval is wider? Explain.

37. Consider the following sample data:

$x$	12	23	11	23	14	21	18	16
$y$	28	43	21	40	33	41	37	32

- Find the sample regression line,  $\hat{y} = b_0 + b_1x$ .
  - Construct the 95% confidence interval for  $E(y)$  if  $x = 15$ .
  - Construct the 95% prediction interval for  $y$  if  $x = 15$ .
38. Consider the following sample data:

$y$	46	51	28	55	29	53	47	36
$x_1$	40	48	29	44	30	58	60	29
$x_2$	13	28	24	11	28	28	29	14

- Find the sample regression equation,  $\hat{y} = b_0 + b_1x_1 + b_2x_2$ .
- Construct the 95% confidence interval for  $E(y)$  if  $x_1$  equals 50 and  $x_2$  equals 20.
- Construct the 95% prediction interval for  $y$  if  $x_1$  equals 50 and  $x_2$  equals 20.

## Applications

39. Using the data in the accompanying table, estimate the model:  $\text{Salary} = \beta_0 + \beta_1\text{Education} + \varepsilon$ , where Salary is measured in \$1,000s and Education refers to years of higher education.

Education	3	4	6	2	5	4	8	0
Salary	40	53	80	42	70	50	110	38

- Construct the 90% confidence interval for the expected salary for an individual who completed 6 years of higher education.
  - Construct the 90% prediction interval for salary for an individual who completed 6 years of higher education.
  - Comment on the difference in the widths of these intervals.
40. **FILE Fertilizer.** A horticulturist is studying the relationship between tomato plant height and fertilizer amount. He uses a sample of 30 to estimate the regression model as  $\text{Height} = \beta_0 + \beta_1\text{Fertilizer} + \varepsilon$ .
- Estimate the regression model to predict the height of a tomato plant that received 3.0 ounces of fertilizer.
  - Use the above prediction to calculate and interpret the 90% confidence interval for the mean height of tomato plants.
  - Calculate and interpret the corresponding 90% prediction interval for the height of an individual tomato plant.

41. With the data in the accompanying table, estimate  $\text{GPA} = \beta_0 + \beta_1\text{GRE} + \varepsilon$ , where GRE is a student's score on the math portion of the Graduate Record Examination score and GPA is the student's grade point average in graduate school.

GRE	700	720	650	750	680	730	740	780
GPA	3.0	3.5	3.2	3.7	3.1	3.9	3.3	3.5

- Construct the 90% confidence interval for the expected GPA for an individual who scored 710 on the math portion of the GRE.
  - Construct the 90% prediction interval for the GPA for an individual who scored 710 on the math portion of the GRE.
42. **FILE Debt\_Payments.** Estimate:  $\text{Debt} = \beta_0 + \beta_1\text{Income} + \varepsilon$ , where Debt is the average debt payments for a household in a particular city (in \$) and Income is the city's median income (in \$1,000s).
- Construct the 95% confidence interval for expected debt payments if income is \$80,000. (Note that income is measured in \$1,000s.)
  - Construct the 95% prediction interval for debt payments if income is \$80,000.
43. **FILE Arlington\_Homes.** Estimate:  $\text{Price} = \beta_0 + \beta_1\text{Sqft} + \beta_2\text{Beds} + \beta_3\text{Baths} + \varepsilon$ , where Price, Sqft, Beds, and Baths refer to home price, square footage, number of bedrooms, and number of bathrooms, respectively. Construct the 95% confidence interval for the expected price of a 2,500-square-foot home in Arlington, Massachusetts, with three bedrooms and two bathrooms. Construct the corresponding prediction interval for an individual home. Interpret both intervals.
44. **FILE Engine\_Overhaul.** The maintenance manager at a trucking company wants to build a regression model to forecast the time until the first engine overhaul based on four explanatory variables: (1) annual miles driven, (2) average load weight, (3) average driving speed, and (4) oil change interval. Based on driver logs and onboard computers, data have been obtained for a sample of 25 trucks.
- Estimate the regression model to predict the time before the first engine overhaul for a truck driven 60,000 miles per year with an average load of 22 tons, an average driving speed of 57 mph, and 18,000 miles between oil changes. (Note that both annual miles driven and oil change interval are measured in 1,000s.)
  - Use the above prediction to calculate and interpret the 90% confidence interval for the mean time before the first engine overhaul.
  - Calculate and interpret the corresponding 90% prediction interval for the time before the first engine overhaul.

## 15.4 MODEL ASSUMPTIONS AND COMMON VIOLATIONS

LO 15.5

So far we have focused on the estimation and the assessment of simple and multiple linear regression models. It is important to understand that the statistical properties of the OLS estimator, as well as the validity of the testing procedures, depend on the assumptions of the classical linear regression model. In this section, we discuss these assumptions. We also address common violations to the assumptions, discuss the consequences when the assumptions are violated, and, where possible, offer some remedies.

Explain the role of the assumptions on the OLS estimators.

### REQUIRED ASSUMPTIONS OF REGRESSION ANALYSIS

1. The regression model given by  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$  is *linear in the parameters*,  $\beta_0, \beta_1, \dots, \beta_k$ .
2. Conditional on  $x_1, x_2, \dots, x_k$ , the error term has an *expected value of zero*, or  $E(\varepsilon) = 0$ . This implies that  $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ .
3. There is no exact linear relationship among the explanatory variables; or in statistical terminology, there is *no perfect multicollinearity*.
4. Conditional on  $x_1, x_2, \dots, x_k$ , the variance of the error term  $\varepsilon$  is the same for all observations; or in statistical terminology, there is *no heteroskedasticity*. The assumption is violated if observations have a *changing variability*.
5. Conditional on  $x_1, x_2, \dots, x_k$ , the error term  $\varepsilon$  is uncorrelated across observations; or in statistical terminology, there is *no serial correlation*. The assumption is violated if *observations are correlated*.
6. The error term  $\varepsilon$  is not correlated with any of the explanatory variables  $x_1, x_2, \dots, x_k$ ; or in statistical terminology, there is *no endogeneity*. In general, this assumption is violated if important *explanatory variables are excluded*.
7. Conditional on  $x_1, x_2, \dots, x_k$ , the error term  $\varepsilon$  is *normally distributed*. This assumption allows us to construct interval estimates and conduct tests of significance. If  $\varepsilon$  is not normally distributed, the interval estimates and the hypothesis tests are valid only for large sample sizes.

Under the assumptions of the classical linear regression model, the OLS estimators have all desired properties. In particular, the OLS estimators of the regression coefficients  $\beta_j$  are unbiased; that is,  $E(b_j) = \beta_j$ . Moreover, among all linear unbiased estimators, they have minimum variations between samples. These desirable properties of the OLS estimators become compromised as one or more model assumptions are violated. Aside from coefficient estimates, the validity of the significance tests is also impacted by the assumptions. For certain violations, the estimated standard errors of the OLS estimators are inappropriate; in these cases it is not possible to make meaningful inferences from the  $t$  and the  $F$  test results.

The assumptions of the classical linear regression model are, for the most part, based on the error term  $\varepsilon$ . Since the residuals, or the observed error term,  $e = y - \hat{y}$ , contain useful information regarding  $\varepsilon$ , it is common to use the residuals to investigate the assumptions. In this section, we will rely on **residual plots** to detect some of the common violations to the assumptions. These graphical plots are easy to use and provide informal analysis of the estimated regression models. Formal tests are beyond the scope of this text.

### RESIDUAL PLOTS

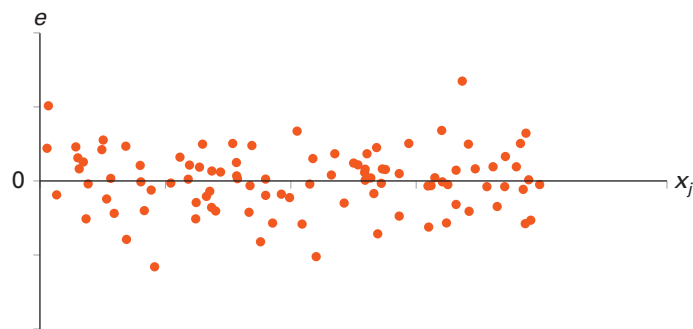
For the regression model,  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$ , the residuals are computed as  $e = y - \hat{y}$ , where  $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$ . These residuals can be plotted sequentially or against an explanatory variable  $x_j$  to look for model inadequacies.

It is common to plot the residuals  $e$  on the vertical axis and the explanatory variable  $x_j$  on the horizontal axis. Such plots are useful for detecting deviations from linearity as well as constant variability. If the regression is based on time series data, we can plot the residuals sequentially to detect if the observations are correlated.

Residual plots can also be used to detect outliers. Recall that outliers are observations that stand out from the rest of the data. For an outlier observation, the resulting residual will appear distinct in a plot; it will stand out from the rest. While outliers can greatly impact the estimates, it is not always clear what to do with them. As mentioned in Chapter 3, outliers may indicate bad data due to incorrectly recorded (or included) observations in the data set. In such cases, the relevant observation should be corrected or simply deleted. Alternatively, outliers may just be due to random variations, in which case the relevant observations should remain. In any event, residual plots help us identify potential outliers so that we can take corrective actions, if needed.

In Figure 15.1, we present a hypothetical residual plot when none of the assumptions has been violated. (Excel computes the residuals and also plots them against all explanatory variables. After choosing **Data > Data Analysis > Regression**, we select *Residuals* and *Residual Plots* in the *Regression* dialog box.)

**FIGURE 15.1**  
Residual plot of a correctly  
specified model



Note that all the points are randomly dispersed around the zero value of the residuals. Also, there is no evidence of outliers since no residual stands out from the rest. As we will see next, any discernible pattern of the residuals indicates that one or more assumptions have been violated.

#### LO 15.6

Describe common violations of the assumptions and offer remedies.

### Common Violation 1: Nonlinear Patterns

Linear regression models are often justified on the basis of their computational simplicity. A simple linear regression model  $y = \beta_0 + \beta_1 x + \varepsilon$  implies that if  $x$  goes up by one unit, we expect  $y$  to change by  $\beta_1$ , irrespective of the value of  $x$ . However, in many applications, the relationship cannot be represented by a straight line and, therefore, must be captured by an appropriate curve. It is always good to rely on economic theory to determine if the linearity assumption is appropriate. We confirm our intuition by analyzing scatterplots or residual plots. The OLS estimates can be quite misleading if there are obvious nonlinear patterns in the data.

#### Detection

We can use residual plots to identify nonlinear patterns. Linearity is justified if the residuals are randomly dispersed across the values of an explanatory variable. A discernible trend in the residuals is indicative of nonlinear patterns.

#### EXAMPLE 15.9

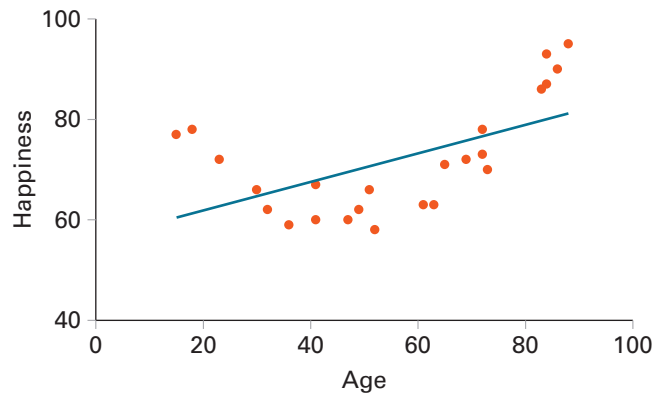
A sociologist wishes to study the relationship between age and happiness. He interviews 24 individuals and collects data on age and happiness, measured on a scale from 0 to 100. A portion of the data is shown in Table 15.13. Examine the linearity assumption in the regression model,  $\text{Happiness} = \beta_0 + \beta_1 \text{Age} + \varepsilon$ .

**TABLE 15.13** Happiness and Age

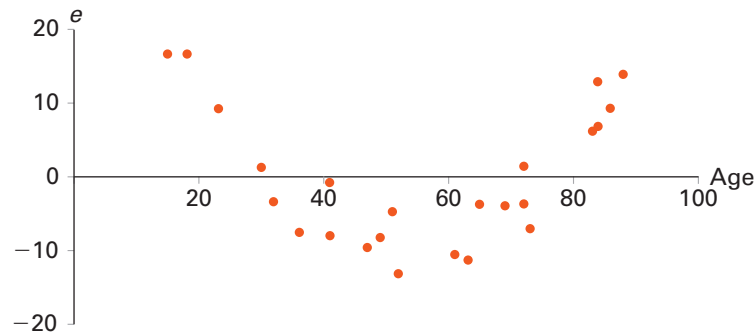
Age	Happiness
49	62
51	66
⋮	⋮
69	72

**FILE**  
Happiness\_Age

**SOLUTION:** We start the analysis with a scatterplot of Happiness against Age. Figure 15.2 shows the scatterplot and the superimposed trend line, which is based on the sample regression equation,  $\widehat{\text{Happiness}} = 56.18 + 0.28\text{Age}$ . It is fairly clear from Figure 15.2 that the linear regression model does not appropriately capture the relationship between Happiness and Age. In other words, the prediction that, every year, happiness of a person increases by 0.28 units is misleading.

**FIGURE 15.2** Scatterplot and the superimposed trendline (Example 15.9)

A residual plot, shown in Figure 15.3, further explores the linearity assumption of the regression model.

**FIGURE 15.3** Residual plot against Age (Example 15.9)

The above residual plot shows that there is an obvious trend with the residuals decreasing until the age of 50 and steadily increasing thereafter. The linear regression model is inappropriate as it underestimates at lower and higher age levels and overestimates in the middle. This result is consistent with a report that shows that happiness initially decreases with age and then increases with age (*The Economist*, December 16, 2010).

### Remedy

Linear regression models are often used as a first pass for most empirical work. In many instances, they provide a very good approximation for the actual relationship. However, if residual plots exhibit strong nonlinear patterns, the inferences made by a linear regression model can be quite misleading. In such instances, we should employ nonlinear regression methods based on simple transformations of the response and the explanatory variables; these methods are discussed in the next chapter.

## Common Violation 2: Multicollinearity

Perfect multicollinearity exists when two or more explanatory variables have an exact linear relationship. Consider the model  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ , where  $y$  is bonus,  $x_1$  is the number of cars sold, and  $x_2$  is the number of cars remaining in the lot. If all car salesmen started with the same inventory, we have a case of *perfect* multicollinearity ( $x_2 = \text{Constant} - x_1$ ) and the model cannot be estimated. However, if  $x_2$  represents the proportion of positive reviews from customers, we have *some* multicollinearity since the number of cars sold and the proportion of positive reviews are likely to be correlated. In most applications, some degree of correlation exists between the explanatory variables.

Multicollinearity makes it difficult to disentangle the separate influences of the explanatory variables on the response variable. If multicollinearity is severe, we may find insignificance of important explanatory variables; some coefficient estimates may even have wrong signs.

### Detection

The detection methods for multicollinearity are mostly informal. The presence of a high  $R^2$  coupled with individually insignificant explanatory variables can indicate multicollinearity. Sometimes researchers examine the correlations between the explanatory variables to detect severe multicollinearity. One such guideline suggests that multicollinearity is severe if the sample correlation coefficient between any two explanatory variables is more than 0.80 or less than  $-0.80$ . Seemingly wrong signs of the estimated regression coefficients may also indicate multicollinearity.

### EXAMPLE 15.10

Examine the multicollinearity issue in a linear regression model that uses median home values as the response variable and median household incomes, per capita incomes, and the proportion of owner-occupied homes as the explanatory variables. A portion of 2010 data for all states in the United States is shown in Table 15.14.

**TABLE 15.14** Home Values and Other Factors

State	Home Value	HH Income	Per Cap Inc	Pct Owner Occ
Alabama	\$117,600	\$42,081	\$22,984	71.1%
Alaska	229,100	66,521	30,726	64.7
⋮	⋮	⋮	⋮	⋮
Wyoming	174,000	53,802	27,860	70.2

SOURCE: 2010 U.S. Census.

**SOLUTION:** We estimate three models to examine the multicollinearity issue; Table 15.15 presents the regression results.

**TABLE 15.15** Summary of Model Estimates (Example 15.10)

Variable	Model 1	Model 2	Model 3
Intercept	417892.04* (0.00)	348187.14* (0.00)	285604.08 (0.08)
HH Income	9.04* (0.00)	7.74* (0.00)	NA
Per Cap Inc	-3.27 (0.31)	NA	13.21* (0.00)
Pct Owner Occ	-8744.30* (0.00)	-8027.90* (0.00)	-6454.08* (0.36)
Adjusted $R^2$	0.8071	0.8069	0.6621

NOTES: The table contains parameter estimates with  $p$ -values in parentheses; \* represents significance at the 5% level. NA denotes not applicable. Adjusted  $R^2$ , reported in the last row, is used for model selection.

**FILE**  
Home\_Values



Model 1 uses all three explanatory variables. Surprisingly, the per capita income variable has a negative estimated coefficient of  $-3.27$  and, with a  $p$ -value of  $0.31$ , is not even statistically significant at the  $5\%$  level. Multicollinearity might be the reason for this surprising result since household income and per capita income are likely to be correlated. We compute the sample correlation coefficient between these two variables as  $0.8582$ , which suggests that multicollinearity is severe. We estimate two more models where one of these collinear variables is removed; Model 2 removes per capita income and Model 3 removes household income. Note that per capita income in Model 3 now exerts a positive and significant influence on home values. Between these two models, Model 2 is preferred to Model 3 because of its higher adjusted  $R^2$  ( $0.8069 > 0.6621$ ). The choice between Model 1 and Model 2 is a little unclear. In general, Model 1, with the highest adjusted  $R^2$  value of  $0.8071$ , is preferred. This is especially so if the purpose of the analysis is to make predictions. However, if the coefficient estimates need to be evaluated, then Model 2 may be the preferred choice.

## Remedy

Inexperienced researchers tend to include too many explanatory variables in their quest not to omit anything important and in doing so may include redundant variables that essentially measure the same thing. When confronted with multicollinearity, a good remedy is to drop one of the collinear variables if we can justify its redundancy. Another option is to obtain more data, since the sample correlation may get weaker as we include more observations. Sometimes it helps to express the explanatory variables differently so that they are not collinear. At times, the best approach may be to *do nothing* when there is a justification to include all explanatory variables. This is especially so if the estimated model yields a high  $R^2$ , which implies that the estimated model is good for prediction as is.

## Common Violation 3: Changing Variability

The assumption of constant variability of observations often breaks down in studies with cross-sectional data. Consider the model  $y = \beta_0 + \beta_1 x + \varepsilon$ , where  $y$  is a household's consumption expenditure and  $x$  is its disposable income. It may be unreasonable to assume that the variability of consumption is the same across a cross-section of household incomes. For example, we would expect higher-income households to have a higher variability in consumption as compared to lower-income households. Similarly, home prices tend to vary more as homes get larger and sales tend to vary more as firm size increases.

In the presence of changing variability, the OLS estimators are still unbiased. However, the estimated standard errors of the OLS estimators are inappropriate. Consequently, we cannot put much faith in the standard  $t$  or  $F$  tests since they are based on these estimated standard errors.

## Detection

We can use residual plots to gauge changing variability. The residuals are generally plotted against each explanatory variable  $x_j$ ; for a multiple regression model, we can also plot them against the predicted value  $\hat{y}$ . There is no violation if the residuals are randomly dispersed across the values of  $x_j$ . On the other hand, there is a violation if the variability increases or decreases over the values of  $x_j$ .

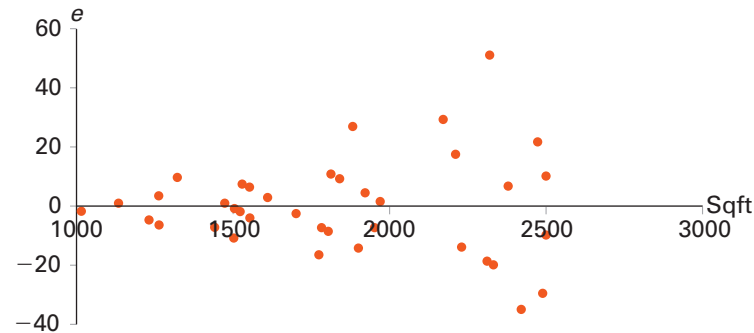
### EXAMPLE 15.11

Consider a simple regression model that relates monthly sales (Sales) from a chain of convenience stores with the square footage (Sqft) of the store. A portion of the data used for the analysis is shown in Table 15.16. Estimate the model and use a residual plot to determine if the observations have a changing variability.

**TABLE 15.16** Sales and Square Footage of Convenience Stores

Sales (in \$1,000s)	Sqft
140	1810
160	2500
⋮	⋮
110	1470

**SOLUTION:** The sample regression is given by  $\widehat{\text{Sales}} = 22.08 + 0.06\text{Sqft}$ . A residual plot of the estimated model is shown in Figure 15.4.

**FIGURE 15.4** Residual plot against square footage (Example 15.11)

Note that the residuals seem to fan out across the horizontal axis. Therefore, we conclude that changing variability is a likely problem in this application relating sales to square footage. This result is not surprising, since you would expect sales to vary more as square footage increases. For instance, a small convenience store is likely to include only bare essentials for which there is a fairly stable demand. A larger store, on the other hand, may include specialty items, resulting in more fluctuation in sales.

## Remedy

As mentioned earlier, in the presence of changing variability, the OLS estimators are unbiased but their estimated standard errors are inappropriate. Therefore, OLS still provides reasonable coefficient estimates, but the  $t$  and the  $F$  tests are no longer valid. This has prompted some researchers to use the OLS estimates along with a correction for the standard errors, called White's correction. Many statistical computer packages routinely make this correction, thus enabling researchers to perform legitimate  $t$  and  $F$  tests. Unfortunately, the current version of Excel does not have the ability to make this correction.

## Common Violation 4: Correlated Observations

When obtaining the OLS estimators, we assume that the observations are uncorrelated. This assumption often breaks down in studies with time series data. Variables such as GDP, employment, and asset returns exhibit business cycles. As a consequence, successive observations are likely to be correlated.

In the presence of correlated observations, the OLS estimators are unbiased, but their estimated standard errors are inappropriate. Generally, these standard errors are distorted downward, making the model look better than it really is with a spuriously high  $R^2$ . Furthermore, the  $t$  and  $F$  tests may suggest that the explanatory variables are individually and jointly significant when this is not true.

## Detection

We can plot the residuals sequentially over time to look for correlated observations. If there is no violation, then the residuals should show no pattern around the horizontal axis. A violation is indicated when a positive residual in one period is followed by positive residuals in the next few periods, followed by negative residuals for a few periods, then positive residuals, and so on. Although not as common, a violation is also indicated when a positive residual is followed by a negative residual, then a positive residual, and so on.

### EXAMPLE 15.12

Consider  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$  where  $y$  represents sales at a sushi restaurant and  $x_1$  and  $x_2$  represent advertising costs and the unemployment rate, respectively. A portion of monthly data from January 2008 to June 2009 is given in Table 15.17. Inspect the behavior of the residuals in order to comment on serial correlation.

**TABLE 15.17** Sales, Advertising Costs, and Unemployment Data for Example 15.12

Month	Year	Sales (in \$1,000s)	Advertising Costs (in \$)	Unemployment Rate (in percent)
January	2008	27.0	550	4.6
February	2008	24.2	425	4.3
⋮	⋮	⋮	⋮	⋮
May	2009	27.4	550	9.1

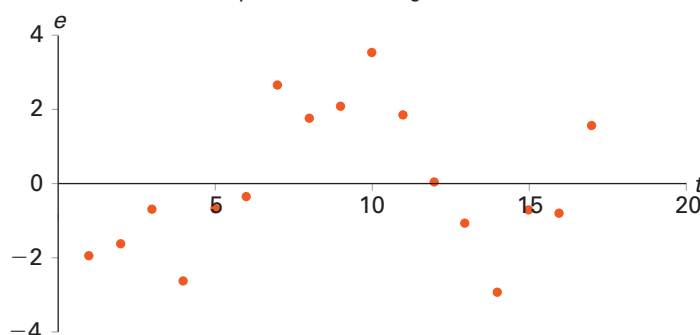
SOURCE FOR THE UNEMPLOYMENT RATE DATA: Development Department, State of California, June 2009.

**FILE**  
Sushi\_Restaurant

**SOLUTION:** The model is estimated as  $\hat{y} = 17.5060 + 0.0266x_1 - 0.6879x_2$ . In order to detect serial correlation, we plot the residuals sequentially against time  $t$ , where  $t$  is given by 1, 2, . . . , 17 for the 17 months of time series data. (In order to construct this residual plot with Excel, we first estimate the model and choose *Residuals* from Excel's *Regression* dialog box. Given the regression output, we select the residual data and choose **Insert** > **Scatter**; choose the option on the top left.)

Figure 15.5 shows a wavelike movement in the residuals over time, first clustering below the horizontal axis, then above the horizontal axis, etc. Given this pattern around the horizontal axis, we conclude that the observations are correlated.

**FIGURE 15.5** Scatterplot of residuals against time  $t$



## Remedy

As mentioned earlier, in the presence of correlated observations, the OLS estimators are unbiased but their standard errors are inappropriate and generally distorted downward, making the model look better than it really is. Therefore, OLS still provides reasonable coefficient estimates, but the  $t$  and the  $F$  tests are no longer valid. This has prompted some researchers to use the OLS estimates but correct the standard errors using the Newey-West procedure. As in the case of changing variability, many statistical computer packages have the capacity to make this correction; unfortunately, the current version of Excel does not have this capability. We can perform legitimate  $t$  and  $F$  tests once the standard errors have been corrected.

## Common Violation 5: Excluded Variables

Another crucial assumption in a linear regression model is that the error term is not correlated with the explanatory variables. In general, this assumption breaks down when important explanatory variables are excluded. If one or more of the relevant explanatory variables are excluded, then the resulting OLS estimators are biased. The extent of the bias depends on the degree of the correlation between the included and the excluded explanatory variables.

Suppose we want to estimate  $y = \beta_0 + \beta_1 x + \varepsilon$ , where  $y$  is salary and  $x$  is years of education. This model excludes innate ability, which is an important ingredient for salary. Since ability is omitted, it gets incorporated in the error term and the resulting error term is likely to be correlated with years of education. Now consider someone who is highly educated and also commands a high salary. The model will associate high salary with education, when, in fact, it may be the person's unobserved high level of innate ability that has raised both education and salary. In sum, this violation leads to unreliable coefficient estimates; some estimates may even have the wrong signs.

## Remedy

It is important that we include all relevant explanatory variables in the regression model. An important first step before running a regression model is to compile a comprehensive list of potential explanatory variables. We can then build down to perhaps a smaller list of explanatory variables using the adjusted  $R^2$  criterion. Sometimes due to data limitations, we are unable to include all relevant variables. For example, innate ability may be an important explanatory variable for a model that explains salary, but we are unable to include it since innate ability is not observable. In such instances, we use a technique called the instrumental variable technique, which is outside the scope of this text.

## Summary

Regression models are an integral part of business statistics. It takes practice to become an effective user of the regression methodology. We should think of regression modeling as an iterative process. We start with a clear understanding of what the regression model is supposed to do. We define the relevant response variable and compile a comprehensive list of potential explanatory variables. The emphasis should be to pick a model that makes economic and intuitive sense and avoid explanatory variables that more or less measure the same thing, thus causing multicollinearity. We then apply this model to data and refine and improve its fit. Specifically, from the comprehensive list, we build down to perhaps a smaller list of explanatory variables using significance tests and goodness-of-fit measures such as the standard error of the estimate and the adjusted  $R^2$ . It is important that we explore residual plots to look for signs of changing variability and correlated observations in cross-sectional and time series studies, respectively. If we identify any of these two violations, we can still trust the point estimates of the regression coefficients. However, we cannot place much faith in the standard  $t$  or  $F$  tests of significance unless we employ the necessary correction.

## EXERCISES 15.4

### Mechanics

45. Using 20 observations, the multiple regression model  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$  was estimated. Excel produced the following relevant results.

	df	SS	MS	F	Significance F	
Regression	2	2.12E+12	1.06E+12	56.5561	3.07E-08	
Residual	17	3.19E+11	1.88E+10			
Total	19	2.44E+12				
	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	-987557	131583	-7.5052	0.0000	-1265173	-709941
$x_1$	29233	32653	0.8952	0.3832	-39660	98125
$x_2$	30283	32645	0.9276	0.3666	-38592	99158

- a. At the 5% significance level, are the explanatory variables jointly significant?
- b. At the 5% significance level, is each explanatory variable individually significant?
- c. What is the likely problem with this model?
46. A simple linear regression,  $y = \beta_0 + \beta_1 x + \varepsilon$ , is estimated with cross-sectional data. The resulting residuals  $e$  along with the values of the explanatory variable  $x$  are shown in the accompanying table.

$x$	1	2	5	7	10	14	15	20	24	30
$e$	-2	1	-3	2	4	-5	-6	8	11	-10

- a. Graph the residuals  $e$  against the values of the explanatory variable  $x$  and look for any discernible pattern.
- b. Which assumption is being violated? Discuss its consequences and suggest a possible remedy.
47. A simple linear regression,  $y = \beta_0 + \beta_1 x + \varepsilon$ , is estimated with time series data. The resulting residuals  $e$  and the time variable  $t$  are shown in the accompanying table.

$t$	1	2	3	4	5	6	7	8	9	10
$e$	-5	-4	-2	3	6	8	4	-5	-3	-2

- a. Graph the residuals against time and look for any discernible pattern.
- b. Which assumption is being violated? Discuss its consequences and suggest a possible remedy.

### Applications

48. **FILE Television.** Numerous studies have shown that watching too much television hurts school grades. Others have argued that television is not necessarily a bad thing for children (*Mail Online*, July 18, 2009). Like books and stories, television not only entertains, it also exposes a child to new information about the world. While watching too much television is harmful, a little bit may actually help. Researcher Matt Castle gathers information on the grade point average (GPA) of 28 middle-school children

and the number of hours of television they watched per week. Examine the linearity assumption in the regression model,  $\text{GPA} = \beta_0 + \beta_1 \text{Hours} + \varepsilon$ .

49. **FILE Work Experience.** Consider the data on salary (in \$) and work experience (in years) of 100 employees in a marketing firm. Estimate the model  $y = \beta_0 + \beta_1 x + \varepsilon$ , where  $y$  is salary and  $x$  is work experience.
- a. Explain why you would be concerned about changing variability in this application.
- b. Use a residual plot to confirm your economic intuition.
50. Consider the results of a survey where students were asked about their GPA and also to break down their typical 24-hour day into study, leisure (including work), and sleep. Consider the model  $\text{GPA} = \beta_0 + \beta_1 \text{Study} + \beta_2 \text{Leisure} + \beta_3 \text{Sleep} + \varepsilon$ .
- a. What is wrong with this model?
- b. Suggest a simple way to reformulate the model.
51. **FILE Ann Arbor Rental.** Consider the monthly rent (Rent) of a home in Ann Arbor, Michigan, as a function of the number of bedrooms (Beds), the number of bathrooms (Baths), and square footage (Sqft).
- a. Estimate:  $\text{Rent} = \beta_0 + \beta_1 \text{Beds} + \beta_2 \text{Baths} + \beta_3 \text{Sqft} + \varepsilon$ .
- b. Which of the explanatory variables might cause changing variability? Explain.
- c. Use residual plots to verify your economic intuition.
52. **FILE Delivery.** Quick2U, a delivery company, would like to standardize its delivery charge model for shipments such that customers will better understand their delivery costs. Three variables are used: (1) distance (one-way), (2) shipment weight, and (3) number of boxes. A sample of 30 recent deliveries is collected; a portion of the data is shown in the accompanying table.

Charge (\$)	Distance (miles)	Weight (lbs)	Boxes (units)
92.50	29	183	1
157.60	96	135	3
⋮	⋮	⋮	⋮
143.00	47	117	7

- a. Estimate the model  $\text{Charge} = \beta_0 + \beta_1 \text{Distance} + \beta_2 \text{Weight} + \beta_3 \text{Boxes} + \varepsilon$  and examine the joint and individual significance of the explanatory variables at the 1% level.
- b. Is there any evidence of multicollinearity?
- c. Examine the residual plots for evidence of heteroskedasticity.
53. **FILE Healthy Living.** Healthy living has always been an important goal for any society. In a recent ad campaign for Walt Disney, First Lady Michelle Obama shows parents and children that eating well and exercising can also be fun (*USA Today*, September 30, 2010). Consider a regression

model that conjectures that fruits and vegetables and regular exercising have a positive effect on health and smoking has a negative effect on health. The sample consists of the percentage of these variables observed in various states in the United States in 2009. A portion of the data is shown in the accompanying table.

State	Healthy (%)	Fruits/Vegetables (%)	Exercise (%)	Smoke (%)
AK	88.7	23.3	60.6	14.6
AL	78.3	20.3	41	16.4
⋮	⋮	⋮	⋮	⋮
WY	87.5	23.3	57.2	15.2

SOURCE: Centers for Disease Control and Prevention.

- Estimate the model  $\text{Healthy} = \beta_0 + \beta_1 \text{Fruits/Vegetables} + \beta_2 \text{Exercise} + \beta_3 \text{Smoke} + \varepsilon$ .
  - Analyze the data to determine if multicollinearity and changing variability are present.
54. **FILE Johnson Johnson.** A capital asset pricing model (CAPM) for Johnson & Johnson (J&J) was discussed in Example 15.3. The model uses the risk-adjusted stock return  $R - R_f$  for J&J as the response variable and the risk-adjusted market return  $R_M - R_f$  as the explanatory variable. Since serial correlation may occur with time series data, it is prudent to inspect the behavior of the residuals. Construct a scatterplot of the residuals against time to comment on correlated observations.
55. **FILE Consumption Quarterly.** The consumption function is one of the key relationships in economics, where consumption  $y$  depends on disposable income  $x$ . Consider the quarterly data for these seasonally adjusted variables, measured in billions of dollars. A portion of the data is shown in the accompanying table.

Date	Consumption (\$ billions)	Disposable Income (\$ billions)
2006:01	9148.2	9705.2
2006:02	9266.6	9863.8
⋮	⋮	⋮
2010:04	10525.2	11514.7

SOURCE: U.S. Department of Commerce.

- Estimate  $\text{Consumption} = \beta_0 + \beta_1 \text{Disposable Income} + \varepsilon$ . Plot the residuals against time to determine if there is a possibility of correlated observations.
  - Discuss the consequences of correlated observations and suggest a possible remedy.
56. **FILE Mowers.** The marketing manager at Turfco, a lawn mower company, believes that monthly sales across all outlets (stores, online, etc.) are influenced by three key variables: (1) outdoor temperature, (2) advertising expenditures, and (3) promotional discounts. A portion of the monthly sales data for the past two years is shown in the accompanying table.

Sales (units)	Temperature (°F)	Advertising (in \$1,000s)	Discount (in percent)
17,235	33	15	5.0
19,854	42	25	5.0
⋮	⋮	⋮	⋮
22,571	44	21	5.0

- Estimate the model  $\text{Sales} = \beta_0 + \beta_1 \text{Temperature} + \beta_2 \text{Advertising} + \beta_3 \text{Discount} + \varepsilon$  and test for the joint and individual significance of the explanatory variables at the 5% level.
- Examine the data for evidence of multicollinearity. Provide two reasons why it might be best to do nothing about multicollinearity in this application.
- Examine the residual plots for evidence of heteroskedasticity.

## WRITING WITH STATISTICS



Ben Leach is a statistician for a Major League Baseball (MLB) team. One aspect of his job is to assess the value of various players. At the moment, Ben's team is in dire need of an outfielder. Management is ready to make an offer to a certain prospect but asks Ben for some input concerning salary. Management believes that a player's batting average (BA), runs batted in (RBI), and years of experience playing professional baseball (Experience) are the most important factors that influence a player's salary. Management is focusing on a player who has played professional baseball for seven years and whose average BA and RBI over this time have been 266 and 50, respectively. Ben collects data on salary (in \$1,000s), BA, RBI, and Experience for 138 outfielders in 2008. Table 15.18 shows a portion of the data.



**TABLE 15.18** Major League Baseball Outfielder Data,  $n = 138$ 

Player	Salary (in \$1,000s)	BA	RBI	Experience
1. Nick Markakis	455	299	87	3
2. Adam Jones	390	261	23	3
⋮	⋮	⋮	⋮	⋮
138. Randy Winn	8,875	288	53	11

NOTES: All data collected from usatoday.com or http://espn.com; BA and RBI are averages over the player's professional life through 2008. For exposition, BA has been multiplied by 1000.

**FILE**  
MLB\_Salary

Ben would like to use information in Table 15.18 to:

1. Summarize Salaries, BAs, RBIs, and Experience of current outfielders. Examine the potential multicollinearity problem.
2. Address management's claim that BA, RBI, and Experience have a statistically significant influence on salary.
3. Evaluate the expected salary for the prospective player, given his values for BA, RBI, and Experience.

In an attempt to assess the factors that influence an outfielder's salary in Major League Baseball (MLB), data were collected from 138 current players. Management believes that an outfielder's salary is best predicted using the outfielder's overall batting average (BA), runs batted in (RBI), and years of experience (Experience) as an MLB player. Table 15.A provides some descriptive statistics on these relevant variables.

**TABLE 15.A** Descriptive Statistics on Salary, BA, RBI, and Experience,  $n = 138$ 

	Salary (in \$1,000s)	BA	RBI	Experience
Mean	3,459	271	43	6
Minimum	390	152	1	1
Maximum	18,623	331	102	20

The average salary of an MLB outfielder in 2008 is a staggering \$3,459,000; however, the minimum salary of \$390,000 and the maximum salary of \$18,623,000 suggest quite a bit of variability in salary. The average outfielder has a BA of 271 with 43 RBIs in a season. Experience of outfielders in 2008 varied from only 1 year to 20 years, with an average of 6 years.

Table 15.B provides regression results from estimating a model where BA, RBI, and Experience are the explanatory variables and Salary is the response variable. All sample correlation coefficients (not reported), between explanatory variables, are less than 0.50, indicating that multicollinearity is not a serious problem in this application.

**TABLE 15.B** Analysis of Salary of Baseball Players

Variable	Coefficient
Intercept	-4769.40 (0.1301)
BA	4.76 (0.6984)
RBI	80.44* (0.0000)
Experience	539.67* (0.0000)
$R^2 = 0.58$	
$F_{(3,133)} = 61.54$ (associated $p$ -value = 0.0000)	

NOTES:  $p$ -values are in parentheses; \* denotes significance at the 5% level.

## Sample Report— Baseball Salaries

The slope coefficients suggest that BA, RBI, and Experience exert a positive influence on Salary. For instance, the slope coefficient of Experience indicates that if an outfielder stays in the major leagues for one additional year, then, on average, his salary will increase by \$539,670, holding BA and RBI constant. The  $p$ -value associated with the value of the  $F_{(3,133)}$  test statistic shows that the explanatory variables are jointly significant at the 5% level. Upon testing the explanatory variables individually, the extremely small  $p$ -values associated with RBI and Experience reveal that these variables have a significant linear relationship with Salary; surprisingly, BA is not significant at the 5% level. The coefficient of determination  $R^2$  shows that 58% of Salary is explained by the estimated regression model, leaving 42% of the variability in Salary unexplained.

Lastly, for an MLB player with seven years' experience and an average BA and RBI of 266 and 50, respectively, the model predicts a salary of \$4,295,320. With 95% confidence, expected salary will lie between \$3,731,360 and \$4,859,280. Perhaps before management makes an offer to the player, the model should consider including other factors that may significantly influence a player's salary. One possible explanatory variable for inclusion is a player's on-base percentage.

## CONCEPTUAL REVIEW

### LO 15.1 Conduct tests of individual significance.

A **test of individual significance** determines whether the explanatory variable  $x_j$  has an individual statistical influence on  $y$ . The value of the test statistic is calculated as  $t_{df} = \frac{b_j - \beta_{j0}}{se(b_j)}$ , where  $df = n - k - 1$ ,  $se(b_j)$  is the standard error of the OLS estimator  $b_j$ , and  $\beta_{j0}$  is the hypothesized value of  $\beta_j$ . If  $\beta_{j0} = 0$ , the value of the test statistic reduces to  $t_{df} = \frac{b_j}{se(b_j)}$ .

Excel and virtually all other statistical packages report a value of the test statistic and its associated  $p$ -value for a two-tailed test that assesses whether the regression coefficient differs from zero.

- If we specify a one-tailed test, then we need to divide the computer-generated  $p$ -value in half.
- If we test whether the coefficient differs from a nonzero value, then we cannot use the value of the computer-generated test statistic and its  $p$ -value.

The  $100(1 - \alpha)\%$  confidence interval for the regression coefficient  $\beta_j$  is given by  $b_j \pm t_{\alpha/2, df} se(b_j)$ , where  $df = n - k - 1$ .

### LO 15.2 Conduct a test of joint significance.

A **test of joint significance** determines whether the explanatory variables  $x_1, x_2, \dots, x_k$  in a multiple regression model have a joint statistically significant influence on  $y$ . The value of the test statistic is calculated as  $F_{(df_1, df_2)} = \frac{SSR/k}{SSE/(n - k - 1)} = \frac{MSR}{MSE}$ , where  $df_1 = k$ ,  $df_2 = n - k - 1$ ,  $SSR$  is the regression sum of squares,  $SSE$  is the sum of squares due to error,  $MSR$  is the mean square regression, and  $MSE$  is the mean square error.

The ANOVA table from computer output provides both the value of the test statistic and its associated  $p$ -value.

### LO 15.3 Conduct a general test of linear restrictions.

When **testing linear restrictions**, the value of the test statistic is calculated as  $F_{(df_1, df_2)} = \frac{(SSE_R - SSE_U)/df_1}{SSE_U/df_2}$ , where  $df_1$  is equal to the number of linear restrictions,  $df_2 = n - k - 1$

where  $k$  is the number of explanatory variables in the unrestricted model;  $SSE_R$  and  $SSE_U$  are the sum of squares due to error of the restricted and unrestricted models, respectively. If the null hypothesis is rejected, we conclude that the linear restrictions are not valid.

#### **LO 15.4 Calculate and interpret confidence intervals and prediction intervals.**

For specific values of  $x_1, x_2, \dots, x_k$ , denoted by  $x_1^0, x_2^0, \dots, x_k^0$ , the  $100(1 - \alpha)\%$  **confidence interval for the expected value of  $y$**  is given by  $\hat{y}^0 \pm t_{\alpha/2, df} se(\hat{y}^0)$  where  $df = n - k - 1$  and  $se(\hat{y}^0)$  is the standard error of  $\hat{y}^0$ . To derive  $\hat{y}^0$  together with  $se(\hat{y}^0)$ , we first estimate a modified regression model where  $y$  is the response variable and the explanatory variables are defined as  $x_1^* = x_1 - x_1^0, x_2^* = x_2 - x_2^0, \dots, x_k^* = x_k - x_k^0$ . The resulting estimate of the intercept and its standard error equal  $\hat{y}^0$  and  $se(\hat{y}^0)$ , respectively.

For specific values of  $x_1, x_2, \dots, x_k$ , denoted by  $x_1^0, x_2^0, \dots, x_k^0$ , the  $100(1 - \alpha)\%$  **prediction interval for an individual value of  $y$**  is given by  $\hat{y} \pm t_{\alpha/2, df} \sqrt{(se(\hat{y}^0))^2 + s_e^2}$ , where  $df = n - k - 1$ ,  $se(\hat{y}^0)$  is the standard error of  $\hat{y}^0$ , and  $s_e$  is the standard error of the estimate.

#### **LO 15.5 Explain the role of the assumptions on the OLS estimators.**

Under the assumptions of the classical linear regression model, OLS provides the best estimates. However, the desirable properties of the OLS estimators become compromised as one or more model assumptions are violated. In addition, for certain violations, it is not possible to make meaningful inferences from the  $t$  and  $F$  test results.

#### **LO 15.6 Describe common violations of the assumptions and offer remedies.**

Residual plots are used to identify model inadequacies; they also help identify outliers. The model is adequate if the residuals are randomly dispersed around the zero value.

Some degree of **multicollinearity** is present in most applications. A high  $R^2$  coupled with insignificant explanatory variables is often indicative of multicollinearity. Multicollinearity is considered serious if the sample correlation coefficient between any two explanatory variables is more than 0.80 or less than  $-0.80$ . We can drop one of the collinear variables if its omission can be justified. We can obtain more data, as that may weaken the correlation. Another option is to express the explanatory variables differently. At times the best approach may be to do nothing, especially if the estimated model yields a high  $R^2$ .

The assumption of **constant variability** often breaks down in cross-sectional studies. The resulting OLS estimators are unbiased, but the standard errors of the OLS estimators are inappropriate, making the standard  $t$  or  $F$  tests invalid. This assumption is violated if the variability of the residuals increases or decreases over the value of an explanatory variable. Researchers often use the OLS estimates along with a correction for the standard errors, called White's correction.

The assumption of **uncorrelated observations** often breaks down in time series studies. The resulting OLS estimators are unbiased but their standard errors are inappropriate. In general, correlated observations make the model look better than it really is with a spuriously high  $R^2$ . Furthermore, the  $t$  and the  $F$  test results may incorrectly suggest significance of the explanatory variables. This assumption is violated if the residuals show a pattern around the horizontal time axis. Researchers often use the OLS estimates along with a correction for the standard errors, using the Newey-West procedure.

It is important that the regression model incorporates all relevant explanatory variables. In the case of **excluded variables**, the OLS estimators are generally biased.

# ADDITIONAL EXERCISES AND CASE STUDIES

## Exercises

57. In an attempt to determine whether or not a linear relationship exists between the price of a home (in \$1,000s) and the number of days it takes to sell the home, a real estate agent collected data from recent sales in his city and estimated the following model:  $\text{Price} = \beta_0 + \beta_1 \text{Days} + \varepsilon$ . A portion of the Excel results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	-491.27	156.94	-3.13	0.0203
Days	6.17	1.19	5.19	0.0020

Specify the hypotheses to determine whether Days is significant in explaining a house's price. At the 5% significance level, what is the conclusion to the test? Explain.

58. **FILE Quotations.** The labor estimation group at Sturdy Electronics, a contract electronics manufacturer of printed circuit boards, wants to simplify the process it uses to quote production costs to potential customers. They have identified the primary drivers for production time (and thus production cost) as being the number of electronic parts that can be machine installed and the number of parts that must be manually installed. Accordingly, they wish to develop a multiple regression model to predict production time, measured as minutes per board, using a random sample of 25 recent product quotations. A portion of the data is shown in the accompanying table.

Production Time	Machine Parts	Manual Parts
9.1	275	14
10.8	446	12
⋮	⋮	⋮
15.5	618	16

- What is the sample regression equation?
- Predict production time for a circuit board with 475 machine-installed components and 16 manually-installed components.
- What proportion of the sample variability in production time is explained by the two explanatory variables?
- At the 5% significance level, are the explanatory variables jointly significant? Are they individually significant?

59. **FILE Happiness\_Age.** A sociologist wishes to study the relationship between happiness and age. He interviews 24 individuals and collects data on age and happiness, measured on a scale from 0 to 100. Estimate:  $\text{Happiness} = \beta_0 + \beta_1 \text{Age} + \varepsilon$ . At the 1% significance level, is Age significant in explaining Happiness? Show the relevant steps of a hypothesis test using the critical value approach.
60. **FILE Home\_Ownership.** The homeownership rate in the U.S. was 67.4% in 2009. In order to determine if homeownership is linked with income, 2009 state level data on the homeownership rate (Ownership) and median household income (Income) were collected. A portion of the data is shown in the accompanying table.

State	Income	Ownership
Alabama	39980	74.1%
Alaska	61604	66.8%
⋮	⋮	⋮
Wyoming	52470	73.8%

SOURCE: [www.census.gov](http://www.census.gov).

- Estimate:  $\text{Ownership} = \beta_0 + \beta_1 \text{Income} + \varepsilon$ .
  - At the 5% significance level, is Income linearly related to Ownership? Show the steps of a hypothesis test using the critical value approach.
  - Construct the 95% confidence interval for the expected value of Ownership if Income is \$50,000.
  - Compare the above confidence interval with the 95% prediction interval for Ownership.
61. **FILE SAT.** A researcher studies the relationship between SAT scores, the test-taker's family income (Income), and his/her grade point average (GPA). Data are collected from 24 students. A portion of the data is shown in the accompanying table.

SAT	Income	GPA
1651	47,000	2.79
1581	34,000	2.97
⋮	⋮	⋮
1940	113,000	3.96

Estimate:  $\text{SAT} = \beta_0 + \beta_1 \text{Income} + \beta_2 \text{GPA} + \varepsilon$ .

- At the 5% significance level, are income and GPA individually significant? Show the relevant steps of each test, using the critical value approach.
- At the 5% significance level, are income and GPA jointly significant? Show the relevant steps of the hypothesis test, using the critical value approach.

- c. Predict SAT if Income is \$80,000 and GPA is 3.5. Use these values for the explanatory variables to construct the 95% prediction interval for the individual SAT score.

62. **FILE Turnover Expense.** George believes that the returns of mutual funds are influenced by annual turnover rates and annual expense ratios. In order to substantiate his claim, he randomly selects eight mutual funds and collects data on each fund's five-year annual return (Return), its annual holding turnover rate (Turnover), and its annual expense ratio (Expense). A portion of the data is shown in the accompanying table.

	Return (%)	Turnover (%)	Expense (%)
American Funds EuroPacific	6.06	41	0.83
Artisan International	2.94	54	1.22
⋮	⋮	⋮	⋮
Royce Value Plus	1.48	42	1.48

SOURCE: All data as of July 31, 2009 from finance.yahoo.com.

- Estimate:  $\text{Return} = \beta_0 + \beta_1 \text{Turnover} + \beta_2 \text{Expense} + \varepsilon$ . Conduct appropriate tests to verify George's theory at the 5% significance level.
  - Discuss the potential problems of multicollinearity and heteroskedasticity.
63. **FILE Crime.** A government researcher examines the factors that influence a city's crime rate. For 41 cities, she collects the crime rate (crimes per 100,000 residents), the poverty rate (in %), the median income (in \$1,000s), the percent of residents younger than 18, and the percent of residents older than 65. A portion of the data is shown in the accompanying table.

Crime	Poverty	Income	Under 18	Over 65
710.6	3.8	58.422	18.3	23.4
1317.7	16.7	48.729	19.0	10.3
⋮	⋮	⋮	⋮	⋮
139.7	3.9	59.445	19.7	16

- Estimate:  $\text{Crime} = \beta_0 + \beta_1 \text{Poverty} + \beta_2 \text{Income} + \beta_3 \text{Under 18} + \beta_4 \text{Over 65} + \varepsilon$ . Discuss the individual and joint significance of the explanatory variables at the 5% significance level.
  - At the 5% level, conduct a partial  $F$  test to determine if the influence of Under 18 is different from that of Over 65.
  - Which explanatory variables are likely to be collinear? Find their sample correlation coefficients to confirm.
64. **FILE Dow 2010.** A research analyst is trying to determine whether a firm's price-earnings (P/E) and price-sales (P/S) ratios can explain the firm's stock

performance over the past year. Generally, a high P/E ratio suggests that investors are expecting higher earnings growth in the future compared to companies with a lower P/E ratio. Investors use the P/S ratio to determine how much they are paying for a dollar of the firm's sales rather than a dollar of its earnings (P/E ratio). In short, the higher the P/E ratio and the lower the P/S ratio, the more attractive the investment. The accompanying table shows a portion of the 30 firms included in the Dow Jones Industrial Average.

DOW Components	Return (in %)	P/E ratio	P/S ratio
3M Co.	4.4	14.37	2.41
Alcoa Inc.	-4.5	11.01	0.78
⋮	⋮	⋮	⋮
Walt Disney Company	16.3	13.94	1.94

SOURCE: The 2010 returns (January 1, 2010–December 31, 2010) were obtained from *The Wall Street Journal*, January 3, 2011; the P/E ratios and the P/S ratios were obtained from finance.yahoo.com on January 20, 2011.

- Estimate:  $\text{Return} = \beta_0 + \beta_1 \text{P/E} + \beta_2 \text{P/S} + \varepsilon$ . Show the regression results in a well-formatted table.
  - Determine whether P/E and P/S are jointly significant at the 5% significance level.
  - Establish whether the explanatory variables are individually significant at the 5% significance level.
  - What is the predicted return for a firm with a P/E ratio of 10 and a P/S ratio of 2? Use these values to construct the 95% confidence interval for the expected return.
65. **FILE Smoking.** A nutritionist wants to understand the influence of income and healthy food on the incidence of smoking. He collects 2009 data on the percentage of smokers in each state in the U.S. and the corresponding median income and the percentage of the population that regularly eats fruits and vegetables. A portion of the data is shown in the accompanying table.

State	Smoke (%)	Fruits/Vegetables (%)	Median Income
AK	14.6	23.3	61604
AL	16.4	20.3	39980
⋮	⋮	⋮	⋮
WY	15.2	23.3	52470

SOURCE: Centers for Disease Control and Prevention and U.S. Census Bureau.

- Estimate:  $\text{Smoke} = \beta_0 + \beta_1 \text{Fruits/Vegetables} + \beta_2 \text{Median Income} + \varepsilon$ .
- At the 5% level of significance, are the explanatory variables individually and jointly significant? Explain.
- Use the sample correlation coefficients to evaluate the potential problem of multicollinearity.



66. **FILE PerCapita.** Consider a regression model for per capita income,  $y$ . The explanatory variables consist of the percentage of the population in the U.S. that is (a) without a high school diploma,  $x_1$  (b) foreign born,  $x_2$  and (c) non-English speaking,  $x_3$ . A portion of the data is shown in the accompanying table.

State	Per capita Income	No High School	Foreign Born	No English
Alabama	22984	18.6%	3.4%	4.9%
Alaska	30726	9.3	72	16.5
⋮	⋮	⋮	⋮	⋮
Wyoming	27860	8.7	3.1	6.7

SOURCE: 2010 U.S. Census

- Estimate the model,  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$  and test for the joint and individual significance of the explanatory variables at the 5% level.
  - What proportion of the sample variability in per capita income is explained by the explanatory variables?
  - Do you suspect multicollinearity in the model? Use sample data to confirm.
67. **FILE MCAS.** A researcher examines the factors that influence student performance. She gathers data on 224 school districts in Massachusetts. The response variable is the students' mean score on a standardized test (Score). She uses

four explanatory variables in her analysis: the student-to-teacher ratio (STR), the average teacher's salary (TSAL), the median household income (INC), and the percentage of single family households (SGL). A portion of the data is shown in the accompanying table.

Score	STR (%)	TSAL (in \$1,000s)	INC (in \$1,000s)	SGL (%)
227.00	19.00	44.01	48.89	4.70
230.67	17.90	40.17	43.91	4.60
⋮	⋮	⋮	⋮	⋮
230.67	19.20	44.79	47.64	5.10

SOURCE: Massachusetts Department of Education and the Census of Population and Housing.

- Estimate:  $\text{Score} = \beta_0 + \beta_1 \text{STR} + \beta_2 \text{TSAL} + \beta_3 \text{INC} + \beta_4 \text{SGL} + \varepsilon$ . Show the regression results in a well-formatted table.
- Suppose you want to test if school input factors, STR and TSAL, are significant in explaining Score. Specify the competing hypotheses. Estimate the restricted model. At the 5% significance level, can you conclude that STR and TSAL are jointly significant?
- Suppose you want to test if socioeconomic factors, INC and SGL, are significant in explaining Score. Specify the competing hypotheses. Estimate the restricted model. At the 5% significance level, can you conclude that INC and SGL are jointly significant?

# CASE STUDIES

**CASE STUDY 15.1** American football is the highest-paying sport on a per-game basis. Given that the quarterback is considered the most important player on the team, he is typically well-compensated. A sports statistician examines the factors that influence a quarterback's salary (Salary). He believes that a quarterback's pass completion rate (PC) is the most important variable affecting Salary. The statistician also wonders how total touchdowns scored (TD) and a quarterback's age (Age) might impact Salary. The statistician collects 2009 data on Salary, PC, TD, and Age. A portion of the data is shown in the accompanying table.

**Data for Case Study 15.1** Quarterback Salary Data, 2009

Name	Salary (in \$ millions)	PC	TD	Age
Philip Rivers	25.5566	65.2	28	27
Jay Cutler	22.0441	60.5	27	26
⋮	⋮	⋮	⋮	⋮
Tony Romo	0.6260	63.1	26	29

SOURCE: USA Today database for salaries; NFL.com for other data.

**FILE**

Quarterback Salaries



In a report, use the sample information to:

1. Estimate and interpret the model:  $\text{Salary} = \beta_0 + \beta_1 \text{PC} + \beta_2 \text{TD} + \beta_3 \text{Age} + \varepsilon$ .
2. Discuss the individual and joint significance of the explanatory variables at the 5% level.
3. Determine whether TD and Age are jointly significant at the 5% significance level.
4. Construct the 95% confidence interval for the expected salary of a quarterback using average values of PCT, TD, and Age.

**CASE STUDY 15.2** Apple Inc. has established a unique reputation in the consumer electronics industry with its development of products such as the iPod, the iPhone, and the iPad. As of May 2010, Apple had surpassed Microsoft as the most valuable company in the world (*The New York Times*, May 26, 2010). Michael Gomez is a stock analyst and wonders if the return on Apple's stock is best modeled using the CAPM model. He collects five years of monthly data, a portion of which is shown in the accompanying table.

**Data for Case Study 15.2** Apple Return Data,  $n = 60$

Date	$R - R_f$	$R_M - R_f$
1/1/2006	4.70	2.21
2/1/2006	-9.65	-0.31
⋮	⋮	⋮
11/1/2010	1.68	2.15

SOURCE: finance.yahoo.com and U.S. Treasury.

**FILE**  
Apple

In a report, use the sample information to:

1. Estimate and interpret CAPM:  $R - R_f = \beta_0 + \beta_1(R_M - R_f) + \varepsilon$ . Search for Apple's reported Beta on the Web and compare it with your estimate.
2. At the 5% significance level, is the stock return for Apple riskier than that of the market? At the 5% significance level, do abnormal returns exist? Explain.
3. Use a residual plot to analyze the potential problem of correlated observations.

**CASE STUDY 15.3** According to a recent report by the government, new home construction fell to an 18-month low in October, 2010 (CNNMoney.com, November 17, 2010). Housing starts, or the number of new homes being built, experienced an 11.7% drop in its seasonally adjusted annual rate. Urmil Singh works for a mortgage company in Madison, Wisconsin. She wants to better understand the quantitative relationship between housing starts, the mortgage rate, and the unemployment rate. She gathers seasonally adjusted monthly data on these variables from 2006:01–2010:12. A portion of the data is shown in the accompanying table.

**Data for Case Study 15.3** Housing Starts and Other Factors,  $n = 60$

Date	Housing Starts (in 1,000s)	Mortgage Rate (%)	Unemployment Rate (%)
2006-01	2273	6.15	4.7
2006-02	2119	6.25	4.8
⋮	⋮	⋮	⋮
2010-12	520	4.71	9.4

SOURCE: Census Bureau and Board of Governors.

**FILE**  
Housing\_Starts

In a report, use the sample information to:

1. Estimate a multiple regression model for housing starts using the mortgage rate and the unemployment rate as the explanatory variables.
2. At the 5% significance level, evaluate the individual and joint significance of the explanatory variables.
3. Discuss the potential problems of multicollinearity and correlated observations in this time series data application.

## APPENDIX 15.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### FILE

*Convenience\_Stores*

#### Residual Plots—Changing Variability

- A. (Replicating Figure 15.4) From the menu choose **Stat > Regression > Regression > Fit Regression Model**.
- B. Next to **Response**, select Sales, and next to **Continuous predictors**, select Sqft. Choose **Graphs**. Under **Residuals Plots**, select **Individual plots**, and under **Residuals versus the variables**, select Sqft.

#### Residual Plots—Correlated Observations

- A. (Replicating Figure 15.5) From the menu choose **Stat > Regression > Regression > Fit Regression Model**.
- B. Next to **Response**, select Sales, and next to **Continuous predictors**, select AdCost and Unempl. Choose **Graphs**. Under **Residuals Plots**, select **Individual plots**, and then select **Residuals versus order**.

#### FILE

*Sushi\_Restaurant*

#### FILE

*Home Values*

#### Assessing Multicollinearity with a Correlation Matrix

(Replicating Example 15.10) From the menu choose **Stat > Basic Statistics > Correlation**. Under **Variables**, select HH Income and Per Cap Inc.

### SPSS

#### FILE

*Convenience\_Stores*

#### Residual Plots—Changing Variability

- A. (Replicating Figure 15.4) From the menu choose **Analyze > Regression > Linear**.
- B. Under **Dependent**, select Sales, and under **Independent(s)**, select Sqft. Choose **Save** and under **Residuals** select **Unstandardized**.
- C. From the menu choose **Graphs > Legacy Dialogs > Scatter/Dot > Simple Scatter**.
- D. Under **Y Axis**, select Unstandardized Residual and under **X Axis**, select Sqft.

#### Residual Plots—Correlated Observations

- A. (Replicating Figure 15.5) Add a column to data labeled “time” and number from 1 to 17.
- B. From the menu choose **Analyze > Regression > Linear**.
- C. Under **Dependent**, select Sales, and under **Independent(s)**, select AdCost and Unempl. Choose **Save** and under **Residuals**, select **Unstandardized**.
- D. From the menu choose **Graphs > Legacy Dialogs > Scatter/Dot > Simple Scatter**.
- E. Under **Y Axis**, select Unstandardized Residual and under **X Axis**, select time.

#### FILE

*Sushi\_Restaurant*

#### FILE

*Home Values*

#### Assessing Multicollinearity with a Correlation Matrix

- A. (Replicating Example 15.10) From the menu choose **Analyze > Correlate > Bivariate**.
- B. Under **Variables**, select HH Income and Per Cap Inc.

## JMP

### Residual Plots—Simple Linear Regression

- A. (Replicating Example 15.4) From the menu choose **Analyze > Fit Y by X**.
- B. Under **Select Columns**, select Sales, and then under **Cast Selected Columns Into Roles** select **Y, Response**. Under **Select Columns**, select Sqft, and then under **Cast Selected Columns Into Roles**, select **X, Factor**.
- C. Click on the red triangle next to **Bivariate Fit Sales by Sqft**, and select **Fit Line**.
- D. Click on the red triangle next to **Linear Fit**, and select **Plot Residuals**.

**FILE**

*Convenience\_Stores*

### Residual Plots—Multiple Regression

- A. (Replicating Figure 15.5) From the menu choose **Analyze > Fit Model**.
- B. Under **Select Columns**, select Sales, and then under **Pick Role Variables**, select **Y**. Under **Select Columns**, select AdCost and Unemp, and then under **Construct Model Effects**, select **Add**.
- C. Click on the red triangle next to **Response Sales**, select **Role Diagnostics > Plot Residual by Row**.

**FILE**

*Sushi\_Restaurant*

### Assessing Multicollinearity with a Correlation Matrix

(Replicating Example 15.10) From the menu choose **Analyze > Multivariate Methods > Multivariate**. Under **Select Columns**, select HH Income and Per Cap Inc. and under **Cast Selected Columns into Roles**, select **Y, Columns**.

**FILE**

*Home Values*

# 16

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 16.1 Use and evaluate polynomial regression models.
- LO 16.2 Use and evaluate log-transformed models.
- LO 16.3 Describe the method used to compare linear with log-transformed models.

# Regression Models for Nonlinear Relationships

Regression analysis is one of the most widely used statistical techniques in business, engineering, and the social sciences. It empirically validates not only whether a relationship exists between variables, but also quantifies the strength of the relationship. So far, we have considered only linear regression models, whether single- or multiple-variable. There are numerous applications where the relationship between the explanatory variable and the response variable cannot be represented by a straight line and, therefore, must be captured by an appropriate curve. In fact, the choice of a functional form is a crucial part of specifying a regression model. In this chapter, we discuss some common nonlinear regression models by making simple transformations of the variables. These transformations include squares and natural logarithms, which capture interesting nonlinear relationships while still allowing easy estimation within the framework of a linear regression model. We use numerical measures to choose between alternative model specifications.





## INTRODUCTORY CASE

### Rental Market in Ann Arbor, Michigan

Real estate investment in college towns continues to promise good returns (*The Wall Street Journal*, September 24, 2010). First, students offer a steady stream of rental demand as cash-strapped public universities are unable to house their students beyond freshman year. Second, this demand is projected to grow as more children of baby boomers head to college. Marcela Treisman works for an investment firm in Michigan. Her assignment is to analyze the rental market in Ann Arbor, which is home to the main campus of the University of Michigan. She knows that with a third of its population consisting of university students, Ann Arbor is consistently rated as one of the best places to live in the United States. Marcela wants to understand what kind of off-campus homes promise good rental income. She gathers data on monthly rent (Rent, in \$) for 2011, along with three characteristics of the home: number of bedrooms (Beds), number of bathrooms (Baths), and square footage (Sqft). A portion of the data is shown in Table 16.1.

**TABLE 16.1** Rental Data for Ann Arbor, Michigan;  $n = 40$

**FILE**  
**AnnArbor**

Rent (in \$)	Beds	Baths	Sqft
645	1	1	500
675	1	1	648
⋮	⋮	⋮	⋮
2400	3	2.5	2700

SOURCE: [www.zillow.com](http://www.zillow.com).

Marcela would like to use the information in Table 16.1 to:

1. Evaluate various models that quantify the relationship between rent and home characteristics.
2. Use model selection criteria to select the most appropriate model.
3. Make predictions for rental income for specific values of home characteristics.

A synopsis of this case is provided at the end of Section 16.2.

Use and evaluate polynomial regression models.

Linear regression models are often justified on the basis of their computational simplicity. An implication of a simple linear regression model,  $y = \beta_0 + \beta_1 x + \varepsilon$ , is that if  $x$  goes up by one unit, we expect  $y$  to change by  $\beta_1$ , irrespective of the value of  $x$ . However, in many applications, the relationship cannot be represented by a straight line and, therefore, must be captured by an appropriate curve. We note that the linearity assumption discussed in Chapter 15 places the restriction of linearity on the parameters and not on the variables. Consequently, we can capture many interesting nonlinear relationships, within the framework of a linear regression model, by simple transformations of the response and/or the explanatory variables.

If you ever studied microeconomics, you may have learned that a firm's (or industry's) average cost curve tends to be "U-shaped." Due to economies of scale, the average cost  $y$  of a firm initially decreases as output  $x$  increases. However, as  $x$  increases beyond a certain point, its impact on  $y$  turns positive. Other applications show the influence of the explanatory variable initially positive but then turning negative, leading to an "inverted U shape." (Mathematically, U-shaped means convex, whereas inverted U-shaped means concave.) The **quadratic regression model** is appropriate when the slope, capturing the influence of  $x$  on  $y$ , changes in magnitude as well as sign.

A quadratic regression model with one explanatory variable is specified as  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ ; we can easily extend it to include multiple explanatory variables. The expression  $\beta_0 + \beta_1 x + \beta_2 x^2$  is the deterministic component of a quadratic regression model. In other words, conditional on  $x$ ,  $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ . This model can easily be estimated as a regression of  $y$  on  $x$  and  $x^2$ .

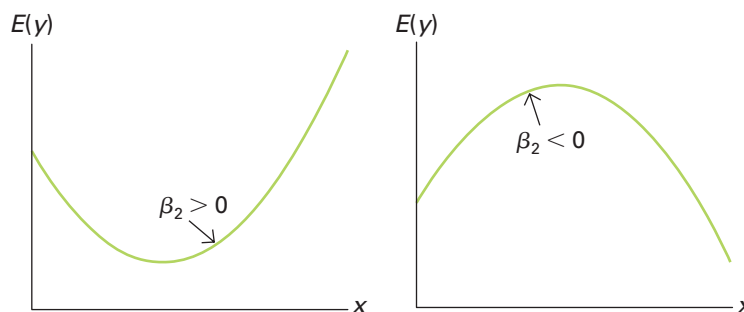
#### THE QUADRATIC REGRESSION MODEL

In a **quadratic regression model**  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ , the coefficient  $\beta_2$  determines whether the relationship between  $x$  and  $y$  is U-shaped ( $\beta_2 > 0$ ) or inverted U-shaped ( $\beta_2 < 0$ ).

**Predictions** with this model are made by  $\hat{y} = b_0 + b_1 x + b_2 x^2$ . It is advisable to use unrounded coefficients (or rounded to at least four decimal places) for making predictions.

As mentioned above, the coefficient  $\beta_2$  determines the shape of the relationship; Figure 16.1 highlights some representative shapes of a quadratic regression model.

**FIGURE 16.1**  
Representative shapes  
of a quadratic  
regression model:  
 $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$



It is important to be able to determine whether a quadratic regression model provides a better fit than the linear regression model. As we learned in Chapter 14, we cannot compare these models on the basis of their respective  $R^2$  values because the quadratic regression model uses one extra parameter than the linear regression model. For comparison purposes, we use adjusted  $R^2$ , which imposes a penalty for the extra parameter.



## EXAMPLE 16.1

Table 16.2 shows a portion of the average cost (in \$) and annual output (in millions of units) for 20 manufacturing firms. We also include a column of  $\text{Output}^2$ , which will be used for estimating the quadratic regression model.

**TABLE 16.2** Average Cost and Output Data for 20 Manufacturing Firms

Average Cost (\$)	Output (millions of units)	Output <sup>2</sup>
9.61	4	16
9.55	5	25
⋮	⋮	⋮
9.62	11	121

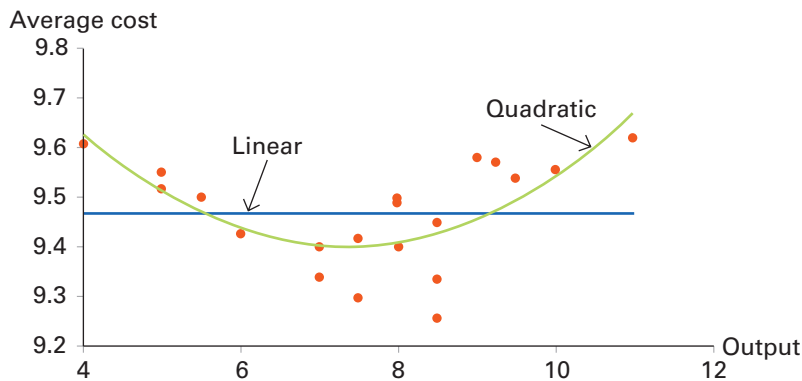
**FILE**  
Cost\_Functions

- Plot average cost (AC) against output.
- Estimate the linear and the quadratic regression models. Determine which model fits the data best.
- Use the best-fitting model to predict the average cost for a firm that produces 7 million units.

### SOLUTION:

- It is always informative to begin with a scatterplot of the response variable against the explanatory variable. Make sure that the response variable is on the vertical axis. Figure 16.2 shows average cost against output. We also superimpose linear and quadratic trends on the scatterplot (in Excel, right-click on the scatterpoints, add Trendline, and choose Linear and Polynomial with Order 2). At lower and higher levels of output, average costs are highest. It appears that the average cost in this industry would best be estimated using a quadratic regression model.

**FIGURE 16.2** Scatterplot of average cost versus output



- The second column of Table 16.3 shows the regression results for the linear regression model:  $AC = \beta_0 + \beta_1 \text{Output} + \varepsilon$ . The linear regression model provides a poor fit, which is not surprising given the scatterplot in Figure 16.2. Not only is Output statistically insignificant, the adjusted  $R^2$  is negative. In order to estimate a quadratic regression model, we have to first create data on the squared Output variable. A portion of these data, computed by squaring Output, are shown in Table 16.2. The third column of Table 16.3 shows the regression results for the quadratic regression model:  $AC = \beta_0 + \beta_1 \text{Output} + \beta_2 \text{Output}^2 + \varepsilon$ . In the quadratic regression model, Output is now significant. In addition, the slope coefficient of  $\text{Output}^2$  is positive and significant, indicating a U-shaped relationship between average cost and output.

**TABLE 16.3** Estimates of the Linear and the Quadratic Regression Models for Example 16.1

Variable	Linear Regression Model	Quadratic Regression Model
Intercept	9.4461* (0.00)	10.5225* (0.00)
Output	0.0029 (0.84)	-0.3073* (0.00)
Output <sup>2</sup>	NA	0.0210* (0.00)
Adjusted $R^2$	-0.0531	0.4540

NOTES: Parameter estimates are in the main body of the table with the  $p$ -values in parentheses; NA denotes not applicable; \* represents significance at the 5% level. The last row presents adjusted  $R^2$  for model comparison.

Given an adjusted  $R^2$  of 0.4540, the quadratic regression model is clearly better than the linear regression model in explaining average cost.

- c. Using the quadratic regression model, the predicted average cost for a firm that produces 7 million units is

$$\widehat{AC} = 10.5225 - 0.3073(7) + 0.0210(7^2) = \$9.40.$$

**Interpretation of coefficients in the quadratic regression model:** It does not make sense to think of  $\beta_1$  in the quadratic regression model of Example 16.1 as being the effect of changing a firm's output, holding the square of the firm's output constant. In nonlinear models, the sample regression equation is best interpreted by calculating, and even graphing, the predicted effect on the response variable of changing one explanatory variable at a time. We will elaborate on this point in Examples 16.2 and 16.3.

**Evaluating the marginal effect of  $x$  on  $y$  in the quadratic regression model:** It is important to evaluate the estimated marginal (partial) effect of the explanatory variable  $x$  on the predicted value of the response variable; that is, evaluate the change in  $\hat{y}$  due to a one unit increase in  $x$ . In a linear regression model,  $y = \beta_0 + \beta_1 x + \varepsilon$ , the marginal effect is constant, estimated by the slope coefficient  $b_1$ . In a quadratic regression model, it can be shown with calculus that the marginal effect of  $x$  on  $\hat{y}$  can be approximated by  $b_1 + 2b_2x$ . This marginal effect, unlike in the case of a linear regression model, depends on the value of  $x$  at which it is evaluated. In addition,  $\hat{y}$  reaches a maximum ( $b_2 < 0$ ) or minimum ( $b_2 > 0$ ) when the marginal effect equals zero. The value of  $x$  at which this happens is obtained from solving the equation  $b_1 + 2b_2x = 0$  as  $x = \frac{-b_1}{2b_2}$ .

## EXAMPLE 16.2

Use the results from estimating the quadratic regression model in Example 16.1 to answer the following questions.

- What is the change in average cost going from an output level of 4 million units to 5 million units?
- What is the change in average cost going from an output level of 8 million units to 9 million units? Compare this result to the result found in part a.
- What is the output level that minimizes average cost?

### SOLUTION:

- a. The predicted average cost for a firm that produces 4 million units is:

$$\widehat{AC} = 10.5225 - 0.3073(4) + 0.0210(4^2) = \$9.63.$$

The predicted average cost for a firm that produces 5 million units is:

$$\widehat{AC} = 10.5225 - 0.3073(5) + 0.0210(5^2) = \$9.51.$$

An increase in output from 4 to 5 million units (a one-unit increase in  $x$ ) results in a \$0.12 decrease (\$9.63 – \$9.51) in predicted average cost.

- b. The predicted average cost for a firm that produces 8 million units is:

$$\widehat{AC} = 10.5225 - 0.3073(8) + 0.0210(8^2) = \$9.41.$$

The predicted average cost for a firm that produces 9 million units is:

$$\widehat{AC} = 10.5225 - 0.3073(9) + 0.0210(9^2) = \$9.46.$$

An increase in output from 8 to 9 million units (a one-unit increase in  $x$ ) results in a \$0.05 increase in predicted average cost. Comparing this result to the one found in part a, we note that a one-unit change in  $x$  depends on the value at which  $x$  is evaluated. A one-unit increase in output from 4 to 5 million units results in a \$0.12 decrease in predicted average cost, while a one-unit increase in output from 8 to 9 million units results in a \$0.05 increase in predicted average cost. Depending on the value at which  $x$  is evaluated, a one-unit change in  $x$  may have a positive or negative influence on  $y$ , and the magnitude of this effect is not constant.

- c. Given  $b_1 = -0.3073$  and  $b_2 = 0.0210$ , the output level that minimizes average cost is  $x = \frac{-b_1}{2b_2} = \frac{-(-0.3073)}{2(0.0210)} = 7.32$  million units. This result is not surprising if we look back at Figure 16.2.

Let's now turn to an example with an inverted U-shaped relationship.

### EXAMPLE 16.3

In the United States, age discrimination is illegal, but its occurrence is hard to prove (*Newsweek*, March 17, 2010). Even without discrimination, it is widely believed that wages of workers decline as they get older. A young worker can expect wages to rise with age only up to a certain point, beyond which wages begin to fall. Ioannes Papadopoulos works in the human resources department of a large manufacturing firm and is examining the relationship between wages (in \$), education, and age. Specifically, he wants to verify the quadratic effect of age on wages. He gathers data on 80 workers in his firm with information on their hourly wage, education, and age. A portion of the data is shown in Table 16.4.

**TABLE 16.4** Data for Example 16.3 on Hourly Wage, Education, and Age;  $n = 80$

Wage (\$)	Education	Age
17.54	12	76
20.93	10	61
⋮	⋮	⋮
23.66	12	49

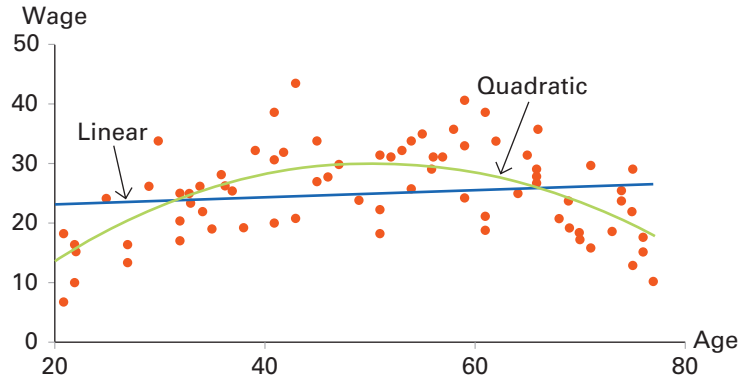
**FILE**  
Wages

- Plot Wage against Age and evaluate whether a linear or quadratic regression model better captures the relationship. Verify your choice by using the appropriate numerical measure.
- Use the appropriate model to predict hourly wages for someone with 16 years of education and age equal to 30, 50, or 70.
- According to the model, at what age will someone with 16 years of education attain the highest wages?

**SOLUTION:**

- a. Figure 16.3 shows a scatterplot of Wage against Age. We superimpose linear and quadratic trends on the scatterplot. It seems that the quadratic regression model provides a better fit for the data as compared to the linear regression model.

**FIGURE 16.3** Scatterplot of Wage versus Age



In order to estimate a quadratic regression model, we first create data on  $\text{Age}^2$ . The relevant portions of the output of the linear regression model,

$\text{Wage} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Age} + \varepsilon$ , and the quadratic regression model,

$\text{Wage} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Age} + \beta_3 \text{Age}^2 + \varepsilon$ , are presented in Table 16.5.

**TABLE 16.5** Estimates of the Linear and the Quadratic Regression Models for Example 16.3

Variable	Linear Regression Model	Quadratic Regression Model
Intercept	2.6381 (0.27)	-22.7219* (0.00)
Education	1.4410* (0.00)	1.2540* (0.00)
Age	0.0472 (0.13)	1.3500* (0.00)
Age <sup>2</sup>	NA	-0.0133* (0.00)
Adjusted $R^2$	0.6088	0.8257

Notes: Parameter estimates are in the main body of the table with the  $p$ -values in parentheses; NA denotes not applicable; \* represents significance at the 5% level. The last row presents adjusted  $R^2$  for model comparison.

The regression results highlight an interesting result. In the linear regression model, Age has an estimated coefficient of only 0.0472, which is not statistically significant ( $p$ -value = 0.13) even at the 10% significance level. However, results change dramatically when  $\text{Age}^2$  is included along with Age. In the quadratic regression model, both of these variables, with  $p$ -values about zero, are statistically significant at any reasonable level. Also, the adjusted  $R^2$  is higher for the quadratic regression model ( $0.8257 > 0.6088$ ), making it a better choice for prediction. This conclusion is consistent with our visual impression from the scatterplot in Figure 16.3, which suggested a weak linear but strong quadratic relationship between age and hourly wage.

- b. From Table 16.5, the estimated regression equation for the quadratic regression model is

$$\widehat{\text{Wage}} = -22.7219 + 1.2540\text{Education} + 1.3500\text{Age} - 0.0133\text{Age}^2.$$

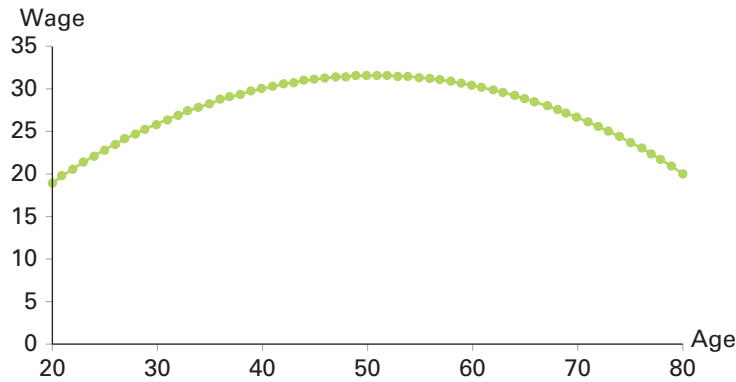
Therefore, the predicted hourly wage for a 30-year-old person with 16 years of education is

$$\widehat{\text{Wage}} = -22.7219 + 1.2540(16) + 1.3500(30) - 0.0133(30^2) = \$25.87.$$

Similarly, the predicted hourly wage for a 50- and a 70-year-old person is \$31.59 and \$26.67, respectively. Note that, consistent with the estimates, the hourly wage increases as a person ages from 30 to 50, but then decreases as a person ages from 50 to 70.

- c. In part b, we predicted the hourly wage for a 30-, 50-, and 70-year-old person with 16 years of education. Therefore, of the three ages considered, a 50-year-old person earns the highest wage. In Figure 16.4, we plot the predicted wage with 16 years of education and vary age from 20 to 80 with increments of 1.

**FIGURE 16.4** Predicted wages with 16 years of education and varying age



In order to determine the optimal age at which wage is maximized, we also solve  $x = \frac{-b_2}{2b_3} = \frac{-(1.3500)}{2(-0.0133)} = 50.75$ . The optimal age at which the wage is maximized is about 51 years, with a wage of about \$31.60. It is worth noting that at a different education level, predicted wages will not be the same, yet the highest wage will still be achieved at the same 51 years of age. We advise you to plot a similar graph with 12 years of education and varying age levels.

The quadratic regression model allows one sign change of the slope capturing the influence of  $x$  on  $y$ . It is a special case of a **polynomial regression model**. Polynomial regression models describe various numbers of sign changes. Sometimes a quadratic regression model with one sign change is referred to as a polynomial regression model of order 2. In fact, a linear regression model is a polynomial regression model of order 1, which, with a constant slope coefficient, does not allow any sign change.

The linear and the quadratic regression models are the most common polynomial regression models. Sometimes, researchers use a polynomial regression model of order 3, also called the **cubic regression model**. The cubic regression model allows for two changes in slope.

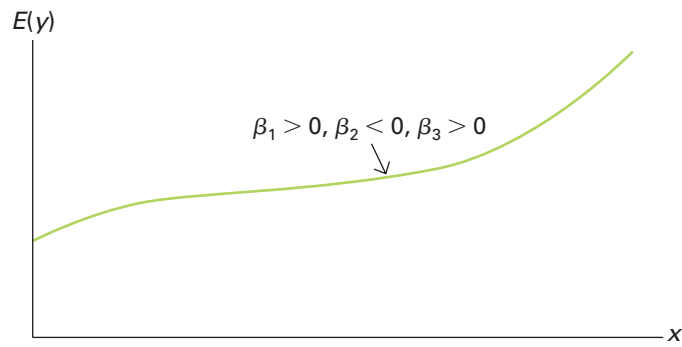
#### THE CUBIC REGRESSION MODEL

A **cubic regression model**,  $y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \epsilon$ , allows two sign changes of the slope capturing the influence of  $x$  on  $y$ .

**Predictions** with this model are made by  $\hat{y} = b_0 + b_1x + b_2x^2 + b_3x^3$ . It is advisable to use unrounded coefficients (or rounded to at least four decimal places) for making predictions.

The expression  $\beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3$  is the deterministic component of a cubic regression model; equivalently, conditional on  $x$ ,  $E(y) = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3$ . The shape of a cubic relationship depends on the coefficients. Figure 16.5 highlights a representative shape of a cubic regression model when  $\beta_1 > 0$ ,  $\beta_2 < 0$ , and  $\beta_3 > 0$ .

**FIGURE 16.5**  
Representative  
shape of a cubic  
regression model:  
 $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$



We often apply the cubic regression model when estimating the total cost curve of a firm. Generally, total costs of a firm increase gradually and then rapidly, as predicted by the law of diminishing returns. In general, the law of diminishing returns states that when increasing amounts of one factor of production (generally labor) are employed in a production process along with a fixed amount of another factor (generally capital), then after some point, the resulting increases in output of the product become smaller and smaller. A common example is adding more workers to a job, such as assembling a car in a factory. At some point, adding more workers will result in inefficiency as workers get in each other's way or wait for access to a machine. Producing one more unit of output will eventually cost increasingly more, due to inputs being used less and less effectively. We can think of Figure 16.5 as a firm's typical total cost curve where  $x$  and  $y$  represent a firm's output and total cost, respectively.

A cubic regression model can easily be estimated within the framework of a linear regression model where we use  $y$  as the response variable and  $x$ ,  $x^2$ , and  $x^3$  as the explanatory variables. It is easy to estimate a cubic regression model after we have created data on  $x^2$  and  $x^3$ . As before, we can compare polynomial models of various orders on the basis of adjusted  $R^2$ .

#### EXAMPLE 16.4

Table 16.6 shows a portion of data on the total cost (in \$1,000s) and output for producing a particular product. We also include a portion of the squared output and cubed output variables; these variables will be used in the estimation process.

- Use a cubic regression model to estimate total cost (TC).
- Predict the total cost if a firm produces 11 units of the product.

**TABLE 16.6** Data for Example 16.4 on Total Cost  $y$  and Output  $x$ ,  $n = 40$

Total Cost (in \$1,000s)	Output	Output <sup>2</sup>	Output <sup>3</sup>
37.49	9	81	729
37.06	7	49	343
⋮	⋮	⋮	⋮
33.92	4	16	64

#### SOLUTION:

- In order to estimate a cubic regression model, we first create data on the squared output and the cubed output variables (see Table 16.6). We then estimate  $TC = \beta_0 + \beta_1 \text{Output} + \beta_2 \text{Output}^2 + \beta_3 \text{Output}^3 + \varepsilon$ . Table 16.7 shows the relevant regression results. All variables are significant at the 5% level. In addition, the adjusted  $R^2$  (not shown) is 0.8551, which is higher than the adjusted  $R^2$  of 0.6452 and 0.8289 for the linear and the quadratic regression models, respectively. We advise you to verify these results.

**FILE**  
Total\_Cost



**TABLE 16.7** Estimates of the Cubic Regression Model for Example 16.4

Variable	Cubic Regression Model
Intercept	17.1836* (0.00)
Output	6.4570* (0.00)
Output <sup>2</sup>	-0.7321* (0.00)
Output <sup>3</sup>	0.0291* (0.01)

Notes:  $p$ -values are in parentheses; \* represents significance at the 5% level.

- b. We calculate the total cost of production for a firm that produces 11 units of the product as  $\hat{TC} = 17.1836 + 6.4570(11) - 0.7321(11^2) + 0.0291(11^3) = 38.37$ , or \$38,370.

## EXERCISES 16.1

### Mechanics

1. Consider the following two estimated models:

$$\hat{y} = 25 + 1.2x$$

$$\hat{y} = 30 + 1.4x - 0.12x^2$$

For each of the estimated models, predict  $y$  when  $x$  equals 5 and 10.

2. Consider the following three models:

$$\hat{y} = 80 + 1.2x$$

$$\hat{y} = 200 + 2.1x - 0.6x^2$$

$$\hat{y} = 100 + 16x - 2.2x^2 + 0.08x^3$$

For each of the estimated models, predict  $y$  when  $x$  equals 10 and 15.

3. Consider the following 10 observations on the response variable  $y$  and the explanatory variable  $x$ .

$y$	13.82	19.06	16.67	13.30	11.77	13.64	18.30	20.78	13.02	16.13
$x$	6	6	5	3	3	12	10	8	5	11

- Plot the above data and then estimate the linear and the quadratic regression models.
  - Use the appropriate numerical measure to justify which model fits the data best.
  - Given the best-fitting model, predict  $y$  for  $x = 4, 8$ , and 12.
  - Find  $x$  at which the quadratic equation reaches a minimum or maximum.
4. Consider the following 10 observations on the response variable  $y$  and the explanatory variable  $x$ .

$y$	9.42	4.88	3.36	3.28	1.67	7.35	6.30	4.67	9.33	5.04
$x$	11	9	5	5	4	10	3	10	11	8

- Plot the above data and estimate the linear and the quadratic regression models.
- Use the appropriate numerical measure to justify which model fits the data best.

- Given the best-fitting model, predict  $y$  for  $x = 4, 6$ , and 12.
  - Find  $x$  at which the quadratic equation reaches a minimum or maximum.
5. Consider the following sample regressions for the linear, the quadratic, and the cubic models along with their respective  $R^2$  and adjusted  $R^2$ .

	Linear	Quadratic	Cubic
Intercept	9.66	10.00	10.06
$x$	2.66	2.75	1.83
$x^2$	NA	-0.31	-0.33
$x^3$	NA	NA	0.26
$R^2$	0.810	0.836	0.896
Adjusted $R^2$	0.809	0.833	0.895

- Predict  $y$  for  $x = 1$  and 2 with each of the estimated models.
  - Select the most appropriate model. Explain.
6. Consider the following sample regressions for the linear, the quadratic, and the cubic models along with their respective  $R^2$  and adjusted  $R^2$ .

	Linear	Quadratic	Cubic
Intercept	19.80	20.08	20.07
$x$	1.35	1.50	1.58
$x^2$	NA	-0.31	-0.27
$x^3$	NA	NA	-0.03
$R^2$	0.640	0.697	0.698
Adjusted $R^2$	0.636	0.691	0.689

- Predict  $y$  for  $x = 2$  and 3 with each of the estimated models.
- Select the most appropriate model. Explain.

## Applications

7. **FILE Television.** Numerous studies have shown that watching too much television hurts school grades. Others have argued that television is not necessarily a bad thing for children (*Mail Online*, July 18, 2009). Like books and stories, television not only entertains, it also exposes a child to new information about the world. While watching too much television is harmful, a little bit may actually help. Researcher Matt Castle gathers information on the grade point average (GPA) of 28 middle school children and the number of hours of television they watched per week. A portion of the data is shown in the accompanying table.

GPA	Hours
3.24	19
3.10	21
⋮	⋮
3.31	4

- Estimate a quadratic regression model where the GPA of middle school children is regressed on hours and hours-squared.
  - Is the quadratic term in this model justified? Explain.
  - Find the optimal number of weekly hours of TV for middle school children.
8. **FILE Crew\_Size.** The project manager at a construction company is evaluating how crew size affects the productivity of framing jobs. He has experimented with varying crew size on a weekly basis over the past 27 weeks and has recorded weekly productivity (in jobs completed). A portion of the data is shown in the accompanying table.

Crew Size (#workers)	Productivity (jobs/week)
2	10
3	12
⋮	⋮
10	12

- Create a scatterplot of the data. Based on the scatterplot alone, what crew size(s) seems optimal?
- Estimate the linear and the quadratic regression models. Evaluate the two models in terms of variable significance and adjusted  $R^2$ . Which model provides the best fit? Provide an intuitive justification for the chosen model.
- Use the best-fitting model to predict how many jobs a crew of 5 could be expected to complete in a week.
- Estimate the cubic regression model. Would it improve the fit as compared to the quadratic regression model?

9. **FILE Bids.** Consider a sample comprised of firms that were targets of tender offers during the period 1978–1985. Conduct an analysis where the response variable represents the number of bids (Bids) received prior to the takeover of the firm. The explanatory variables include the bid premium (Premium) and firm size (Size). It is generally believed that a high initial bid premium, defined as the percentage excess of the firm's stock price, would deter subsequent bids. Moreover, while tender offers for large firms are likely to receive more media coverage and thereby attract the attention of opportunistic bidders, it also is a wealth constraint to potential bidders. A portion of the data is shown in the accompanying table.

Bids	Premium	Size (in \$ billions)
3	1.1905	0.7668
1	1.0360	0.1625
⋮	⋮	⋮
2	1.0329	3.4751

SOURCE: Compustat and *The Wall Street Journal* Index.

- Estimate the model,  $\text{Bids} = \beta_0 + \beta_1 \text{Premium} + \beta_2 \text{Size} + \beta_3 \text{Size}^2 + \varepsilon$ .
  - Justify the inclusion of the quadratic term in the model.
  - Find the predicted number of bids for a firm that has a bid premium of 1.2 and firm size of \$4, 8, 12, and 16 billion, respectively. What firm size is likely to get the highest number of bids?
10. **FILE Inspection.** A lead inspector at ElectroTech, an electronics assembly shop, wants to convince management that it takes longer, on a per-component basis, to inspect large devices with many components than it does to inspect small devices because it is difficult to keep track of which components have already been inspected. To prove her point, she has collected data from the last 25 devices. A portion of the data is shown in the accompanying table.

Number of Components on Device	Inspection Time (seconds)
32	84
13	49
⋮	⋮
23	70

- Create a scatterplot of the data. Does the lead inspector's claim seem credible?
- Estimate the linear, quadratic, and cubic regression models. Evaluate each model in terms of variable significance and adjusted  $R^2$ . Which model provides the best fit?
- Use the best model to predict the time required to inspect a device with 35 components.

11. **FILE Debt Payments.** You collect data on 26 metropolitan areas to analyze average monthly debt payments in terms of income and the unemployment rate. A portion of the data is shown in the accompanying table.

Metropolitan Area	Income (in \$1,000s)	Unemployment	Debt
Washington, D.C.	\$103.50	6.3%	\$1,285
Seattle	81.70	8.5	1,135
⋮	⋮	⋮	⋮
Pittsburgh	63.00	8.3	763

SOURCE: eFannieMae.com; bls.com; and Experian.com.

- Estimate the model  $\text{Debt} = \beta_0 + \beta_1 \text{Inc} + \beta_2 \text{Unemp} + \varepsilon$ . Is unemployment significant at the 5% level?
- You are told that the unemployment rate might have a quadratic influence on monthly debt payments. Provide an intuitive justification for this claim.
- Estimate  $\text{Debt} = \beta_0 + \beta_1 \text{Inc} + \beta_2 \text{Unemp} + \beta_3 \text{Unemp}^2 + \varepsilon$  to determine if Unemp and  $\text{Unemp}^2$  are jointly significant at the 5% level.

## 16.2 REGRESSION MODELS WITH LOGARITHMS

LO 16.2

In the preceding section, we squared and/or cubed the explanatory variable in order to capture nonlinearities between the response variable and the explanatory variables. Another commonly used transformation is based on the natural logarithm. You may recall from your math courses that the natural logarithmic function is the inverse of the exponential function. It is useful to briefly review exponential and logarithmic functions before using them in regression models.

Use and evaluate log-transformed models.

The exponential function is defined as

$$y = \exp(x) = e^x,$$

where  $e \approx 2.718$  is a constant and  $x$  is the function argument. We can use Excel or a calculator to compute, for example,  $e^2 = 7.39$ , or  $e^5 = 148.41$ .

The inverse of the exponential function is the natural logarithm (or simply, log); that is, the logarithm with the base  $e \approx 2.718$ . In other words,

$$\text{if } y = e^x, \text{ then } \ln(y) = x,$$

where  $\ln(y)$  is the natural log of  $y$ . For example, if  $y = e^2 = 7.39$ , then  $\ln(y) = \ln(7.39) = 2$ . Similarly, if  $y = e^5 = 148.41$ , then  $\ln(y) = \ln(148.41) = 5$ . Since  $\exp(\ln(x)) = x$ , the exponential function is sometimes referred to as the anti-log function. Finally, the log of a negative or zero value is not defined. Therefore, we can log-transform only those values that are positive.

As mentioned earlier, in many applications, linearity is not justifiable. For instance, consider an estimated linear regression of annual food expenditure  $y$  on annual income  $x$ :  $\hat{y} = 9000 + 0.20x$ . An estimated slope coefficient value of  $b_1 = 0.20$  implies that a \$1,000 increase in annual income would lead to a \$200 increase in annual food expenditure, irrespective of whether the income increase is from \$20,000 to \$21,000 or from \$520,000 to \$521,000. Since we would expect the impact to be smaller at high income levels, it may be more meaningful to analyze what happens to food expenditure as income increases by a certain percentage rather than by some dollar amount.

Logarithms convert changes in variables into percentage changes, and many relationships are naturally expressed in terms of percentages. For instance, it is common to log-transform variables such as incomes, house prices, and sales. On the other hand, variables such as age, experience, and scores are generally expressed in their original form. We rely both on economic intuition as well as statistical measures to find the appropriate form for the variables.

We first illustrate log models with only one explanatory variable, which we later extend to a multiple regression model.

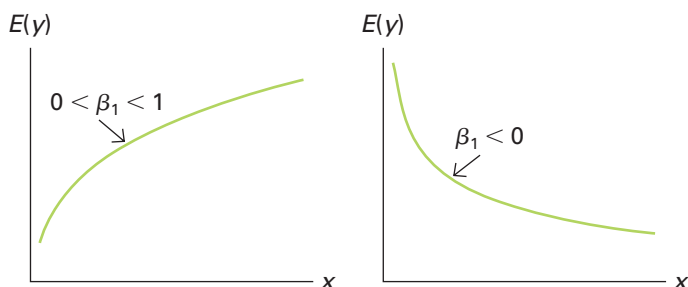
## A Log-Log Model

In a **log-log model**, both the response variable and the explanatory variable are transformed into natural logs. We can write this model as

$$\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon,$$

where  $\ln(y)$  is the log-transformed response variable and  $\ln(x)$  is the log-transformed explanatory variable. With these transformations, the relationship between  $y$  and  $x$  is captured by a curve whose shape depends on the sign and magnitude of the slope coefficient  $\beta_1$ . Figure 16.6 shows a couple of representative shapes of a log-log regression model.

**FIGURE 16.6**  
Representative shapes of a  
log-log model:  
 $\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$



For  $0 < \beta_1 < 1$ , the log-log model implies a positive relationship between  $x$  and  $E(y)$ ; as  $x$  increases,  $E(y)$  increases at a slower rate. This may be appropriate in the earlier example, where we expect food expenditure to react positively to changes in income, but with the impact diminishing at higher income levels. If  $\beta_1 < 0$ , it suggests a negative relationship between  $x$  and  $E(y)$ ; as  $x$  increases,  $E(y)$  decreases at a slower rate. Finally,  $\beta_1 > 1$  implies a positive and increasing relationship between  $x$  and  $y$ ; this case is not shown in Figure 16.6. For any application, the estimated value of  $\beta_1$  is determined by the data.

Note that while the log-log regression model is nonlinear in the variables, it is still linear in the coefficients, thus satisfying the requirement of the linear regression model. The only requirement is that we have to first transform both variables into logs before running the regression. We should also point out that in a log-log regression model, the slope coefficient  $\beta_1$  measures the percentage change in  $y$  for a small percentage change in  $x$ . In other words,  $\beta_1$  is a measure of elasticity. For instance, if  $y$  represents the quantity demanded of a particular good and  $x$  is its unit price, then  $\beta_1$  measures the price elasticity of demand, a parameter of considerable economic interest. Suppose  $\beta_1 = -1.2$ ; then a 1% increase in the price of this good is expected to lead to a 1.2% decrease in its quantity demanded.

Finally, even though the response variable is transformed into logs, we still make predictions in regular units. Given  $\ln(\hat{y}) = b_0 + b_1 \ln(x)$ , you may be tempted to use the anti-log function, to make predictions in regular units as  $\hat{y} = \exp(\ln(\hat{y})) = \exp(b_0 + b_1 \ln(x))$ , where  $b_0$  and  $b_1$  are the coefficient estimates. However, this transformation is known to systematically underestimate the expected value of  $y$ . One relatively simple correction is to make predictions as  $\hat{y} = \exp(b_0 + b_1 \ln(x) + s_e^2/2)$ , where  $s_e$  is the standard error of the estimate from the log-log model. This correction is easy to implement since virtually all statistical packages report  $s_e$ .

### THE LOG-LOG REGRESSION MODEL

A **log-log model** is specified as  $\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$ , and  $\beta_1$  measures the approximate percentage change in  $E(y)$  when  $x$  increases by 1%.

**Predictions** with this model are made by  $\hat{y} = \exp(b_0 + b_1 \ln(x) + s_e^2/2)$ , where  $b_0$  and  $b_1$  are the coefficient estimates and  $s_e$  is the standard error of the estimate. It is advisable to use unrounded coefficients (or rounded to at least four decimal places) for making predictions.

### EXAMPLE 16.5

Refer back to the food expenditure example where  $y$  represents expenditure on food and  $x$  represents income. Let the sample regression be  $\ln(\hat{y}) = 3.64 + 0.50\ln(x)$  with the standard error of the estimate  $s_e = 0.18$ .

- What is the predicted food expenditure for an individual whose income is \$20,000?
- What is the predicted food expenditure if income increases to \$21,000?
- Interpret the slope coefficient,  $b_1 = 0.50$ .

**SOLUTION:** For this log-log model, we make predictions as  $\hat{y} = \exp(b_0 + b_1\ln(x) + s_e^2/2)$ .

- For income  $x = 20000$ , we predict food expenditure as  $\hat{y} = \exp(3.64 + 0.50 \times \ln(20000) + 0.18^2/2) = 5475$ .
- For  $x = 21000$ , we find  $\hat{y} = \exp(3.64 + 0.50 \times \ln(21000) + 0.18^2/2) = 5610$ .
- In the log-log model, a slope coefficient of  $b_1 = 0.50$  implies that a 1% increase in income will lead to a 0.5% increase in predicted food expenditure. Here, as income increases from \$20,000 to \$21,000, or by 5%, the predicted food expenditure increases from \$5,475 to \$5,610, or by about 2.5%. This is consistent with the elasticity interpretation of the slope coefficient; that is, a 5% increase in income will lead to approximately a 2.5% ( $= 5 \times 0.5$ ) increase in the predicted food expenditure.

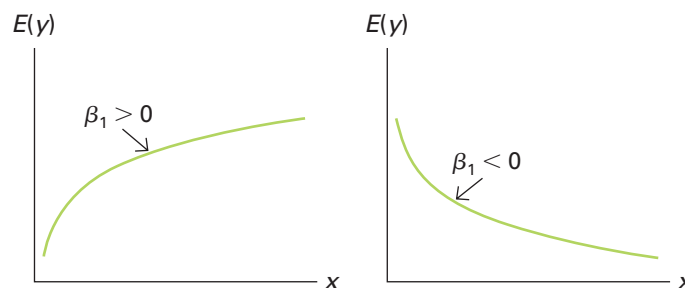
## The Logarithmic Model

A log-log specification transforms all variables into logs. It is also common to employ a **semi-log** model, in which not all variables are transformed into logs. We will discuss two types of semi-log models in the context of simple regression. A semi-log model that transforms only the explanatory variable is called the **logarithmic model** and a semi-log model that transforms only the response variable is called the **exponential model**. We can have many variants of semi-log models when we extend the analysis to include multiple explanatory variables.

The logarithmic model is defined as

$$y = \beta_0 + \beta_1 \ln(x) + \varepsilon.$$

Like the log-log model, this model implies that an increase in  $x$  will lead to an increase ( $\beta_1 > 0$ ) or decrease ( $\beta_1 < 0$ ) in  $E(y)$  at a decreasing rate. These models are especially attractive when only the explanatory variable is better captured in percentages. Figure 16.7 highlights some representative shapes of this model. Since the log-log and the logarithmic model can allow similar shapes, the choice between the two models can be tricky. We will compare models later in this section.



**FIGURE 16.7**  
Representative shapes of a  
logarithmic model:  
 $y = \beta_0 + \beta_1 \ln(x) + \varepsilon$

In the logarithmic model, the response variable is specified in regular units but the explanatory variable is transformed into logs. Therefore,  $\beta_1/100$  measures the approximate unit change in  $E(y)$  when  $x$  increases by 1%. For example, if  $\beta_1 = 5000$ , then a 1% increase in  $x$  leads to a 50 unit ( $= 5000/100$ ) increase in  $E(y)$ . Since the response variable is already specified in regular units, no further transformation is necessary when making predictions.

#### THE LOGARITHMIC MODEL

A **logarithmic model** is specified as  $y = \beta_0 + \beta_1 \ln(x) + \varepsilon$ , and  $\beta_1/100$  measures the approximate change in  $E(y)$  when  $x$  increases by 1%.

**Predictions** with this model are made by  $\hat{y} = b_0 + b_1 \ln(x)$ , where  $b_0$  and  $b_1$  are the coefficient estimates. It is advisable to use unrounded coefficients (or rounded to at least four decimal places) for making predictions.

#### EXAMPLE 16.6

Continuing with the earlier example of food expenditure, let the estimated regression be  $\hat{y} = 12 + 566 \ln(x)$ .

- What is the predicted food expenditure for an individual whose income is \$20,000?
- What is the predicted food expenditure if income increases to \$21,000?
- Interpret the slope coefficient,  $b_1 = 566$ .

**SOLUTION:** For this logarithmic model, we make predictions as  $\hat{y} = b_0 + b_1 \ln(x)$ .

- For income  $x = 20000$ , we predict food expenditure as  $\hat{y} = 12 + 566 \times \ln(20000) = 5617$ .
- For  $x = 21000$ , we find  $\hat{y} = 12 + 566 \times \ln(21000) = 5645$ .
- With a 5% increase in income from \$20,000 to \$21,000, the predicted food expenditure increases from \$5,617 to \$5,645, or by about \$28. This is consistent with the interpretation of the slope coefficient; that is,  $b_1 = 566$  implies that a 5% increase in income will lead to approximately a \$28 ( $= \frac{5 \times 566}{100}$ ) increase in predicted food expenditure.

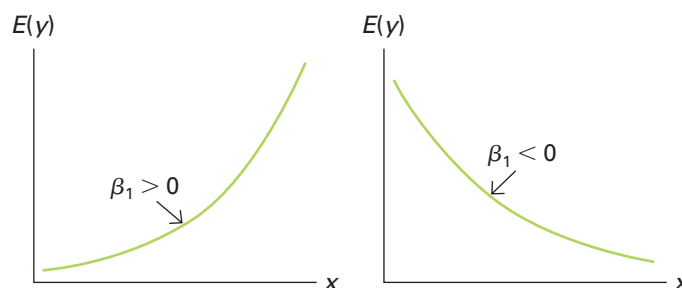
### The Exponential Model

Unlike the logarithmic model, in which we were interested in finding the unit change in  $E(y)$  for a 1% increase in  $x$ , the **exponential model** allows us to estimate the percent change in  $E(y)$  when  $x$  increases by one unit. The exponential model is defined as

$$\ln(y) = \beta_0 + \beta_1 x + \varepsilon.$$

Figure 16.8 shows some representative shapes of this model.

**FIGURE 16.8**  
Representative shapes of an  
exponential model:  
 $\ln(y) = \beta_0 + \beta_1 x + \varepsilon$





For an exponential model,  $\beta_1 \times 100$  measures the approximate percentage change in  $E(y)$  when  $x$  increases by one unit. For example, a value of  $\beta_1 = 0.05$  implies that a one-unit increase in  $x$  leads to a 5% ( $= 0.05 \times 100$ ) increase in  $E(y)$ . In applied work, we often see this model used to describe the rate of growth of certain economic variables, such as population, employment, wages, productivity, and the gross national product (GNP). As in the case of a log-log model, we make a correction for making predictions, since the response variable is measured in logs.

### THE EXPONENTIAL MODEL

An **exponential model** is specified as  $\ln(y) = \beta_0 + \beta_1 x + \varepsilon$ , and  $\beta_1 \times 100$  measures the approximate percentage change in  $E(y)$  when  $x$  increases by one unit.

**Predictions** with this model are made by  $\hat{y} = \exp(b_0 + b_1 x + s_e^2/2)$ , where  $b_0$  and  $b_1$  are the coefficient estimates and  $s_e$  is the standard error of the estimate. It is advisable to use unrounded coefficients (or rounded to at least four decimal places) for making predictions.

### EXAMPLE 16.7

Continuing again with the example of expenditure on food, let the estimated regression be  $\ln(\hat{y}) = 7.60 + 0.00005x$  with the standard error of the estimate,  $s_e = 0.20$ .

- What is the predicted expenditure on food for an individual whose income is \$20,000?
- What is the predicted food expenditure if income increases to \$21,000?
- Interpret the slope coefficient,  $b_1 = 0.00005$ .

**SOLUTION:** For this exponential model, we make predictions as  $\hat{y} = \exp(b_0 + b_1 x + s_e^2/2)$ .

- For income  $x = 20000$ , we predict food expenditure as  $\hat{y} = \exp(7.60 + 0.00005 \times 20000 + 0.20^2/2) = 5541$ .
- For  $x = 21000$ , we find  $\hat{y} = \exp(7.60 + 0.00005 \times 21000 + 0.20^2/2) = 5825$ .
- With a \$1,000 increase in income, the predicted food expenditure increases from \$5,541 to \$5,825, or by about 5%. This result is consistent with the interpretation of the slope coefficient; that is,  $b_1 = 0.00005$  implies that a \$1,000 increase in income, will lead to approximately a 5% ( $= 1000 \times 0.00005 \times 100$ ) increase in predicted food expenditure.

While these log models are easily estimated within the framework of a linear regression model, care must be exercised in making predictions and interpreting the estimated slope coefficient. When interpreting the slope coefficient, keep in mind that logs essentially convert changes in variables into percentage changes. Table 16.8 summarizes the results.

**TABLE 16.8** Summary of the Linear, Log-Log, Logarithmic, and Exponential Models

Model	Predicted Value	Estimated Slope Coefficient
$y = \beta_0 + \beta_1 x + \varepsilon$	$\hat{y} = b_0 + b_1 x$	$b_1$ measures the change in $\hat{y}$ when $x$ increases by one unit.
$\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$	$\hat{y} = \exp(b_0 + b_1 \ln(x) + s_e^2/2)$	$b_1$ measures the approximate percentage change in $\hat{y}$ when $x$ increases by 1%.
$y = \beta_0 + \beta_1 \ln(x) + \varepsilon$	$\hat{y} = b_0 + b_1 \ln(x)$	$b_1/100$ measures the approximate change in $\hat{y}$ when $x$ increases by 1%.
$\ln(y) = \beta_0 + \beta_1 x + \varepsilon$	$\hat{y} = \exp(b_0 + b_1 x + s_e^2/2)$	$b_1 \times 100$ measures the approximate percentage change in $\hat{y}$ when $x$ increases by one unit.

It is advisable to use unrounded coefficients (or rounded to at least four decimal places) for making predictions.

## EXAMPLE 16.8

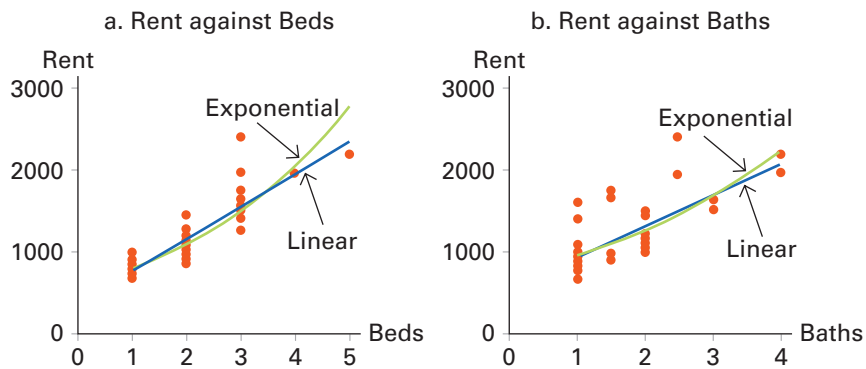
The objective outlined in the introductory case was to evaluate the influence of the number of bedrooms (Beds), the number of bathrooms (Baths), and the square footage (Sqft) on monthly rent (Rent). Use the Ann Arbor rental data in Table 16.1 to answer the following questions.

- Plot rent against each of the three explanatory variables and evaluate whether the relationship is best captured by a line or a curve. Identify variables that may require a log-transformation.
- Estimate the linear and the relevant log models to predict rent for a 1,600-square-foot home with three bedrooms and two bathrooms.

**SOLUTION:** Given the nature of Beds and Baths, we will specify these variables only in regular units. We will, however, consider log-transformations for Rent and Sqft, since their changes are often expressed in percentages.

- In Figure 16.9, we plot Rent against (a) Beds and (b) Baths and superimpose linear and exponential curves (recall that an exponential model log-transforms only the response variable).

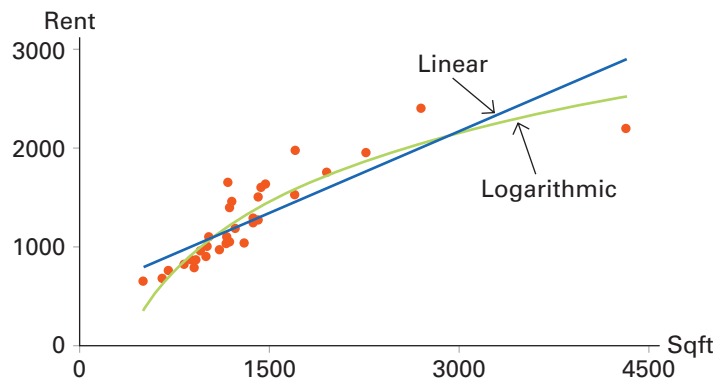
**FIGURE 16.9** Comparing Rent against (a) Beds and (b) Baths



It is hard to tell from Figure 16.9 whether the relationship between Rent and Beds or Rent and Baths is better captured by a line or a curve. We will use formal numerical measures for the selection.

We now plot Rent against Sqft in Figure 16.10.

**FIGURE 16.10** Comparing Rent against Sqft



Here it appears that the relationship between Rent and Sqft is better captured by a curve than a line. Figure 16.10 shows that a logarithmic model that log-transforms Sqft fits the data better than the linear model, suggesting that as square

footage increases, rent increases at a decreasing rate. In other words, the increase in Rent is higher when Sqft increases from 1,000 to 2,000 than from 2,000 to 3,000. Two other models worth considering are the exponential model, where only Rent is log-transformed, and a log-log model, where both Rent and Sqft are log-transformed. In order to avoid a “cluttered” figure, these curves are not superimposed on the scatterplot; however, we will formally evaluate all models.

- b. While the preceding visual tools are instructive, we evaluate four models and use numerical measures to select the most appropriate model for prediction.

$$\text{Model 1: Rent} = \beta_0 + \beta_1 \text{Beds} + \beta_2 \text{Baths} + \beta_3 \text{Sqft} + \varepsilon$$

$$\text{Model 2: Rent} = \beta_0 + \beta_1 \text{Beds} + \beta_2 \text{Baths} + \beta_3 \ln(\text{Sqft}) + \varepsilon$$

$$\text{Model 3: } \ln(\text{Rent}) = \beta_0 + \beta_1 \text{Beds} + \beta_2 \text{Baths} + \beta_3 \text{Sqft} + \varepsilon$$

$$\text{Model 4: } \ln(\text{Rent}) = \beta_0 + \beta_1 \text{Beds} + \beta_2 \text{Baths} + \beta_3 \ln(\text{Sqft}) + \varepsilon$$

In order to estimate these models, we first log-transform Rent and Sqft; see the last two columns of Table 16.9.

**TABLE 16.9** Transforming Rent and Sqft into Logs

Rent	Beds	Baths	Sqft	ln(Rent)	ln(Sqft)
645	1	1	500	6.4693	6.2146
675	1	1	648	6.5147	6.4739
⋮	⋮	⋮	⋮	⋮	⋮
2400	3	2.5	2700	7.7832	7.9010

In Models 1 and 2, we use Rent as the response variable with Beds and Baths, along with Sqft in Model 1 and ln(Sqft) in Model 2, as the explanatory variables. Similarly, in Models 3 and 4, we use ln(Rent) as the response variable with Beds and Baths, along with Sqft in Model 3 and ln(Sqft) in Model 4, as the explanatory variables. Model estimates are summarized in Table 16.10.

**TABLE 16.10** Regression Results for Example 16.8

	Response Variable: Rent		Response Variable: ln(Rent)	
	Model 1	Model 2	Model 3	Model 4
Intercept	300.4116* (0.00)	−3,909.7415* (0.00)	6.3294* (0.00)	3.3808* (0.00)
Beds	225.8100* (0.00)	131.7781* (0.04)	0.2262* (0.00)	0.1246* (0.01)
Baths	89.2661 (0.12)	36.4255 (0.49)	0.0831 (0.06)	0.0254 (0.51)
Sqft	0.2096* (0.03)	NA	0.0001 (0.36)	NA
ln(Sqft)	NA	675.2648* (0.00)	NA	0.4742* (0.00)
$s_e$	193.1591	172.2711	0.1479	0.1262
$R^2$	0.8092	0.8482	0.8095	0.8613

NOTES: Parameter estimates are followed with the  $p$ -values in parentheses; NA denotes not applicable; \* represents significance at the 5% level.

For the most part, the number of bedrooms and the square footage of the house are statistically significant at the 5% level, while the number of bathrooms is insignificant. We use the model results to predict rent for a 1,600-square-foot home with three bedrooms and two bathrooms. In order to make a prediction with Models 3 and 4, which are both based on ln(Rent), we will add the correction term  $s_e^2/2$ .

$$\begin{aligned} \text{Model 1: } \widehat{\text{Rent}} &= 300.4116 + 225.8100(3) + 89.2661(2) + 0.2096(1600) \\ &= 1492 \end{aligned}$$

$$\text{Model 2: } \widehat{\text{Rent}} = -3909.7415 + 131.7781(3) + 36.4255(2) + 675.2648 \times \ln(1600) = 1540$$

$$\text{Model 3: } \widehat{\text{Rent}} = \exp(6.3294 + 0.2262(3) + 0.0831(2) + 0.0001(1600) + 0.1479^2/2) = 1549$$

$$\text{Model 4: } \widehat{\text{Rent}} = \exp(3.3808 + 0.1246(3) + 0.0254(2) + 0.4742 \times \ln(1600) + 0.1262^2/2) = 1498$$

The predicted rent ranges from \$1,492 in Model 1 to \$1,549 in Model 3. We would like to know which model provides the best prediction, as we discuss next.

### LO 16.3

Describe the method used to compare linear with log-transformed models.

## Comparing Linear and Log-Transformed Models

As seen in Example 16.8, it is often not clear which regression model is best suited for an application. While we can use economic intuition and scatterplots for direction, we also justify our selection on the basis of numerical measures. In Chapter 14, we introduced  $R^2$  to compare models based on the same number of explanatory variables; we compared adjusted  $R^2$  if the number of explanatory variables was different. Such comparisons are valid only when the response variable of the competing models is the same. Since  $R^2$  measures the percentage of sample variations of the response variable explained by the model, we cannot compare the percentage of explained variations of  $y$  with that of  $\ln(y)$ . Comparing models based on the computer-generated  $R^2$  is like comparing apples with oranges. For a valid comparison, we need to compute the percentage of explained variations of  $y$  even though the estimated model uses  $\ln(y)$  as the response variable. To do this, it will help to revisit the formula for calculating  $R^2$ . Recall that an easy way to compute  $R^2$  is by squaring the sample correlation coefficient of  $y$  and  $\hat{y}$ .

### REVISITING THE CALCULATION OF THE COEFFICIENT OF DETERMINATION, $R^2$

The coefficient of determination,  $R^2$ , can be computed as  $R^2 = (r_{y\hat{y}})^2$  where  $r_{y\hat{y}}$  is the sample correlation coefficient between  $y$  and  $\hat{y}$ .

Example 16.9 elaborates on the method.

### EXAMPLE 16.9

Revisit the four regression models in Example 16.8 and determine which model is best suited for making predictions.

**SOLUTION:** From Table 16.10, Model 4 has the highest computer-generated  $R^2$  value of 0.8613. However, this does not mean that Model 4 is necessarily the best, since  $R^2$  is based on Rent for Models 1 and 2 and on  $\ln(\text{Rent})$  for Models 3 and 4. Therefore, while we can infer that Model 2 is superior to Model 1 ( $0.8482 > 0.8092$ ) and Model 4 is superior to Model 3 ( $0.8613 > 0.8095$ ), we cannot directly compare Models 2 and 4 based on the computer-generated  $R^2$ . For a valid comparison, we compute  $R^2$  for Model 4 from scratch; that is,  $R^2$  is based on  $y$ , even though it uses  $\ln(y)$  for estimation.

For Model 4, we can compute  $\widehat{\text{Rent}} = \exp(b_0 + b_1\text{Beds} + b_2\text{Baths} + b_3\ln(\text{Sqft}) + s_e^2/2)$  for the given sample values of the explanatory variables. For example, for the first sample observation, with Beds = 1, Baths = 1, and Sqft = 500, the predicted rent is computed as

$$\widehat{\text{Rent}} = \exp(3.3808 + 0.1246(1) + 0.0254(1) + 0.4742 \times \ln(500) + 0.1262^2/2) = \$656.$$

Excel is useful in performing these calculations—it provides the predicted values for the  $\ln(\widehat{\text{Rent}})$  if we check *Residuals* in the *Regression* dialog box. Since Excel provides  $\ln(\widehat{\text{Rent}})$  for Model 4, we can easily compute the predicted rent as  $\widehat{\text{Rent}} = \exp(\ln(\widehat{\text{Rent}}) + s_e^2/2)$ . In Table 16.11, we present a portion of these calculations, using  $y$  to represent Rent; we used unrounded values in these calculations.

**TABLE 16.11** Predicted Rent for Model 4

$y$	$\ln(\widehat{y})$	$\hat{y} = \exp(\ln(\widehat{y}) + 0.1262^2/2)$
645	6.4778	655.7334
675	6.6007	742.5210
$\vdots$	$\vdots$	$\vdots$
2400	7.5648	1944.5760

We use Excel's CORREL function to calculate the correlation between  $y$  and  $\hat{y}$  as  $r_{y\hat{y}} = 0.8691$ . (The values for  $y$  and  $\hat{y}$  are shown in columns 1 and 3 of Table 16.11.) We square the sample correlation coefficient to calculate the coefficient of determination,  $R^2 = (0.8691)^2 = 0.7553$ . We can now compare this value with the computer-generated value for Model 2. We conclude that Model 2 is better suited for making predictions, since  $0.8482 > 0.7553$ .

## SYNOPSIS OF INTRODUCTORY CASE

The recession-resistance of campus towns has prompted many analysts to call investment in off-campus student housing a smart choice (*The Wall Street Journal*, September 24, 2010). First, there is a stable source of demand in college towns, as cash-strapped public universities are unable to house all students. Second, this demand may actually improve due to a projected increase in college enrollment. In this study, Ann Arbor, which is home to the main campus of the University of Michigan, is used to study rental opportunities. Four regression models analyze the monthly rent (Rent) on the basis of the number of bedrooms (Beds), the number of bathrooms (Baths), and the square footage (Sqft) of off-campus houses. Nonlinearities between the variables are captured by transforming Rent and/or Sqft into natural logs. The coefficient of determination  $R^2$ , computed in the original units, is used to select the best model. The selected model is estimated as  $\widehat{\text{Rent}} = -3,909.74 + 131.78\text{Beds} + 36.43\text{Baths} + 675.26\ln(\text{Sqft})$ . The bedroom coefficient implies that for every additional bedroom, the monthly rent is predicted to go up by about \$132, holding other factors constant. Similarly, for every 1% increase in square footage, the monthly rent is predicted to increase by about \$6.75 ( $675/100$ ). This sample regression model can also be used to make predictions for rent. For example, a 1,000-square-foot house with two bedrooms and one bathroom is predicted to rent for \$1,055. Similarly, a 1,600-square-foot house with three bedrooms and two bathrooms is predicted to rent for \$1,540. These results are useful to any investor interested in off-campus housing in Ann Arbor.



## EXERCISES 16.2

### Mechanics

12. Consider the following four estimated models:

$$\hat{y} = 500 - 4.2x$$

$$\hat{y} = 1370 - 280 \ln(x)$$

$$\ln(\hat{y}) = 8.4 - 0.04x; s_e = 0.13$$

$$\ln(\hat{y}) = 8 - 0.8 \ln(x); s_e = 0.11$$

- Interpret the slope coefficient in each of these estimated models.
  - For each model, what is the predicted change in  $y$  when  $x$  increases by 1%, from 100 to 101?
13. Consider the following estimated models:
- $$\hat{y} = 10 + 4.4x$$
- $$\hat{y} = 2 + 23 \ln(x)$$
- $$\ln(\hat{y}) = 3.0 + 0.1x; s_e = 0.07$$
- $$\ln(\hat{y}) = 2.6 + 0.6 \ln(x); s_e = 0.05$$
- Interpret the slope coefficient in each of these estimated models.
  - For each model, what is the predicted change in  $y$  when  $x$  increases by 5%, from 10 to 10.5?
14. Consider the sample regressions for the linear, the logarithmic, the exponential, and the log-log models. For each of the estimated models, predict  $y$  when  $x$  equals 100.

	Response Variable: $y$		Response Variable: $\ln(y)$	
	Model 1	Model 2	Model 3	Model 4
Intercept	240.42	-69.75	1.58	0.77
$x$	4.68	NA	0.05	NA
$\ln(x)$	NA	162.51	NA	1.25
$s_e$	83.19	90.71	0.12	0.09

15. Consider the sample regressions for the linear, the logarithmic, the exponential, and the log-log models. For each of the estimated models, predict  $y$  when  $x$  equals 50.

	Response Variable: $y$		Response Variable: $\ln(y)$	
	Model 1	Model 2	Model 3	Model 4
Intercept	18.52	-6.74	1.48	1.02
$x$	1.68	NA	0.06	NA
$\ln(x)$	NA	29.96	NA	0.96
$s_e$	23.92	19.71	0.12	0.10

16. Consider the following 10 observations of  $y$  and  $x$ .

$y$	22	10	23	24	8	19	20	21	23	22
$x$	12	5	15	16	4	8	11	12	18	16

- Plot the accompanying data to choose between the linear and logarithmic models.
- Justify your choice using the appropriate numerical measure.
- Use the selected model to predict  $y$  for  $x = 10$ .

17. Consider the following 10 observations of  $y$  and  $x$ .

$y$	34	8	12	23	11	27	18	11	11	21
$x$	22	2	11	19	2	21	18	11	19	20

- Plot the above data to choose between the linear and the exponential model.
- Justify your choice using  $R^2$  defined in terms of  $y$ .
- With the best-fitting model, predict  $y$  for  $x = 20$ .

### Applications

18. An economist is interested in examining how an individual's cigarette consumption ( $C$ ) may be influenced by the price for a pack of cigarettes ( $P$ ) and the individual's annual income ( $I$ ). Using data from 50 individuals, she estimates a log-log model and obtains the following regression results.

$$\ln(\hat{C}) = 3.90 - 1.25 \ln(P) + 0.18 \ln(I)$$

$$p\text{-values} = (0.0000) \quad (0.0045) \quad (0.3996)$$

- Interpret the value of the elasticity of demand for cigarettes with respect to price.
  - At the 5% significance level, is the price elasticity of demand statistically significant?
  - Interpret the value of the income elasticity of demand for cigarettes.
  - At the 5% significance level, is the income elasticity of demand statistically significant? Is this result surprising? Explain.
19. **FILE BMI.** According to the *World Health Organization*, obesity has reached epidemic proportions globally. While obesity has generally been linked with chronic disease and disability, researchers argue that it may also affect wages. Body Mass Index (BMI) is a widely used weight measure that also adjusts for height. A person is considered normal weight if BMI is between 18.5 to 25, overweight if BMI is between 25 to 30, and obese if BMI is over 30. The accompanying table shows a portion of data on the salary of 30 college-educated men with their respective BMI.

Salary (in \$1,000s)	BMI
34	33
43	26
:	:
45	21



- Estimate a linear model with salary as the response variable and BMI as the explanatory variable. What is the estimated salary of a college-educated man with a BMI of 25? With a BMI of 30?
  - Estimate an exponential model using log of salary as the response variable and BMI as the explanatory variable. What is the estimated salary of a college-educated man with a BMI of 25? With a BMI of 30?
  - Which of the above two models is more appropriate for this application? Use  $R^2$  for comparison.
20. **FILE** **Dexterity.** A manufacturing manager uses a dexterity test on 20 current employees in order to predict watch production based on time to completion (in seconds). A portion of the data is shown below.

Time (seconds)	Watches per Shift
513	23
608	19
⋮	⋮
437	20

- Estimate the linear model:  $\text{Watches} = \beta_0 + \beta_1 \text{Time} + \varepsilon$ . Interpret the slope coefficient. If the time required to complete the dexterity test is 550 seconds, what is the predicted watch production?
  - Estimate the logarithmic model:  $\text{Watches} = \beta_0 + \beta_1 \ln(\text{Time}) + \varepsilon$ . Interpret the slope coefficient. If the time required to complete the dexterity test is 550 seconds, what is the predicted watch production?
  - Which model provides a better fit? Explain.
21. **FILE** **Wine Pricing.** Professor Orley Ashenfelter of Princeton University is a pioneer in the field of wine economics. He claims that, contrary to old orthodoxy, the quality of wine can be explained mostly in terms of weather conditions. Wine romantics accuse him of undermining the whole wine-tasting culture. In an interesting co-authored paper that appeared in *Chance* magazine in 1995, he ran a multiple regression model where quality, measured by the prices that wines fetch at auctions, is used as the response variable  $y$ . The explanatory variables used in the analysis were the average temperature in Celsius  $x_1$ , the amount of winter rain  $x_2$ , the amount of harvest rain  $x_3$ , and the years since vintage  $x_4$ . A portion of the data is shown in the accompanying table.

Price	Temperature	Winter Rain	Harvest Rain	Vintage
1.4448	17.12	600	160	31
1.8870	16.73	690	80	30
⋮	⋮	⋮	⋮	⋮
1.1457	16.00	578	74	3

SOURCE: [www.liquidasset.com](http://www.liquidasset.com).

- Estimate the linear model:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ . What is the predicted price if  $x_1 = 16$ ,  $x_2 = 600$ ,  $x_3 = 120$ , and  $x_4 = 20$ ?
  - Estimate the exponential model:  $\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ . What is the predicted price if  $x_1 = 16$ ,  $x_2 = 600$ ,  $x_3 = 120$ , and  $x_4 = 20$ ?
  - Use  $R^2$  to select the appropriate model for prediction.
22. **FILE** **Electricity Cost.** The facility manager at a pharmaceutical company wants to build a regression model to forecast monthly electricity cost (Cost). Three main variables are thought to influence electricity cost: (1) average outdoor temperature (Temp), (2) working days per month (Days), and (3) tons of product produced (Tons). A portion of the past year's monthly data is shown in the accompanying table.

Cost (\$)	Temp (°F)	Days	Tons
24100	26	24	80
23700	32	21	73
⋮	⋮	⋮	⋮
26000	39	22	69

- Estimate the linear model:  $\text{Cost} = \beta_0 + \beta_1 \text{Temp} + \beta_2 \text{Days} + \beta_3 \text{Tons} + \varepsilon$ . What is the predicted electricity cost in a month during which the average outdoor temperature is 65°, there are 23 working days, and 76 tons are produced?
  - Estimate the exponential model:  $\ln(\text{Cost}) = \beta_0 + \beta_1 \text{Temp} + \beta_2 \text{Days} + \beta_3 \text{Tons} + \varepsilon$ . What is the predicted electricity cost in a month during which the average outdoor temperature is 65°, there are 23 working days, and 76 tons are produced?
  - Based on  $R^2$ , which model provides the better fit?
23. **FILE** **Davis Rental.** Chad Dobson has heard about the positive outlook for real estate investment in college towns. He is interested in investing in Davis, California, which houses one of the University of California campuses. He uses [zillow.com](http://zillow.com) to access data on 2011 monthly rents for 27 houses, along with three characteristics of the home: number of bedrooms (Beds), number of bathrooms (Baths), and square footage (Sqft). A portion of the data is shown in the accompanying table.

Rent	Beds	Baths	Sqft
2950	4	4	1453
2400	4	2	1476
⋮	⋮	⋮	⋮
744	2	1	930

SOURCE: [www.zillow.com](http://www.zillow.com).

- Estimate a linear model that uses rent and an exponential model that uses log of rent as the response variable.
  - Compute the predicted rent for a 1,500-square-foot house with three bedrooms and two bathrooms for the linear and the exponential models (ignore the significance tests).
  - Use  $R^2$  to select the appropriate model for prediction.
24. **FILE Life\_Expectancy.** Life expectancy at birth is the average number of years that a person is expected to live. There is a huge variation in life expectancies between countries, with the highest being in Japan, and the lowest in some African countries. An important factor for such variability is the availability of suitable health care. One measure of a person's access to health care is the people-to-physician ratio. We expect life expectancy to be lower for countries where this ratio is high. The accompanying table lists a portion of life expectancy of males and females in 40 countries and their corresponding people-to-physician ratio.

Country	Male Life Expectancy	Female Life Expectancy	People/Physician
Argentina	67	74	370
Bangladesh	54	53	6,166
⋮	⋮	⋮	⋮
Zaire	52	56	23,193

SOURCE: *The World Almanac and Book Facts*, 1993.

- Construct a scatterplot of female life expectancy against the people-to-physician ratio. Superimpose a linear trend and a logarithmic trend to determine the appropriate model.
- Estimate a simple linear regression model with life expectancy of females as the response variable and the people-to-physician ratio as the explanatory variable. What happens to life expectancy of females as the people-to-physician ratio decreases from 1,000 to 500?

- Estimate a logarithmic regression model with the natural log of the people-to-physician ratio as the explanatory variable. What happens to the life expectancy of females as the people-to-physician ratio decreases from 1,000 to 500?
  - Use  $R^2$  to determine which of the preceding two models is more appropriate.
25. **FILE Life\_Expectancy.** Use the data in Exercise 24 to answer the same four questions regarding life expectancy of males. Who is more likely to benefit from adding more physicians to the population? Explain.
26. **FILE Production\_Function.** Economists often examine the relationship between the inputs of a production function and the resulting output. A common way of modeling this relationship is referred to as the Cobb–Douglas production function. This function can be expressed as  $\ln(Q) = \beta_0 + \beta_1 \ln(L) + \beta_2 \ln(K) + \varepsilon$ , where  $Q$  stands for output,  $L$  for labor, and  $K$  for capital. The accompanying table lists a portion of data relating to the U.S. agricultural industry in the year 2004.

State	Output	Labor	Capital
AL	3.1973	2.7682	3.1315
AR	7.7006	4.9278	4.7961
⋮	⋮	⋮	⋮
WY	1.2993	1.6525	1.5206

SOURCE: [www.ers.usda.gov/Data/AgProductivity](http://www.ers.usda.gov/Data/AgProductivity); see Tables 3, 8, 10. Values in table are indices.

Estimate  $\ln(Q) = \beta_0 + \beta_1 \ln(L) + \beta_2 \ln(K) + \varepsilon$ .

- What is the predicted change in output if labor increases by 1%, holding capital constant?
- Holding capital constant, can we conclude at the 5% level that a 1% increase in labor will increase the output by more than 0.5%?

## WRITING WITH STATISTICS



Numerous attempts have been made to relate happiness to various factors. Since there is no unique way to quantify happiness, researchers generally rely on surveys to capture a subjective assessment of well-being. One recent study relates happiness with age and finds that holding everything else constant, people seem to be least happy when they are in their mid- to upper-40s (*The Economist*, December 16, 2010). Perhaps with greater age comes maturity that contributes to a better sense of overall well-being. With regard to the influence of money, a study from Princeton University's Woodrow Wilson School suggests that

money does buy happiness, but its effect diminishes as incomes rise above \$75,000 a year (*Time Magazine*, September 6, 2010). Perhaps people do not need more than \$75,000 to do what matters most to their emotional well-being, such as spending time with friends and family and meeting their basic food, health, and leisure needs. Nick Fisher is a young business school graduate who is fascinated by these reports. He decides to collect his own data to better comprehend and also verify the results of these studies. He surveys working adults in his hometown and inputs information on the respondent's self-assessed happiness on a scale of 0 to 100, along with age and family income. A portion of the data is shown in Table 16.12.

**TABLE 16.12** Happiness, Age, and Income Data,  $n = 100$ .

Respondent	Happiness	Age	Family Income
1	69	49	52000
2	83	47	123000
$\vdots$	$\vdots$	$\vdots$	$\vdots$
100	79	31	105000

**FILE**  
Happiness

Nick would like to use the above sample information to:

1. Find the appropriate functional form to capture the influence of age and family income on happiness.
2. With a family income of \$80,000, calculate happiness associated with varying levels of age.
3. For a 60-year-old working adult, compute happiness associated with varying levels of family income.

In a survey of 100 working adults, respondents were asked to report their age and family income, as well as rate their happiness on a scale of 0 to 100. This report summarizes the analysis of several regression models that examine the influence of age and income on the perceived happiness of respondents. The models used various transformations to capture interesting nonlinearities suggested by recent research reports. For example, one such report shows that people get happier as they get older, despite the fact that old age is associated with a loss of hearing, vision, and muscle tone (*The New York Times*, May 31, 2010). In addition, while people start out feeling pretty good about themselves in their 20s, their self-assessed happiness deteriorates until around age 50 and then improves steadily thereafter. In order to quantify this possible quadratic effect, both age and age-squared variables are used for the regression. Also, the natural log of income is considered in order to capture the possible diminishing effect on happiness of incomes above \$75,000 (*Time Magazine*, September 6, 2010). The results of the various regression models are summarized in Table 16.A.

**TABLE 16.A** Regression Results

	Model 1	Model 2	Model 3	Model 4
Intercept	49.1938* (0.00)	118.5285* (0.00)	-81.0939* (0.00)	-13.3021 (0.39)
Age	0.2212* (0.00)	-2.4859* (0.00)	0.2309* (0.00)	-2.4296* (0.00)
Age-squared	NA	0.0245* (0.00)	NA	0.0241* (0.00)
Income	0.0001* (0.00)	0.0001* (0.00)	NA	NA
ln(Income)	NA	NA	12.6761* (0.00)	12.7210* (0.00)
Adjusted $R^2$	0.4863	0.6638	0.5191	0.6907

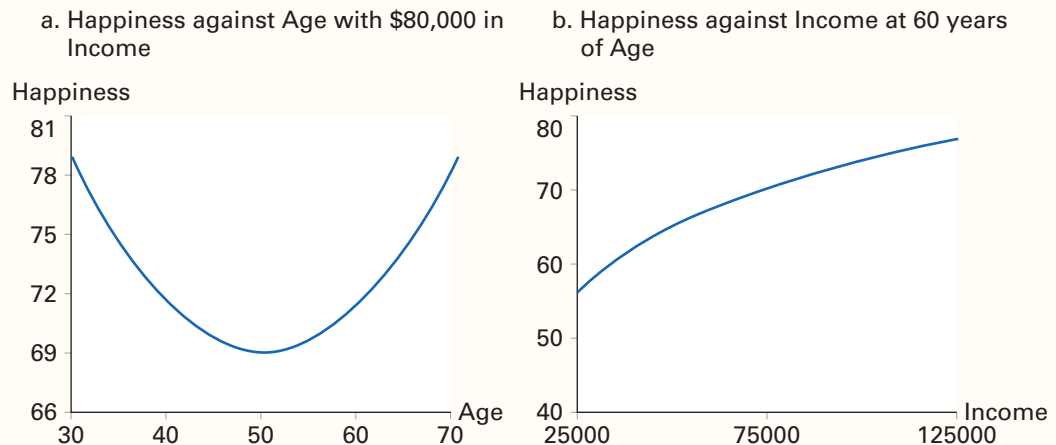
NOTES: Parameter estimates are in the top portion of the table with the  $p$ -values in parentheses; NA denotes not applicable; \* represents significance at the 5% level. The last row presents the adjusted  $R^2$  values for model comparison.

## Sample Report— Understanding Happiness

Model 4 was selected as the most appropriate model because it has the highest adjusted  $R^2$  value of 0.6907. The estimated parameters of this model were used to make predictions. For instance, with family income equal to \$80,000, the predicted happiness for a 30-, 50-, and 70-year-old is 79.09, 69.00, and 78.17, respectively. Note that these results are consistent with those suggesting that happiness first decreases and then increases with age. Specifically, using the estimated coefficients for Age, a person is least happy at 50.4 years of age. These results are shown graphically in Figure 16.A(a) where Happiness is plotted against Age, holding Income fixed at \$80,000.

The regression results were also used to analyze the income effect. For instance, for a 60-year-old, the predicted happiness with family income of \$50,000, \$75,000, and \$100,000 is 65.20, 70.36, and 74.02, respectively. Note that there is a greater increase in Happiness when income increases from \$50,000 to \$75,000 than when it increases from \$75,000 to \$100,000. These results are shown in Figure 16.A(b) where predicted Happiness is plotted against Income, holding Age fixed at 60 years. Overall, the results support recent research findings.

**FIGURE 16.A** Predicted Happiness using Model 4 regression results



## CONCEPTUAL REVIEW

### LO 16.1 Use and evaluate polynomial regression models.

In a **quadratic regression model**,  $y = \beta_0 + \beta_1x + \beta_2x^2 + \varepsilon$ , the sign of the coefficient  $\beta_2$  determines whether the relationship between  $x$  and  $E(y)$  is U-shaped ( $\beta_2 > 0$ ) or inverted U-shaped ( $\beta_2 < 0$ ). **Predictions** are made by  $\hat{y} = b_0 + b_1x + b_2x^2$ .

In a quadratic regression model, the marginal effect of  $x$  on  $\hat{y}$  is approximated by  $b_1 + 2b_2x$ ; so this effect depends on the value of  $x$ . The quadratic equation reaches a maximum (if  $b_2 < 0$ ) or minimum (if  $b_2 > 0$ ) at  $x = \frac{-b_1}{2b_2}$ .

A **cubic regression model**,  $y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \varepsilon$ , allows two sign changes of the slope capturing the influence of  $x$  on  $E(y)$ . **Predictions** are made by  $\hat{y} = b_0 + b_1x + b_2x^2 + b_3x^3$ .

It is advisable to use unrounded coefficients (or rounded to at least four decimal places) for making predictions. We compare polynomial regression models of various orders on the basis of adjusted  $R^2$ .

### LO 16.2 Use and evaluate log-transformed models.

Many interesting nonlinear relationships can be captured by transforming the response and/or the explanatory variables into natural logs. These regression models are summarized as follows:

In a **log-log model**,  $\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$ , the slope coefficient  $\beta_1$  measures the approximate percentage change in  $E(y)$  when  $x$  increases by 1%. **Predictions** are made by  $\hat{y} = \exp(b_0 + b_1 \ln(x) + s_e^2/2)$ , where  $b_0$  and  $b_1$  are the coefficient estimates and  $s_e$  is the standard error of the estimate.

In a **logarithmic model**,  $y = \beta_0 + \beta_1 \ln(x) + \varepsilon$ ,  $\beta_1/100$  measures the approximate change in  $E(y)$  when  $x$  increases by 1%. **Predictions** are made by  $\hat{y} = b_0 + b_1 \ln(x)$  where  $b_0$  and  $b_1$  are the coefficient estimates.

In an **exponential model**,  $\ln(y) = \beta_0 + \beta_1 x + \varepsilon$ ,  $\beta_1 \times 100$  measures the approximate percentage change in  $E(y)$  when  $x$  increases by one unit. **Predictions** are made by  $\hat{y} = \exp(b_0 + b_1 x + s_e^2/2)$ , where  $b_0$  and  $b_1$  are the coefficient estimates and  $s_e$  is the standard error of the estimate.

It is advisable to use unrounded coefficients (or rounded to at least four decimal places) for making predictions.

**LO 16.3** Describe the method used to compare linear with log-transformed models.

We use the coefficient of determination  $R^2$  to compare models that employ the same number of explanatory variables and use adjusted  $R^2$  if the number of explanatory variables differs. Such comparisons are valid only when the response variable of the competing models is the same. In other words, we cannot compare the percentage of explained variations of  $y$  with that of  $\ln(y)$ . For a valid comparison, for any model that uses  $\ln(y)$  as the response variable, we compute  $R^2$  as  $R^2 = (r_{y\hat{y}})^2$ , where  $r_{y\hat{y}}$  is the sample correlation coefficient between  $y$  and  $\hat{y}$ .

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

27. **FILE** *Quarterback\_Salaries*. A sports enthusiast wants to examine the factors that influence a quarterback's salary (Salary). In particular, he wants to assess the influence of the pass completion rate (PC), the total touchdowns scored (TD), and a quarterback's age (Age) on Salary. He uses 2009 data, a portion of which is shown in the accompanying table.

Name	Salary (in \$ millions)	PC	TD	Age
Philip Rivers	25.5566	65.2	28	27
Jay Cutler	22.0441	60.5	27	26
⋮	⋮	⋮	⋮	⋮
Tony Romo	0.6260	63.1	26	29

SOURCE: USA Today database for salaries; NFL.com for other data.

- Estimate and interpret the model:  $\text{Salary} = \beta_0 + \beta_1 \text{PC} + \beta_2 \text{TD} + \beta_3 \text{Age} + \varepsilon$ . Show that this model is preferable to a model that uses log of salary as the response variable.
- Consider the quadratic effect of Age by adding  $\text{Age}^2$  in the regression. Use a partial  $F$  test to determine the joint statistical significance of Age and  $\text{Age}^2$ .

28. **FILE** *Fertilizer2*. A horticulturist is studying the relationship between tomato plant height and fertilizer amount. Thirty tomato plants grown in similar conditions were subjected to various amounts of fertilizer over a four-month period, and then their heights were measured. A portion of the results is shown in the accompanying table.

Fertilizer Amount (ounces)	Tomato Plant Height (inches)
1.9	20.4
5.0	29.1
⋮	⋮
3.1	36.4

- Estimate the linear regression model:  
 $\text{Height} = \beta_0 + \beta_1 \text{Fertilizer} + \varepsilon$ .
- Estimate the quadratic regression model:  
 $\text{Height} = \beta_0 + \beta_1 \text{Fertilizer} + \beta_2 \text{Fertilizer}^2 + \varepsilon$ . Find the fertilizer amount at which the height reaches a minimum or maximum.
- Use the best-fitting model to predict, after a four-month period, the height of a tomato plant that received 3.0 ounces of fertilizer.



29. **FILE** *Arlington\_Homes*. A realtor examines the factors that influence the price of a house. He collects data on the prices for 36 single-family homes in Arlington, Massachusetts, sold in the first quarter of 2009. For explanatory variables, he uses the house's square footage (Sqft), as well as its number of bedrooms (Beds) and bathrooms (Baths). A portion of the data is shown in the accompanying table.

Price	Sqft	Beds	Baths
840000	2768	4	3.5
822000	2500	4	2.5
⋮	⋮	⋮	⋮
307500	850	1	1

SOURCE: NewEnglandMoves.com.

- Estimate the linear model:  $\text{Price} = \beta_0 + \beta_1 \text{Sqft} + \beta_2 \text{Beds} + \beta_3 \text{Baths} + \varepsilon$ . Estimate the exponential model:  $\ln(\text{Price}) = \beta_0 + \beta_1 \text{Sqft} + \beta_2 \text{Beds} + \beta_3 \text{Baths} + \varepsilon$ .
  - Interpret the slope coefficients of the estimated models.
  - Use the coefficient of determination to choose the preferred model.
30. **FILE** *Circuit\_Boards*. The operators manager at an electronics company believes that the time required for workers to build a circuit board is not necessarily proportional to the number of parts on the board. He wants to develop a regression model to predict time based on part quantity. He has collected data for the last 25 boards. A portion of this data is shown in the accompanying table.

Time (minutes)	Parts
30.8	62
9.8	32
⋮	⋮
29.8	60

- Estimate the linear regression model to predict time as a function of number of parts (Parts). Then estimate the quadratic regression model to predict time as a function of Parts and Parts squared. Provide an intuitive justification for the quadratic term.
- Evaluate the two models in terms of variable significance ( $\alpha = 0.05$ ) and adjusted  $R^2$ .
- Use the best-fitting model to predict how long it would take to build a circuit board consisting of 48 parts.

31. **FILE** *Smoking*. A nutritionist wants to understand the influence of income and healthy food on the incidence of smoking. He collects 2009 data on the percentage of smokers in each state in the U.S. and the corresponding median income and the percentage of the population that regularly eats fruits and vegetables. A portion of the data is shown in the accompanying table.

State	Smoke (%)	Fruits/ Vegetables (%)	Median Income
AK	14.6	23.3	61604
AL	16.4	20.3	39980
⋮	⋮	⋮	⋮
WY	15.2	23.3	52470

SOURCE: Centers for Disease Control and Prevention and U.S. Census Bureau.

- Estimate:  $\text{Smoke} = \beta_0 + \beta_1 \text{Fruits/Vegetables} + \beta_2 \text{Median Income} + \varepsilon$ .
  - Compare this model with a model that log-transforms the median income variable.
32. **FILE** *Savings\_Rate*. The accompanying table shows a portion of the monthly data on the personal savings rate (Savings) and personal disposable income (Income) in the U.S. from January 2007 to November 2010.

Date	Savings (%)	Income (\$ billions)
2007-01	2.2	10198.2
2007-02	2.3	10252.9
⋮	⋮	⋮
2010-11	5.5	11511.9

SOURCE: Bureau of Economic Analysis.

- Compare the linear model,  $\text{Savings} = \beta_0 + \beta_1 \text{Income} + \varepsilon$ , with a log-log model,  $\ln(\text{Savings}) = \beta_0 + \beta_1 \ln(\text{Income}) + \varepsilon$ .
  - Interpret the estimated slope coefficient of both models.
  - Which is the preferred model? Explain.
33. **FILE** *Inventory\_Cost*. The inventory manager at a warehouse distributor wants to predict inventory cost based on order quantity. She thinks it may be a nonlinear relationship since its two primary components move in opposite directions: (1) order processing cost (costs of procurement personnel, shipping, transportation), which *decreases* as order quantity increases (due to fewer orders needed), and (2) holding cost (costs of capital, facility, warehouse personnel, equipment), which *increases* as order quantity increases (due to more inventory held). She has collected monthly



inventory costs and order quantities for the past 36 months. A portion of the data is shown in the accompanying table.

Inventory Cost (\$1,000s)	Order Quantity (units)
844	54.4
503	52.1
⋮	⋮
870	55.5

- Create a scatterplot of inventory cost as a function of quantity. Superimpose a linear trendline and quadratic trendline.
  - Estimate the linear regression model to predict inventory cost as a function of order quantity. Then estimate the quadratic regression model to predict inventory cost as a function of order quantity and order quantity squared.
  - Evaluate the two models in terms of significance tests ( $\alpha = 0.05$ ) and adjusted  $R^2$ .
  - Use the best-fitting model to predict monthly inventory cost for an order quantity of 800 units.
34. **FILE Learning Curve.** Learning curves are used in production operations to estimate the

time required to complete a repetitive task as an operator gains experience. Suppose a production manager has compiled 30 time values for a particular operator as she progressed down the learning curve during the first 100 units. A portion of this data is shown in the accompanying table.

Unit Number	Time per Unit (minutes)
3	18.30
5	17.50
⋮	⋮
100	5.60

- Create a scatterplot of time per unit against units built. Superimpose a linear trendline and a logarithmic trendline to determine visually the best-fitting model.
- Estimate a simple linear regression model and a logarithmic regression model for explaining time per unit using unit number as the explanatory variable.
- Based on  $R^2$ , use the best-fitting model to predict the time that was required for the operator to build Unit 50.

## CASE STUDIES

**CASE STUDY 16.1** Executive compensation has risen dramatically beyond the rising levels of an average worker's wage over the years. This has been a hot topic for discussion, especially with the crisis in the financial sector and the controversy over the federal bailout. The government is even considering a cap on high-flying salaries for executives (*The New York Times*, February 9, 2009). Consider the following portion of data that link total compensation of the 455 highest-paid CEOs in 2006 with two performance measures (industry-adjusted return on assets, ROA, and industry-adjusted stock return) and the firm's size (Total Assets).

**Data for Case Study 16.1** Executive Compensation and Other Factors,  $n = 455$

Compensation (in \$ million)	Adj ROA	Adj Return	Total Assets (in \$ millions)
16.58	2.53	-0.15	20917.5
26.92	1.27	0.57	32659.5
⋮	⋮	⋮	⋮
2.30	0.45	0.75	44875.0

SOURCE: SEC website and Compustat.

**FILE**  
*Exec\_Comp*

In a report, use the sample information to:

1. Estimate two models where each model uses Compensation as the response variable and Adj ROA and Adj Return as the explanatory variables along with Total Assets in Model 1 and natural log of Total Assets in Model 2.
2. Use the preferred model to predict compensation given the average values of the explanatory variables.

**CASE STUDY 16.2** A British survey just revealed that the New York Yankees baseball team pays their players, on average, more than any other team in the world (<http://sportsillustrated.cnn.com>, April 7, 2010). Brendan Connolly, a statistician for a Major League Baseball (MLB) team, wants to elaborate on the salary of baseball players. Excluding pitchers from his analysis, he believes that a baseball player's batting average (BA), runs batted in (RBI), and years of experience playing professional baseball (Experience) are the most important factors that influence a player's salary. Further, he believes that salaries rise with experience only up to a point, beyond which they begin to fall; in other words, experience has a quadratic effect on salaries. Brendan collects data on salary (in \$1,000s), BA, RBI, and experience for 138 outfielders in 2008. A portion of the data is shown in the accompanying table.

**FILE**  
MLB\_Salary

**Data for Case Study 16.2** Major League Baseball Outfielder Data,  $n = 138$

Player	Salary (in \$1,000s)	BA	RBI	Experience
1. Nick Markakis	455	299	87	3
2. Adam Jones	390	261	23	3
⋮	⋮	⋮	⋮	⋮
138. Randy Winn	8875	288	53	11

Notes: All data collected from [usatoday.com](http://usatoday.com) or [espn.com](http://espn.com); BA and RBI are averages over the player's professional life through 2008. For exposition, BA has been multiplied by 1,000.

In a report, use the sample information to:

1. Estimate a quadratic regression model using Salary as the response variable and BA, RBI, Experience, and Experience<sup>2</sup> as the explanatory variables.
2. Compare the above quadratic regression model with a linear model that uses BA, RBI, and Experience as the explanatory variables.

**CASE STUDY 16.3** According to a recent report by the government, new home construction fell to an 18-month low in October, 2010 (CNNMoney.com, November 17, 2010). Housing starts, or the number of new homes being built, experienced an 11.7% drop in the seasonally adjusted annual rate. Beena Singh works for a mortgage company in Madison, Wisconsin. She wants to better understand the quantitative relationship between housing starts, the mortgage rate, and the unemployment rate. She gathers monthly data on these variables from 2006:01–2010:12. A portion of the data is shown in the accompanying table.

**FILE**  
Housing\_Starts

**Data for Case Study 16.3** Housing Starts and Other Factors,  $n = 60$

Date	Housing Starts (in 1,000s)	Mortgage Rate (%)	Unemployment Rate (%)
2006–01	2273	6.15	4.7
2006–02	2119	6.25	4.8
⋮	⋮	⋮	⋮
2010–12	520	4.71	9.4

Source: Census Bureau and Board of Governors.

In a report, use the sample information to:

1. Construct scatterplots to quantify the relationship of housing starts with the mortgage rate and the unemployment rate.
2. Estimate a linear and an exponential regression model and use numerical measures to select the most appropriate model for prediction.
3. Discuss the potential problems of correlated observations in this time series data application.

## APPENDIX 16.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### Estimating Polynomial Regression Models

- A. (Replicating Example 16.3) To create the variable  $\text{Age}^2$ , from the menu select **Calc > Calculator**. After **Store result in variable**, enter  $\text{AgeSq}$ . After **Expression**, select  $\text{Age}$ , select  $*$ , and select  $\text{Age}$ .
- B. Estimate the regression model using the standard commands.

**FILE**  
*Wages*

#### Estimating Logarithmic Regression Models

- A. (Replicating Example 16.8, Model 4) To create the variable  $\ln(\text{Rent})$ , from the menu select **Calc > Calculator**. After **Store result in variable**, enter  $\ln(\text{Rent})$ . Under **Functions**, select **Natural log**, then select  $\text{Rent}$ . Repeat these steps to create the variable  $\ln(\text{Sqft})$ .
- B. Estimate the regression model using the standard commands.

**FILE**  
*AnnArbor*

#### Calculating $R^2$ Based on $y$ Rather than $\ln(y)$

- A. (Replicating Example 16.9) Create the variables  $\ln(\text{Rent})$  and  $\ln(\text{Sqft})$ .
- B. From the menu choose **Stat > Regression > Regression > Fit Regression Model**. After **Responses**, select  $\ln(\text{Rent})$ , and after **Continuous predictors**, select  $\text{Bed}$ ,  $\text{Bath}$ , and  $\ln(\text{Sqft})$ . Choose **Storage**, and then select **Fits**.
- C. From the menu select **Calc > Calculator**. After **Store result in variable**, enter  $\hat{y}$ . After **Expression**, enter  $\text{Exp}(\text{FITS1} + 0.1262 * 0.1262 / 2)$ . (Recall that 0.1262 is the standard error of the estimate.)
- D. From the menu select **Stat > Basic Statistics > Correlation**. Under **Variables**, select  $\text{Rent}$  and  $\hat{y}$ . Square the correlation coefficient 0.869 to obtain  $R^2$ .

**FILE**  
*AnnArbor*

### SPSS

#### Estimating Polynomial Regression Models

- A. (Replicating Example 16.3) To create the variable  $\text{Age}^2$ , from the menu select **Transform > Compute Variables**. Under **Target Variable**, enter

**FILE**  
*Wages*

AgeSqu. In the **Numeric Expression** dialog box, select Age, select \*, and select Age.

- B. Estimate the regression model using the standard commands.

### Estimating Logarithmic Regression Models

**FILE**  
AnnArbor

- A. (Replicating Example 16.8, Model 4) To create the variable  $\ln(\text{Rent})$ , from the menu select **Transform > Compute Variables**. Under **Target Variable**, enter  $\ln(\text{Rent})$ . Under **Function group**, select **Arithmetic**, and under **Functions and Special Variables** double-click on **Ln**. Under **Numeric Expression**, select  $\text{Rent}$ . Repeat these steps to calculate  $\ln(\text{Sqft})$ .
- B. Estimate the regression model using the standard commands.

### Calculating $R^2$ Based on $y$ Rather than $\ln(y)$

**FILE**  
AnnArbor

- A. (Replicating Example 16.9) Create the variables  $\ln(\text{Rent})$  and  $\ln(\text{Sqft})$ .
- B. From the menu select **Analyze > Regression > Linear**.
- C. Under **Dependent**, select  $\ln(\text{Rent})$ , and under **Independent(s)**, select  $\text{Bed}$ ,  $\text{Bath}$ , and  $\ln(\text{Sqft})$ . Choose **Save** and select **Predicted Values – Unstandardized**.
- D. From the menu select **Transform > Compute Variables**. Under **Target Variable**, enter  $\hat{y}$ . Under **Numeric Expression**, input  $\text{EXP}(\text{PRE\_1} + 0.1262 ** 2/2)$ . (Recall that 0.1262 is the standard error.)
- E. From the menu, select **Analyze > Correlate > Bivariate**. Under **Variables**, select  $\text{Rent}$  and  $\hat{y}$ . Square the correlation coefficient 0.869 to obtain  $R^2$ .

## JMP

### Estimating Polynomial Regression Models

**FILE**  
Wages

- A. (Replicating Example 16.3) To create the variable  $\text{Age}^2$ , right-click on a new column and select **New Column**, and label it  $\text{AgeSqu}$ . Right-click on  $\text{AgeSqu}$ , and select **Formula**. Under **Table Columns**, select  $\text{Age}$ , select  $\times$ , and select  $\text{Age}$ .
- B. Estimate the regression model using the standard commands.

### Estimating Logarithmic Regression Models

**FILE**  
AnnArbor

- A. (Replicating Example 16.8, Model 4) To create the variable  $\ln(\text{Rent})$ , right-click on a new column and select **New Column**, and label it  $\ln(\text{Rent})$ . Right-click on  $\ln(\text{Rent})$ , and select **Formula**. Under **Functions (grouped)**, select **Transcendental > Log**, and then select  $\text{Rent}$  in the bracket. Repeat these steps to create the variable  $\ln(\text{Sqft})$ .
- B. Estimate the regression model using the standard commands.

### Calculating $R^2$ Based on $y$ Rather than $\ln(y)$

**FILE**  
AnnArbor

- A. (Replicating Example 16.9) Create the variables  $\ln(\text{Rent})$  and  $\ln(\text{Sqft})$ .
- B. From the menu choose **Analyze > Fit Model**.
- C. Under **Select Columns**, select  $\ln(\text{Rent})$ , and then under **Pick Role Variables** select **Y**. Under **Select Columns**, select  $\text{Bed}$ ,  $\text{Bath}$ , and  $\ln(\text{Sqft})$ , and then under **Construct Model Effect**, select **Add**.
- D. In the red triangle next to **Response  $\ln(\text{Rent})$** , select **Save Columns > Predicted Values**. A new column named Predicted  $\ln(\text{Rent})$  should appear in the JMP spreadsheet.

- E. In the JMP spreadsheet, right-click on a new column, select **New Column**, and enter  $\hat{y}$  as the column name. Right-click on  $\hat{y}$ , input **Formula > Transcendental > Exp**. In the bracket, input  $\text{Predicted } \ln(\text{Rent}) + 0.1262 \times 0.1262/2$ . (Recall that 0.1262 is the standard error.)
- F. From the menu, select **Analyze > Multivariate Methods > Multivariate**.
- G. Under **Select Columns**, select Rent and  $\hat{y}$ , and under **Cast Selected Columns into Roles**, select **Y, Columns**. Square the correlation coefficient 0.869 to obtain  $R^2$ .

# 17

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 17.1 Use dummy variables to represent qualitative explanatory variables.
- LO 17.2 Test for differences between the categories of a qualitative variable.
- LO 17.3 Use dummy variables to capture interactions between qualitative and quantitative explanatory variables.
- LO 17.4 Use a linear probability model to estimate a binary response variable.
- LO 17.5 Use a logit model to estimate a binary response variable.

# Regression Models with Dummy Variables

Up until now, regression analysis has allowed us to answer questions such as: How much does an extra year of education contribute to salary? What is the contribution of advertisement expenditures on sales of electronic goods? Is a student's SAT score a good predictor of his/her college GPA? All of these questions use response and explanatory variables that are quantitative. There are other important applications that use qualitative variables representing two or more categories. For instance, we may want answers to questions such as: Do women get paid as much as men for the same work? Are sales of electronic goods higher in the 4th quarter than in the other quarters? What is the influence of family income on the probability of buying a house? In order to answer these questions, the regression analysis must incorporate qualitative response and/or explanatory variables. This chapter examines these kinds of situations, using methods called dummy variable models and binary choice probability models.





## INTRODUCTORY CASE

### Is There Evidence of Wage Discrimination?

Three female Seton Hall professors recently learned in a court decision that they could pursue their lawsuit alleging that the university paid better salaries to younger instructors and male professors ([www.nj.com](http://www.nj.com), November 23, 2010). Numerous studies have focused on salary differences between men and women, whites and blacks, and young and old. Mary Schweitzer works in the human resources department at a large liberal arts college. After the Seton Hall news, the college asked her to test for both gender and age discrimination in salaries. Mary gathered information on the annual salaries (in \$1,000s) of 42 professors, along with their experience (in years), gender (male or female), and age (under 60 years old or at least 60 years old). A portion of the data is shown in Table 17.1.

**TABLE 17.1** Salary and Other Information on 42 Professors

**FILE**  
Professor

Individual	Salary (in \$1,000s)	Experience (in years)	Gender	Age
1	67.50	14	Male	Under
2	53.51	6	Male	Under
⋮	⋮	⋮	⋮	⋮
42	73.06	35	Female	Over

Mary would like to use the sample information in Table 17.1 to:

1. Determine whether salary differs by a fixed amount between males and females.
2. Determine whether there is evidence of age discrimination in salaries.
3. Determine whether the salary difference between males and females increases with experience.

A synopsis of this case is provided in Section 17.2.

## 17.1 DUMMY VARIABLES

Up until now, the explanatory variables used in the regression applications have been **quantitative**; in other words, they assume meaningful numerical values. For example, in Chapter 14 we used income and unemployment (both quantitative variables) to explain variations in consumer debt. In empirical work, however, it is common to include some explanatory variables that are **qualitative**. Although qualitative variables can be described by several categories, they are commonly described by only two categories. Examples include gender (male or female), homeownership (own or do not own), shipment (rejected or not rejected), and admission (yes or no).

### QUANTITATIVE VERSUS QUALITATIVE VARIABLES IN REGRESSION

Explanatory variables employed in a regression can be **quantitative** or **qualitative**. Quantitative variables assume meaningful numerical values, whereas qualitative variables represent categories.

Given the professor salary data in the introductory case, we can estimate the model as  $\hat{y} = 48.83 + 1.15x$  where  $y$  represents salary (in \$1,000s) and  $x$  is the usual quantitative variable, representing experience (in years). The sample regression equation implies that the predicted salary increases by about \$1,150 ( $1.15 \times 1,000$ ) for every year of experience. Arguably, in addition to experience, variations in salary are also caused by qualitative explanatory variables such as gender (male or female) and age (under or over 60 years).

### Qualitative Variables with Two Categories

A qualitative variable with two categories can be associated with a **dummy variable**, also referred to as an **indicator variable**. A dummy variable  $d$  is defined as a variable that assumes a value of 1 for one of the categories and 0 for the other. For example, in the case of a dummy variable categorizing gender, we can define 1 for males and 0 for females. Alternatively, we can define 1 for females and 0 for males, with no change in inference. Sometimes we define a dummy variable by converting a quantitative variable to a qualitative variable. In the introductory case the qualitative variable age (under or over 60 years) was actually defined from the quantitative variable age. Similarly, in studying teen behavior, we may have access to quantitative information on age, but we can generate a dummy variable that equals 1 for ages between 13 and 19 and 0 otherwise.

### A DUMMY VARIABLE

A **dummy variable**  $d$  is defined as a variable that takes on values of 1 or 0. It is commonly used to describe a qualitative variable with two categories.

#### LO 17.1

Use dummy variables to represent qualitative explanatory variables.

For the sake of simplicity, we will first consider a model containing one quantitative explanatory variable and one dummy variable. As we will see shortly, the model can easily be extended to include additional variables.

Consider the following model:

$$y = \beta_0 + \beta_1x + \beta_2d + \varepsilon,$$

where  $x$  is a quantitative variable and  $d$  is a dummy variable with values of 1 or 0. We can use sample data to estimate the model as

$$\hat{y} = b_0 + b_1x + b_2d.$$

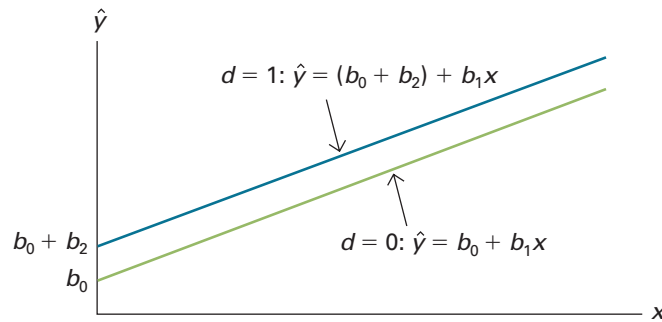
For a given  $x$  and  $d = 1$ , we can compute the predicted value as

$$\hat{y} = b_0 + b_1x + b_2 = (b_0 + b_2) + b_1x.$$

Similarly, for  $d = 0$ ,

$$\hat{y} = b_0 + b_1x.$$

Observe that the two regression lines,  $\hat{y} = (b_0 + b_2) + b_1x$  and  $\hat{y} = b_0 + b_1x$ , have the same slope  $b_1$ . Thus, the sample regression equation  $\hat{y} = b_0 + b_1x + b_2d$  accommodates two parallel lines; that is, the dummy variable  $d$  affects the intercept but not the slope. The difference between the intercepts is  $b_2$  when  $d$  changes from 0 to 1. Figure 17.1 shows the two regression lines when  $b_2 > 0$ .



**FIGURE 17.1**  
Using  $d$  for an  
intercept shift

### EXAMPLE 17.1

The objective outlined in the introductory case is to determine if there is any gender or age discrimination at a large liberal arts college. Use the data in Table 17.1 to answer the following questions.

- Estimate  $y = \beta_0 + \beta_1x + \beta_2d_1 + \beta_3d_2 + \varepsilon$ , where  $y$  is the annual salary (in \$1,000s) of a professor,  $x$  is the number of years of experience,  $d_1$  is a gender dummy variable that equals 1 if the professor is male and 0 otherwise, and  $d_2$  is an age dummy variable that equals 1 if the professor is 60 years of age or older and 0 otherwise.
- Compute the predicted salary of a 50-year-old male professor with 10 years of experience. Compute the predicted salary of a 50-year-old female professor with 10 years of experience. Discuss the impact of gender on predicted salary.
- Compute the predicted salary of a 65-year-old female professor with 10 years of experience. Discuss the impact of age on predicted salary.

#### SOLUTION:

- To estimate the model in part a, we first convert the qualitative variables in Table 17.1 to their respective gender and age dummy variables,  $d_1$  and  $d_2$ . A portion of the converted data is shown in Table 17.2.

**TABLE 17.2** Generating  $d_1$  and  $d_2$  from the Data in Table 17.1

$y$	$x$	$d_1$ (Gender)	$d_2$ (Age)
67.50	14	1	0
53.51	6	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
73.06	35	0	1

Table 17.3 shows a portion of the regression results.

**TABLE 17.3** Regression Results for Example 17.1

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p-Value</i>
Intercept	40.61	3.69	11.00	0.00
$x$	1.13	0.18	6.30	0.00
$d_1$ (Gender)	13.92	2.87	4.86	0.00
$d_2$ (Age)	4.34	4.64	0.94	0.36

The estimated model is  $\hat{y} = 40.61 + 1.13x + 13.92d_1 + 4.34d_2$ .

- b.** The predicted salary of a 50-year-old male professor ( $d_1 = 1$  and  $d_2 = 0$ ) with 10 years of experience ( $x = 10$ ) is

$$\hat{y} = 40.61 + 1.13(10) + 13.92(1) + 4.34(0) = 65.83, \text{ or } \$65,830.$$

The corresponding salary of a 50-year-old female professor ( $d_1 = 0$  and  $d_2 = 0$ ) is

$$\hat{y} = 40.61 + 1.13(10) + 13.92(0) + 4.34(0) = 51.91, \text{ or } \$51,910.$$

The predicted difference in salary between a male and a female professor with 10 years of experience is \$13,920 ( $65,830 - 51,910$ ). This difference can also be inferred from the estimated coefficient 13.92 of the gender dummy variable  $d_1$ . Note that the salary difference does not change with experience. For instance, the predicted salary of a 50-year-old male with 20 years of experience is \$77,130. The corresponding salary of a 50-year-old female is \$63,210, for the same difference of \$13,920.

- c.** For a 65-year-old female professor with 10 years of experience, the predicted salary is

$$\hat{y} = 40.61 + 1.13(10) + 13.92(0) + 4.34(1) = 56.25, \text{ or } \$56,250.$$

Prior to any statistical testing, it appears that an older female professor earns, on average, \$4,340 ( $56,250 - 51,910$ ) more than a younger female professor with the same experience.

## LO 17.2

Test for differences between the categories of a qualitative variable.

Dummy variables are treated just like other explanatory variables; that is, all statistical tests discussed in Chapter 15 remain valid. In particular, we can examine whether a particular dummy variable is statistically significant by using the standard  $t$  test. Here, the statistical significance indicates that the response variable depends on the two categories of the dummy variable.

### TESTING THE SIGNIFICANCE OF DUMMY VARIABLES

In a model,  $y = \beta_0 + \beta_1x + \beta_2d_1 + \beta_3d_2 + \varepsilon$ , we can perform the  $t$  test to determine the significance of each dummy variable.

## EXAMPLE 17.2

Refer to the regression results in Table 17.3.

- Determine whether there is a difference in salary depending on gender at the 5% significance level.
- Determine whether an older professor's salary differs from a younger professor's salary at the 5% significance level.

**SOLUTION:**

- a. In order to test for a salary difference between male and female professors, we set up the hypotheses as  $H_0: \beta_2 = 0$  against  $H_A: \beta_2 \neq 0$ . Given a value of the  $t_{df}$  test statistic of 4.86 with a  $p$ -value  $\approx 0$ , we reject the null hypothesis and conclude that the gender dummy variable is statistically significant at the 5% level. We conclude that male and female professors do not make the same salary, holding other variables constant.
- b. Here the hypotheses take the form  $H_0: \beta_3 = 0$  against  $H_A: \beta_3 \neq 0$ . Given a value of the  $t_{df}$  test statistic of 0.94 with a  $p$ -value = 0.36, we cannot reject the null hypothesis. At the 5% significance level, we cannot conclude that age discrimination exists with respect to a professor's salary.

We now turn our attention to selecting the preferred model for the analysis. Regression results are summarized in Table 17.4.

**TABLE 17.4** Summary of Model Estimates

Variable	Model 1	Model 2	Model 3
Intercept	48.83* (0.00)	39.43* (0.00)	40.61* (0.00)
Experience	1.15* (0.00)	1.24* (0.00)	1.13* (0.00)
Gender	NA	13.89* (0.01)	13.92* (0.00)
Age	NA	NA	4.34 (0.36)
Adjusted $R^2$	0.5358	0.7031	0.7022

Notes: The table contains parameter estimates with  $p$ -values in parentheses; NA denotes not applicable;

\* represents significance at the 5% level; adjusted  $R^2$ , reported in the last row, is used for model selection.

Model 1 uses only the quantitative variable, Experience. In addition to Experience, Model 2 includes a dummy variable, Gender, and Model 3 includes Experience and two dummy variables, Gender and Age. This raises an important question: which of the above three models should we use for making predictions? As discussed in Chapter 14, we usually rely on adjusted  $R^2$  to compare models with different numbers of explanatory variables. Based on the adjusted  $R^2$  values of the models, reported in the last row of Table 17.4, we select Model 2 as the preferred model because it has the highest adjusted  $R^2$  value of 0.7031. This is consistent with the test results that showed that Gender is significant, but Age is not significant, at the 5% level.

## Qualitative Variables with Multiple Categories

So far we have used dummy variables to describe qualitative variables with only two categories, such as gender with males and females. Sometimes, a qualitative variable may be defined by more than two categories. In such cases we use multiple dummy variables to capture all categories. For example, the mode of transportation used to commute may be described by three categories: Public Transportation, Driving Alone, and Car Pooling. We can then define two dummy variables  $d_1$  and  $d_2$ , where  $d_1$  equals 1 for Public Transportation, 0 otherwise, and  $d_2$  equals 1 for Driving Alone, 0 otherwise. For this three-category case, we need to define only two dummy variables; Car Pooling is indicated when  $d_1 = d_2 = 0$ .

Consider the following regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \varepsilon,$$



where  $y$  represents commuting expenditure,  $x$  represents distance to work, and  $d_1$  and  $d_2$  represent the Public Transportation and Driving Alone dummy variables, respectively. We can use sample data to estimate the model as

$$\hat{y} = b_0 + b_1x + b_2d_1 + b_3d_2.$$

For  $d_1 = 1, d_2 = 0$  (Public Transportation),  $\hat{y} = b_0 + b_1x + b_2 = (b_0 + b_2) + b_1x$ .

For  $d_1 = 0, d_2 = 1$  (Driving Alone),  $\hat{y} = b_0 + b_1x + b_3 = (b_0 + b_3) + b_1x$ .

For  $d_1 = d_2 = 0$  (Car Pooling),  $\hat{y} = b_0 + b_1x$ .

Here we use Car Pooling as the reference category in the estimated regression line with the intercept  $b_0$ . The intercept changes to  $(b_0 + b_2)$  for Public Transportation and  $(b_0 + b_3)$  for Driving Alone. Therefore, we account for all three categories with just two dummy variables.

Given the intercept term, we exclude one of the dummy variables from the regression, where the excluded variable represents the reference category against which the others are assessed. If we include as many dummy variables as there are categories, then their sum will equal one. For instance, if we add a third dummy  $d_3$  that equals 1 to denote Car Pooling, then for all observations,  $d_1 + d_2 + d_3 = 1$ . This creates the problem called perfect multicollinearity, a topic discussed in Chapter 15; recall that such a model cannot be estimated. This situation is sometimes referred to as the **dummy variable trap**.

#### AVOIDING THE DUMMY VARIABLE TRAP

Assuming that the linear regression model includes an intercept, the number of dummy variables representing a qualitative variable should be **one less than the number of categories** of the variable.

### EXAMPLE 17.3

A recent article suggests that Asian-Americans face serious discrimination in the college admissions process (*The New York Times*, February 8, 2012). Specifically, Asian students need higher standardized test scores than white students for college admission. Another report suggests that colleges are eager to recruit Hispanic students who are generally underrepresented in applicant pools (*USA TODAY*, February 8, 2010). A researcher from the Center for Equal Opportunity wants to determine if SAT scores of admitted students at a large state university differed by ethnic background. She collects data on 200 admitted students with their SAT scores and ethnic background. A portion of the data is shown in Table 17.5.

**TABLE 17.5** SAT Scores and Ethnic Background;  $n = 200$

Individual	SAT	White	Black	Asian	Hispanic
1	1515	1	0	0	0
2	1530	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
200	1614	1	0	0	0

- Estimate the model  $y = \beta_0 + \beta_1d_1 + \beta_2d_2 + \beta_3d_3 + \varepsilon$ , where  $y$  represents a student's SAT score;  $d_1$  equals 1 if the student is white, 0 otherwise;  $d_2$  equals 1 if the student is black, 0 otherwise; and  $d_3$  equals 1 if the student is Asian, 0 otherwise. Note that the reference category is Hispanic.

#### FILE

SAT\_Ethnicity



- b. What is the predicted SAT score for an Asian student? For a Hispanic student?
- c. Do SAT scores vary by ethnic background at the 5% significance level? Explain.

**SOLUTION:**

- a. We report a portion of the regression results of this model in Table 17.6.

**TABLE 17.6** Regression Results for Example 17.3

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p-Value</i>
Intercept	1388.89	9.36	148.44	0.00
$d_1$ (White)	201.14	12.91	15.59	0.00
$d_2$ (Black)	-31.45	22.19	-1.42	0.16
$d_3$ (Asian)	264.86	17.86	14.83	0.00

- b. For an Asian student, we set  $d_1 = 0$ ,  $d_2 = 0$ ,  $d_3 = 1$  and calculate  $\hat{y} = 1388.89 + 264.86 = 1653.75$ . Thus, the predicted SAT score for an Asian student is approximately 1654. The predicted SAT score for a Hispanic student ( $d_1 = d_2 = d_3 = 0$ ) is  $\hat{y} = 1388.89$ , or approximately 1389.
- c. Since the  $p$ -values corresponding to  $d_1$  and  $d_3$  are approximately zero, we conclude at the 5% level that the SAT scores of admitted white and Asian students are different from those of Hispanic students. However, with a  $p$ -value of 0.16, we cannot conclude that the SAT scores of admitted black and Hispanic students are statistically different.

### EXAMPLE 17.4

Use the SAT data in Table 17.5 to reformulate the model to determine, at the 5% level of significance, if the SAT scores of white students are lower than the SAT scores of Asian students. As in Example 17.3, we must consider all ethnic categories for the analysis.

**SOLUTION:** We note that the regression results reported in Table 17.6 cannot be used to determine if the SAT scores of white students are lower than the SAT scores of Asian students. In order to conduct the relevant test, we must use either Asians or whites as the reference category against which the other one is assessed. We estimate the model as  $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \varepsilon$ , where  $d_1$  and  $d_2$  are again dummy variables corresponding to the categories of white and black students, respectively, but now  $d_3$  equals 1 if the student is Hispanic; 0 otherwise. Here the reference category is Asian. We report a portion of the regression results of this model in Table 17.7.

**TABLE 17.7** Regression Results for Example 17.4

	<i>Coefficient</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p-Value</i>
Intercept	1653.75	15.21	108.72	0.00
$d_1$ (White)	-63.71	17.62	-3.62	0.00
$d_2$ (Black)	-296.31	25.22	-11.75	0.00
$d_3$ (Hispanic)	-264.86	17.86	-14.83	0.00

For an Asian student, we set  $d_1 = d_2 = d_3 = 0$  to find the predicted SAT score as 1653.75, which is the same as derived earlier. In fact, we can show that all predicted SAT scores are identical to those found in Example 17.3. This shows that the choice of the reference category does not matter for making predictions.

The results in Table 17.7, however, can be used to determine if the SAT scores of white students are lower than the SAT scores of Asian students. We specify the hypothesis for a left-tailed test as  $H_0: \beta_1 \geq 0$  against  $H_A: \beta_1 < 0$  and use the critical value approach for the test. Given  $n = 200$  and  $k = 3$ , we find  $df = n - k - 1 = 196$ . At  $\alpha = 0.05$ , we reject the null hypothesis since the value of the test statistic,  $t_{196} = -3.62$ , is less than the critical value,  $-t_{0.05, 196} = -1.653$ . Therefore, we conclude that the SAT scores of admitted white students are less than the SAT scores of admitted Asian students at the 5% significance level.

## EXERCISES 17.1

### Mechanics

- Consider a linear regression model where  $y$  represents the response variable,  $x$  is a quantitative explanatory variable, and  $d$  is a dummy variable. The model is estimated as

$$\hat{y} = 14.8 + 4.4x - 3.8d.$$

- Interpret the dummy variable coefficient.
  - Compute  $\hat{y}$  for  $x = 3$  and  $d = 1$ .
  - Compute  $\hat{y}$  for  $x = 3$  and  $d = 0$ .
- Consider a linear regression model where  $y$  represents the response variable and  $d_1$  and  $d_2$  are dummy variables. The model is estimated as  $\hat{y} = 160 + 15d_1 + 32d_2$ .
    - Compute  $\hat{y}$  for  $d_1 = 1$  and  $d_2 = 1$ .
    - Compute  $\hat{y}$  for  $d_1 = 0$  and  $d_2 = 0$ .
  - Using 50 observations, the following regression output is obtained from estimating  $y = \beta_0 + \beta_1x + \beta_2d_1 + \beta_3d_2 + \varepsilon$ .

	Coefficients	Standard Error	t Stat	p-Value
Intercept	-0.61	0.23	-2.75	0.0074
$x$	3.12	1.04	3.01	0.0034
$d_1$	-13.22	15.65	-0.85	0.4006
$d_2$	5.35	1.25	4.27	0.0000

- Compute  $\hat{y}$  for  $x = 250$ ,  $d_1 = 1$ , and  $d_2 = 0$ ; then compute  $\hat{y}$  for  $x = 250$ ,  $d_1 = 0$ , and  $d_2 = 1$ .
- Interpret  $d_1$  and  $d_2$ . Are both dummy variables individually significant at the 5% level? Explain.

### Applications

- An executive researcher wants to better understand the factors that explain differences in salaries for marketing majors. He decides to estimate two models:  $y = \beta_0 + \beta_1d_1 + \varepsilon$  (Model 1) and  $y = \beta_0 + \beta_1d_1 + \beta_2d_2 + \varepsilon$  (Model 2). Here  $y$  represents salary,  $d_1$  is a dummy variable that equals 1 for male employees, and  $d_2$  is a dummy variable that equals 1 for employees with an MBA.

- What is the reference group in Model 1?
  - What is the reference group in Model 2?
  - In the above models, would it matter if  $d_1$  equaled 1 for female employees?
- House price  $y$  is estimated as a function of the square footage of a house  $x$  and a dummy variable  $d$  that equals 1 if the house has ocean views. The estimated house price, measured in \$1,000s, is given by  $\hat{y} = 118.90 + 0.12x + 52.60d$ .
    - Compute the predicted price of a house with ocean views and square footage of 2,000 and 3,000, respectively.
    - Compute the predicted price of a house without ocean views and square footage of 2,000 and 3,000, respectively.
    - Discuss the impact of ocean views on the house price.
  - FILE Urban.** A sociologist is studying the relationship between consumption expenditures  $y$  of families in the United States, family income  $x$ , and whether or not the family lives in an urban or rural community. She collects data on 50 families, a portion of which is shown in the accompanying table.

Consumption (\$)	Income (\$)	Community
62,336	87,534	Rural
60,076	94,796	Urban
:	:	:
59,055	100,908	Urban

- Estimate  $y = \beta_0 + \beta_1x + \beta_2d + \varepsilon$  where the dummy variable  $d$  equals 1 for urban families. Use the estimated model to predict the consumption expenditure of urban families with an income of \$80,000. What is the corresponding consumption expenditure of rural families?
- Estimate  $y = \beta_0 + \beta_1x + \beta_2d + \varepsilon$  where the dummy variable  $d$  equals 1 for rural families. Use the estimated model to predict the consumption

- expenditure of urban families with an income of \$80,000. What is the corresponding consumption expenditure of rural families?
- c. Interpret the results of the preceding two models.
7. **FILE IPO.** One of the theories regarding initial public offering (IPO) pricing is that the initial return  $y$  (change from offer to open price) on an IPO depends on the price revision  $x$  (change from pre-offer to offer price). Another factor that may influence the initial return is whether or not it is a high-tech firm. The following table shows a portion of the data on 264 IPO firms from January 2001 through September 2004.

Initial Return (%)	Price Revision (%)	High-Tech?
33.93	71.4	No
18.68	-26.39	No
⋮	⋮	⋮
0.08	-29.41	Yes

SOURCE: [www.ipohome.com](http://www.ipohome.com); [www.nasdaq.com](http://www.nasdaq.com).

- a. Estimate  $y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$  where the dummy variable  $d$  equals 1 for firms that are high-tech. Use the estimated model to predict the initial return of a high-tech firm with a 10% price revision. Find the corresponding predicted return of a firm that is not high-tech.
- b. Estimate  $y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$  where the dummy variable  $d$  equals 1 for firms that are not high-tech. Use the estimated model to predict the initial return of a high-tech firm with a 10% price revision. Find the corresponding predicted return of a firm that is not high-tech.
- c. In the above two models, determine if the dummy variable is significant at the 5% level.
8. **FILE BMI.** According to the *World Health Organization*, obesity has reached epidemic proportions globally. While obesity has generally been linked with chronic disease and disability, researchers argue that it may also affect salaries. In other words, the body mass index (BMI) of an employee is a predictor for salary. (A person is considered overweight if his/her BMI is at least 25 and obese if BMI exceeds 30.) The accompanying table shows a portion of salary data for 30 college-educated men with their respective BMI and a dummy variable that equals 1 for a white man and 0 otherwise.

Salary (\$1,000)	BMI	Race
34	33	1
43	26	1
⋮	⋮	⋮
45	21	1

- a. Estimate a model for Salary using BMI and Race as the explanatory variables. Determine if BMI influences salary at the 5% level of significance.
- b. What is the estimated salary of a white college-educated man with a BMI of 30? Compute the corresponding salary of a nonwhite man.
9. **FILE Wage.** A researcher wonders whether males get paid more, on average, than females at a large firm. She interviews 50 employees and collects data on each employee's hourly wage (Wage), years of higher education (EDUC), experience (EXPER), age (AGE), and gender. The GENDER dummy variable equals 1 if male, 0 if female. A portion of the data is shown in the accompanying table.

Wage	EDUC	EXPER	AGE	GENDER
\$37.85	11	2	40	1
21.72	4	1	39	0
⋮	⋮	⋮	⋮	⋮
24.18	8	11	64	0

- a. Estimate:  $\text{Wage} = \beta_0 + \beta_1 \text{EDUC} + \beta_2 \text{EXPER} + \beta_3 \text{AGE} + \beta_4 \text{GENDER} + \varepsilon$ .
- b. Predict the hourly wage of a 40-year-old male employee with 10 years of higher education and 5 years experience. Predict the hourly wage of a 40-year-old female employee with the same qualifications.
- c. Interpret the estimated coefficient for GENDER. Is the variable GENDER significant at the 5% level? Do the data suggest that gender discrimination exists at this firm?
10. **FILE Nicknames.** In the United States, baseball has always been a favorite pastime and is rife with statistics and theories. While baseball purists may disagree, to an applied statistician no topic in baseball is too small or hypothesis too unlikely. In a recent paper, researchers at Wayne State University showed that major league players who have nicknames live  $2\frac{1}{2}$  years longer than those without them (*The Wall Street Journal*, July 16, 2009). The following table shows a portion of data on the lifespan (Years) of a player and a Nickname dummy variable that equals 1 if the player had a nickname and 0 otherwise.

Years	Nickname
74	1
62	1
⋮	⋮
64	0

- a. Create two subsamples, with one consisting of players with a nickname and the other one without a nickname. Calculate the average longevity for each subsample.

- b. Estimate a linear regression model of Years on the Nickname dummy variable. Compute the predicted longevity of players with and without a nickname.
- c. Conduct a one-tailed test at a 5% level to determine if players with a nickname live longer.

11. **FILE SAT.** The SAT has gone through many revisions over the years. In 2005, a new writing section was introduced that includes a direct writing measure in the form of an essay. People argue that female students generally do worse on math tests but better on writing tests. Therefore, the new section may help reduce the usual male lead on the overall average SAT score (*Washington Post*, August 30, 2006). Consider the following portion of data on 20 students who took the SAT test last year. Information includes each student's score on the writing and math sections of the exam. Also included are the student's GPA and the values of a Gender dummy variable with 1 for females and 0 for males.

Writing	Math	GPA	Gender
620	600	3.44	0
570	550	3.04	0
⋮	⋮	⋮	⋮
540	520	2.84	0

- a. Estimate a linear regression model with Writing as the response variable and GPA and Gender as the explanatory variables.
  - b. Compute the predicted writing score for a male student with a GPA of 3.5. Repeat the computation for a female student.
  - c. Perform a test to determine if there is a statistically significant gender difference in writing scores at a 5% level.
12. **FILE SAT.** Use the data described in Exercise 11 to estimate a linear regression model with Math as the response variable and GPA and Gender as the explanatory variables.
    - a. Compute the predicted math score for a male student with a GPA of 3.5. Repeat the computation for a female student.
    - b. Perform a test to determine if there is a statistically significant gender difference in math scores at the 5% level.
  13. **FILE Ice Cream.** A manager at an ice cream store is trying to determine how many customers to expect on any given day. Overall business has been relatively steady over the past several years, but the customer count seems to have ups and downs. He collects data over 30 days and records the number of customers, the high temperature (degrees Fahrenheit), and whether the day fell on a weekend (1 equals weekend, 0 otherwise). A portion of the data is shown in the accompanying table.

Customers	Temperature	Weekend
376	75	0
433	78	0
⋮	⋮	⋮
401	68	0

- a. Estimate:  $\text{Customers} = \beta_0 + \beta_1 \text{Temperature} + \beta_2 \text{Weekend} + \varepsilon$ .
  - b. How many customers should the manager expect on a Sunday with a forecasted high temperature of 80°?
  - c. Interpret the estimated coefficient for Weekend. Is it significant at the 5% level? How might this affect the store's staffing needs?
14. In an attempt to "time the market," a financial analyst studies the quarterly returns of a stock. He uses the model  $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \varepsilon$  where  $y$  is the quarterly return of a stock,  $d_1$  is a dummy variable that equals 1 if quarter 1 and 0 otherwise,  $d_2$  is a dummy variable that equals 1 if quarter 2 and 0 otherwise, and  $d_3$  is a dummy variable that equals 1 if quarter 3 and 0 otherwise. The following table shows a portion of the regression results.

	Coefficients	Standard Error	t Stat	p-Value
Intercept	10.62	5.81	1.83	0.08
$d_1$	-7.26	8.21	-0.88	0.38
$d_2$	-1.87	8.21	-0.23	0.82
$d_3$	-9.31	8.21	-1.13	0.27

- a. Given that there are four quarters in a year, why doesn't the analyst include a fourth dummy variable in his model? What is the reference category?
  - b. At the 5% significance level, are the dummy variables individually significant? Explain. Is the analyst able to obtain higher returns depending on the quarter?
  - c. Explain how you would reformulate the model to determine if the quarterly return is higher in quarter 2 than in quarter 3, still accounting for all quarters.
15. **FILE Industry.** The issues regarding executive compensation have received extensive media attention. The government is even considering a cap on high-flying salaries for executives (*The New York Times*, February 9, 2009). Consider a regression model that links executive compensation with the total assets of the firm and the firm's industry. Dummy variables are used to represent four industries: Manufacturing Technology  $d_1$ , Manufacturing Other  $d_2$ , Financial Services  $d_3$ , and Nonfinancial Services  $d_4$ . A portion of the data for the 455 highest-paid CEOs in 2006 is shown in the accompanying table.

Compensation (in \$ million)	Assets (in \$ millions)	$d_1$	$d_2$	$d_3$	$d_4$
16.58	20,917.5	1	0	0	0
26.92	32,659.5	1	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2.30	44,875.0	0	0	1	0

SOURCE: SEC website and Compustat.

- Estimate the model:  $y = \beta_0 + \beta_1x + \beta_2d_1 + \beta_3d_2 + \beta_4d_3 + \varepsilon$ , where  $y$  and  $x$  denote compensation and total assets, respectively. Here the reference category is nonfinancial services industry.
  - Interpret the estimated coefficients.
  - Use a 5% level of significance to determine which industries, relative to the nonfinancial services industry, have a different executive compensation.
  - Reformulate the model to determine, at the 5% significance level, if compensation is higher in Manufacturing Other than in Manufacturing Technology. Your model must account for assets and all industry types.
16. **FILE Retail.** A government researcher is analyzing the relationship between retail sales and the gross national product (GNP). He also wonders whether there are significant differences in retail sales related to the quarters of the year. He collects 10 years of quarterly data. A portion is shown in the accompanying table.

Year	Quarter	Retail Sales (in \$ millions)	GNP (in \$ billions)	$d_1$	$d_2$	$d_3$	$d_4$
2001	1	696,048	9,740.5	1	0	0	0
	2	753,211	9,983.5	0	1	0	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
2009	4	985,649	14,442.8	0	0	0	1

SOURCE: Retail sales obtained from [www.census.gov](http://www.census.gov); GNP obtained from <http://research.stlouisfed.org>.

- Estimate  $y = \beta_0 + \beta_1x + \beta_2d_1 + \beta_3d_2 + \beta_4d_3 + \varepsilon$  where  $y$  is retail sales,  $x$  is GNP,  $d_1$  is a dummy variable that equals 1 if quarter 1 and 0 otherwise,  $d_2$  is a dummy variable that equals 1 if quarter 2 and 0 otherwise,

and  $d_3$  is a dummy variable that equals 1 if quarter 3 and 0 otherwise. Here the reference category is quarter 4.

- Predict retail sales in quarters 2 and 4 if GNP equals \$13,000 billion.
  - Which of the quarterly sales are significantly different from those of the 4th quarter at the 5% level?
  - Reformulate the model to determine, at the 5% significance level, if sales differ between quarter 2 and quarter 3. Your model must account for all quarters.
17. **FILE QuickFix.** The general manager of QuickFix, a chain of quick-service, no-appointment auto repair shops, wants to develop a model to forecast monthly vehicles served at any particular shop based on four factors: garage bays, population within 5-mile radius, interstate highway access (1 = convenient access, 0 = otherwise), and time of year (1 = winter, 0 = otherwise). He believes that, all else equal, shops near an interstate will service more vehicles and that more vehicles will be serviced in the winter due to battery and tire issues. A sample of 19 locations has been obtained. A portion of the data is shown in the accompanying table.

Vehicles Served (per month)	Garage Bays	Population in 5-Mile Radius (000's)	Interstate Access?	Time of Year?
200	3	15	0	0
351	3	22	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
464	6	74	1	1

- Estimate the regression equation relating vehicles serviced to the four explanatory variables.
- Interpret each of the slope coefficients.
- At the 5% significance level, are the explanatory variables jointly significant? Are they individually significant? What about at the 10% significance level?
- What proportion of the variability in vehicles served is explained by the four explanatory variables?
- Predict vehicles serviced in July for a particular location with 5 garage bays, a population of 40,000, and convenient interstate access.

## 17.2 INTERACTIONS WITH DUMMY VARIABLES

LO 17.3

So far we have used a dummy variable  $d$  to allow for a shift in the intercept. In other words,  $d$  allows the predicted  $y$  to differ between the two categories of a qualitative variable by a fixed amount across the values of  $x$ . We can also use  $d$  to create an **interaction variable**, which allows the predicted  $y$  to differ between the two categories of a qualitative variable by a varying amount across the values of  $x$ . The interaction variable is a product term  $xd$  that captures the interaction between a quantitative variable  $x$  and a dummy variable  $d$ .

Use dummy variables to capture interactions between qualitative and quantitative explanatory variables.

Together, the variables  $d$  and  $xd$  allow the intercept as well as the slope of the estimated linear regression line to vary between the two categories of a qualitative variable.

Consider the following regression model:

$$y = \beta_0 + \beta_1x + \beta_2d + \beta_3xd + \varepsilon.$$

We can use sample data to estimate the model as

$$\hat{y} = b_0 + b_1x + b_2d + b_3xd.$$

For a given  $x$  and  $d = 1$ , we can compute the predicted value as

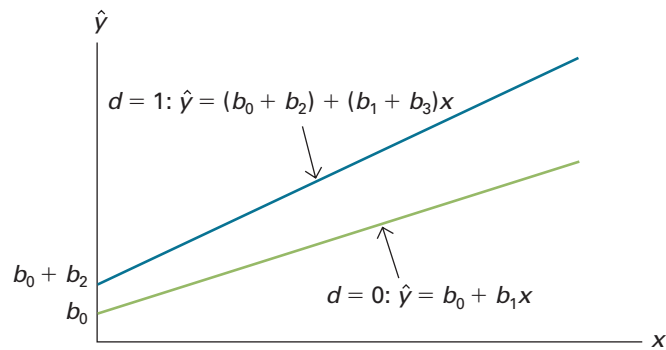
$$\hat{y} = b_0 + b_1x + b_2 + b_3x = (b_0 + b_2) + (b_1 + b_3)x.$$

Similarly, for  $d = 0$ ,

$$\hat{y} = b_0 + b_1x.$$

The use of the dummy variable  $d$  along with the interaction variable  $xd$  affects the intercept as well as the slope of the estimated regression line. Note that the estimated intercept  $b_0$  and slope  $b_1$  when  $d = 0$  shift to  $(b_0 + b_2)$  and  $(b_1 + b_3)$ , respectively, when  $d = 1$ . Figure 17.2 shows a shift in the intercept and the slope of the estimated regression line when  $d = 0$  changes to  $d = 1$ , given  $b_2 > 0$  and  $b_3 > 0$ .

**FIGURE 17.2**  
Using  $d$  and  $xd$  for  
intercept and  
slope shifts



Prior to estimation, we use sample data to generate two variables,  $d$  and  $xd$ , which we use along with other explanatory variables in the regression. Tests of significance are performed as before.

#### TESTING THE SIGNIFICANCE OF DUMMY AND INTERACTION VARIABLES

In a model  $y = \beta_0 + \beta_1x + \beta_2d + \beta_3xd + \varepsilon$ , we can perform  $t$  tests for the individual significance of the dummy  $d$  and the interaction  $xd$ . Similarly, we can perform the partial  $F$  test for the joint significance of  $d$  and  $xd$ .

#### EXAMPLE 17.5

In Section 17.1, we estimated a regression model to test for gender and age discrimination in salaries. We found that the number of years of experience  $x$  and the gender dummy variable  $d_1$  were significant in explaining salary differences; however, the age dummy variable  $d_2$  was insignificant. In an attempt to refine the



model explaining salary, we drop  $d_2$  and estimate three models using the data from Table 17.1, where  $y$  represents annual salary (in \$1,000s).

$$\text{Model 1: } y = \beta_0 + \beta_1x + \beta_2d_1 + \varepsilon$$

$$\text{Model 2: } y = \beta_0 + \beta_1x + \beta_2xd_1 + \varepsilon$$

$$\text{Model 3: } y = \beta_0 + \beta_1x + \beta_2d_1 + \beta_3xd_1 + \varepsilon$$

- Estimate and interpret each of the three models.
- Select the most appropriate model, based on an objective model selection criterion.
- Use the selected model to predict salaries for males and females over various years of experience.

#### SOLUTION:

- In order to estimate the three models, we first generate data on both  $d_1$  and  $xd_1$ ; Table 17.8 shows a portion of the data.

**TABLE 17.8** Generating  $d_1$  and  $xd_1$  from the Data in Table 17.1

$y$	$x$	$d_1$	$xd_1$
67.50	14	1	$14 \times 1 = 14$
53.51	6	1	$6 \times 1 = 6$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
73.06	35	0	$35 \times 0 = 0$

A portion of the regression results is summarized in Table 17.9.

**TABLE 17.9** Summary of Model Estimates

	Model 1	Model 2	Model 3
Intercept	39.43* (0.00)	47.07* (0.00)	49.42* (0.00)
Experience $x$	1.24* (0.00)	0.85* (0.00)	0.76* (0.00)
Gender Dummy $d_1$	13.89* (0.01)	NA	-4.00 (0.42)
Interaction Variable $xd_1$	NA	0.77* (0.00)	0.93* (0.00)
Adjusted $R^2$	0.7031	0.7923	0.7905

NOTES: The top portion of the table contains parameter estimates with  $p$ -values in parentheses; NA denotes not applicable; \* represents significance at the 5% level; Adjusted  $R^2$ , reported in the last row, is used for model selection.

Model 1 uses a gender dummy variable  $d_1$  to allow salaries between males and females to differ by a fixed amount, irrespective of experience. It is estimated as  $\hat{y} = 39.43 + 1.24x + 13.89d_1$ . Since  $d_1$  is associated with a  $p$ -value of 0.01, we conclude at the 5% level that  $d_1$  has a statistically significant influence on salary. The estimated model implies that, on average, males earn \$13,890 ( $13.89 \times 1,000$ ) more than females at all levels of experience.

Model 2 uses an interaction variable  $xd_1$  to allow the difference in salaries between males and females to vary with experience. It is estimated as  $\hat{y} = 47.07 + 0.85x + 0.77xd_1$ . Since  $xd_1$  is associated with a  $p$ -value  $\approx 0.00$ ,

we conclude that it is statistically significant at the 5% level. With every extra year of experience, the estimated difference in salaries between males and females increases by \$770 ( $0.77 \times 1,000$ ).

Model 3 uses  $d_1$  along with  $xd_1$  to allow a fixed as well as a varying difference in salaries between males and females. The estimated regression equation is  $\hat{y} = 49.42 + 0.76x - 4.00d_1 + 0.93xd_1$ . Interestingly, with a  $p$ -value of 0.42, the variable  $d_1$  is no longer statistically significant at the 5% level. However, the variable  $xd_1$  is significant, suggesting that with every extra year of experience, the estimated difference in salaries between males and females increases by \$930 ( $0.93 \times 1,000$ ).

- b. While Model 1 shows that the gender dummy variable  $d_1$  is significant and Model 2 shows that the interaction variable  $xd_1$  is significant, Model 3 provides somewhat conflicting results. This raises an important question: which model should we trust? It is not uncommon to contend with such scenarios in business applications. As discussed earlier, we usually rely on adjusted  $R^2$  to compare models that have a different number of explanatory variables. Based on the adjusted  $R^2$  values of the models, reported in the last row of Table 17.9, we select Model 2 as the preferred model because it has the highest value of 0.7923.
- c. In order to interpret the results further, we use Model 2 to estimate salaries with varying levels of experience, for both males and females. For example, with 10 years of experience, the predicted salary for males ( $d_1 = 1$ ) is

$$\hat{y} = 47.07 + 0.85(10) + 0.77(10 \times 1) = 63.27, \text{ or } \$63,270.$$

The corresponding predicted salary for females ( $d_1 = 0$ ) is

$$\hat{y} = 47.07 + 0.85(10) + 0.77(10 \times 0) = 55.57, \text{ or } \$55,570.$$

Therefore, with 10 years of experience, the salary difference between males and females is about \$7,700. Predicted salaries for both males and females at other levels of experience are presented in Table 17.10.

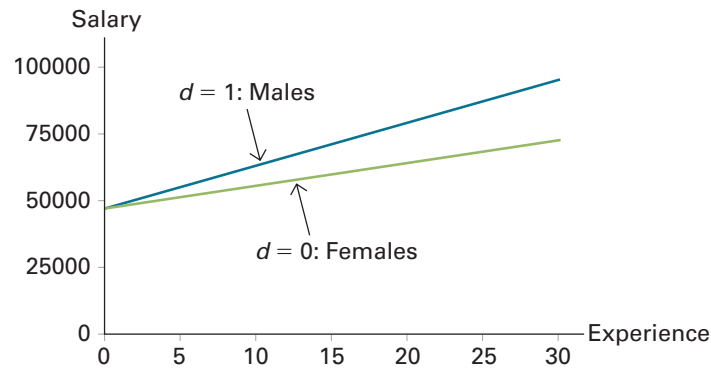
**TABLE 17.10** Estimated Salaries at Various Levels of Experience

Experience	Males	Females	Difference
1	\$48,690	\$47,920	\$770
2	50,310	48,770	1,540
3	51,930	49,620	2,310
4	53,550	50,470	3,080
5	55,170	51,320	3,850
10	63,270	55,570	7,700
15	71,370	59,820	11,550
20	79,470	64,070	15,400
25	87,570	68,320	19,250
30	95,670	72,570	23,100

Note that as experience increases, the salary difference between males and females becomes wider. For instance, the difference is \$7,700 with 10 years of experience. However, the difference increases to \$19,250 with 25 years of experience. This is consistent with the inclusion of the interaction variable in Model 2.

The shift in the slope, implied by the predicted salaries in Table 17.10, is shown in Figure 17.3.

**FIGURE 17.3** Predicted salaries of male and female professors



## SYNOPSIS OF INTRODUCTORY CASE

A recent lawsuit brought against Seton Hall University by three female professors alleges that the university engages in both age and gender discrimination with respect to salaries ([www.nj.com](http://www.nj.com), November 23, 2010). Another large university wonders if the same can be said about its practices. Information is collected on the annual salaries (in \$1,000s) of 42 professors, along with their experience (in years), gender (male or female), and age (whether he/she is 60 years old or older). A regression of salary against experience, a gender dummy variable, and an age dummy variable reveals that the gender of the professor is significant in explaining variations in salary, but the professor's age is not significant.



In an attempt to refine the model describing salary, various models are estimated that remove the age dummy variable, but use the gender dummy variable to allow both fixed and changing effects on salary. The sample regression line that best fits the data does not include the gender dummy variable for a fixed effect. However, the interaction variable, defined as the product of gender and experience, is significant at any reasonable level, implying that males make about \$770 more than females for every year of experience. While the estimated difference in salaries between males and females is only \$770 with 1 year of experience, the difference increases to \$19,250 with 25 years of experience. In sum, the findings suggest that salaries do indeed differ by gender, and this difference increases with every extra year of experience.

## EXERCISES 17.2

### Mechanics

18. Consider a linear regression model where  $y$  represents the response variable and  $x$  and  $d$  are the explanatory variables;  $d$  is a dummy variable assuming values 0 or 1. A model with the dummy variable  $d$  and the interaction variable  $xd$  is estimated as  $\hat{y} = 5.2 + 0.9x + 1.4d + 0.2xd$ .
  - a. Compute  $\hat{y}$  for  $x=10$  and  $d=1$ .
  - b. Compute  $\hat{y}$  for  $x=10$  and  $d=0$ .

19. Using 20 observations, the following regression output is obtained from estimating  $y = \beta_0 + \beta_1x + \beta_2d + \beta_3xd + \varepsilon$ .

	Coefficients	Standard Error	t Stat	p-Value
Intercept	13.56	3.31	4.09	0.0009
$x$	4.62	0.56	8.31	0.0000
$d$	-5.15	4.97	-1.04	0.3156
$xd$	2.09	0.79	2.64	0.0178

- Compute  $\hat{y}$  for  $x = 10$  and  $d = 1$ ; then compute  $\hat{y}$  for  $x = 10$  and  $d = 0$ .
- Are the dummy variable  $d$  and the interaction variable  $xd$  individually significant at the 5% level? Explain.

## Applications

20. The annual salary of an employee  $y$  (in thousands of dollars) is estimated as a function of years of experience  $x$ ; a dummy variable  $d$  that equals 1 for college graduates and 0 for those graduating high school but not college; and a product of this dummy variable and experience,  $xd$ . The estimated salary is given by  $\hat{y} = 30.3 + 1.2x + 15.5d + 2.0xd$ .

- What is the predicted salary of a college graduate who has 5 years of experience? What is the predicted salary of a college graduate who has 15 years of experience?
- What is the predicted salary of a non-college graduate who has 5 years of experience? What is the predicted salary of a non-college graduate who has 15 years of experience?
- Discuss the impact of a college degree on salary.

21. House price  $y$  is estimated as a function of the square footage of a house  $x$ ; a dummy variable  $d$  that equals 1 if the house has ocean views and 0 otherwise; and a product of this dummy variable and the square footage  $xd$ . The estimated house price, measured in \$1,000s, is given by  $\hat{y} = 80 + 0.12x + 40d + 0.01xd$ .

- Compute the predicted price of a house with ocean views and square footage of 2,000 and 3,000, respectively.
- Compute the predicted price of a house without ocean views and square footage of 2,000 and 3,000, respectively.
- Discuss the impact of ocean views on the house price.

22. **FILE Urban.** A sociologist is looking at the relationship between consumption expenditures  $y$  of families in the United States, family income  $x$ , and whether or not the family lives in an urban or rural community (Urban = 1 if urban, 0 otherwise). She collects data on 50 families across the United States, a portion of which is shown in the accompanying table.

Consumption (\$)	Income (\$)	Urban
62336	87534	0
60076	94796	1
$\vdots$	$\vdots$	$\vdots$
59055	100908	1

- Estimate a linear model without a dummy variable:  $y = \beta_0 + \beta_1x + \varepsilon$ . Compute the predicted consumption expenditures of a family with income of \$75,000.

- Include a dummy variable  $d$  to predict consumption for Income = \$75,000 in urban and rural communities.
- Include a dummy variable  $d$  and an interaction variable  $xd$  to predict consumption for Income = \$75,000 in urban and rural communities.
- Which of the preceding models is most suitable for the data? Explain.

23. **FILE BMI.** According to the *World Health Organization*, obesity has reached epidemic proportions globally. While obesity has generally been linked with chronic disease and disability, researchers argue that it may also affect wages. In other words, the body mass index (BMI) of an employee is a predictor for salary. (A person is considered overweight if his/her BMI is at least 25 and obese if BMI exceeds 30.) The accompanying table shows a portion of salary data (in \$1,000s) for 30 college-educated men with their respective BMI and a dummy variable that represents 1 for a white man and 0 otherwise.

Salary	BMI	White
34	33	1
43	26	1
$\vdots$	$\vdots$	$\vdots$
45	21	1

- Estimate a model for Salary with BMI and White as the explanatory variables.
- Reestimate the model with BMI, White, and a product of BMI and White as the explanatory variables.
- Which of the models is most suitable? Explain. Use this model to estimate the salary for a white college-educated man with a BMI of 30. Compute the corresponding salary for a nonwhite man.

24. **FILE Pick\_Errors.** The distribution center for an online retailer has been experiencing quite a few "pick errors" (i.e., retrieving the wrong item). Although the warehouse manager thinks most errors are due to inexperienced workers, she believes that a training program also may help to reduce them. Before sending all employees to training, she examines data from a pilot study of 30 employees (half attended training and the other half did not) to decide whether the training program will be effective. A portion of the data is shown in the accompanying table.

Pick Errors (year to date)	Years of Experience	Attended Training? (0=no, 1=yes)
13	9	0
3	27	0
$\vdots$	$\vdots$	$\vdots$
4	24	1

- Estimate two linear models:  
 $\text{Errors} = \beta_0 + \beta_1 \text{Exper} + \beta_2 \text{Train} + \varepsilon$ , and  
 $\text{Errors} = \beta_0 + \beta_1 \text{Exper} + \beta_2 \text{Train} + \beta_3 \text{Exper} \times \text{Train} + \varepsilon$ .
- Using adjusted  $R^2$ , summarize the fit results of the two models and specify which terms are significant at the 5% level. Which model is preferable in terms of fit?
- Using the chosen model, by how much will pick errors be reduced, on average, for employees who attend the training program?
- Use the chosen model to predict the number of pick errors for an employee with 10 years of experience who attended the training program, and for an employee with 20 years of experience who did not attend the training program.
- Give a practical interpretation for the positive interaction coefficient.

25. **FILE IPO.** One of the theories regarding initial public offering (IPO) pricing is that the initial return (change from offer to open price) on an IPO depends on the price revision (change from pre-offer to offer price). Another factor that may influence the initial return is a high-tech dummy that equals 1 for high-tech firms and 0 otherwise. The following table shows a portion of data on 264 IPO firms from January 2001 through September 2004.

Initial Return (%)	Price Revision (%)	High-Tech Dummy
33.93	71.4	0
18.68	-26.39	0
⋮	⋮	⋮
0.08	-29.41	1

SOURCE: [www.ipohome.com](http://www.ipohome.com), [www.nasdaq.com](http://www.nasdaq.com).

- Estimate a model with the initial return as the response variable and the price revision and the high-tech dummy variable as the explanatory variables.

- Reestimate the model with price revision along with the dummy variable and the product of the dummy variable and the price revision.
- Which of these models is the preferred model? Explain. Use this model to estimate the initial return for a high-tech firm with a 15% price revision. Compute the corresponding initial return for a firm that is not high-tech.

26. **FILE Savings.** The following table shows a portion of monthly data on the personal savings rate (Savings) and the personal disposable income (Income) in the U.S. from January 2007 to November 2010.

Date	Savings (%)	Income (\$ billions)
2007-01	2.2	10198.2
2007-02	2.3	10252.9
⋮	⋮	⋮
2010-11	5.5	11511.9

SOURCE: Bureau of Economic Analysis.

- Estimate and interpret a log-log model,  $\ln(\text{Savings}) = \beta_0 + \beta_1 \ln(\text{Income}) + \varepsilon$ . What is the predicted percentage change in savings when personal disposable income increases by 1%?
- Suppose we want to test whether or not there has been a structural shift due to the financial crisis that erupted in the fall of 2008. Consider a dummy variable  $d$  that assumes a value 0 before August 2008 and a value of 1 starting August 2008 onwards. Estimate:  $\ln(\text{Savings}) = \beta_0 + \beta_1 \ln(\text{Income}) + \beta_2 d + \beta_3 \ln(\text{Income}) \times d + \varepsilon$ . What is the predicted percentage change in savings when personal disposable income increases by 1% prior to August 2008? What is the predicted percentage change starting in August 2008 onward?
- At the 5% significance level, conduct the partial  $F$  test to determine whether or not  $\beta_2$  and  $\beta_3$  are jointly significant. Has there been a structural shift?

## 17.3 BINARY CHOICE MODELS

So far we have considered regression models where dummy (binary) variables are used as explanatory variables. In this section, we analyze models where the variable of interest—the response variable—is binary. The consumer choice literature is replete with applications such as whether or not to buy a house, join a health club, or go to graduate school. At the firm level, managers make decisions such as whether or not to distribute dividends, hire people, or launch a new product. In all such applications the response variable is binary, where one of the choices can be designated as 1 and the other as 0. Usually,

this choice can be related to a host of factors—the explanatory variables. For instance, whether or not a family buys a house depends on explanatory variables such as household income, mortgage rates, and so on.

BINARY CHOICE MODELS

Regression models that use a dummy (binary) variable as the response variable are called **binary choice** models. They are also referred to as **discrete choice** models or **qualitative response** models.

LO 17.4

Use a linear probability model to estimate a binary response variable.

The Linear Probability Model

Consider a simple linear regression model  $y = \beta_0 + \beta_1x + \varepsilon$  where  $y$  is a binary variable; we can easily extend the model to include multiple explanatory variables. A linear regression model applied to a binary response variable is called a **linear probability model (LPM)**. While we know that the relationship implied by this model is linear, it may not be obvious why it is also called a probability model. Recall that in the above simple linear regression model, the expression  $\beta_0 + \beta_1x$  is its deterministic component, which is the expected value of  $y$  for a given value of  $x$ . In other words, conditional on  $x$ ,  $E(y) = \beta_0 + \beta_1x$ . Here, since  $y$  is a discrete random variable with possible values 0 and 1, its expected value conditional on  $x$  can also be computed as  $E(y) = 0 \times P(y = 0) + 1 \times P(y = 1) = P(y = 1)$ , where  $P(y = 1)$  is often referred to as the probability of success. Therefore,  $E(y) = P(y = 1) = \beta_0 + \beta_1x$ . In other words, we can write  $y = \beta_0 + \beta_1x + \varepsilon = P(y = 1) + \varepsilon$ , where  $P(y = 1)$ , or simply  $P$ , is a linear function of the explanatory variable.

A LINEAR PROBABILITY MODEL

A **linear probability model (LPM)** is specified as  $y = \beta_0 + \beta_1x + \varepsilon = P(y = 1) + \varepsilon$ , where  $y$  assumes a 0 or 1 value and  $P(y = 1)$  is the probability of success. Predictions with this model are made by  $\hat{P} = \hat{y} = b_0 + b_1x$  where  $b_0$  and  $b_1$  are the estimates of the population parameters  $\beta_0$  and  $\beta_1$ .

EXAMPLE 17.6

The subprime mortgage crisis has forced financial institutions to be extra stringent in granting mortgage loans. Thirty recent mortgage applications are obtained to analyze the mortgage approval rate. The response variable  $y$  equals 1 if the mortgage loan is approved, 0 otherwise. It is believed that approval depends on the percentage of the down payment  $x_1$  and the percentage of income-to-loan amount  $x_2$ . Table 17.11 shows a portion of the data. Estimate and interpret the linear probability model,  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon$ . Predict the approval probability for representative applicants.

TABLE 17.11 Mortgage Application Data (Example 17.6)

Approval	Down Payment (%)	Income-to-Loan (%)
1	16.35	49.94
1	34.43	56.16
⋮	⋮	⋮
0	17.85	26.86

FILE  
Mortgage



**SOLUTION:** Table 17.12 shows a portion of the regression results. The estimated regression equation is  $\hat{P} = \hat{y} = -0.8682 + 0.0188x_1 + 0.0258x_2$ . With  $p$ -values of 0.0120 and 0.0003, respectively, both explanatory variables exert a positive and statistically significant influence on loan approval at a 5% level. Also,  $b_1 = 0.0188$  implies that a 1-percentage-point increase in down payment increases the approval probability by 0.0188, or by 1.88%. Similarly, a 1-percentage-point increase in the income-to-loan ratio increases the approval probability by 0.0258 or by 2.58%.

**TABLE 17.12** LPM Model Results for Example 17.6

	Coefficients	Standard Error	t Stat	p-Value
Intercept	-0.8682	0.2811	-3.0889	0.0046
Down Payment (%)	0.0188	0.0070	2.6945	0.0120
Income-to-Loan (%)	0.0258	0.0063	4.1070	0.0003

We can use this estimated model to predict the approval probability for any applicant. Consider an applicant who puts 20% down ( $x_1 = 20$ ), and has an income of \$60,000 and a loan amount of \$200,000 ( $x_2 = (60/200) \times 100 = 30$ ). The predicted approval probability for this applicant is  $\hat{P} = -0.8682 + 0.0188(20) + 0.0258(30) = 0.2818$ . Similarly, with 30% down and the same income-to-ratio of 30%,  $\hat{P} = -0.8682 + 0.0188(30) + 0.0258(30) = 0.4698$ . In other words, as down payment increases by 10 percentage points, the predicted probability of approval increases by 0.1880 ( $= 0.4698 - 0.2818$ ), which is essentially the estimated slope, 0.0188, multiplied by 10. The estimated slope coefficient for the percentage of income-to-loan variable can be interpreted similarly.

Although it is easy to estimate and interpret a linear probability model, it has some shortcomings. The major shortcoming is that it can produce predicted probabilities that are greater than 1 or less than 0. For instance, for a down payment of 60%, with the same income-to-loan ratio of 30%, we get a predicted mortgage approval rate of  $\hat{P} = -0.8682 + 0.0188(60) + 0.0258(30) = 1.0338$ , a probability greater than one! Similarly, for a down payment of 5%, the model predicts a negative probability,  $\hat{P} = -0.8682 + 0.0188(5) + 0.0258(30) = -0.0002$ . Furthermore, the linearity of the relationship may also be questionable. For instance, we would expect a big increase in the probability of loan approval if the applicant makes a down payment of 30% instead of 20%. This increase in probability is likely to be much smaller if the same 10-percentage-point increase in down payment is from 60% to 70%. An LPM cannot differentiate between these two scenarios. For these reasons, we introduce the logit model, which is a more appropriate probability model for binary choice variables.

## The Logit Model

Recall that in an LPM model with a single explanatory variable,  $y = \beta_0 + \beta_1x + \varepsilon$ , the following relationship is implied:  $P = \beta_0 + \beta_1x$ . Here, the influence of  $x$  on  $P$ , captured by the slope  $\beta_1$ , is constant. As mentioned earlier, the limitation of the LPM model is that for any given slope, we can find some value of  $x$  for which the predicted probability is outside the  $[0,1]$  interval. We basically want a nonlinear specification that constrains the predicted probability between 0 and 1.

Consider the specification

$$P = \frac{\exp(\beta_0 + \beta_1x)}{1 + \exp(\beta_0 + \beta_1x)},$$

### LO 17.5

Use a logit model to estimate a binary response variable.

where  $\exp(\beta_0 + \beta_1 x) = e^{\beta_0 + \beta_1 x}$  and  $e \approx 2.718$ . This specification is the cumulative distribution function of the logistic distribution. Thus, the resulting regression model is called a logistic model, or simply a **logit model**. The logit model ensures that the probability is between 0 and 1 for all values of  $x$ .

The logit model cannot be estimated with standard ordinary least squares (OLS) procedures. Instead, we rely on the method of **maximum likelihood estimation (MLE)** to estimate a logit model. While the MLE of the logit model is not supported by Excel, it can easily be estimated with most statistical packages, including Minitab, JMP, and SPSS. The theory of MLE is beyond the scope of this book; however, given the relevance of the logit model in numerous business applications, it is important to be able to interpret the estimated logit model.

#### THE LOGIT MODEL

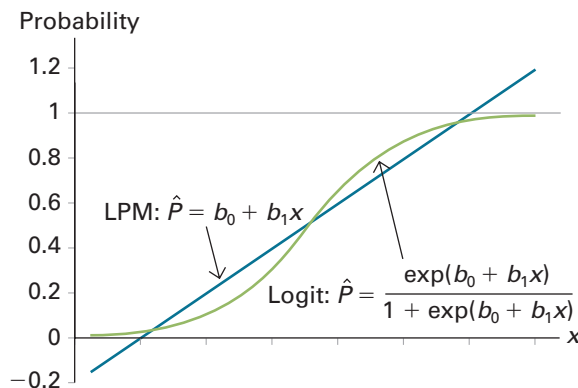
The **logit model** can be estimated with standard statistical packages. Predictions with this model are made by

$$\hat{P} = \frac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)},$$

where  $b_0$  and  $b_1$  are the estimates of the population parameters  $\beta_0$  and  $\beta_1$ .

Figure 17.4 highlights the relationship between the predicted probability  $\hat{P}$  and the explanatory variable  $x$  for an LPM and a logit model, given  $b_1 > 0$ . Note that in an LPM, the probability falls below 0 for small values of  $x$  and exceeds 1 for large values of  $x$ . The probabilities implied by a logit model, however, are always constrained in the  $[0, 1]$  interval. (For ease of exposition, we use the same notation to refer to the coefficients in the LPM and logit model. We note, however, that these coefficients and their estimates have a different meaning depending on which model we are referencing.)

**FIGURE 17.4**  
Predicted probabilities  
with an LPM and  
a logit model



It is important to be able to interpret the regression coefficients of a logit model. In an LPM, the interpretation of a regression coefficient is straightforward. For instance, if the estimated LPM is  $\hat{P} = -0.20 + 0.03x$ , it implies that for every 1-unit increase in  $x$ , the predicted probability  $\hat{P}$  increases by 0.03. We note that  $\hat{P}$  increases by 0.03, whether  $x$  increases from 10 to 11 or from 20 to 21.

Now consider the estimated logit model,  $\hat{P} = \frac{\exp(-2.10 + 0.18x)}{1 + \exp(-2.10 + 0.18x)}$ . Since the regression coefficient  $b_1 = 0.18$  is positive, we can infer that  $x$  exerts a positive influence on  $\hat{P}$ . However, the exact impact based on the estimated regression coefficient is not obvious. A useful method to interpret the estimated coefficient is to highlight the changing impact of  $x$  on  $\hat{P}$ . For instance, given  $x = 10$ , we compute the predicted probability as  $\hat{P} = \frac{\exp(-2.10 + 0.18 \times 10)}{1 + \exp(-2.10 + 0.18 \times 10)} = 0.43$ . Similarly, for  $x = 11$ , the predicted probability is  $\hat{P} = 0.47$ . Therefore, as  $x$  increases by one unit from 10 to 11, the predicted probability increases by 0.04. However, the increase in  $\hat{P}$  will not be the same if  $x$  increases from 20 to 21. We can show that  $\hat{P}$  increases from 0.82 when  $x = 20$  to 0.84 when  $x = 21$ , for a smaller increase of 0.02.

### EXAMPLE 17.7

There is a declining interest among teenagers in pursuing a career in science (*U.S. News & World Report*, May 23, 2011). In a recent survey, 50% of high school students showed no interest in the sciences. An educator wants to determine if a student's interest in science is linked with the student's GPA. She estimates a logit model where the choice of field (1 for choosing science, 0 otherwise) depends on the student's GPA. She uses Minitab to produce the logit regression results shown in Table 17.13.

**TABLE 17.13** Logit Regression Results for Example 17.7

Predictor	Coef	SE	Z	P
Constant	-4.4836	1.5258	-2.938	0.0033
GPA	1.5448	0.4774	3.236	0.0012

- Use a 5% level of significance to determine if GPA has a statistically significant influence on the probability of pursuing a career in science.
- Compute and interpret the probability that a student will pursue a career in science given a GPA of 3.0, 3.5, and 4.0.

#### SOLUTION:

- In order to determine the significance of GPA, we specify the competing hypotheses as  $H_0: \beta_1 = 0$  against  $H_A: \beta_1 \neq 0$ . Since the  $p$ -value = 0.0012 is less than  $\alpha = 0.05$ , we reject  $H_0$  and conclude that GPA influences the probability that a student pursues a career in science. (In maximum likelihood estimation, the significance tests are valid only with large samples. Consequently, we conduct the  $z$  test, in place of the usual  $t$  test, to evaluate the statistical significance of a coefficient.)
- Since the estimated regression coefficient for GPA is positive ( $b_1 = 1.5448$ ), it suggests that GPA exerts a positive influence on the predicted probability of pursuing a career in science. For a student with a GPA = 3.0, we compute the predicted probability as

$$\hat{P} = \frac{\exp(-4.4836 + 1.5448 \times 3.0)}{1 + \exp(-4.4836 + 1.5448 \times 3.0)} = 0.54.$$

Similarly, we compute the predicted probabilities for a student with GPA = 3.5 and GPA = 4.0 as 0.72 and 0.84, respectively. Note that the predicted probability increases by 0.18 (= 0.72 – 0.54) as GPA increases from 3.0 to 3.5. The increase is only 0.12 (= 0.84 – 0.72) when GPA increases from 3.5 and 4.0.

**FILE**  
Mortgage

### EXAMPLE 17.8

Let us revisit Example 17.6. Estimate and interpret a logit model for mortgage approval  $y$  based on the applicant's percentage of down payment  $x_1$  and the applicant's percentage of income-to-loan ratio  $x_2$ . Make predictions of the approval probability for representative applicants. Compare the results of the estimated logit model with those of the estimated linear probability model (LPM).

**SOLUTION:** We again use Minitab to estimate the logit model; Table 17.14 shows a portion of the regression output.

**TABLE 17.14** Logit Regression Results for Example 17.8

Predictor	Coef	SE	Z	P
Constant	–9.3671	3.1960	–2.9309	0.0034
Down Payment (%)	0.1349	0.0640	2.1074	0.0351
Income to Loan (%)	0.1782	0.0646	2.7577	0.0058

The estimated probability equation is computed as

$$\hat{P} = \frac{\exp(-9.3671 + 0.1349x_1 + 0.1782x_2)}{1 + \exp(-9.3671 + 0.1349x_1 + 0.1782x_2)}.$$

As in the case of the linear probability model, both explanatory variables exert a positive and statistically significant influence on loan approval at a 5% level, given positive estimated coefficients and  $p$ -values of 0.0351 and 0.0058, respectively.

We can use the estimated model to predict approval probabilities for any applicant. For instance, for an individual with  $x_1 = 20$  and  $x_2 = 30$ , the predicted approval probability is

$$\hat{P} = \frac{\exp(-9.3671 + 0.1349 \times 20 + 0.1782 \times 30)}{1 + \exp(-9.3671 + 0.1349 \times 20 + 0.1782 \times 30)} = 0.2103.$$

Table 17.15 provides predicted probabilities based on the linear probability model, estimated in Example 17.6, and the above logit model for selected values of  $x_1$  given  $x_2 = 30$ .

**TABLE 17.15** Predicted Probabilities with a LPM versus a Logit Model

Down Payment (%) $x_1$	Income to Loan Amount (%) $x_2$	LPM	Logit Model
5	30	–0.0002	0.0340
20	30	0.2818	0.2103
30	30	0.4698	0.5065
60	30	1.0338	0.9833

As discussed earlier, with a linear probability model, the predicted probabilities can be negative or greater than one. The probabilities based on a logit model always stay between zero and one for all possible values of the explanatory variables. Therefore, whenever possible, it is preferable to use the logit model over the linear probability model for binary choice models.

## EXERCISES 17.3

### Mechanics

27. Consider a binary response variable  $y$  and an explanatory variable  $x$  that varies between 0 and 4. The linear model is estimated as  $\hat{y} = -1.11 + 0.54x$ .

- Compute the estimated probability for  $x = 2$  and  $x = 3$ .
- For what values of  $x$  is the estimated probability negative or greater than one?

28. Consider a binary response variable  $y$  and an explanatory variable  $x$  that varies between 0 and 50. The linear probability model is estimated as  $\hat{y} = 0.92 - 0.02x$ .

- Compute the estimated probability for  $x = 25$  and  $x = 40$ .
- For what values of  $x$  is the estimated probability negative?

29. Consider a binary response variable  $y$  and an explanatory variable  $x$ . The following table contains the parameter estimates of the linear probability model (LPM) and the logit model, with the associated  $p$ -values shown in parentheses.

Variable	LPM	Logit
Constant	-0.72 (0.04)	-6.2 (0.04)
$x$	0.05 (0.06)	0.26 (0.02)

- Test for the significance of the intercept and the slope coefficients at a 5% level in both models.
- What is the predicted probability implied by the linear probability model for  $x = 20$  and  $x = 30$ ?
- What is the predicted probability implied by the logit model for  $x = 20$  and  $x = 30$ ?

30. Consider a binary response variable  $y$  and an explanatory variable  $x$ . The following table contains the parameter estimates of the linear probability model (LPM) and the logit model, with the associated  $p$ -values shown in parentheses.

Variable	LPM	Logit
Constant	-0.40 (0.03)	-4.50 (0.01)
$x$	0.32 (0.04)	1.54 (0.03)

- Use both models to predict the probability of success as  $x$  varies from 1 to 5 with increments of 1.
- Comment on the suitability of the linear probability model in modeling binary outcomes.

31. Consider a binary response variable  $y$  and two explanatory variables  $x_1$  and  $x_2$ . The following table contains the

parameter estimates of the linear probability model (LPM) and the logit model, with the associated  $p$ -values shown in parentheses.

Variable	LPM	Logit
Constant	-0.40 (0.03)	-2.20 (0.01)
$x_1$	0.32 (0.04)	0.98 (0.06)
$x_2$	-0.04 (0.01)	-0.20 (0.01)

- Comment on the significance of the variables.
- What is the predicted probability implied by the linear probability model for  $x_1 = 4$  with  $x_2$  equal to 10 and 20?
- What is the predicted probability implied by the logit model for  $x_1 = 4$  with  $x_2$  equal to 10 and 20?

32. Using 30 observations, the following regression output is obtained from estimating the linear probability model  $y = \beta_0 + \beta_1x + \varepsilon$ .

	Coefficients	Standard Error	$t$ Stat	$p$ -Value
Intercept	1.31	0.31	4.17	0.0002
$x$	-0.04	0.01	-2.67	0.0125

- What is the predicted probability when  $x = 20$ ?
- Is  $x$  significant at the 5% level?

33. Using 30 observations, the following output was obtained when estimating the logit model.

Predictor	Coef	SE	Z	P
Constant	-0.188	0.083	2.27	0.024
$x$	3.852	1.771	2.18	0.030

- What is the predicted probability when  $x = 0.40$ ?
- Is  $x$  significant at the 5% level?

34. Using 40 observations, the following output was obtained when estimating the logit model.

Predictor	Coef	SE	Z	P
Constant	1.609	1.405	1.145	0.252
$x_1$	-0.194	0.143	-1.357	0.177
$x_2$	0.202	0.215	0.940	0.348
$x_3$	0.223	0.086	2.593	0.010

- What is the predicted probability when  $x_1 = 15$ ,  $x_2 = 10$ , and  $x_3 = -2$ ?
- At the 5% significance level, which of the explanatory variables are significant?

## Applications

35. **FILE Purchase.** Annabel, a retail analyst, has been following Under Armour, Inc., the pioneer in the compression-gear market. Compression garments are meant to keep moisture away from a wearer's body during athletic activities in warm and cool weather. Annabel believes that the Under Armour brand attracts a younger customer, whereas the more established companies, Nike and Adidas, draw an older clientele. In order to test her belief, she collects data on the age of the customers and whether or not they purchased Under Armour (1 for Under Armour, 0 otherwise). A portion of the data is shown in the accompanying table.

Under Armour	Age
1	30
0	19
⋮	⋮
1	24

- Estimate a linear probability model using Under Armour as the response variable and age as the explanatory variable.
  - Compute the predicted probability of an Under Armour purchase for a 20-year-old customer and a 30-year-old customer.
  - Test Annabel's belief that the Under Armour brand attracts a younger customer at the 5% level.
36. **FILE Purchase.** Redo exercise 35 using a logit model.
- Compute the predicted probability of an Under Armour purchase for a 20-year-old customer and a 30-year-old customer.
  - Test Annabel's belief that the Under Armour brand attracts a younger customer at the 5% level.
37. **FILE Health.** According to the National Coalition on Health Care, there has been a steady decline in the proportion of Americans who have health insurance. The rising insurance premiums have made it difficult for small employers to offer insurance and those that do offer insurance are contributing a smaller share of the premium. As a result, an increasing number of Americans do not have health insurance because they cannot afford it. Consider a portion of data in the following table relating to insurance coverage (1 for coverage, 0 for no coverage) for 30 working individuals in Atlanta, Georgia. Also included in the table is the percentage of the premium paid by the employer and the individual's income (in \$1,000s).

Insurance	Premium Percentage (in %)	Income (in \$1,000s)
1	0	88
0	0	60
⋮	⋮	⋮
0	60	60

- Analyze a linear probability model for insurance coverage with premium percentage and income used as the explanatory variables.
  - Consider an individual with an income of \$60,000. What is the probability that she has insurance coverage if her employer contributes 50% of the premium? What if her employer contributes 75% of the premium?
38. **FILE Health.** Redo exercise 37 using a logit model. Consider an individual with an income of \$60,000. What is the probability that she has insurance coverage if her employer contributes 50% of the premium? What if her employer contributes 75% of the premium?
39. **FILE Divorce.** According to a recent estimate, the divorce rate in England has fallen to a 26-year low (*The Guardian*, August 29, 2008). However, it is documented that the rate of divorce is more than twice as high for men and women aged 25 to 29. John Haddock is a sociologist from Sussex University who wants to analyze the divorce rate based on the individual's age, family income, and the number of children that the couple has. He collects data on 30 individuals in a small town near Brighton, a portion of which is shown in the accompanying table.

Divorce	Age	Income (in £1,000s)	Children
0	1	19	3
0	0	46	3
⋮	⋮	⋮	⋮
0	0	26	0

- Estimate and interpret a linear probability model where divorce (1 for divorce; 0 otherwise) depends on age (1 if 25–29 years old; 0 otherwise), family income (in £1,000s), and the number of children.
  - Do the data support the article's claim that the divorce rate is higher for those aged 25–29 years old? Explain.
  - Use the above estimates to predict the probability of divorce for an individual who is 27 years old, has £60,000 of family income and one child. Recalculate the probability with three children.
40. **FILE Divorce.** Redo exercise 39 using the logit model.
- Do the data support the article's claim that the divorce rate is higher for those aged 25–29 years old? Explain.
  - Use the above estimates to predict the probability of divorce for an individual who is 27 years old, has £60,000 of family income, and one child. Recalculate the probability with three children.



## WRITING WITH STATISTICS

During the 2009–2010 NBA season, the Los Angeles Lakers had the highest offensive production throughout the league. Led by Kobe Bryant, the Lakers beat the Boston Celtics in game seven of the championships for the 2010 NBA title. Jaqueline Thomsen, an amateur statistician, would like to examine the factors that led to the Lakers' success. Specifically, Jaqueline wishes to predict the likelihood of a Lakers win as a function of field goal percentage (FG), rebounds (Rebounds), and turnovers (Turnovers). The probability of winning should be positively influenced by FG and Rebounds but negatively affected by Turnovers. In addition, she wonders if there is a home court advantage in that playing at home significantly influences the team's chances of winning. Table 17.16 shows a portion of data on the Lakers' 82-game regular season.



**TABLE 17.16** Statistics on the Los Angeles Lakers' 2009–2010 Regular Season

Game	Win/Loss	FG %	Rebounds	Turnovers	Home/Away
1	Win	41.2	47	16	Home
2	Loss	39.5	40	19	Home
⋮	⋮	⋮	⋮	⋮	⋮
82	Loss	39.5	49	14	Away

SOURCE: [www.nba.com](http://www.nba.com).

**FILE**  
Lakers

Jaqueline would like to use the above sample information to:

1. Choose an appropriate model to predict the probability of winning.
2. Determine whether there is a home court advantage.
3. Predict the probability of winning if the Lakers are playing at home or away, with average values of FG, Rebounds, and Turnovers.

With the highest offensive production throughout the league during the 2009–2010 season, it is not surprising that the Los Angeles Lakers won the 2010 NBA championship. Other teams might benefit if they could unravel the factors that led to the Lakers' success. In an attempt to examine the factors that influence a team's chances of winning, regression analysis is conducted on the Lakers' 82-game regular season. The response variable is Win (equals 1 for a win, 0 otherwise) and the explanatory variables include:

- The team's field goal percentage (FG),
- The number of rebounds,
- The number of turnovers, and
- A "home" dummy variable that equals 1 for a home game and 0 otherwise.

The probability of winning should be positively influenced by FG and rebounds, but negatively affected by turnovers. In addition, if there truly is a home court advantage, then playing at home should positively influence the team's chances of winning.

Two models are evaluated that link the probability of winning with the explanatory variables: the linear probability model and the logit model. The parameter estimates of both models are shown in Table 17.A.

**Sample  
Report—  
Predicting the  
Probability  
of Winning**

**TABLE 17.A** Regression Results of the Linear Probability Model and the Logit Model

Response Variable: Win (equals 1 if Lakers win, 0 otherwise)		
	LPM	Logit
Constant	-2.391* (0.00)	-28.76* (0.00)
FG	0.047* (0.00)	0.49* (0.00)
Rebounds	0.019* (0.00)	0.17* (0.02)
Turnovers	-0.004 (0.68)	-0.04 (0.71)
Home	0.232* (0.01)	1.82* (0.02)

NOTES: Parameter estimates of both models are presented with  $p$ -values in parentheses; \* represents significance at the 5% level.

The linear probability model is estimated as  $\widehat{\text{Win}} = -2.391 + 0.047\text{FG} + 0.019\text{Rebounds} - 0.004\text{Turnovers} + 0.232\text{Home}$ . All signs on the estimated slope coefficients are as expected; that is, the field goal percentage, the number of rebounds, and playing at home exert a positive influence on the chances of winning; the number of turnovers suggests a negative relationship with the response variable. Upon testing the explanatory variables individually, the extremely small  $p$ -values associated with FG, Rebounds, and Home reveal that these variables have a significant relationship with the probability of winning; Turnovers is not significant at the 5% level. The slope coefficient of Home indicates that the likelihood of winning increases by approximately 23% if the Lakers play at home. While the results of the linear probability model seem reasonable, some values of the explanatory variables may yield predicted probabilities that are either negative or greater than one. In order to avoid this possibility, the logit model is preferred.

The logit model is estimated as

$$\widehat{\text{Win}} = \frac{\exp(-28.76 + 0.49\text{FG} + 0.17\text{Rebounds} - 0.04\text{Turnovers} + 1.82\text{Home})}{1 + \exp(-28.76 + 0.49\text{FG} + 0.17\text{Rebounds} - 0.04\text{Turnovers} + 1.82\text{Home})}.$$

As in the case of the linear probability model, FG, Rebounds, and Home are again individually significant at the 5% level; thus, the significance of Home supports the belief of a home field advantage. Over the 82-game season, the averages for field goal percentage, the number of rebounds, and the number of turnovers were 45%, 44, and 13, respectively. If the Lakers are playing an “average” game away from home, then the model predicts a 56.2% probability of winning. However, if they are playing an “average” game at home, then their probability of winning jumps to 88.8%. In sum, the home court advantage overwhelmingly puts the likelihood of success in their favor.

## CONCEPTUAL REVIEW

### LO 17.1 Use dummy variables to represent qualitative explanatory variables.

A **dummy variable**  $d$  is defined as a variable that takes on values of 1 or 0. Dummy variables are used to represent categories of a qualitative variable. The number of dummy variables needed should be one less than the number of categories of the variable.

A regression model with a quantitative variable  $x$  and a dummy variable  $d$  is specified as  $y = \beta_0 + \beta_1x + \beta_2d + \varepsilon$ . The dummy variable  $d$  allows the predicted  $y$  to differ between the two categories of a qualitative variable by a fixed amount across the values of  $x$ . We estimate this model to make predictions as  $\hat{y} = (b_0 + b_2) + b_1x$  for  $d = 1$  and as  $\hat{y} = b_0 + b_1x$  for  $d = 0$ .

**LO 17.2 Test for differences between the categories of a qualitative variable.**

Using  $y = \beta_0 + \beta_1x + \beta_2d + \varepsilon$ , we can perform a standard  $t$  test to determine whether difference exists between two categories.

**LO 17.3 Use dummy variables to capture interactions between qualitative and quantitative explanatory variables.**

A regression model with a dummy variable  $d$ , a quantitative variable  $x$ , and an interaction variable  $xd$  is specified by  $y = \beta_0 + \beta_1x + \beta_2d + \beta_3xd + \varepsilon$ . We estimate this model to make predictions as  $\hat{y} = (b_0 + b_2) + (b_1 + b_3)x$  for  $d = 1$ , and as  $\hat{y} = b_0 + b_1x$  for  $d = 0$ . The interaction variable  $xd$  allows the predicted  $y$  to differ between the two categories of a qualitative variable by a varying amount across the values of  $x$ . In addition to performing a  $t$  test to determine the individual significance of  $d$  or  $xd$ , we can also implement the partial  $F$  test to determine their joint significance.

**LO 17.4 Use a linear probability model to estimate a binary response variable.**

Models that use a dummy (binary) variable as the response variable are called binary choice models. A **linear probability model (LPM)** is specified as  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k + \varepsilon = P(y = 1) + \varepsilon$ , where  $y$  assumes values of 0 or 1 and  $P(y = 1)$  is the probability of success.

Predictions with this model are made by  $\hat{P} = \hat{y} = b_0 + b_1x + b_2x_2 + \cdots + b_kx_k$ , where  $b_0, b_1, b_2, \dots, b_k$  are the estimates.

The major shortcoming of the LPM is that it can produce predicted probabilities that are greater than one or less than zero.

**LO 17.5 Use a logit model to estimate a binary response variable.**

A **logit model** can be estimated with standard statistical packages. Predictions with this model are made by  $\hat{P} = \frac{\exp(b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k)}{1 + \exp(b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k)}$ , where  $b_0, b_1, b_2, \dots, b_k$  are the estimates. The estimated model ensures that the predicted probability of the binary response variable falls between zero and one.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

41. **FILE Magellan.** A financial analyst would like to determine whether the return on Fidelity's Magellan mutual fund varies depending on the quarter; that is, if there is a seasonal component describing return. He collects 10 years of quarterly return data. A portion is shown in the accompanying table.

Year	Quarter	Return	$d_1$	$d_2$	$d_3$
2000	1	4.85	1	0	0
2000	2	-3.96	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2009	4	4.06	0	0	0

Source: <http://finance.yahoo.com>.

- a. Estimate  $y = \beta_0 + \beta_1d_1 + \beta_2d_2 + \beta_3d_3 + \varepsilon$ , where  $y$  is Magellan's quarterly return,  $d_1$  is a dummy

variable that equals 1 if quarter 1 and 0 otherwise,  $d_2$  is a dummy variable that equals 1 if quarter 2 and 0 otherwise, and  $d_3$  is a dummy variable that equals 1 if quarter 3 and 0 otherwise.

- Interpret the slope coefficients of the dummy variables.
- Predict Magellan's stock return in quarters 2 and 4.

42. **FILE Hiring.** In a seminal study, researchers documented race-based hiring in the Boston and Chicago labor markets (*American Economic Review*, September 2004). They sent out identical resumes to employers, half with traditionally African-American names and the other half with traditionally Caucasian names. Interestingly, there was a 53% difference in call-back rates between the two groups of people. A research fellow at an institute in Santa Barbara decides to repeat the same experiment with names along with age in the Los Angeles labor market. She repeatedly sends out resumes for sales positions in the city that are identical except for the difference in the names and ages of the applicants. She also records the call-back rate for each candidate. The accompanying table shows a portion of data on call-back rate (%), age, and a Caucasian dummy that equals 1 for a Caucasian-sounding name.

Call-back	Age	Caucasian
12	60	1
9	56	0
⋮	⋮	⋮
15	38	0

- Estimate a linear regression model with call-back as the response variable, and age and the Caucasian dummy variable as the explanatory variables.
  - Compute the call-back rate for a 30-year-old applicant with a Caucasian-sounding name. What is the corresponding call-back rate for a non-Caucasian?
  - Conduct a test for race discrimination at the 5% significance level.
43. An analyst studies quarterly data on the relationship between retail sales ( $y$ , in \$ millions), gross national product ( $x$ , in \$ billions), and a quarterly dummy  $d$  that equals 1 if the sales are for the 4th quarter; 0 otherwise. He estimates the model  $y = \beta_0 + \beta_1x + \beta_2d + \beta_3xd + \varepsilon$ . Relevant regression results are shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-Value
Intercept	186553.3	56421.1	3.31	0.0021
$x$	55.0	4.6	12.08	0.0000
$d$	112605.8	117053.0	0.96	0.3424
$xd$	-4.7	9.3	-0.50	0.6178

- Interpret the dummy variable,  $d$ . Is it significant at the 5% level?
- Interpret the interaction variable. Is it significant at the 5% level?

44. **FILE Overweight.** According to the U.S. Department of Health and Human Services, African-American women have the highest rates of being overweight compared to other groups in the U.S. Individuals are considered overweight if their body mass index (BMI) is 25 or greater. The following table shows a portion of data on BMI of 120 individuals and the corresponding gender and race dummy variables.

BMI	Female	Black
28.70	0	1
28.31	0	0
⋮	⋮	⋮
24.90	0	1

Note: Female = 1 for females and 0 for males; Black = 1 for African Americans and 0 otherwise.

- Estimate the model,  $BMI = \beta_0 + \beta_1\text{Female} + \beta_2\text{Black} + \beta_3(\text{Female} \times \text{Black}) + \varepsilon$ , to predict the BMI for white males, white females, black males, and black females.
  - Is the difference between white females and white males statistically significant at the 5% level?
  - Is the difference between white males and black males statistically significant at the 5% level?
45. **FILE Longevity.** According to the Center for Disease Control and Prevention, life expectancy at age 65 in America is about 18.7 years. Medical researchers have argued that while excessive drinking is detrimental to health, drinking a little alcohol every day, especially wine, may be associated with an increase in life expectancy. Others have also linked longevity with income and gender. The accompanying table shows a portion of data relating to the length of life after 65, average income (in \$1,000s) at a retirement age of 65, a "woman" dummy, and the average number of alcoholic drinks consumed per day.

Life	Income (in \$1,000s)	Woman	Drinks
19.00	64	0	1
19.30	43	1	3
⋮	⋮	⋮	⋮
20.24	36	1	0

- Use the data to model life expectancy at 65 on the basis of Income, Woman, and Drinks.
- Conduct a one-tailed test at  $\alpha = 0.01$  to determine if women live longer than men.

- c. Predict the life expectancy at 65 of a man with an income of \$40,000 and an alcoholic consumption of two drinks per day; repeat the prediction for a woman.

46. **FILE Shifts.** The manager of a diner wants to reevaluate his staffing needs depending on variations in customer traffic during the day. He collects data on the number of customers served along with four dummy variables representing the morning, afternoon, evening, and night shifts. The dummy variable Morning equals 1 if the information was from the morning shift and 0 otherwise; other dummy variables are defined similarly. The accompanying table shows a portion of the data.

Customers	Morning	Afternoon	Evening	Night
99	0	0	0	1
148	0	1	0	0
⋮	⋮	⋮	⋮	⋮
111	0	1	0	0

- Estimate a regression model using the number of customers as the response variable and the shift dummy variables as the explanatory variables.
  - What is the predicted number of customers served during the morning, afternoon, evening, and night shifts?
  - Estimate the appropriate model to determine, at the 5% significance level, if the diner is busier in the afternoon compared to other times.
47. **FILE Study.** A researcher in the education department wants to determine if the number of hours that business students study per week at a state university varies by quarter. He conducts a survey where business students are asked how much they study per week in each of the three quarters. He defines a dummy variable Fall that equals 1 if the survey was conducted in the fall quarter and 0 otherwise. The dummy variables Winter and Spring are defined similarly. The accompanying table shows a portion of the data for 120 students.

Study Hours	Fall	Winter	Spring
15	0	0	1
16	0	1	0
⋮	⋮	⋮	⋮
14	0	0	1

- Estimate the appropriate model to determine, at the 5% significance level, if students study the least in the spring quarter.
- Find the predicted number of hours that students study per week in the fall and spring quarters.

48. **FILE Compensation.** To encourage performance, loyalty, and continuing education, the human resources department at a large company wants to develop a regression-based compensation model for mid-level managers based on three variables: (1) business unit-profitability, (2) years with company, and (3) graduate degree in relevant field. The accompanying table shows a portion of data collected for 36 managers.

Compensation (\$/year)	Business-Unit Profit (\$000/year)	Years with Company	Graduate Degree? (1=yes, 0=no)
118,100	4,500	37	1
90,800	5,400	5	1
⋮	⋮	⋮	⋮
85,000	4,200	29	0

- a. Estimate the following model for compensation:

$$y = \beta_0 + \beta_1 \text{ Profit} + \beta_2 \text{ Years} + \beta_3 \text{ Grad} + \beta_4 \text{ Profit} \times \text{Years} + \beta_5 \text{ Profit} \times \text{Grad} + \beta_6 \text{ Years} \times \text{Grad} + \varepsilon.$$

- At the 5% significance level, is the overall regression model significant?
  - Which predictor variables and interaction terms are significant at  $\alpha = 0.05$ ?
  - Use the (full) model to determine compensation for a manager having 15 years with the company, a graduate degree, and a business-unit profit of \$4,800(000) last year.
49. **FILE Assembly.** Assembly line work can be tedious and repetitive. Therefore, it is not suited for everybody. Consequently, a production manager is developing a binary choice regression model to predict whether a newly hired worker will stay in the job for at least one year. Three explanatory variables will be used: (1) age, (2) gender, and (3) whether the new hire has worked on an assembly line before. Records have been obtained for the past 32 assembly line workers hired. A portion of the data is shown in the accompanying table.

Still in Job One Year Later? (1=yes, 0=no)	Age	Gender (1=female, 0=male)	Worked on Assembly Line Before? (1=yes, 0=no)
0	35	1	0
0	26	1	0
⋮	⋮	⋮	⋮
1	38	0	1

- Estimate a linear probability model in which being in the job one year later depends on age, gender, and whether the new hire has worked on an assembly line before. Use this model to predict the probability that (1) a 45-year-old female who has not worked on an assembly line before will still be in the job one year later and (2) a



35-year-old male who has worked on an assembly line will still be in the job one year later.

- b. Estimate a logit model where being in the job one year later depends on age, gender, and whether the new hire has worked on an assembly line before. Use this model to predict the probability that (1) a 45-year-old female who has not worked on an assembly line before will still be in the job one year later and (2) a 35-year-old male who has worked on an assembly line will still be in the job one year later.
- c. At  $\alpha = 0.05$ , compare the significance of the parameters in both models. What do the significance results imply from a practical standpoint?

50. **FILE SetonHall.** Seton Hall University is a Roman Catholic university situated in New Jersey, with easy access to New York City. Like most universities, it uses SAT scores and high school GPA as primary criteria for admission. The accompanying table shows a portion of data concerning information on admission (1 for admission and 0 otherwise), SAT score, and GPA for 30 students who had recently applied to Seton Hall.

Admission	SAT	GPA
1	1700	3.39
1	2020	2.65
⋮	⋮	⋮
0	1300	2.47

- a. Estimate the linear probability model where admission is a function of the SAT score and high school GPA. Analyze the significance of the variables at the 5% level.
- b. Use these estimates to predict the probability of admission for an individual with a GPA of 3.5 and an SAT score of 1700.

- c. Reestimate the probability of admission for an individual with a GPA of 3.5 and an SAT score of 1800.

51. **FILE SetonHall.** Redo exercise 50 using the logit model.
52. **FILE Parole.** More and more parole boards are using risk assessment tools when trying to determine an individual's likelihood of returning to crime (*The Boston Globe*, February 20, 2011). Most of these models are based on a range of character traits and biographical facts about an individual. Many studies have found that older people are less likely to re-offend than younger ones. In addition, once released on parole, women are not likely to re-offend. A sociologist collects data on 20 individuals who were released on parole two years ago. She notes if he/she committed another crime over the last two years (crime equals 1 if crime committed, 0 otherwise), the individual's age at the time of release, and the gender of the individual (gender equals 1 if male, 0 otherwise). The accompanying table shows a portion of the data.

Crime	Age	Gender
1	25	1
0	42	1
⋮	⋮	⋮
0	30	1

- a. Estimate the linear probability model where crime depends on age and gender.
  - b. Are the results consistent with the claims of other studies with respect to age and gender?
  - c. Predict the probability of a 25-year-old male parolee committing another crime; repeat the prediction for a 25-year-old female parolee.
53. **FILE Parole.** Redo exercise 52 using the logit model.

## CASE STUDIES

**CASE STUDY 17.1** A recent study examined “sidewalk rage” in an attempt to find insight into anger’s origins and offer suggestions for anger-management treatments (*The Wall Street Journal*, February 15, 2011). “Sidewalk ragers” tend to believe that pedestrians should behave in a certain way. For instance, slower pedestrians should keep to the right or should step aside to take a picture. If pedestrians violate these “norms,” then ragers feel that the “violaters” are breaking the rules of civility. Since anger is associated with a host of negative health consequences, psychologists suggest developing strategies to quell the rage. One possible strategy is to avoid slow walkers. A portion of the study looked at the average speed of walkers (feet per second) in Lower Manhattan and found that average speeds differ when the pedestrian is distracted by other activities (smoking, talking on a cell phone, tourism, etc.) or exhibits other traits (elderly, obese, etc.). Sample



data were obtained from 50 pedestrians in Lower Manhattan. Each pedestrian's speed was calculated (feet per second). In addition, it was noted if the pedestrian was smoking (equalled 1 if smoking, 0 otherwise), was a tourist (equalled 1 if tourist, 0 otherwise), was elderly (equalled 1 if over 65 years old, 0 otherwise), and/or was obese (equalled 1 if obese, 0 otherwise). Each pedestrian is associated with no more than one of these four characteristics/traits. The accompanying table shows a portion of the data.

**Data for Case Study 17.1** Pedestrian Speeds with Defining Characteristics/Traits

Speed	Smoking	Tourist	Elderly	Obese
3.76	0	1	0	0
3.82	0	1	0	0
⋮	⋮	⋮	⋮	⋮
5.02	0	0	0	0

**FILE**  
PedSpeeds

In a report, use the sample information to:

1. Estimate  $\text{Speed} = \beta_0 + \beta_1\text{Smoking} + \beta_2\text{Tourist} + \beta_3\text{Elderly} + \beta_4\text{Obese} + \varepsilon$ .
2. Interpret the slope coefficient of tourist. Interpret the intercept. Predict the speed of an elderly pedestrian. Predict the speed of an obese pedestrian.
3. Are the explanatory variables jointly significant in explaining speed at the 5% significance level? Are all explanatory variables individually significant at the 5% level? What type of pedestrian should a “sidewalk rager” avoid?

**CASE STUDY 17.2** Jack Sprague is the relocation specialist for a real estate firm in the town of Arlington, Massachusetts. He has been working with a client who wishes to purchase a single-family home in Arlington. After seeing the information that Jack provided, the client is perplexed by the variability of home prices in Arlington. She is especially puzzled by the premium that a colonial house commands. (A colonial house is a style dating back to the time of the American colonies, with a simple rectangular structure and a peaked roof.) Despite Jack's eloquent explanations, it seems that the client will not be satisfied until she understands the quantitative relationship between house prices and house characteristics. Jack decides to use a multiple regression model to provide the client with the necessary information. He collects data on the prices for 36 single-family homes in Arlington sold in the first quarter of 2009. Also included in the data is the information on square footage, the number of bedrooms, the number of bathrooms, and whether or not the house is a colonial (1 for colonial; 0 otherwise). A portion of the data is shown in the accompanying table.

**Data for Case Study 17.2** Sales Information of Single-Family Homes in Arlington, MA

Price	Square feet	Bedrooms	Baths	Colonial
\$840,000	2,768	4	3.5	1
822,000	2,500	4	2.5	1
⋮	⋮	⋮	⋮	⋮
307,500	850	1	1	0

**FILE**  
Arlington

SOURCE: NewEnglandMoves.com.

In a report, use the sample information to:

1. Estimate and interpret three models, where  $d$  is the colonial dummy variable.  
 Model 1:  $\text{Price} = \beta_0 + \beta_1\text{Sqft} + \beta_2\text{Beds} + \beta_3\text{Baths} + \beta_4d + \varepsilon$ .  
 Model 2:  $\text{Price} = \beta_0 + \beta_1\text{Sqft} + \beta_2\text{Beds} + \beta_3\text{Baths} + \beta_4(\text{Sqft} \times d) + \varepsilon$ .  
 Model 3:  $\text{Price} = \beta_0 + \beta_1\text{Sqft} + \beta_2\text{Beds} + \beta_3\text{Baths} + \beta_4d + \beta_5(\text{Sqft} \times d) + \varepsilon$ .
2. Choose which model is more reliable in predicting the price of a house. Provide at least one reason for your choice. Are price differences between colonial homes versus other styles fixed and/or changing at the 5% significance level?

3. Use this model to make predictions for a colonial home versus other styles, given the average values of the explanatory variables.

**CASE STUDY 17.3** The Chartered Financial Analyst (CFA®) designation is fast becoming a requirement for serious investment professionals. Although it requires successfully completing three levels of grueling exams, it also promises great careers with lucrative salaries. Susan Wayne works as a research analyst at Fidelity Investments. She is thinking about taking the CFA exam in the summer and wants to understand why the recent pass rate for Level I has been under 40%. She firmly believes that those who were good students in college have a better chance of passing. She has also been told that work experience helps. She has access to the information on 30 Fidelity employees who took the test last year, including their success on the exam (1 for pass, 0 for fail), their college GPA, and years of work experience. A portion of the data is shown in the accompanying table.

**Data for Case Study 17.3** Information on Individuals Who Took CFA Exam

**FILE**  
CFA

Pass	GPA	Experience
1	3.64	12
0	3.16	5
⋮	⋮	⋮
0	2.64	4

In a report, use the sample information to:

1. Analyze a linear probability model to explain the probability of success. Predict the probability of passing the CFA exam for a candidate with various values of college GPA and years of experience.
2. Analyze the logit model to explain the probability of success. Predict the probability of passing the CFA exam for a candidate with various values of college GPA and years of experience.
3. Which model is more reliable for predicting the probability of passing the CFA exam? Provide at least one reason for your choice.

## APPENDIX 17.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Where a data file is specified, copy and paste it into the relevant software spreadsheet prior to following the commands.

**FILE**  
Professor

Even though some software packages automatically adjust categorical variables into dummy variables in a regression model, we find it easier to make the adjustment while in Excel. For example, suppose we have the *Professor* data and want to convert the Gender variable into a dummy variable. Label a new column as d1 and in the cell directly below d1, input “= IF(D2 =“Male”, 1, 0)”. (Note that D2 is the cell below the column labeled Gender.) Highlight this cell and all other cells that need to be converted, and from the menu choose **Home > Fill > Down**. Convert the Age variable in a similar manner. Import the reformat- ted data into one of the software spreadsheets, and estimate the regression model using the standard commands.

## Minitab

### Estimating a Logit Model

(Replicating Example 17.8) From the menu choose **Stat > Regression > Binary Logistic Regression > Fit binary logistic regression**. Select **Response in response/frequency format**, and after **Response** select *y*. After **Continuous predictors**, select *x*<sub>1</sub> and *x*<sub>2</sub>. Choose **Results**, and after **Display of results**, select **Expanded tables**.

**FILE**  
*Mortgage*

## SPSS

### Estimating a Logit Model

(Replicating Example 17.8) From the menu select **Analyze > Regression > Binary Logistic**. Under **Dependent**, select *y* and under **Covariates**, select *x*<sub>1</sub> and *x*<sub>2</sub>.

**FILE**  
*Mortgage*

## JMP

### Estimating a Logit Model

- A. (Replicating Example 17.8) Right-click on *y*, and under **Data Type**, select **Character**, and under **Modeling Type**, select **Nominal**.
- B. From the menu choose **Analyze > Fit Model**.
- C. Under **Select Columns**, select *y*, and then under **Pick Role Variables**, select **Y**. Under **Select Columns**, select *x*<sub>1</sub> and *x*<sub>2</sub>, and then under **Construct Model Effect**, select **Add**. Note that under **Personality**, you now see **Nominal Logistic** rather than **Standard Least Squares**. (Note: By default, JMP fits the 0 response. Thus, the results have opposite signs from those in the text.)

**FILE**  
*Mortgage*

# 18

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 18.1 Distinguish among the various models used in forecasting.
- LO 18.2 Use smoothing techniques to make forecasts.
- LO 18.3 Use trend regression models to make forecasts.
- LO 18.4 Calculate and interpret seasonal indices and use them to seasonally adjust a time series.
- LO 18.5 Use decomposition analysis to make forecasts.
- LO 18.6 Use trend regression models with seasonal dummy variables to make forecasts.
- LO 18.7 Use causal forecasting models to make forecasts.

# Time Series and Forecasting

Forecasting is an important aspect of statistical analysis, providing guidance for decisions in all areas of business. Examples include forecasting product sales, product defects, the inflation rate, the price of a financial asset, or a company's cash flows. In fact, the success of any business or government agency depends on the ability to accurately forecast many vital variables. Sound forecasts not only improve the quality of business plans, but also help identify and evaluate potential risks. The field of forecasting has developed rapidly over the last few decades, with some approaches requiring highly sophisticated techniques. In this chapter we focus on some of the easier approaches, which nevertheless provide a flavor and insight into this fascinating field. In particular, we use simple smoothing techniques for making forecasts when short-term fluctuations in the data represent random departures from the overall pattern with no discernible trend or seasonal fluctuations. Special forecasting methods are introduced when trend and seasonal fluctuations are present in the data. We will also explain a regression approach for forecasting.



## INTRODUCTORY CASE

### Nike Revenue Forecast

Chad Moriarty, a research analyst at a small investment firm, is evaluating Nike Inc.'s performance by analyzing the firm's revenues. Some analysts argue that Nike's revenue may slow down due to the global economic crisis and increased competition from emerging brands. Others believe that with a strong and free cash flow, Nike will likely survive this current environment and emerge stronger as some of the weaker competitors get squeezed. Chad fully understands that nobody really knows how well this Oregon-based sportswear company will perform in a softening global economy. However, he believes that Nike's past performance will aid in predicting its future performance. He collects quarterly data on Nike's revenue for the fiscal years 1999 through 2008, where data for fiscal year 1999, for example, refer to the time period from June 1, 1998 through May 31, 1999. A portion of the data is shown in Table 18.1.

**TABLE 18.1** Quarterly Revenue for Nike, Inc. (in millions \$)

**FILE**  
*Nike\_Revenues*

Year	Quarters Ended			
	August 31	November 30	February 28	May 31
1999	2,505	1,913	2,177	2,182
2000	2,501	2,060	2,162	2,273
⋮	⋮	⋮	⋮	⋮
2008	4,655	4,340	4,544	5,088

NOTES: All data retrieved from Annual Reports for Nike, Inc.

Chad would like to use the information in Table 18.1 to:

1. Determine whether revenue exhibits any sort of trend.
2. Determine whether revenue exhibits a significant seasonal component.
3. Forecast revenue for fiscal year 2009.

A synopsis of this case is provided at the end of Section 18.4.



Distinguish among the various models used in forecasting.

In this chapter, we focus our attention on **time series** data. Observations of any variable recorded over time in sequential order are considered a time series. The time period can be expressed in terms of a year, a quarter, a month, a week, a day, or even an hour. Examples of time series include the *hourly* volume of stocks traded on the New York Stock Exchange (NYSE) on four consecutive days; the number of *daily* visitors that frequent the Statue of Liberty over the month of June; the *monthly* sales for a retailer over a five-year period; and the growth rate of a country over the past 15 years.

A **time series** is a set of sequential observations of a variable over time.

Let  $y_1, y_2, \dots, y_T$  represent a sample of  $T$  observations of a variable of interest  $y$  with  $y_t$  denoting the value of  $y$  at time  $t$ . With time series data, it is customary to use the notation  $T$ , instead of  $n$ , to represent the number of sample observations and to use a subscript  $t$  to identify time. For instance, if the number of daily visitors (in 1,000s) to the Statue of Liberty over five days are 100, 94, 98, 110, 102, then  $y_1 = 100, y_2 = 94, \dots, y_5 = 102$ .

## Forecasting Methods

Forecasting methods are broadly classified as **quantitative** or **qualitative**. Qualitative forecasting procedures are based on the judgment of the forecaster, who uses prior experience and expertise to make forecasts. On the other hand, quantitative forecasting uses a formal model along with historical data for the variable of interest.

Qualitative forecasting is especially attractive when past data are either not available or are misleading. For instance, a manager may use qualitative forecasts when she attempts to project sales for a brand new product, or when a major structural change in market conditions has rendered previous data obsolete. Similarly, an economist may use qualitative forecasts of credit flow resulting from a newly introduced stimulus package by the federal government.

Although attractive in certain scenarios, qualitative forecasts are often criticized on the ground that they are prone to some well-documented biases such as optimism and overconfidence. Decisions based on the judgment of an overly optimistic manager may prove costly to the business. Also, qualitative forecasting is difficult to document and its quality is totally dependent on the judgment and skill of the forecaster. Two people with access to similar information may offer different qualitative forecasts.

In this chapter, we focus on quantitative forecasting. Formal quantitative models have been used extensively to forecast variables such as sales, inflation, and housing starts. These models are further split up into **causal** and **noncausal** models. Causal methods are based on a regression framework, where the variable of interest is related to a single or multiple explanatory variables. In other words, forecasts are “caused” by the known values of the explanatory variables. Noncausal models, also referred to as purely time series models, do not present any explanation of the mechanism generating the variable of interest and simply provide a method for projecting historical data. Despite the lack of theory, noncausal models can provide sound forecasts. However, they provide no guidance on the likely effects of changes in policy (explanatory) variables. Both types of quantitative forecasting methods are discussed in this chapter, although the emphasis is on noncausal methods.

### TYPES OF FORECASTING METHODS

Forecasting methods are broadly classified as **quantitative** or **qualitative**. Quantitative forecasting models are further divided into **causal** and **noncausal** models. Noncausal models are also referred to as purely time series models.



## Model Selection Criteria

Numerous models can be used to make a forecast, with each model well-suited to capture a particular feature of the time series. It would be easy to choose the right model if we knew which feature truly describes the given series. Unfortunately, the truth is almost never known. Because we do not know *a priori* which of the competing models is likely to provide the best forecast, it is common to consider various models. **Model selection** is one of the most important steps in forecasting. Therefore, it is important to understand model selection criteria before we even introduce any of the formal models.

Two types of model selection criteria are used to compare the performance of competing models. These are broadly defined as **in-sample criteria** and **out-of-sample criteria**. These criteria give rise to two important questions: How well does a model explain the given sample data? And how well does a model make out-of-sample forecasts? Ideally, the chosen model is best in terms of its in-sample predictability and its out-of-sample forecasting ability. In this chapter we will focus on in-sample criteria.

Let  $y_t$  denote the value of the series at time  $t$  and let  $\hat{y}_t$  represent its forecast. This in-sample forecast is also referred to as the **predicted** or **fitted value**. For every forecasting model, the sample forecast is likely to differ from the actual series. In other words,  $\hat{y}_t$  will not equal  $y_t$ . Recall that we define  $e_t = y_t - \hat{y}_t$  as the residual. All in-sample model selection criteria compare competing models on the basis of these residuals.

The in-sample forecast  $\hat{y}_t$  is also called the **predicted** or **fitted value** of  $y_t$ . As always, the **residuals** are computed as  $e_t = y_t - \hat{y}_t$ .

In the earlier chapters on regression, we used the coefficient of determination  $R^2$  as a goodness-of-fit measure. We cannot use  $R^2$  in noncausal models because many of them do not use a regression model framework. Instead, a commonly used measure for the comparison of competing forecasting models is the **mean square error (MSE)**, which is the sum of squares due to error (residual) divided by the number of observations  $n$  for which the residuals are available.<sup>1</sup> As we will see shortly, it is not uncommon for  $n$  to be less than the number of observations  $T$  in the series. Another measure is the **mean absolute deviation (MAD)**, which is the mean of the absolute residuals. The preferred model will have the lowest *MSE* and *MAD*.

### MODEL SELECTION CRITERIA

The **mean square error (MSE)** and the **mean absolute deviation (MAD)** are computed as

$$MSE = \frac{\sum (y_t - \hat{y}_t)^2}{n} = \frac{\sum e_t^2}{n} \quad \text{and} \\ MAD = \frac{\sum |y_t - \hat{y}_t|}{n} = \frac{\sum |e_t|}{n},$$

where  $n$  is the number of residuals used in the computation. We choose the model with the lowest *MSE* and *MAD*.

In the following sections we employ various forecasting models and compare them on the basis of these goodness-of-fit measures.

<sup>1</sup>Here the *MSE* formula is different from the one defined in the context of regression analysis in Chapter 15, where the sum of squares due to error, *SSE*, was divided by the appropriate degrees of freedom.

Use smoothing techniques to make forecasts.

Time series generally consist of **systematic** and **unsystematic** patterns. Systematic patterns are caused by a set of identifiable components, whereas unsystematic patterns by definition are difficult to identify. Three identifiable components occur in systematic patterns: the trend, the seasonal, and the cyclical components. In this section we focus on applications where the time series is described primarily by unsystematic patterns. In the following sections we discuss systematic patterns.

Unsystematic patterns are caused by the presence of a **random error** term. As mentioned earlier, a time series is a sequence of observations that are ordered in time. Inherently, any data collected over time is likely to exhibit some form of random variation. For instance, the checkout time at a campus bookstore or weekly sales at a convenience store encounter random variations for no apparent reason.

#### TIME SERIES PATTERNS

Time series consist of **systematic** and **unsystematic** patterns. Systematic patterns are caused by the **trend**, the **seasonal**, and the **cyclical** components. Unsystematic patterns are difficult to identify and are caused by the presence of a **random error** term.

A simple plot of the time series provides insights into its components. A jagged appearance, caused by abrupt changes in the series, indicates random variations. Smoothing techniques are often employed to reduce the effect of random fluctuations. These techniques can also be used to provide forecasts if short-term fluctuations represent random departures from the structure, with no discernible systematic patterns. These techniques are especially attractive when forecasts of multiple variables need to be updated frequently. For example, consider a manager of a convenience store who has to update the inventories of numerous items on a weekly basis. It is not practical in such situations to develop complex forecasting models. We discuss two distinct smoothing techniques: the **moving average** and the **exponential smoothing** techniques.

### Moving Average Methods

Due to its simplicity, the moving average method ranks among the most popular techniques for processing time series. The method is based on computing the average from a fixed number  $m$  of the most recent observations. For instance, a 3-period moving average is formed by averaging the three most recent observations. The term “moving” is used because as a new observation becomes available, the average is updated by including the newest and dropping the oldest observation.

#### CALCULATING A MOVING AVERAGE

An  $m$ -period moving average is computed as

$$\text{Moving Average} = \frac{\text{Sum of the } m \text{ most recent observations}}{m}.$$

Here, we focus on the calculation of odd-numbered moving averages, such as 3-period, 5-period, and so on. In Section 18.4, we will use even-numbered moving averages to extract the seasonal component of a time series.

#### EXAMPLE 18.1

According to the Energy Information Administration, the United States consumes about 21 million barrels (882 million gallons) of petroleum each day. About half of this consumption is in the form of gasoline. Table 18.2 shows a portion of weekly

U.S. finished motor gasoline production, measured in thousands of barrels per day.

- Construct a 3-period moving average series for the data.
- Plot production and its corresponding 3-period moving average and comment on any differences.
- Using the 3-period moving average series, forecast gasoline production on May 29, 2009 (week 22).
- Calculate the mean square error,  $MSE$ , and the mean absolute deviation,  $MAD$ .

**TABLE 18.2** Weekly U.S. Finished Motor Gasoline Production

Date	Week	Production (1,000s of barrels/day)
January 2, 2009	1	9,115
January 9, 2009	2	8,813
January 16, 2009	3	8,729
January 23, 2009	4	8,660
January 30, 2009	5	8,679
⋮	⋮	⋮
May 8, 2009	19	8,710
May 15, 2009	20	8,735
May 22, 2009	21	9,378

SOURCE: Energy Information Administration.

**FILE**  
Gas\_Production

**SOLUTION:**

- For notational simplicity, let production be denoted by  $y_t$  and the corresponding moving average be denoted by  $\bar{y}_t$ . We form a 3-period moving average series by averaging all sets of three consecutive values of the original series. The first value of a 3-period moving average is calculated as

$$\bar{y}_2 = \frac{y_1 + y_2 + y_3}{3} = \frac{9,115 + 8,813 + 8,729}{3} = 8,885.67.$$

We designate this value  $\bar{y}_2$  because it represents the average in weeks 1 through 3. The next moving average, representing the average in weeks 2 through 4, is

$$\bar{y}_3 = \frac{y_2 + y_3 + y_4}{3} = \frac{8,813 + 8,729 + 8,660}{3} = 8,734.00.$$

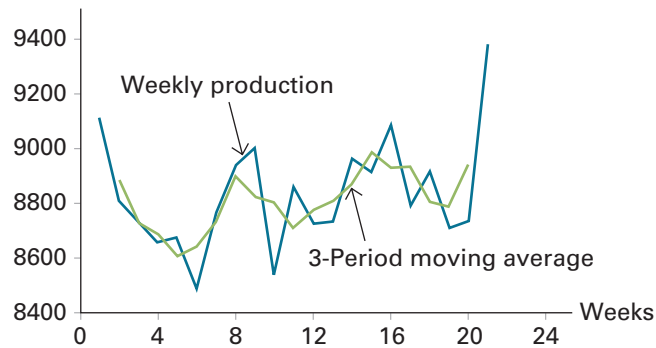
Other values of  $\bar{y}_t$  are calculated similarly and are presented in column 3 of Table 18.3. Note that we lose one observation at the beginning and one at the end of the 3-period moving average series  $\bar{y}_t$ . (If it were a 5-period moving average, we would lose two observations at the beginning and two at the end.)

**TABLE 18.3** 3-Period Moving Averages, Forecasts, and Residuals

Week (1)	$y$ (2)	$\bar{y}$ (3)	$\hat{y}$ (4)	$e = y - \hat{y}$ (5)	$e^2$ (6)	$ e $ (7)
1	9,115	—	—	—	—	—
2	8,813	8,885.67	—	—	—	—
3	8,729	8,734.00	—	—	—	—
4	8,660	8,689.33	8,885.67	-225.67	50,925.44	225.67
5	8,679	8,610.33	8,734.00	-55.00	3,025.00	55.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	8,710	8,787.67	8,932.00	-222.00	49,284.00	222.00
20	8,735	8,941.00	8,806.00	-71.00	5,041.00	71.00
21	9,378	—	8,787.67	590.33	348,493.44	590.33
Total					953,509.78	3,312.67

- b. In Figure 18.1, we plot production and its corresponding 3-period moving average against weeks. Note that the original production series has a jagged appearance, suggesting the presence of an important random component of the series. The series of moving averages, on the other hand, presents a much smoother picture.

**FIGURE 18.1** Weekly production and 3-period moving average



- c. As mentioned earlier, if the primary component of the series is random variations, we can use moving averages to generate forecasts. Since  $\bar{y}_2$  represents the average in weeks 1 through 3, it is the most updated estimate of the series prior to period 4. Therefore, with a 3-period moving average,  $\hat{y}_4 = \bar{y}_2$  where  $\hat{y}_4$  is the in-sample forecast for period 4. Similarly,  $\hat{y}_5 = \bar{y}_3$  is the forecast for period 5, where  $\bar{y}_3$  is the average in weeks 2 through 4, and so on. These forecasts, derived as  $\hat{y}_t = \frac{y_{t-3} + y_{t-2} + y_{t-1}}{3}$ , are shown in column 4 of Table 18.3. Following this simple process, we compute the out-of-sample forecast in week 22 as

$$\hat{y}_{22} = \bar{y}_{20} = \frac{y_{19} + y_{20} + y_{21}}{3} = \frac{8,710 + 8,735 + 9,378}{3} = 8,941.$$

Therefore, our forecast for gasoline production on May 29, 2009 (week 22) is 8,941 thousand barrels. One potential weakness when using the moving average technique is that all future forecasts take on the same value as the first out-of-sample forecast; that is, the forecast for week 23 is also 8,941 thousand barrels.

- d. To calculate the mean square error, *MSE*, and the mean absolute deviation, *MAD*, we first compute the residuals as  $e_t = y_t - \hat{y}_t$ , shown in column 5 of Table 18.3. These residuals are squared (see column 6) and then summed to compute *MSE* as

$$MSE = \frac{\sum e_t^2}{n} = \frac{953,509.78}{18} = 52,973.$$

The absolute values of the residuals, presented in column 7, are used to compute *MAD* as

$$MAD = \frac{\sum |e_t|}{n} = \frac{3,312.67}{18} = 184.$$

While it is difficult to interpret the numerical values of *MSE* and *MAD*, they are useful in comparing alternative models.

## Exponential Smoothing Methods

Although the moving average approach is popular, it has some shortcomings. First, the choice of the order  $m$  is arbitrary, although we can use trial and error to choose the value of  $m$  that results in the smallest *MSE* and *MAD*. Second, it may not be appropriate to give equal weight to all recent  $m$  observations. Whereas the moving average method weighs all recent observations equally, the method called exponential smoothing assigns exponentially decreasing weights as the observations get older. As in the case of moving averages,

exponential smoothing is a procedure for continually revising a forecast in the light of more recent observations.

Let  $A_t$  denote the estimated level of the series at time  $t$ , where  $A_t$  is defined as

$$A_t = \alpha y_t + \alpha(1 - \alpha)y_{t-1} + \alpha(1 - \alpha)^2 y_{t-2} + \alpha(1 - \alpha)^3 y_{t-3} + \dots, \text{ where } 0 \leq \alpha \leq 1.$$

That is,  $A_t$  is simply a weighted average of exponentially declining weights, with  $\alpha$  dictating the speed of decline. For example, with  $\alpha = 0.8$ ,

$$A_t = 0.8y_t + 0.16y_{t-1} + 0.032y_{t-2} + 0.0064y_{t-3} + \dots$$

Similarly, with  $\alpha = 0.2$ ,

$$A_t = 0.2y_t + 0.16y_{t-1} + 0.128y_{t-2} + 0.1024y_{t-3} + \dots$$

Note that the speed of decline is higher when  $\alpha = 0.8$  as compared to when  $\alpha = 0.2$ .

Using algebra, it can be shown that the initial equation simplifies to

$$A_t = \alpha y_t + (1 - \alpha)A_{t-1}.$$

We generally use this representation to define the formula for exponential smoothing. Because  $A_t$  represents the most updated level at time  $t$ , we can use it to make a one-period-ahead forecast as  $\hat{y}_{t+1} = A_t$ .

#### CALCULATING AN EXPONENTIALLY SMOOTHED SERIES

The **exponential smoothing** procedure continually updates the level of the series as

$$A_t = \alpha y_t + (1 - \alpha)A_{t-1},$$

where  $\alpha$  represents the speed of decline. **Forecasts** are made as  $\hat{y}_{t+1} = A_t$ .

In order to implement this method, we need to determine  $\alpha$  and the initial value of the series  $A_1$ . Typically, the initial value is set equal to the first value of the time series, that is,  $A_1 = y_1$ ; the choice of the initial value is less important if the number of observations is large. The optimal value for  $\alpha$  is determined by a trial-and-error method. We evaluate various values of  $\alpha$  and choose the one that results in the smallest *MSE* and *MAD*.

#### EXAMPLE 18.2

Revisit the **Gas\_Production** data on weekly U.S. finished motor gasoline production, measured in thousands of barrels per day.

- Construct the exponentially smoothed series with  $\alpha = 0.20$  and  $A_1 = y_1$ .
- Plot production and its corresponding exponentially smoothed series against weeks. Comment on any differences.
- Using the exponentially smoothed series, forecast gasoline production on May 29, 2009 (week 22).
- Calculate *MSE* and *MAD*. Compare these values with those obtained using the 3-period moving average method.

#### SOLUTION:

- In Column 3 of Table 18.4, we present sequential estimates of  $A_t$  with the initial value  $A_1 = y_1 = 9,115$ . We use  $A_t = \alpha y_t + (1 - \alpha)A_{t-1}$  to continuously update the level with  $\alpha = 0.2$ . For instance, for periods 2 and 3 we calculate:

$$\begin{aligned} A_2 &= 0.20(8,813) + 0.80(9,115) = 9,054.60, \text{ and} \\ A_3 &= 0.20(8,729) + 0.80(9,054.60) = 8,989.48. \end{aligned}$$

All other estimates of  $A_t$  are found in a like manner.

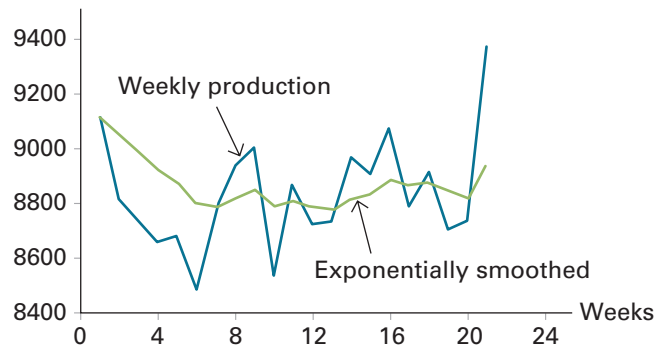
FILE

Gas\_Production

**TABLE 18.4** Exponentially Smoothed Series with  $\alpha = 0.20$ , Forecasts, and Residuals

Week (1)	$y$ (2)	$A_t$ (3)	$\hat{y}$ (4)	$e = y - \hat{y}$ (5)	$e^2$ (6)	$ e $ (7)
1	9,115	9,115.00	—	—	—	—
2	8,813	9,054.60	9,115.00	-302.00	91,204.00	302.00
3	8,729	8,989.48	9,054.60	-325.60	106,015.36	325.60
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
21	9,378	8,933.29	8,822.11	555.89	309,011.71	555.89
Total					1,153,160.28	3,995.40

- b. In Figure 18.2, we plot production and its corresponding exponentially smoothed series against weeks. As mentioned earlier, while the original series has the jagged appearance, the exponentially smoothed series removes most of the sharp points and, much like the moving average series, presents a much smoother picture.

**FIGURE 18.2** Weekly production and exponentially smoothed series

- c. Forecasts given by  $\hat{y}_{t+1} = A_t$  are presented in column 4 of Table 18.4. For instance, for period 2,  $\hat{y}_2 = A_1 = 9,115$ . Similarly,  $A_2 = 9,054.60$  is the forecast for  $\hat{y}_3$ . Therefore, the forecast for gasoline production on May 29, 2009, computed as  $\hat{y}_{22} = A_{21}$ , equals 8,933.29 thousand barrels. As with the moving average method, any further out-of-sample forecasts also assume this same value; for instance,  $\hat{y}_{23} = 8,933.29$  thousand barrels.
- d. In columns 5, 6, and 7 we present the residuals, their squares, and their absolute values, respectively. We compute model selection measures as

$$MSE = \frac{\sum e_t^2}{n} = \frac{1,153,160.28}{20} = 57,658 \quad \text{and}$$

$$MAD = \frac{\sum |e_t|}{n} = \frac{3,995.40}{20} = 200.$$

The moving average model employed in Example 18.1 outperforms, as it yields a lower  $MSE$  of 52,973 and a lower  $MAD$  of 184 than the exponential model. Note that we used the residuals from Weeks 4 to 21 with moving averages and Weeks 2 to 21 with exponential smoothing. For a fair comparison, we recalculated  $MSE = 53,108$  and  $MAD = 187$  with exponential smoothing, using the residuals only from Weeks 4 to 21. The moving average method still outperforms.

There is nothing sacrosanct about  $\alpha = 0.2$ ; we used this value primarily to illustrate the exponential smoothing procedure. As we noted earlier, it is common to evaluate various values of  $\alpha$  and choose the one that produces the smallest  $MSE$  and  $MAD$  for forecasting. In order to illustrate how  $\alpha$  is chosen, we generate the  $MSE$  and  $MAD$  with  $\alpha$  values ranging from 0.1 to 0.9 with increments of 0.1. The results are summarized in Table 18.5.



**TABLE 18.5** Various Values of  $\alpha$  and the Resulting *MSE* and *MAD*

$\alpha$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
<i>MSE</i>	66,906	57,658	54,368	53,364	53,470	54,200	55,394	57,047	59,235
<i>MAD</i>	209	200	196	194	191	188	185	184	188

Here, the choice of  $\alpha$  depends on whether we employ *MSE* or *MAD* for comparison, with *MSE* suggesting  $\alpha = 0.4$  and *MAD* suggesting  $\alpha = 0.8$ . In instances where *MSE* and *MAD* give conflicting results, it is common to choose the procedure with the smallest *MSE*; we make this choice because *MSE* penalizes larger deviations more harshly due to squaring. Therefore, we choose  $\alpha$  equal to 0.4 since it has the smallest *MSE* of 53,364. In this application, the moving average model still outperforms the exponential smoothing model, as measured by its lower *MSE* and *MAD* values.

## Using Excel for Moving Averages and Exponential Smoothing

Excel easily calculates moving averages. From the menu choose **Data > Data Analysis > Moving Average > OK** in order to activate the *Moving Average* dialog box. Click on the box next to *Input Range* and select the relevant time series  $y_t$ . Next to *Interval*, enter the value for  $m$ ; for example, if you want to calculate a 3-period moving average, enter 3. Then click **OK**.

Excel also calculates an exponentially smoothed series. From the menu choose **Data > Data Analysis > Exponential Smoothing > OK** in order to activate the *Exponential Smoothing* dialog box. Click on the box next to *Input Range* and select the relevant time series  $y_t$ . Select the box next to *Damping Factor*. If we want to construct an exponentially smoothed series with  $\alpha = 0.2$ , then for *Damping Factor* we enter  $1 - \alpha = 1 - 0.2 = 0.8$ . Then click **OK**.

## EXERCISES 18.2

### Mechanics

1. **FILE Exercise 18.1.** Consider the following sample data, consisting of 10 observations.

$t$	1	2	3	4	5	6	7	8	9	10
$y_t$	11	12	9	12	10	8	11	12	10	9

- Construct a 3-period moving average and plot it along with the actual series. Comment on smoothing.
  - Use the 3-period moving average to make forecasts and compute the resulting *MSE* and *MAD*.
  - Make a forecast for period 11.
2. **FILE Exercise 18.2.** Consider the following sample data, consisting of 20 observations.

$t$	1	2	3	4	5	6	7	8	9	10
$y_t$	27	35	38	33	34	39	40	38	48	35
$t$	11	12	13	14	15	16	17	18	19	20
$y_t$	37	38	35	44	40	37	30	39	34	45

- Construct a 5-period moving average and plot it along with the actual series. Comment on smoothing.
  - Use the 5-period moving average to make forecasts and compute the resulting *MSE* and *MAD*.
  - Make a forecast for period 21.
3. **FILE Exercise 18.3.** Consider the following sample data, consisting of 20 observations.

$t$	1	2	3	4	5	6	7	8	9	10
$y_t$	12.9	12.6	11.1	14.8	11.9	12.9	12.1	13.6	11.9	9.0
$t$	11	12	13	14	15	16	17	18	19	20
$y_t$	8.9	9.3	13.3	10.7	13.5	15.1	11.3	13.6	12.4	13.0

- Plot the above series and discuss the presence of random variations.
- Use the exponential smoothing method to make forecasts with  $\alpha = 0.2$ . Compute the resulting *MSE* and *MAD*.

- c. Repeat the process with  $\alpha = 0.4$ .
  - d. Use the appropriate value of  $\alpha$  to make a forecast for period 21.
4. **FILE Exercise 18.4.** Consider the following sample data, consisting of 20 observations.

$t$	1	2	3	4	5	6	7	8	9	10
$y_t$	14	17	12	16	18	16	15	19	23	23
$t$	11	12	13	14	15	16	17	18	19	20
$y_t$	18	19	19	21	21	25	23	26	23	20

- a. Use the 3-period moving average to make forecasts and compute the resulting *MSE* and *MAD*.
- b. Use the exponential smoothing method to make forecasts with  $\alpha = 0.4$ . Compute the resulting *MSE* and *MAD*.
- c. Use the preferred method to make a forecast for period 21.

## Applications

5. **FILE Rock\_Music.** Rock 'n' roll is a form of music that evolved in the United States and quickly spread to the rest of the world. The interest in rock music, like any other genre, has gone through ups and downs over the years. The Recording Industry Association of America (RIAA) reports consumer trends on the basis of annual data on genre, format, age, and gender of purchasers and place of purchase. The accompanying table lists a portion of the percentage (share) of total shipment of music that falls in the category of rock music from 1991–2008.

Year	Share
1991	34.8
1992	31.6
⋮	⋮
2008	31.8

SOURCE: [www.riaa.com](http://www.riaa.com).

- a. Plot the series and discuss the presence of random variations.
  - b. Use a 3-period moving average to make a forecast for the share of rock music in 2009.
  - c. Use a 5-period moving average to make a forecast for the share of rock music in 2009.
  - d. Use the *MSE* to pick the appropriate moving average for making a forecast for the share of rock 'n' roll music in 2009.
6. **FILE Rock\_Music.** Use the data for the share of total shipment of music that falls in the category of rock music from 1991–2008.
- a. Make a forecast for the share of rock music in 2009 using the exponential smoothing method with  $\alpha = 0.4$ .

- b. Make a forecast for the share of rock music in 2009 using the exponential smoothing method with  $\alpha = 0.6$ .
- c. Use the *MSE* to pick the appropriate speed of decline for making a forecast for the share of rock music in 2009.

7. **FILE Poverty\_Rate.** According to the Census Bureau, the number of people below the poverty level has been steadily increasing (CNN, September 16, 2010). This means many families are finding themselves there for the first time. The following table shows a portion of the percent of families in the United States who are below the poverty level from 1986–2009.

Year	Poverty Rate
1986	10.9
1987	10.7
⋮	⋮
2009	11.1

SOURCE: U.S. Census Bureau.

- a. Plot the series and comment on its shape.
  - b. Use a 3-period moving average to make in-sample forecasts. Compute the resulting *MSE* and *MAD*.
  - c. Use the exponential smoothing method to make in-sample forecasts with  $\alpha = 0.6$ . Compute the resulting *MSE* and *MAD*.
  - d. Choose the appropriate model to make a forecast of the poverty rate in 2010.
8. **FILE S&P\_Price.** Consider the following table, which shows a portion of the closing prices of the S&P 500 Index for 21 trading days in November 2010.

Date	S&P Price
1-Nov	1184.38
2-Nov	1193.57
⋮	⋮
30-Nov	1180.55

SOURCE: [finance.yahoo.com](http://finance.yahoo.com).

- a. Use a 3-period moving average to make a price forecast for December 1, 2010.
- b. Use the exponential smoothing method to make a price forecast for December 1, 2010. Use  $\alpha = 0.4$ .
- c. Which of the above smoothing methods results in a lower *MSE*?
- d. You find out that the actual S&P 500 closing price on December 1, 2010 was 1,206.07. Was the forecast performance of the two methods consistent with their in-sample performance in part c?

9. **FILE Unemployment\_Inflation.** The accompanying table shows a portion of monthly data on seasonally adjusted inflation and unemployment rates in the United States from January 2009 to November 2010.

Year	Month	Unemployment	Inflation
2009	Jan	7.7	0.3
⋮	⋮	⋮	⋮
2010	Nov	9.8	0.1

SOURCE: Bureau of Labor Statistics.

- Use a 3-period moving average and exponential smoothing with  $\alpha = 0.6$  to make in-sample forecasts for unemployment. Use the more appropriate smoothing method to forecast unemployment for December 2010.
- Use a 3-period moving average and exponential smoothing with  $\alpha = 0.6$  to make in-sample forecasts for inflation. Use the more appropriate smoothing method to forecast inflation for December 2010.

## 18.3 TREND FORECASTING MODELS

LO 18.3

The smoothing techniques discussed in the preceding section are used when the time series represent random fluctuations with no discernible trend or seasonal fluctuations. When trend and seasonal variations are present in the time series, we need to use special models for the analysis. In this section we focus on trend analysis, which extracts long-term upward or downward movements of the series.

Use trend regression models to make forecasts.

### The Linear Trend

We can estimate a linear trend using the regression techniques described in earlier chapters. Let  $y_t$  be the value of the response variable at time  $t$ . Here we use  $t$  as the explanatory variable corresponding to consecutive time periods, such as 1, 2, 3, and so on. Example 18.3 shows how to use this model to make forecasts.

#### THE LINEAR TREND MODEL

A **linear trend model** is specified as  $y_t = \beta_0 + \beta_1 t + \varepsilon_t$ , where  $y_t$  is the value of the series at time  $t$ . The estimated model is used to make **forecasts** as  $\hat{y}_t = b_0 + b_1 t$ , where  $b_0$  and  $b_1$  are the coefficient estimates.

### EXAMPLE 18.3

The United States continues to increase diversity, with more than a third of its population belonging to a minority group (CNN.com, May 14, 2009). Hispanics are the fastest-growing minority segment, comprising one out of six residents in the country. Table 18.6 shows a portion of data relating to the number as well as the median income of Hispanic households from 1975 through 2007.

**TABLE 18.6** Number and Median Income of Hispanics, 1975–2007

Year	Number (in 1,000s)	Median Income
1975	2,948	\$8,865
1976	3,081	9,569
⋮	⋮	⋮
2007	13,339	38,679

SOURCE: United States Census Bureau.

**FILE**

*Hispanic Characteristics*

- a. Use the sample data to estimate the linear trend model for the number (Regression 1) and the median income (Regression 2) of Hispanic households. Interpret the slope coefficients.
- b. Forecast the number and the median income of Hispanic households in 2008.

**SOLUTION:** In order to estimate the linear trend models, it is advisable to first relabel the 33 years of observations from 1 to 33. In other words, we make the explanatory variable  $t$  assume values 1, 2, ..., 33 rather than 1975, 1976, ..., 2007. Table 18.7 shows a portion of the data.

**TABLE 18.7** Generating the Time Variable  $t$

Year	$t$	Number (in 1,000s)	Median Income
1975	1	2,948	\$8,865
1976	2	3,081	9,569
$\vdots$	$\vdots$	$\vdots$	$\vdots$
2007	33	13,339	38,679

The regression results are presented in Table 18.8.

**TABLE 18.8** Regression Results for Example 18.3

	Response Variable: Number (Regression 1)	Response Variable: Income (Regression 2)
	Coefficients	Coefficients
Intercept	1657.8428* (0.00)	7796.9186* (0.00)
$t$	327.4709* (0.00)	887.7249* (0.00)

NOTES: Parameter estimates are followed by the  $p$ -values in parentheses; \* represents significance at 5%.

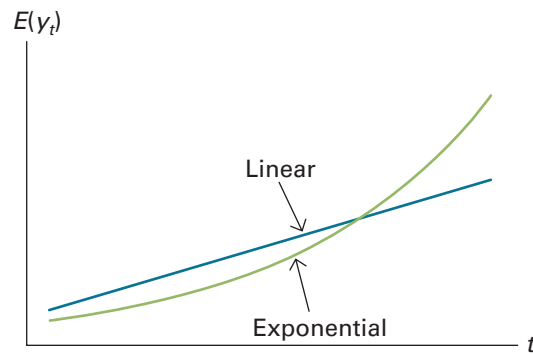
- a. The slope coefficient in Regression 1 implies that the number of Hispanic households has grown, on average, by approximately 327 (thousand) each year. Regression 2 shows that the median income for Hispanic households has grown, on average, by approximately \$888 each year. The slope coefficients in both regressions are significant at any level, since the  $p$ -values approximate zero in each case.
- b. Using the estimates from Regression 1, we forecast the number of Hispanic households in 2008 ( $t = 34$ ) as  $1,657.8428 + 327.4709(34) = 12,792$  (in 1,000s). Similarly, using the estimates from Regression 2, the forecast for the median income of Hispanic households in 2008 is  $7,796.9186 + 887.7249(34) = \$37,980$ . Forecasts for other years can be computed similarly.

## The Exponential Trend

A linear trend model by definition uses a straight line to capture the trend, thus implying that for each period, the value of the series changes by a fixed amount. For example, in Example 18.3 we concluded that the median income of Hispanic households grows by approximately \$888 each year. The **exponential trend model** is attractive when the increase in the series gets larger over time. Figure 18.3 compares linear and exponential trends. While both graphs have positive slopes, the exponential trend, unlike the linear trend, allows the series to grow by an increasing amount for each time period.

Recall from Chapter 16 that we estimate an exponential model as  $\ln(y_t) = \beta_0 + \beta_1 t + \varepsilon_t$ . In order to estimate this model, we first generate the series in natural logs,  $\ln(y_t)$ ,

and then run a regression of  $\ln(y_t)$  on  $t$ . Since the response variable is measured in logs, we make forecasts in regular units as  $\hat{y}_t = \exp(b_0 + b_1t + s_e^2/2)$  where  $s_e$  is the standard error of the estimate. As discussed in Chapter 16, if we make the forecast using  $\exp(b_0 + b_1t)$ , then  $\hat{y}$  systematically underestimates the expected value of  $y$ ; the inclusion of  $s_e^2/2$  in the forecast equation resolves this problem.

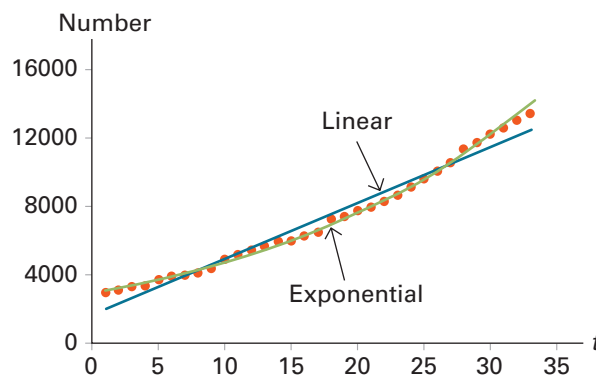


**FIGURE 18.3**  
Linear and exponential trends

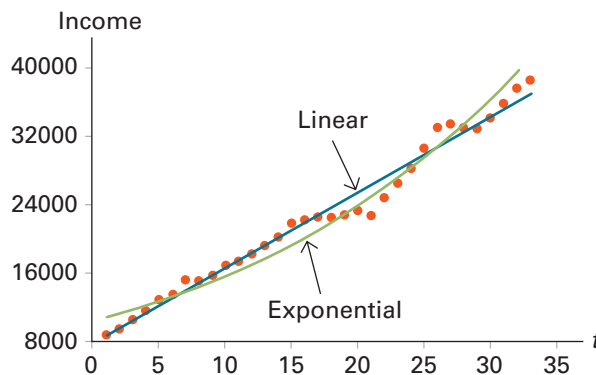
#### THE EXPONENTIAL TREND MODEL

An **exponential trend model** is specified as  $\ln(y_t) = \beta_0 + \beta_1t + \varepsilon_t$ , where  $\ln(y_t)$  is the natural log of  $y_t$ . The estimated model is used to make **forecasts** as  $\hat{y}_t = \exp(b_0 + b_1t + s_e^2/2)$ , where  $b_0$  and  $b_1$  are the coefficient estimates and  $s_e$  is the standard error of the estimate.

It is always advisable to inspect the data visually as a first step. Graphs offer an informal way to gauge whether a linear or an exponential trend provides a better fit. Figures 18.4 and 18.5 are scatterplots of the number and the median income of Hispanic households from 1975 through 2007. These plots are based on the *Hispanic\_Characteristics* data. We relabel the 33 years of annual observations from 1 to 33 and also superimpose the linear and the exponential trends to the data.



**FIGURE 18.4**  
Number of Hispanic households with superimposed trends



**FIGURE 18.5**  
Median income of Hispanic households with superimposed trends

It appears that while the median income follows a linear trend, the number of Hispanics seems to grow exponentially.

**FILE**

*Hispanic\_Characteristics*

**EXAMPLE 18.4**

- Revisit the *Hispanic\_Characteristics* data to estimate the exponential trend model for both the number (Regression 1) and the median income (Regression 2) of Hispanic households. Interpret the slope coefficients.
- Forecast the number as well as the median income of Hispanic households in 2008.
- Use formal model selection criteria to decide whether the linear or the exponential model is more appropriate for the series.

**SOLUTION:** In order to estimate the exponential model, we first transform both series to natural logs. Table 18.9 shows a portion of the data where the variables (number and income) are transformed into natural logs. The table also includes the explanatory variable  $t$  relabeled from 1 to 33.

**TABLE 18.9** Generating the Natural Log of the Series (Example 18.4)

Year	$t$	Number	Income	$\ln(\text{Number})$	$\ln(\text{Income})$
1975	1	2,948	\$8,865	7.9889	9.0899
1976	2	3,081	\$9,569	8.0330	9.1663
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2007	33	13,339	38,679	9.4984	10.5631

Relevant regression results for the exponential regression models are presented in Table 18.10.

**TABLE 18.10** Regression Results for Example 18.4

	Response Variable: Log of Number (Regression 1)	Response Variable: Log of Income (Regression 2)
	Coefficients	Coefficients
Intercept	7.9706* (0.00)	9.2517* (0.00)
$t$	0.0479* (0.00)	0.0418* (0.00)
$s_e$	0.0311	0.0775

NOTES: Parameter estimates are followed by the  $p$ -values in parentheses; \* represents significance at the 5% level. The last row shows the standard error of the estimate  $s_e$ .

- Consistent with the results of the linear trend models, the exponential trend models suggest that the number and the median income of Hispanic households are trending upward, since both regressions show positive slope coefficients: 0.0479 for Regression 1 and 0.0418 for Regression 2. In addition, the slope coefficients are highly significant at any level since the  $p$ -values approximately zero in both regressions.
- We make forecasts as  $\hat{y}_t = \exp(b_0 + b_1 t + s_e^2/2)$ . In order to forecast the number of Hispanic households for 2008 ( $t = 34$ ), we compute

$$\hat{y}_{34} = \exp(7.9706 + 0.0479(34) + 0.0311^2/2) = 14,760.$$

Similarly, the forecast for Hispanic median income is computed as

$$\hat{y}_{34} = \exp(9.2517 + 0.0418(34) + 0.0775^2/2) = \$43,300.$$



Forecasts for other years can be computed similarly. Note that 2008 forecasts with the exponential trend model are higher than those with the linear trend model.

Note: Whenever possible, it is advisable to use unrounded values for making forecasts in an exponential model because even a small difference, when exponentiated, can make a big difference in the forecast.

- c. We compute  $\hat{y}$  for the exponential model in regular units and not in natural logs. The resulting  $\hat{y}$  also enables us to compare the linear and the exponential models in terms of *MSE* and *MAD*. In Table 18.11, we present a portion of the series  $\hat{y}_t$ , along with  $y_t$ , for both models; we did these calculations in Excel with unrounded values for the estimates. We then compute  $MSE = \frac{\sum e_t^2}{n}$  and  $MAD = \frac{\sum |e_t|}{n}$  where  $e_t = y_t - \hat{y}_t$ . While these calculations are not reported, the *MSE* and *MAD* values for the linear and the exponential models are shown in the last two rows of Table 18.11.

**TABLE 18.11** Analysis of Linear and Exponential Trend Models

t	Number of Hispanics y			Income of Hispanics y		
	y	$\hat{y}$ (Linear)	$\hat{y}$ (Exponential)	y	$\hat{y}$ (Linear)	$\hat{y}$ (Exponential)
1	2,948	1,985.31	3,037.97	8,865	8,684.64	10,899.65
2	3,081	2,312.79	3,186.97	9,569	9,572.37	11,364.43
⋮	⋮	⋮	⋮	⋮	⋮	⋮
33	13,339	12,464.38	14,061.05	38,679	37,091.84	41,471.14
MSE		281,255	45,939		1,508,369	2,210,693
MAD		460	155		881	1,289

The exponential trend model appears to be better suited to describe the number of Hispanic households, since it has a lower *MSE* and *MAD* than the linear trend model. On the other hand, median income is better described by the linear trend model. These findings are consistent with our earlier analysis with Figures 18.4 and 18.5. Therefore, we use the exponential trend model to forecast the number of Hispanic households in 2008 as 14,760. The linear trend model is used to forecast the median income of Hispanic households in 2008 as \$37,980.

## Polynomial Trends

Sometimes a time series reverses direction, due to any number of circumstances. A common polynomial function that allows for curvature in the series is a **quadratic trend model**. This model describes one change in the direction of a series and is estimated as

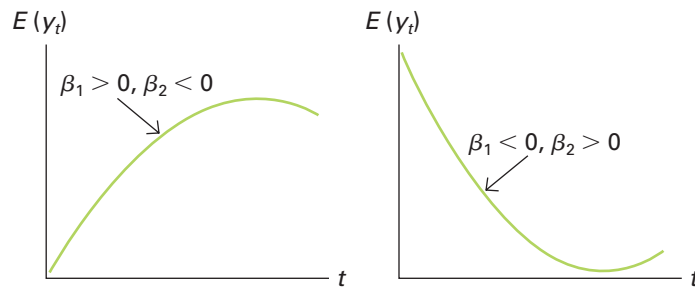
$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t.$$

The coefficient  $\beta_2$  determines whether the trend is U-shaped or inverted U-shaped. Figure 18.6 depicts possible trends of a quadratic model.

In order to estimate the quadratic trend model, we generate  $t^2$ , which is simply the square of  $t$ . Then we run a multiple regression model that uses  $y$  as the response variable and both  $t$  and  $t^2$  as the explanatory variables. The estimated model is used to make forecasts as

$$y_t = b_0 + b_1 t + b_2 t^2.$$

**FIGURE 18.6**  
Representative shapes  
of a quadratic trend

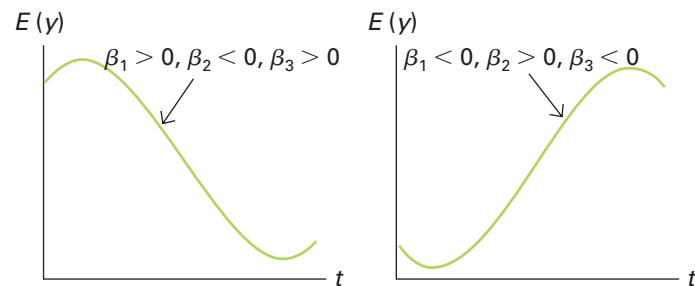


Higher-order polynomial functions can be estimated similarly. For instance, consider a cubic trend model specified as

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \varepsilon_t.$$

The cubic trend model allows for two changes in the direction of a series. Figure 18.7 presents possible shapes of a cubic model.

**FIGURE 18.7**  
Representative shapes  
of a cubic trend



In the cubic trend model, we basically generate two additional variables,  $t^2$  and  $t^3$ , for the regression. A multiple regression model is run that uses  $y$  as the response variable and  $t$ ,  $t^2$ , and  $t^3$  as the explanatory variables. The estimated model is used to make forecasts as  $\hat{y}_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3$ .

While we use the *MSE* and the *MAD* of in-sample forecast errors to compare the linear and the exponential models, we cannot use them to compare the linear, quadratic, and cubic trend models. The reason is that the values of *MSE* and *MAD* are always lowest for the highest-order polynomial trend model, since the values decrease as we estimate additional parameters. The problem is similar to that of the coefficient of determination  $R^2$  discussed in earlier chapters. When comparing polynomial trend models, we use adjusted  $R^2$ , which imposes a penalty for over-parameterization.

#### THE POLYNOMIAL TREND MODEL

A polynomial trend model of order  $q$  is estimated as

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \dots + \beta_q t^q + \varepsilon_t.$$

This model specializes to a linear trend model, quadratic trend model, and cubic trend model for  $q = 1, 2$ , and  $3$ , respectively. The estimated model is used to make forecasts as  $\hat{y}_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3 + \dots + b_q t^q$ , where  $b_0, b_1, \dots, b_q$  are the coefficient estimates. We use **adjusted  $R^2$**  to compare polynomial trend models with different orders.

A good application of the polynomial trend model is used in the Writing with Statistics section of this chapter.

## EXERCISES 18.3

### Mechanics

10. Consider the following estimated trend models. Use them to make a forecast for  $t = 21$ .
- Linear Trend:  $\hat{y} = 13.54 + 1.08t$
  - Quadratic Trend:  $\hat{y} = 18.28 + 0.92t - 0.01t^2$
  - Exponential Trend:  $\ln(\hat{y}) = 1.8 + 0.09t$ ;  $s_e = 0.01$
11. **FILE Exercise 18.11.** Consider the following table, consisting of 20 observations of the variable  $y$  and time  $t$ .

$t$	$y$	$t$	$y$	$t$	$y$	$t$	$y$
1	3.01	6	4.94	11	7.28	16	14.16
2	3.13	7	6.10	12	9.04	17	14.85
3	4.19	8	5.91	13	9.49	18	16.77
4	5.07	9	6.56	14	12.12	19	18.07
5	4.46	10	7.29	15	13.15	20	19.99

- Plot the series along with the superimposed linear and exponential trends. Which trend model do you think describes the data well?
  - Estimate a linear trend model and an exponential trend model for the sample. Validate your guess from the graphs by comparing the  $MSE$  and  $MAD$ .
12. **FILE Exercise 18.12.** Consider the following table, consisting of 20 observations of the variable  $y$  and time  $t$ .

$t$	$y$	$t$	$y$	$t$	$y$	$t$	$y$
1	10.32	6	13.84	11	16.95	16	16.26
2	12.25	7	14.39	12	16.18	17	16.77
3	12.31	8	14.40	13	17.22	18	17.10
4	13.00	9	15.05	14	16.71	19	16.91
5	13.15	10	14.99	15	16.64	20	16.79

- Plot the series along with the superimposed linear and quadratic trends. Which trend model do you think describes the data well?
- Estimate a linear trend model and a quadratic trend model. Validate your guess from the graphs by comparing their adjusted  $R^2$ .

### Applications

13. Despite the growth in digital entertainment, the nation's 400 amusement parks have managed to hold on to visitors, as the following data show:

Year	Visitors (in millions)
2000	317
2001	319
2002	324
2003	322
2004	328
2005	335
2006	335
2007	341

SOURCE: International Association of Amusement Parks and Attractions.

- Plot the series. Does the linear trend model or the exponential trend model fit the series best?
  - Estimate both models. Verify your answer in part a by comparing the  $MSE$  of the models.
  - Given the model of best fit, make a forecast for visitors to amusement parks in 2008 and 2009.
14. The potentially deadly 2009 Swine Flu outbreak was due to a new flu strain of subtype H1N1 not previously reported in pigs. When the World Health Organization declared a pandemic, the virus continued to spread in the United States, causing illness along with regular seasonal influenza viruses. Consider the following 2009 weekly data on total Swine Flu cases in the United States, reported by the Centers for Disease Control and Prevention (CDC).

Week	Total	Week	Total
17	1,190	22	2,904
18	2,012	23	3,024
19	1,459	24	3,206
20	2,247	25	1,829
21	2,280	26	1,093

SOURCE: www.cdc.gov.

- Plot the series. Estimate the linear and the quadratic trend models. Use their adjusted  $R^2$  to choose the preferred model.
  - Given the preferred model, make a forecast for the number of Swine Flu cases in the U.S. for week 27.
15. **FILE Recording Industry.** Rapid advances in technology have had a profound impact on the United States recording industry (*The New York Times*, July 28, 2008). While cassette tapes gave vinyl records strong competition, they were subsequently eclipsed by the introduction of the compact disc (CD) in the early 1980s. Lately, the CD, too, has been in rapid decline, primarily because of Internet music stores. The following data show a portion of year-end shipment statistics on the three formats of the United States recording industry, in particular, the manufacturers' unit shipments, in millions, of vinyl, cassettes, and CDs from 1991 to 2008.

Year	Vinyl	Cassettes	CDs
1991	4.8	360.1	333.3
1992	2.3	366.4	407.5
⋮	⋮	⋮	⋮
2008	2.9	0.1	384.7

SOURCE: www.riaa.com.

- a. Plot the series for cassettes. Estimate the quadratic and the cubic trend models for this series. Make a forecast with the chosen model for 2009.
  - b. Plot the series for CDs. Estimate the linear and the quadratic trend models for this series. Make a forecast with the chosen model for 2009.
16. **FILE California\_Unemployment.** Consider the following table, which lists a portion of the monthly unemployment rates (seasonally adjusted) in California from 2007–2010.

Year	Month	Unemployment Rate (%)
2007	Jan	4.9
2007	Feb	5.0
⋮	⋮	⋮
2010	Dec	12.5

SOURCE: Bureau of Labor Statistics.

- a. Plot the above series. Which polynomial trend model do you think is most appropriate?
- b. Verify your answer by formally comparing the linear, the quadratic, and the cubic trend models.
- c. Forecast the unemployment rate in California for January 2011.

## 18.4 TREND AND SEASONALITY

As mentioned earlier, time series generally consist of systematic and unsystematic patterns. Smoothing methods prove useful for forecasting a series that is described primarily by an unsystematic component. Systematic patterns are identified by the trend, seasonal, and cyclical components. In the preceding section, we focused on trend. We now turn our attention to the other components of systematic patterns.

The **seasonal** component typically represents repetitions over a one-year period. Time series consisting of weekly, monthly, or quarterly observations tend to exhibit seasonal variations that repeat year after year. For instance, every year, sales of retail goods increase during the Christmas season, and the number of vacation packages goes up during the summer. The **cyclical** component represents wavelike fluctuations or business cycles, often caused by expansion and contraction of the economy. Unlike the well-defined seasons, the length of a business cycle varies, as fluctuations may last for several months or years. In addition, even the magnitude of business cycles varies over time. Because cycles vary in length and amplitude, they are difficult to capture with historical data. For these reasons, we ignore the cyclical component in this text and refer the reader to advanced books for details.

In this section we will make forecasts based on the seasonal as well as the trend components of a series. Note that some economic series are available in a seasonally adjusted format. In such instances, we only have to focus on trend.

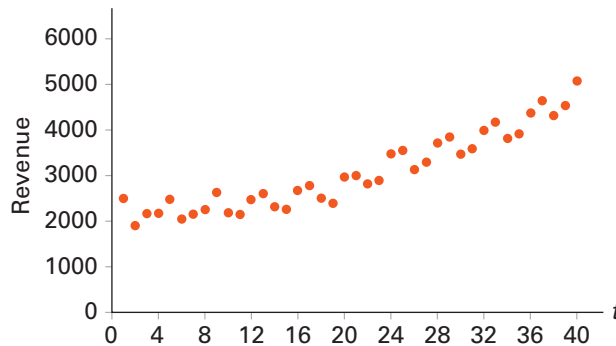
### Decomposition Analysis

Let  $T$ ,  $S$ , and  $I$  represent the trend, the seasonal, and the irregular or random components, respectively, of the time series  $y_t$ . With seasonal data, it is necessary to model seasonality along with trend. Consider the following multiplicative model, which relates sample observations to  $T$ ,  $S$ , and  $I$ :

$$y_t = T_t \times S_t \times I_t.$$

We can use sample information to extract the trend and the seasonal components of the series and then project them into the future. Let  $\hat{T}_t$  and  $\hat{S}_t$  denote the estimated trend and seasonal components, respectively. Note that by their very nature, random variations cannot be identified. Therefore, we set  $\hat{I}_t = 1$  and make forecasts as  $\hat{y}_t = \hat{T}_t \times \hat{S}_t$ . This process is often referred to as **decomposition analysis**. Alternatively, we can use a multiple regression model to simultaneously estimate trend along with **seasonal dummy variables**. We first elaborate on decomposition analysis, then discuss the use of seasonal dummy variables.

In the introductory case to this chapter, we considered the *Nike\_Revenues* data on Nike's quarterly revenue from 1999 through 2008. Figure 18.8 is a scatterplot of the data, where we have relabeled the 10 years of quarterly observations from 1 to 40. The graph highlights some important characteristics of Nike's revenue. First, there is a persistent upward movement in the data. Second, the trend does not seem to be linear and is likely to be better captured by an exponential model. Third, a seasonal pattern repeats itself year after year. For instance, revenue is consistently higher in the first and fourth quarters as compared to the second and third quarters.



**FIGURE 18.8**  
Scatterplot of Nike's quarterly revenue (in millions \$)

### Extracting Seasonality

In this section, we employ moving averages to separate the effect of trend from seasonality. Given quarterly data, we use a 4-period moving average by averaging all sets of four consecutive quarterly values of the series. Using the earlier notation, let  $y_t$  denote revenue at time  $t$  and let  $\bar{y}_t$  represent its corresponding 4-period moving average. We calculate the first moving average as

$$\bar{y}_{2.5} = \frac{y_1 + y_2 + y_3 + y_4}{4} = \frac{2,505 + 1,913 + 2,177 + 2,182}{4} = 2,194.25.$$

We designate the first moving average  $\bar{y}_{2.5}$  because it represents the average in quarters 1 through 4. The next two moving averages, representing the average in quarters 2 through 5, and quarters 3 through 6, respectively, are

$$\bar{y}_{3.5} = \frac{y_2 + y_3 + y_4 + y_5}{4} = \frac{1,913 + 2,177 + 2,182 + 2,501}{4} = 2,193.25, \text{ and}$$

$$\bar{y}_{4.5} = \frac{y_3 + y_4 + y_5 + y_6}{4} = \frac{2,177 + 2,182 + 2,501 + 2,060}{4} = 2,230.00.$$

Other moving averages are calculated similarly.

We note two important points at the outset. First, a 4-period moving average basically represents a quarterly average in one year. The first moving average uses all four quarters of 1999, the second uses three quarters of 1999 and one of 2000, and so on. Since all four quarters are represented in every 4-period moving average, it eliminates seasonal fluctuations in the series. Second, while it is appropriate to designate a moving average in the middle quarter, there is no middle quarter in the original series. For instance, the moving average is represented by  $\bar{y}_{2.5}$ ,  $\bar{y}_{3.5}$ ,  $\bar{y}_{4.5}$ , etc. when the original series is  $y_1$ ,  $y_2$ ,  $y_3$ , etc.

In order to represent an even-period moving average, we rely on the **centered moving average, CMA**, which is essentially the average of two consecutive moving averages. In the above example, the first 4-period centered moving average is formed as an average of the first two 4-period moving averages. In other words,

$$\bar{y}_3 = \frac{\bar{y}_{2.5} + \bar{y}_{3.5}}{2} = \frac{2,194.25 + 2,193.25}{2} = 2,193.75.$$

Note that this centered moving average  $\bar{y}_3$  is not designated in the middle quarter and corresponds with  $y_3$  of the actual series. Similarly,

$$\bar{y}_4 = \frac{\bar{y}_{3.5} + \bar{y}_{4.5}}{2} = \frac{2,193.25 + 2,230.00}{2} = 2,211.63.$$

### LO 18.4

Calculate and interpret seasonal indices and use them to seasonally adjust a time series.

The CMA series,  $\bar{y}_t$ , is shown in column 4 of Table 18.12. (We should mention here that Excel calculates  $\bar{y}_{2.5}$ ,  $\bar{y}_{3.5}$ ,  $\bar{y}_{4.5}$ , ..., when calculating a 4-quarter moving average. We then need to prompt Excel to “center” the values by creating another column that calculates the average between each pairing.)

**TABLE 18.12** Analysis of Seasonal Data

Period	$t$	$y$	Centered Moving Average: $\bar{y}$	Ratio-to-Moving Average: $y/\bar{y}$
1999:01	1	2,505	—	—
1999:02	2	1,913	—	—
1999:03	3	2,177	2,193.75	0.9924
1999:04	4	2,182	2,211.63	0.9866
⋮	⋮	⋮	⋮	⋮
2008:01	37	4,655	4,403.38	1.0571
2008:02	38	4,340	4,568.63	0.9500
2008:03	39	4,544	—	—
2008:04	40	5,088	—	—

We note that  $\bar{y}_t$  eliminates seasonal variations and also some random variations; that is,  $\bar{y}_t$  represents a series that only includes the trend component  $T_t$ . Heuristically, since  $y_t = T_t \times S_t \times I_t$  and  $\bar{y}_t = T_t$ , when we divide  $y_t$  by  $\bar{y}_t$ , we are left with  $S_t \times I_t$ . This series,  $y_t/\bar{y}_t$ , is called the **ratio-to-moving average** and is presented in column 5 of Table 18.12.

In Table 18.13, we rearrange  $y_t/\bar{y}_t$  by quarter from 1999–2008. Note that each quarter has multiple ratios, where each ratio corresponds to a different year. For instance,  $y_t/\bar{y}_t$  for the third quarter is 0.9924 in 1999, 0.9541 in 2000, and so on. In this example, each quarter has nine ratios. We use the arithmetic average (sometimes statisticians prefer to use the median) to determine a common value for each quarter. By averaging, we basically cancel out the random component and extract the seasonal component of the series. We refer to this summary measure as the **unadjusted seasonal index** for the quarter. For instance, the average of the ratios for the first quarter is calculated as  $(1.1225 + \dots + 1.0571)/9 = 1.0815$ .

**TABLE 18.13** Computation of Seasonal Indices

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
1999	—	—	0.9924	0.9866
2000	1.1225	0.9206	0.9541	0.9881
⋮	⋮	⋮	⋮	⋮
2008	1.0571	0.9500	—	—
Unadjusted Seasonal Index	1.0815	0.9413	0.9379	1.0377
Adjusted Seasonal Index	1.0819	0.9417	0.9383	1.0381

For quarterly data, the seasonal indices must add up to 4 (the number of seasons or  $m$ ) so that the average is one. In order to ensure this requirement, we multiply each unadjusted seasonal index by 4 and divide by the sum of the four unadjusted seasonal indices. In this case, the “multiplier” equals

$$\text{Multiplier} = \frac{4}{1.0815 + 0.9413 + 0.9379 + 1.0377} = 1.0004.$$

Therefore, the adjusted seasonal index for quarter 1 is calculated as  $1.0815(1.0004) = 1.0819$ . This is our final estimate of the seasonal index rounded to the 4th decimal



place—it is referred to as the **adjusted seasonal index**. Table 18.13 shows adjusted seasonal indices for each quarter. Note that the average of the adjusted seasonal indices equals one.

Let us interpret these adjusted seasonal indices. There is no seasonality if all indices equal their average of one. On the other hand, if the seasonal index for a quarter is greater (less) than one, it implies that the observations in the given quarter are greater (less) than the average quarterly values. In the preceding example, the adjusted seasonal index of 1.0819 for quarter 1 implies that Nike's revenue is about 8.19% higher in the first quarter as compared to the average quarterly revenue. The adjusted seasonal index for the second quarter is 0.9417, suggesting that revenues are about 5.83% lower than the average quarterly revenue. Other adjusted seasonal indices are interpreted similarly.

#### CALCULATING A SEASONAL INDEX

1. Calculate the **moving average, MA** (if  $m$  is odd, where  $m$  is the number of seasons), or the **centered moving average, CMA** (if  $m$  is even), of the series. We represent  $MA$  or  $CMA$  by  $\bar{y}_t$ .
2. Compute the **ratio-to-moving average** as  $y_t/\bar{y}_t$ .
3. Find the average of  $y_t/\bar{y}_t$  for each season. This average is referred to as the **unadjusted seasonal index**.
4. Multiply each unadjusted seasonal index by  $m/(\text{Sum of the unadjusted seasonal indices})$ . The resulting value, referred to as the **adjusted seasonal index**, is the final estimate for the seasonal index.

### Extracting Trend

In order to extract the trend from a time series, we first eliminate seasonal variations by dividing the original series  $y_t$  by its corresponding adjusted seasonal index  $\hat{S}_t$ . Here  $\hat{S}_t$  represents four quarters of adjusted seasonal indices, repeated over the years. The seasonally adjusted series,  $(y_t/\hat{S}_t)$ , is shown in column 5 of Table 18.14.

#### LO 18.5

Use decomposition analysis to make forecasts.

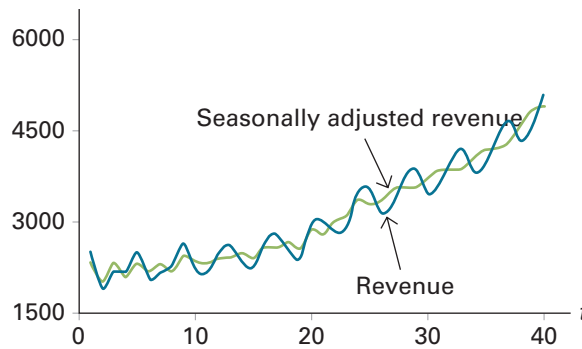
**TABLE 18.14** Creating Seasonally Adjusted Series

Period	$t$	Unadjusted Series: $y$	Seasonal Index: $\hat{S}$	Seasonally Adjusted Series: $y/\hat{S}$
1999:01	1	2,505	1.0819	2,315.29
1999:02	2	1,913	0.9417	2,031.39
1999:03	3	2,177	0.9383	2,320.24
1999:04	4	2,182	1.0381	2,101.97
2000:01	5	2,501	1.0819	2,311.59
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2008:04	40	5,088	1.0381	4,901.38

As noted earlier, the adjusted seasonal index of 1.0819 implies that Nike's revenue in the first quarter is about 8.19% higher than average quarterly revenue. Therefore, without the seasonal effect, the revenue would be lower. For instance, the revenue of 2,505 million in 1999:01 is only 2,315.29 million once it has been seasonally adjusted (we used unrounded values in the calculations). In Figure 18.9, we plot revenue along with seasonally adjusted revenue.

The seasonally adjusted series is free of seasonal variations, thus highlighting the long-term movement (trend) of the series. Figure 18.9 also confirms that the exponential trend model is better suited for the data than is the linear trend model.

**FIGURE 18.9**  
Regular and  
seasonally adjusted  
quarterly revenue



In order to obtain the **seasonally adjusted series**, we divide the original series by its corresponding seasonal index ( $y_t/\hat{S}_t$ ). We use the appropriate trend model on the seasonally adjusted series to extract the trend component  $\hat{T}_t$ .

We estimate the exponential trend model where the response variable is the natural log of the seasonally adjusted revenues  $\ln(y_t/\hat{S}_t)$  and the explanatory variable  $t$  assumes values 1, 2, ..., 40, representing 10 years of quarterly data from 1999 to 2008. The relevant portion of the regression output is given in Table 18.15. We encourage you to compare the results of this model with the linear trend model, using model selection criteria.

**TABLE 18.15** Exponential Regression Results on Seasonally Adjusted Data

	Coefficients	Standard Error	t Stat	p-Value
Intercept	7.5571	0.0185	407.52	0.00
$t$	0.0218	0.0008	27.63	0.00
Standard error of the estimate $s_e$ equals 0.0575.				

Note that the slope coefficient is positive and highly significant ( $p$ -value  $\approx 0$ ). We forecast trend for the seasonally adjusted revenue as  $\hat{T}_t = \exp(b_0 + b_1t + s_e^2/2)$ . Therefore, a trend forecast for the seasonally adjusted revenue in the first quarter of 2009 is computed as

$$\hat{T}_{2009:01} = \hat{T}_{41} = \exp(7.5571 + 0.0218(41) + 0.0575^2/2) = 4,687.02.$$

Similarly, the seasonally adjusted trend for other quarters is computed as

$$\hat{T}_{2009:02} = \hat{T}_{42} = \exp(7.5571 + 0.0218(42) + 0.0575^2/2) = 4,790.32,$$

$$\hat{T}_{2009:03} = \hat{T}_{43} = \exp(7.5571 + 0.0218(43) + 0.0575^2/2) = 4,895.90, \text{ and}$$

$$\hat{T}_{2009:04} = \hat{T}_{44} = \exp(7.5571 + 0.0218(44) + 0.0575^2/2) = 5,003.80.$$

As mentioned earlier, whenever possible, it is preferable to use unrounded estimates in deriving forecasts with an exponential trend model.

### Forecasting with Decomposition Analysis

Now that we have estimated the trend and the seasonal components, we recompose them to make forecasts for Nike's quarterly revenue in 2009. We basically multiply the trend forecast of the seasonally adjusted series with the appropriate seasonal index. This gives us the forecast as  $\hat{y}_t = \hat{T}_t \times \hat{S}_t$ . The trend forecast of Nike's seasonally adjusted revenue in the first quarter of 2009 is  $\hat{T}_{41} = 4,687.02$ . We also have  $\hat{S}_{41} = 1.0819$  since  $t = 41$  represents the first quarter for which we expect the revenue to be 8.19% higher. Therefore, we derive the forecast as

$$\hat{y}_{2009:01} = \hat{y}_{41} = \hat{T}_{41} \times \hat{S}_{41} = 4,687.02 \times 1.0819 = \$5,070.89 \text{ million.}$$

Similarly, the forecast of Nike's revenue for the remaining quarters of 2009 is computed as

$$\begin{aligned}\hat{y}_{2009:02} &= \hat{y}_{42} = \hat{T}_{42} \times \hat{S}_{42} = 4,790.32 \times 0.9417 = \$4,511.04 \text{ million,} \\ \hat{y}_{2009:03} &= \hat{y}_{43} = \hat{T}_{43} \times \hat{S}_{43} = 4,895.90 \times 0.9383 = \$4,593.82 \text{ million, and} \\ \hat{y}_{2009:04} &= \hat{y}_{44} = \hat{T}_{44} \times \hat{S}_{44} = 5,003.80 \times 1.0381 = \$5,194.44 \text{ million.}\end{aligned}$$

### EXAMPLE 18.5

A tourism specialist uses decomposition analysis to examine hotel occupancy rates for Bar Harbor, Maine. She collects quarterly data for the past five years ( $n = 20$ ) and finds that the linear trend model best captures the trend of the seasonally adjusted series:  $\hat{T}_t = 0.60 + 0.0049t$ . In addition, she calculates quarterly indices as  $\hat{S}_1 = 0.53$ ,  $\hat{S}_2 = 0.90$ ,  $\hat{S}_3 = 1.40$ , and  $\hat{S}_4 = 1.17$ .

- Interpret the first and third quarterly indices.
- Forecast next year's occupancy rates.

#### SOLUTION:

- The first quarter's index of 0.53 implies that the occupancy rate in the first quarter is 47% below the average quarterly occupancy rate, whereas the occupancy rate in the third quarter is 40% above the average quarterly occupancy rate.
- We calculate the next year's quarterly occupancy rates as  $\hat{y}_t = \hat{T}_t \times \hat{S}_t$  or

$$\begin{aligned}\text{Year 6, Quarter 1: } \hat{y}_{21} &= (0.60 + 0.0049(21)) \times 0.53 = 0.3725. \\ \text{Year 6, Quarter 2: } \hat{y}_{22} &= (0.60 + 0.0049(22)) \times 0.90 = 0.6370. \\ \text{Year 6, Quarter 3: } \hat{y}_{23} &= (0.60 + 0.0049(23)) \times 1.40 = 0.9978. \\ \text{Year 6, Quarter 4: } \hat{y}_{24} &= (0.60 + 0.0049(24)) \times 1.17 = 0.8396.\end{aligned}$$

## Seasonal Dummy Variables

With the method of **seasonal dummy variables**, we estimate a trend forecasting model that includes dummy variables that capture each season. In Chapter 17, we learned how dummy variables are used to describe a qualitative variable with two or more categories. Recall that a dummy variable is a binary variable that equals 1 for one of the categories and 0 for the other. Here we use dummy variables to describe seasons. For quarterly data, we need to define only three dummy variables. Let  $d_1$ ,  $d_2$ , and  $d_3$  be the dummy variables for the first three quarters, using the fourth quarter as reference. Therefore for quarter 1, we use  $d_1 = 1$ ,  $d_2 = 0$ , and  $d_3 = 0$ . Similarly, for quarter 2,  $d_1 = 0$ ,  $d_2 = 1$ , and  $d_3 = 0$ , for quarter 3,  $d_1 = 0$ ,  $d_2 = 0$ , and  $d_3 = 1$ , and for quarter 4,  $d_1 = 0$ ,  $d_2 = 0$ , and  $d_3 = 0$ .

The linear and the exponential trend models with seasonal dummy variables are summarized in the following caption. We have focused on seasonal dummy variables using quarterly data, but the models can be easily modified to capture other types of seasonal data, such as monthly data. In the caption, we have removed the  $t$  subscript to simplify the notation.

#### TREND MODELS WITH SEASONAL DUMMY VARIABLES

A **linear trend model with seasonal dummy variables** is specified as

$$y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 t + \varepsilon.$$

**Forecasts** based on the estimated model are as follows:

$$\begin{aligned}\text{Quarter 1 } (d_1 = 1, d_2 = 0, d_3 = 0): \hat{y}_t &= (b_0 + b_1) + b_4 t \\ \text{Quarter 2 } (d_1 = 0, d_2 = 1, d_3 = 0): \hat{y}_t &= (b_0 + b_2) + b_4 t \\ \text{Quarter 3 } (d_1 = 0, d_2 = 0, d_3 = 1): \hat{y}_t &= (b_0 + b_3) + b_4 t \\ \text{Quarter 4 } (d_1 = 0, d_2 = 0, d_3 = 0): \hat{y}_t &= b_0 + b_4 t\end{aligned}$$

#### LO 18.6

Use trend regression models with seasonal dummy variables to make forecasts.

An **exponential trend model with seasonal dummy variables** is specified as

$$\ln(y) = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 t + \varepsilon.$$

**Forecasts** based on the estimated model are as follows:

$$\text{Quarter 1 } (d_1 = 1, d_2 = 0, d_3 = 0): \hat{y}_t = \exp((b_0 + b_1) + b_4 t + s_e^2/2)$$

$$\text{Quarter 2 } (d_1 = 0, d_2 = 1, d_3 = 0): \hat{y}_t = \exp((b_0 + b_2) + b_4 t + s_e^2/2)$$

$$\text{Quarter 3 } (d_1 = 0, d_2 = 0, d_3 = 1): \hat{y}_t = \exp((b_0 + b_3) + b_4 t + s_e^2/2)$$

$$\text{Quarter 4 } (d_1 = 0, d_2 = 0, d_3 = 0): \hat{y}_t = \exp(b_0 + b_4 t + s_e^2/2)$$

## EXAMPLE 18.6

Revisit the *Nike\_Revenues* data considered in the introductory case. Use the seasonal dummy variable model to make a forecast for Nike's quarterly revenue in 2009.

**SOLUTION:** Given quarterly data, we first construct relevant variables for the regression. Table 18.16 specifies seasonal dummy variables, along with the time variable  $t$ .

**TABLE 18.16** Constructing Seasonal Dummy Variables (Example 18.6)

Period	$y$	$\ln(y)$	$t$	$d_1$	$d_2$	$d_3$
1999:01	2,505	7.8260	1	1	0	0
1999:02	1,913	7.5564	2	0	1	0
1999:03	2,177	7.6857	3	0	0	1
1999:04	2,182	7.6880	4	0	0	0
2000:01	2,501	7.8245	5	1	0	0
2000:02	2,060	7.6305	6	0	1	0
2000:03	2,162	7.6788	7	0	0	1
2000:04	2,273	7.7289	8	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2008:03	4544	8.4216	39	0	0	1
2008:04	5,088	8.5346	40	0	0	0

As in the case of decomposition analysis, we use the exponential model to capture trend:  $\ln(y) = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 t + \varepsilon$ . Relevant estimates of the regression model are presented in Table 18.17.

**TABLE 18.17** Regression Results for Example 18.6

	Coefficients	Standard Error	$t$ Stat	$p$ -Value
Intercept	7.5929	0.0261	290.57	0.00
$d_1$	0.0501	0.0268	1.87	0.07
$d_2$	-0.1036	0.0267	-3.87	0.00
$d_3$	-0.0985	0.0267	-3.69	0.00
$t$	0.0218	0.0008	26.53	0.00
The standard error of the estimate $s_e$ equals 0.0597.				

The estimated equation, with values rounded to the 4th decimal place, is  $\hat{y} = \exp(7.5929 + 0.0501d_1 - 0.1036d_2 - 0.0985d_3 + 0.0218t + 0.0597^2/2)$ . The coefficients for seasonal dummy variables indicate that the revenue is about 5% higher in the first quarter and about 10% lower in the second and third quarters as compared to the fourth quarter. As discussed in Chapter 16, for the exponential model, the estimated coefficient of 0.0218 suggests that the predicted quarterly increase in revenue is about 2.18%.

### FILE

*Nike\_Revenues*

For 2009:01, we use  $d_1 = 1$ ,  $d_2 = 0$ ,  $d_3 = 0$ ,  $t = 41$  to forecast Nike's revenue as  $\hat{y}_{41} = \exp(7.5929 + 0.0501 + 0.0218(41) + 0.0597^2/2) = \$5,108.10$  million. Similarly, for 2009:02, we use  $d_1 = 0$ ,  $d_2 = 1$ ,  $d_3 = 0$ ,  $t = 42$  to determine  $\hat{y}_{42} = \exp(7.5929 - 0.1036 + 0.0218(42) + 0.0597^2/2) = \$4,476.88$  million.

Forecasts for 2009:03 and 2009:04 yield \$4,598.94 million and \$5,186.85 million, respectively. As before, whenever possible, it is advisable to make forecasts with unrounded values.

As emphasized earlier, we always use model selection criteria to choose the appropriate forecasting model. In Table 18.18 we present the *MSE* and *MAD* based on the residuals,  $e_t = y_t - \hat{y}_t$ , with decomposition analysis and seasonal dummy variables; we did these calculations in Excel with unrounded values for the estimates. We encourage you to replicate these results.

**TABLE 18.18** In-Sample Model Selection Criteria

Model	<i>MSE</i>	<i>MAD</i>
Decomposition Analysis	24,353.95	118.10
Seasonal Dummy Variables	24,843.48	121.32

We select the decomposition analysis method to make forecasts because it has the lower *MSE* and *MAD* of in-sample forecast errors. Therefore, the quarterly forecasts for 2009, as derived earlier, are \$5,070.89, \$4,511.04, \$4,593.82, and \$5,194.44 million, respectively. This results in a sum of \$19,370.19 million or \$19.37 billion for fiscal year 2009.

## SYNOPSIS OF INTRODUCTORY CASE

Nike, Inc., is the world's leading supplier and manufacturer of athletic shoes, apparel, and sports equipment. Its revenue in the fiscal year ending May 31, 2008, was \$18.627 billion. While some analysts argue that a slow-down of Nike's revenue may occur due to the global economic crisis and increased competition from emerging brands, others believe that with its strong cash flow, Nike will emerge even stronger than before as its competitors get squeezed. This report analyzes the quarterly revenue of Nike from 1999–2008 to make a forecast for fiscal year 2009.

The detailed analysis of the data suggests significant trend and seasonal components in Nike's revenue. For each fiscal year, the revenue is generally higher in the first quarter (June 1–August 31) and the fourth quarter (March 1–May 31). This result is not surprising given that these quarters cover summer and spring seasons, when people most often participate in outdoor activities. Based on various model selection criteria, the decomposition method is chosen as the preferred forecasting technique. It provides forecasts by multiplying the exponential trend estimate with the corresponding seasonal index. The quarterly revenue forecasts for 2009 are \$5,070.89, \$4,511.04, \$4,593.82, and \$5,194.44 million, respectively, resulting in \$19.37 billion for fiscal year 2009. Interestingly, this forecast, based on a time series analysis of revenue, is extremely close to the actual revenue of \$19.20 billion reported by Nike.



## EXERCISES 18.4

### Mechanics

17. Six years of quarterly data of a seasonally adjusted series are used to estimate a linear trend model as  $\hat{T}_t = 128.20 + 1.06t$ . In addition, quarterly seasonal indices are calculated as  $\hat{S}_1 = 0.93$ ,  $\hat{S}_2 = 0.88$ ,  $\hat{S}_3 = 1.14$ , and  $\hat{S}_4 = 1.05$ .
  - a. Interpret the first and fourth quarterly indices.
  - b. Make a forecast for all four quarters of next year.
18. Eight years of quarterly data of a seasonally adjusted series are used to estimate an exponential trend model as  $\ln(\hat{T}_t) = 2.80 + 0.03t$  with a standard error of the estimate,  $s_e = 0.08$ . In addition, quarterly seasonal indices are calculated as  $\hat{S}_1 = 0.94$ ,  $\hat{S}_2 = 1.08$ ,  $\hat{S}_3 = 0.86$ , and  $\hat{S}_4 = 1.12$ .
  - a. Interpret the third and fourth quarterly indices.
  - b. Make a forecast for all four quarters of next year.
19. Ten years of monthly data of a seasonally adjusted series are used to estimate a linear trend model as  $\hat{T}_t = 24.50 + 0.48t$ . In addition, seasonal indices for January and February are calculated as 1.04 and 0.92, respectively. Make a forecast for the first two months of next year.
20. **FILE Exercise 18.20.** Consider the following 20 observations, representing quarterly information for 5 years.

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
1	8.37	12.78	8.84	15.93
2	10.03	12.48	8.91	24.81
3	9.61	9.65	15.93	22.00
4	8.80	11.45	6.79	10.16
5	7.46	10.58	13.35	19.77

- a. Calculate the 4-quarter centered moving average.
    - b. Calculate the ratio-to-moving average.
    - c. Calculate and interpret the seasonal indices for quarters 1 and 4.
  21. **FILE Exercise 18.21.** Consider the following monthly observations for 5 years.
- | Year | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1    | 13  | 17  | 15  | 12  | 32  | 15  | 21  | 17  | 33  | 15  | 34  | 21  |
| 2    | 26  | 10  | 14  | 19  | 17  | 14  | 27  | 30  | 25  | 15  | 18  | 27  |
| 3    | 24  | 11  | 19  | 23  | 19  | 18  | 31  | 18  | 15  | 31  | 34  | 27  |
| 4    | 23  | 22  | 17  | 24  | 17  | 24  | 15  | 19  | 33  | 19  | 16  | 39  |
| 5    | 17  | 19  | 25  | 30  | 18  | 17  | 19  | 36  | 19  | 26  | 21  | 35  |
- a. Calculate the 12-month centered moving average.
    - b. Calculate the ratio-to-moving average.
    - c. Calculate and interpret the seasonal indices for April and November.
  22. **FILE Exercise 18.22.** Consider the following 20 observations, representing quarterly information for 5 years.

Year	Q1	Q2	Q3	Q4
1	6.49	7.34	7.11	10.82
2	7.04	7.92	7.69	11.71
3	7.62	8.58	8.34	12.68
4	8.25	9.29	9.02	13.74
5	8.94	10.08	9.78	14.88

- a. Plot the above series and discuss its trend and seasonal components.
  - b. Use decomposition analysis to make forecasts with the exponential trend and seasonal indices. Compute the mean square errors of in-sample forecast errors.
  - c. Estimate an exponential trend with seasonal dummies model. Compute the mean square errors of in-sample forecast errors.
  - d. Use the appropriate model to make forecasts for all four quarters of the next year.
23. **FILE Exercise 18.23.** Consider a portion of monthly sales data for 5 years for a growing firm.

Year	Month	Sales
1	Jan	345
1	Feb	322
:	:	:
5	Dec	10,745

- a. Construct the seasonal indices for the data.
  - b. Plot the seasonally adjusted series to recommend the appropriate trend model.
  - c. Use the trend and seasonal estimates to make forecasts for the next two months.
24. **FILE Exercise 18.23.** Using the data from exercise 23, estimate (a) a linear trend model with seasonal dummies, and (b) an exponential trend model with seasonal dummies. Which of the two models has a lower *MSE* and *MAD*? Use the appropriate model to make forecasts for the next two months.

### Applications

25. **FILE Sales Data.** Hybrid cars have gained popularity because of their fuel economy and the uncertainty regarding the price of gasoline. All automakers, including the Ford Motor Co., have planned to significantly expand their hybrid vehicle lineup (CNN.com, November 9, 2005). The following table contains quarterly sales of Ford and Mercury hybrid cars.

Year	Q1	Q2	Q3	Q4
2006	6,192	5,663	4,626	5,645
2007	5,149	6,272	5,196	6,545
2008	5,467	5,235	3,160	7,007
2009	5,337			

SOURCE: Internal Revenue Service, United States Department of Treasury.



- a. Plot the preceding series. Comment on the trend and seasonal variation in the sales of hybrid cars.
  - b. Compute and interpret the seasonal indices.
26. **FILE Treasury Bonds.** Consider a portion of monthly return data on 20-year Treasury Bonds from 2006–2010.

Year	Month	Return (%)
2006	Jan	4.65
2006	Feb	4.73
⋮	⋮	⋮
2010	Dec	4.16

SOURCE: Federal Reserve Bank of Dallas.

- a. Plot the above series and discuss its seasonal variations.
  - b. Construct the seasonal indices for the data.
  - c. Estimate a linear trend model to the seasonally adjusted series.
  - d. Use the trend and seasonal estimates to make forecasts for the first three months of 2011.
27. **FILE Treasury Bonds.** Using the data from exercise 26, estimate a linear trend model with seasonal dummy variables to make forecasts for the first three months of 2011.
28. **FILE Expenses.** The controller of a small construction company is attempting to forecast expenses for the next year. He collects quarterly data on expenses (in \$1,000s) over the past 5 years, a portion of which is shown in the accompanying table.

Year	Quarter	Expenses
2006	1	\$2,136
2006	2	2,253
⋮	⋮	⋮
2010	4	3,109

- a. Estimate a linear trend model with seasonal dummy variables and compute the *MSE* and *MAD* of in-sample forecast errors.
  - b. Estimate an exponential trend model with seasonal dummy variables and compute the resulting *MSE* and *MAD*.
  - c. Which model is more appropriate? With this model, forecast expenses for year 2011.
29. **FILE Blockbuster.** Blockbuster Inc. has lately faced challenges by the growing online market (CNNMoney.com, March 3, 2009). Its revenue from rental stores has been sagging as customers are increasingly getting

their movies through the mail or high-speed Internet connections. The following table contains a portion of the total revenue from rentals of all formats of movies at Blockbuster Inc. (in millions of dollars).

Year	Quarter	Revenue
2001	1	\$1.403683
2001	2	1.287625
⋮	⋮	⋮
2008	4	1.097712

- a. Compute seasonal indices.
  - b. Fit linear and quadratic trend models to the seasonally adjusted data. Which model do you prefer?
  - c. Use decomposition analysis to make quarterly forecasts for 2009.
30. **FILE Blockbuster.** Use the data from exercise 29 to:
- a. Estimate a linear trend model with seasonal dummy variables.
  - b. Estimate a quadratic trend model with seasonal dummy variables.
  - c. Use the appropriate model to make quarterly forecasts for 2009.
31. **FILE Consumer Sentiment.** The following table lists a portion of the University of Michigan's Consumer Sentiment Index. This index is normalized to have a value of 100 in 1965 and is used to record changes in consumer morale.

Year	Month	Consumer Sentiment
2005	Jan	95.5
2005	Feb	91.2
⋮	⋮	⋮
2010	Oct	67.7

SOURCE: Federal Reserve Bank of St. Louis.

- a. Plot and interpret the series.
  - b. Construct and interpret seasonal indices for the series.
  - c. Estimate linear, quadratic and cubic trend models to the seasonally adjusted data. Select the best fitting model.
  - d. Use decomposition analysis to make a forecast for November and December of 2010.
32. **FILE Consumer Sentiment.** Using the data from exercise 31, fit an appropriate polynomial trend model along with seasonal dummies to make a forecast for November and December of 2010.

Use causal forecasting models to make forecasts.

So far, we have discussed noncausal, or purely time series, models. These models do not offer any explanation of the mechanism generating the variable of interest and simply provide a method for projecting historical data. Although this method can be effective, it provides no guidance on the likely effects of changes in policy (explanatory) variables. **Causal forecasting models** are based on a regression framework, where the explanatory variables influence the outcome of the response variable. For instance, consider the following simple linear regression model:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

Here  $y$  is the response variable caused by the explanatory variable  $x$ . Let the sample observations be denoted by  $y_1, y_2, \dots, y_T$  and  $x_1, x_2, \dots, x_T$ , respectively. Once we have estimated this model, we can make a one-step-ahead forecast as

$$\hat{y}_{T+1} = b_0 + b_1 x_{T+1}.$$

Multi-step-ahead forecasts can be made similarly. This causal approach works only if we know, or can predict, the future value of the explanatory variable  $x_{T+1}$ . For instance, let sales  $y$  be related to expenditure on advertisement  $x$ . We can forecast sales  $\hat{y}_{T+1}$  only if we know the advertisement budget,  $x_{T+1}$ , for the next period.

### Lagged Regression Models

For forecasting, sometimes we use a causal approach with lagged values of  $x$  and  $y$  as explanatory variables. For instance, consider the model,

$$y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t,$$

where  $\beta_1$  represents the slope of the lagged explanatory variable  $x$ . Note that if we have  $T$  sample observations, the estimable sample will consist of  $T - 1$  observations, where  $y_2, y_3, \dots, y_T$  are matched with  $x_1, x_2, \dots, x_{T-1}$ . Here a one-step-ahead forecast is easily made as

$$\hat{y}_{T+1} = b_0 + b_1 x_T.$$

This forecast is not conditional on any predicted value of the explanatory variables, since  $x_T$  is its last known sample value. We can generalize this model to include more lags of  $x$ . For example, we can specify a two-period lagged regression model as  $y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \varepsilon_t$ . A one-step-ahead forecast is now made as  $\hat{y}_{T+1} = b_0 + b_1 x_T + b_2 x_{T-1}$ .

Another popular specification for causal forecasting uses lagged values of the response variable as an explanatory variable. For instance, consider the model

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t,$$

where the parameter  $\beta_1$  represents the slope of the lagged response variable  $y$ . This regression is also referred to as an **autoregressive model** of order one, or simply an AR(1). Higher-order autoregressive models can be constructed similarly. A one-period-ahead forecast is made as

$$\hat{y}_{T+1} = b_0 + b_1 y_T.$$

Finally, we can also use lagged values of both  $x$  and  $y$  as the explanatory variables. For instance, consider

$$y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 y_{t-1} + \varepsilon_t.$$

Here, a one-period-ahead forecast is made as

$$\hat{y}_{T+1} = b_0 + b_1 x_T + b_2 y_T.$$

## EXAMPLE 18.7

Table 18.19 shows a portion of data on net private housing units sold (in 1,000s), and real per-capita gross domestic product (in \$1,000s). Let Housing denote housing units sold and GDP denote real per-capita gross domestic product. Estimate the following three models and use the most suitable model to make a forecast for housing units sold in 2009.

$$\text{Model 1: Housing}_t = \beta_0 + \beta_1 \text{GDP}_{t-1} + \varepsilon_t.$$

$$\text{Model 2: Housing}_t = \beta_0 + \beta_1 \text{Housing}_{t-1} + \varepsilon_t.$$

$$\text{Model 3: Housing}_t = \beta_0 + \beta_1 \text{GDP}_{t-1} + \beta_2 \text{Housing}_{t-1} + \varepsilon_t.$$

**TABLE 18.19** Housing and GDP Data

Year	Housing Units Sold	Real Per-Capita GDP
1981	436	23.007
1982	412	22.346
1983	623	23.146
⋮	⋮	⋮
2008	509	38.399

SOURCE: The Department of Commerce.

**FILE**  
Housing\_Units

**SOLUTION:** In order to estimate these models, we first have to lag Housing and GDP. Table 18.20 uses a portion of the data to show lagged values.

**TABLE 18.20** Generating Lagged Values

Year	Housing <sub>t</sub>	GDP <sub>t</sub>	GDP <sub>t-1</sub>	Housing <sub>t-1</sub>
1981	436	23.007	—	—
1982	412	22.346	23.007	436
1983	623	23.146	22.346	412
⋮	⋮	⋮	⋮	⋮
2008	509	38,399	38.148	776

Note that we lose one observation for the regression, since we do not have information on the lagged values for 1981. Table 18.21 summarizes the regression results of the three models.

**TABLE 18.21** Model Evaluation with Causal Forecasting

Parameters	Model 1	Model 2	Model 3
Constant	-112.9093 (0.59)	122.7884 (0.14)	300.0187 (0.07)
GDP <sub>t-1</sub>	29.0599* (0.00)	NA	-10.9084 (0.20)
Housing <sub>t-1</sub>	NA	0.8433* (0.00)	1.0439* (0.00)
Adjusted R <sup>2</sup>	0.3974	0.7264	0.7340

NOTES: The top portion of the table contains parameter estimates with *p*-values in parentheses; NA denotes not applicable; the symbol \* denotes significance at the 5% level.

As discussed earlier, it is preferable to compare competing multiple regression models in terms of adjusted  $R^2$  since it appropriately penalizes for the excessive use of explanatory variables. We choose Model 3 because it has the highest adjusted  $R^2$  of 0.7340. The forecast for 2009 is made as:

$$\begin{aligned}\widehat{\text{Housing}}_{2009} &= b_0 + b_1 \text{GDP}_{2008} + b_2 \text{Housing}_{2008} \\ &= 300.0187 - 10.9084(38.399) + 1.0439(509) = 412.49.\end{aligned}$$

Therefore, we forecast that about 412,490 net private housing units will be sold in 2009.

## EXERCISES 18.5

### Mechanics

33. **FILE Exercise 18.33.** Consider the following portion of data on the response variable  $y$  and the explanatory variable  $x$ .

$t$	$y$	$x$
1	27.96	29.88
2	30.15	21.07
$\vdots$	$\vdots$	$\vdots$
12	24.47	26.41

- a. Estimate  $y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$  and  $y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \varepsilon_t$ .
- b. Use the appropriate model to make a one-step-ahead forecast ( $t = 13$ ) for  $y$ .
34. **FILE Exercise 18.34.** Consider the following portion of data on the variable  $y$ .

$t$	$y$
1	29.32
2	30.96
$\vdots$	$\vdots$
24	48.58

- a. Estimate an autoregressive model of order 1,  $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ , to make a one-step-ahead forecast ( $t = 25$ ) for  $y$ .
- b. Estimate an autoregressive model of order 2,  $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t$ , to make a one-step-ahead forecast ( $t = 25$ ) for  $y$ .
- c. Which of the above models is more appropriate for forecasts? Explain.
35. **FILE Exercise 18.35.** Consider the following portion of data on  $y$  and  $x$ .

$t$	$y$	$x$
1	18.23	17.30
2	19.82	16.05
$\vdots$	$\vdots$	$\vdots$
12	22.75	13.66

- a. Estimate  $y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$  to make a one-step-ahead forecast for period 13.
- b. Estimate  $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$  to make a one-step-ahead forecast for period 13.
- c. Which of the above models is more appropriate for forecasts? Explain.
36. **FILE Exercise 18.36.** Consider the following portion of data on  $y$  and  $x$ .

$t$	$y$	$x$
1	56.96	9,171.61
2	57.28	9,286.56
$\vdots$	$\vdots$	$\vdots$
12	51.99	9,217.94

- a. Estimate  $y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$ .
- b. Estimate  $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ .
- c. Estimate  $y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 y_{t-1} + \varepsilon_t$ .
- d. Use the most suitable model to make a one-step-ahead forecast ( $t = 13$ ) for  $y$ .

### Applications

37. Hiroshi Sato, an owner of a sushi restaurant in San Francisco, has been following an aggressive marketing campaign to thwart the effect of rising unemployment rates on business. He used monthly data on sales (\$1,000s), advertising costs (\$), and the unemployment rate (%) from January 2008 to May 2009 to estimate the following sample regression equation:
- $$\text{Sales}_t = 17.51 + 0.03\text{Advertising Costs}_{t-1} - 0.69\text{Unemployment Rate}_{t-1}.$$
- a. Hiroshi had budgeted \$620 toward advertising costs in May 2009. Make a forecast for Sales for June 2009 if the unemployment rate in May 2009 was 9.1%.
- b. What will be the forecast if he raises his advertisement budget to \$700?
- c. Reevaluate the above forecasts if the unemployment rate was 9.5% in May 2009.
38. **FILE Phillips Curve.** The Phillips curve is regarded as a reliable tool for forecasting inflation. It captures the inverse relation between the rate of unemployment and the rate of inflation; the lower the unemployment in an economy, the higher is the inflation rate. Consider the following portion of monthly data on seasonally adjusted inflation and unemployment rates in the United States from January 2009 to November 2010.

Year	Month	Unemployment	Inflation
2009	Jan	7.7	0.3
2009	Feb	8.2	0.4
$\vdots$	$\vdots$	$\vdots$	$\vdots$
2010	Nov	9.8	0.1

SOURCE: Bureau of Labor Statistics.

- a. Estimate two models, of order 1 and 2, using unemployment as the response variable and lagged inflation as the explanatory variable(s). Should you use either model for forecasting unemployment? Explain.
- b. Estimate autoregressive models of order 1 and 2 on unemployment. Choose the appropriate model to make a forecast of unemployment for December 2010.
39. **FILE HD\_DOW.** A research analyst at an investment firm is attempting to forecast the daily stock price of Home Depot, using causal models. The following table shows a portion of the daily adjusted closing prices of Home Depot

$y$  and the Dow Jones Industrial Average  $x$  from August 14, 2009, to August 31, 2009.

$t$	$y$	$x$
August 14	26.92	9,321.40
August 17	25.89	9,135.34
$\vdots$	$\vdots$	$\vdots$
August 31	27.07	9,496.28

SOURCE: www.finance.yahoo.

Estimate three models: (a)  $y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$ , (b)  $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ , and (c)  $y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 y_{t-1} + \varepsilon_t$ . Use the most suitable model to forecast Home Depot's daily stock price for September 1, 2009.

## WRITING WITH STATISTICS

An important indicator of an economy is its inflation rate, which is generally defined as the percentage change in the consumer price index over a specific period of time. It is well documented that high inflation rates lead to a decline in the real value of money, which in turn can discourage investment and savings. The task of keeping the inflation rate within desired limits is entrusted to monetary authorities who use various policy instruments to control it. However, their actions depend primarily on their ability to gauge inflationary pressures accurately, or in other words, to correctly forecast the inflation rate.

Pooja Nanda is an economist working for *Creative Thinking*, a well-regarded policy institute based in Washington, DC. She has been given the challenging task of forecasting inflation for January 2009. She has access to seasonally adjusted monthly inflation rates in the United States from January 2007 to December 2008, a portion of which is shown in Table 18.22.



**TABLE 18.22** Seasonally Adjusted Monthly Inflation Rates

Date	Inflation
Jan-07	0.2
Feb-07	0.4
$\vdots$	$\vdots$
Dec-08	-0.8

SOURCE: Bureau of Labor Statistics.

**FILE**  
Inflation-Rates

Pooja would like to use the sample information to:

1. Evaluate various polynomial trend models for the inflation rate.
2. Use the best-fitting trend model to forecast the inflation rate for January 2009.

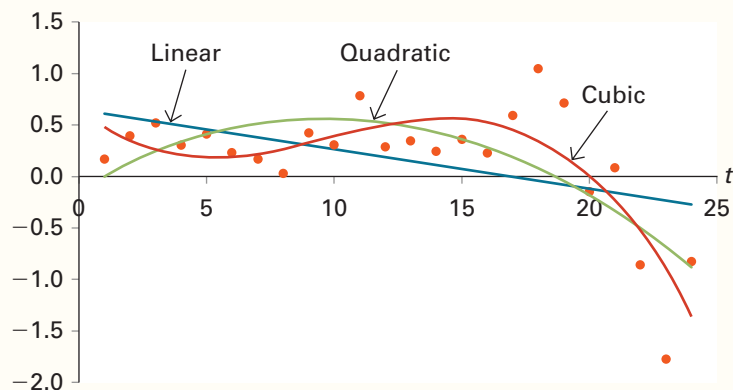
Economists generally agree that high levels of inflation are caused by the money supply growing faster than the rate of economic growth. During high inflationary pressures, monetary authorities decrease the money supply, thereby raising short-term interest rates. Sometimes they also have to contend with deflation, or a prolonged reduction in the level of prices. As prices fall, consumers tend to delay purchases until prices fall further, which in turn can depress overall economic activity.

**Sample Report—**  
**Forecasting**  
**the Monthly**  
**Inflation Rate**

The global economic crisis that began in the summer/fall of 2008 raised deflationary fears, with rapidly rising unemployment rates and capital markets in turmoil. An increase in price levels in 2007 was followed by a decrease in 2008. This report does not focus on the effectiveness of monetary policy. Instead, a forecast of the inflation rate is made from a noncausal perspective by simply projecting historical data. Seasonality is not a concern, since the inflation data are already seasonally adjusted.

A simple plot of the inflation rate from January 2007 to December 2008 is shown in Figure 18.A. In order to gauge whether a linear or nonlinear trend is appropriate, various trend models are superimposed on the inflation rate scatterplot. The exponential trend is not included, since the log of the inflation rate is not defined for nonpositive inflation rates.

**FIGURE 18.A** Scatterplot of inflation (in percent) and superimposed trends



Interestingly, the implied forecasts seem to vary widely between competing models. Although Figure 18.A suggests that the cubic model accurately captures the changing trend of the inflation rate over the last 24 months, this finding must be supplemented with formal model selection criteria.

Three trend models are estimated, where  $y_t$  represents the inflation rate and  $t$  represents the relabeled monthly observations from 1 (January 2007) through 24 (December 2008).

$$\text{Linear Model: } y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

$$\text{Quadratic Model: } y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$$

$$\text{Cubic Model: } y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \varepsilon_t$$

Table 18.A presents parameter estimates of the three models. Also included in the table are the adjusted  $R^2$  for model comparison, which suitably penalizes over-parameterization.

**TABLE 18.A** Analysis of the Linear, Quadratic, and Cubic Trend Models

Variable	Linear	Quadratic	Cubic
Constant	0.6478* (0.009)	-0.1352 (0.647)	0.6568 (0.078)
$t$	-0.0383* (0.024)	0.1424* (0.015)	-0.2031 (0.106)
$t^2$	NA	-0.0072* (0.002)	0.0266* (0.026)
$t^3$	NA	NA	-0.0009* (0.006)
Adjusted $R^2$	0.1757	0.4506	0.6106

NOTES: The top portion of the table contains parameter estimates with  $p$ -values in parentheses. NA denotes not applicable. The asterisk designates significance at the 5% level. The last row of the table contains adjusted  $R^2$  for model comparison.



Consistent with the informal graphical analysis, the cubic trend provides the best sample fit, as it has the highest adjusted  $R^2$  of 0.6106. Therefore, the estimated cubic trend model is used with unrounded estimates to derive the forecast for June 2009 as

$$\hat{y}_{30} = 0.6568 - 0.2031 \times 25 + 0.0266 \times 25^2 - 0.0009 \times 25^3 = -1.86.$$

Inflation forecasts are widely regarded as key inputs for implementing monetary policy. In this report, we employ basic trend models to project historical data on inflation. We note that the resulting forecasts ignore the likely effects of policy changes.

## CONCEPTUAL REVIEW

### LO 18.1 Distinguish among the various models used in forecasting.

Observations of any variable recorded over time in sequential order are considered a **time series**. The purpose of any forecasting model is to forecast the outcome of a time series at time  $t$ , or  $\hat{y}_t$ . Forecasting methods are broadly classified as **quantitative** or **qualitative**. While qualitative forecasts are based on prior experience and the expertise of the forecaster, quantitative forecasts use a formal model, along with historical data for the variable of interest. Quantitative forecasting models are further divided into **causal** and **noncausal models**. Causal methods are based on a regression framework, where the variable of interest is related to a single or multiple explanatory variables. Noncausal models, also referred to as purely time series models, do not present any explanation of the mechanism generating the variable of interest and simply provide a method for projecting historical data.

The in-sample forecast  $\hat{y}_t$  is also called the **predicted** or **fitted** value of  $y_t$ . The in-sample forecast errors or the residuals are computed as  $e_t = y_t - \hat{y}_t$ . We use the residuals, computed as  $e_t = y_t - \hat{y}_t$ , to compute  $MSE = \frac{\sum e_t^2}{n}$  and  $MAD = \frac{\sum |e_t|}{n}$ , where  $n$  is the number of residuals used in the computation. When selecting among various models, we choose the one with the smallest  $MSE$  and  $MAD$ . If  $MSE$  and  $MAD$  provide conflicting results, then we choose the model with the smallest  $MSE$ .

### LO 18.2 Use smoothing techniques to make forecasts.

Time series consist of **systematic** and **unsystematic** patterns. Systematic patterns are caused by the **trend**, the **seasonal**, and the **cyclical** components. Unsystematic patterns are difficult to identify and are caused by the presence of a **random error** term. **Smoothing techniques** are employed to provide forecasts if short-term fluctuations represent random departures from the structure with no discernible systematic patterns.

A **moving average** is the average from a fixed number of the  $m$  most recent observations. We use moving averages to make forecasts as  $\hat{y}_t = \frac{y_{t-m} + y_{t-m+1} + \dots + y_{t-1}}{m}$ .

**Exponential smoothing** is a weighted average approach where the weights decline exponentially as they become more distant. The exponential smoothing procedure continually updates the level of the series as  $A_t = \alpha y_t + (1 - \alpha)A_{t-1}$ , where  $\alpha$  represents the speed of decline. Forecasts are made as  $\hat{y}_{t+1} = A_t$ .

### LO 18.3 Use trend regression models to make forecasts.

For a time series that grows by a fixed amount for each time period, we use the **linear trend model**,  $y_t = \beta_0 + \beta_1 t + \varepsilon_t$ . We estimate this model to make forecasts as  $\hat{y}_t = b_0 + b_1 t$ , where  $b_0$  and  $b_1$  are the coefficient estimates.

For a time series that grows by an increasing amount for each time period, we use the **exponential trend model**,  $\ln(y_t) = \beta_0 + \beta_1 t + \varepsilon_t$ , where  $\ln(y_t)$  is the natural log of the series. We estimate this model to make forecasts as  $\hat{y}_t = \exp(b_0 + b_1 t + s_e^2/2)$ , where  $b_0$  and  $b_1$  are the coefficient estimates and  $s_e$  is the standard error of the estimate.

A **polynomial trend model** of order  $q$  is estimated as  $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \dots + \beta_q t^q + \varepsilon_t$ . This model specializes to a linear trend model for  $q = 1$ , to a quadratic trend model for  $q = 2$ , and to a cubic trend model for  $q = 3$ . We estimate this model to make forecasts as  $\hat{y}_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3 + \dots + b_q t^q$ , where  $b_0, b_1, \dots, b_q$  are the coefficient estimates.

It is always informative to start a trend analysis with a scatterplot of the series. We compare the linear and the exponential trend models on the basis of their *MSE* and *MAD*. We use adjusted  $R^2$  to compare various orders of the polynomial trend model.

#### LO 18.4 Calculate and interpret seasonal indices and use them to seasonally adjust a time series.

**Centered moving averages, CMAs**, are often employed to separate the effect of trend from seasonality. We use a centered 4-period moving average for quarterly data and a centered 12-period moving average for monthly data. The **ratio-to-moving average** is calculated as  $y_t/\bar{y}_t$ , where  $\bar{y}_t = CMA_t$ . For each season, there are many estimates of the ratio-to-moving average. The **unadjusted seasonal index** is computed by averaging these ratios over the years. A minor adjustment ensures that the average of **adjusted seasonal indices**  $\hat{S}$  equals one. A **seasonally adjusted** series is computed as  $y_t/\hat{S}_t$ . We use the appropriate trend model on the seasonally adjusted series to extract  $\hat{T}_t$ .

#### LO 18.5 Use decomposition analysis to make forecasts.

We let  $T$ ,  $S$ , and  $I$  represent the trend, the seasonal, and the random components, respectively, of the series  $y_t$ . Using **decomposition analysis**, we decompose or isolate the individual components of the time series to make forecasts. Forecasts are made as  $\hat{y}_t = \hat{T}_t \times \hat{S}_t$ , where  $\hat{T}_t$  and  $\hat{S}_t$  represent the estimated trend and the seasonal index, respectively.

#### LO 18.6 Use trend regression models with seasonal dummy variables to make forecasts.

As an alternative to decomposition analysis, we use a multiple regression model to simultaneously estimate trend along with **seasonal dummy variables**. With quarterly data, a **linear** trend model with seasonal dummy variables is specified as  $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 t + \varepsilon$ . Forecasts based on the estimated model are  $\hat{y}_t = (b_0 + b_1) + b_4 t$  (Quarter 1),  $\hat{y}_t = (b_0 + b_2) + b_4 t$  (Quarter 2),  $\hat{y}_t = (b_0 + b_3) + b_4 t$  (Quarter 3), and  $\hat{y}_t = b_0 + b_4 t$  (Quarter 4). An **exponential** trend model with seasonal dummy variables is specified as  $\ln(y) = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 t + \varepsilon$ . Forecasts based on the estimated model are  $\hat{y}_t = \exp((b_0 + b_1) + b_4 t + s_e^2/2)$  (Quarter 1),  $\hat{y}_t = \exp((b_0 + b_2) + b_4 t + s_e^2/2)$  (Quarter 2),  $\hat{y}_t = \exp((b_0 + b_3) + b_4 t + s_e^2/2)$  (Quarter 3), and  $\hat{y}_t = \exp(b_0 + b_4 t + s_e^2/2)$  (Quarter 4). Forecasts with monthly data can be made similarly.

#### LO 18.7 Use causal forecasting models to make forecasts.

**Causal forecasting models** are based on a regression framework. A forecast with a simple regression model,  $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ , can be made as  $\hat{y}_{T+1} = b_0 + b_1 x_{T+1}$  only if the future value of the explanatory variable  $x_{T+1}$  is known. Sometimes researchers use a causal approach with lagged values of  $x$  and  $y$  for making forecasts. For instance, we can estimate  $y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$  to make a forecast as  $\hat{y}_{T+1} = b_0 + b_1 x_T$ . Similarly, we can estimate  $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$  to make a forecast as  $\hat{y}_{T+1} = b_0 + b_1 y_T$ . We can also estimate a combined model  $y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 y_{t-1} + \varepsilon_t$  to make a forecast as  $\hat{y}_{T+1} = b_0 + b_1 x_T + b_2 y_T$ .

# ADDITIONAL EXERCISES AND CASE STUDIES

## Exercises

40. **FILE Yields.** The U.S. housing market remains fragile despite historically low mortgage rates (AARP, July 2, 2010). Since the rate on 30-year mortgages is tied to the 10-year yield on Treasury bonds, it is important to be able to predict this yield accurately. The accompanying table shows a portion of the 10-year yield on Treasury bonds (in %) for 21 trading days in November 2010.

Date	Yield (in %)
1-Nov	2.63
2-Nov	2.59
⋮	⋮
30-Nov	2.80

SOURCE: finance.yahoo.com.

- Use a 3-period moving average to make a forecast for December 1, 2010.
  - Use the exponential smoothing method to make a forecast for December 1, 2010. Use  $\alpha = 0.5$ .
  - Which of these smoothing methods has a better in-sample performance?
  - The actual 10-year yield on December 1, 2010 was 2.96. Was the forecast performance of the two methods consistent with their in-sample performance in part c?
41. **FILE Country\_Rap.** The following table lists a portion of the percentage (share) of total shipment of music that falls in the category of country and rap/hip-hop rock music from 1990–2008.

Year	Country (in %)	Rap/Hip-hop (in %)
1990	9.6	8.5
1991	12.8	10.0
⋮	⋮	⋮
2008	11.9	10.7

SOURCE: www.riaa.com.

- Plot each of the above series and comment on the respective trend.
- Estimate a linear, a quadratic, and a cubic trend model for the share of country music in the United States. Use adjusted  $R^2$  to choose the preferred model and with this model make a forecast for 2009.
- Estimate a linear, a quadratic, and a cubic trend model for the share of rap/hip-hop music in the United States. Use adjusted  $R^2$  to choose the preferred model and with this model make a forecast for 2009.

42. **FILE Tourism\_Spending.** Tourism was hit hard by the international financial crisis that began in the fall of 2008. According to the Bureau of Economic Analysis (December 20, 2010), tourism spending has picked up, but it still remains below its peak, which occurred in 2007. The accompanying table shows a portion of seasonally adjusted data on real tourism spending (in millions of \$).

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2004	664,924	672,678	675,262	679,410
2005	683,989	690,226	695,753	700,037
⋮	⋮	⋮	⋮	⋮
2010	676,929	682,681	695,917	NA

SOURCE: U.S. Bureau of Economic Analysis.

- Construct seasonal indices for the data.
  - Plot the seasonally adjusted series. Estimate the cubic trend model on the seasonally adjusted series.
  - Use the seasonal and trend components to forecast tourism spending for the fourth quarter of 2010 and the first three quarters of 2011.
43. **FILE GasPrice\_Forecast.** Consider the following portion of monthly data on price per gallon of unleaded regular gasoline in the U.S from January 2009 to December 2010.

Year	Month	Price Per Gallon
2009	Jan	\$1.79
2009	Feb	1.92
⋮	⋮	⋮
2010	Dec	2.99

SOURCE: U.S. Energy Information Administration.

- Plot the above series to identify an appropriate polynomial trend model. You may ignore seasonality.
  - Compare the adjusted  $R^2$  of the linear, the quadratic, and the cubic trend models.
  - Use the appropriate model to make a forecast for the price of regular unleaded gasoline for January and February of 2011.
44. **FILE Loans.** Consider the following portion of data on real estate loans granted by FDIC-insured Commercial Banks in the United States (in billions of U.S. dollars, base = 2007) from 1972 to 2007.

Year	Loans
1972	489.27
1973	567.26
⋮	⋮
2007	3,604.03

SOURCE: www2.fdic.gov.

- Plot the series and comment on the growth of real estate loans.
- Estimate the linear and exponential trend models for real estate loans. Compare the models in terms of their mean square errors. Use the preferred model to make a forecast for real estate loans in year 2008.
- Compare the in-sample performance of the preferred model used in part b with an autoregressive model of order one, AR(1). Make a forecast for real estate loans in year 2008 with this model.

45. **FILE Inventory\_Sales.** While U.S. inventory levels remain low, there is a slight indication of an increase in the U.S. business inventory-to-sales ratio, due to higher sales (*The Wall Street Journal*, December 15, 2010). The accompanying table shows a portion of seasonally adjusted inventory-to-sales ratios from January 2008 to October 2010.

Year	Month	Inventory-to-Sales
2008	Jan	1.28
2008	Feb	1.30
⋮	⋮	⋮
2010	Oct	1.27

SOURCE: U.S. Department of Commerce.

- Plot the above series. Which polynomial trend model do you think is most appropriate?
- Verify your answer by formally comparing the linear, quadratic, and cubic trend models.
- Make a forecast for the inventory-to-sales ratio for November and December of 2010.

46. **FILE Revenue\_Miles.** Revenue passenger-miles are calculated by multiplying the number of paying passengers by the distance flown in thousands. The accompanying table shows a portion of monthly data on revenue passenger-miles (in millions) from January 2006 through September 2010.

Year	Month	Revenue Passenger-Miles
2006	Jan	43.1652
2006	Feb	44.0447
⋮	⋮	⋮
2010	Sep	43.7704

SOURCE: Bureau of Transportation Statistics.

- Plot the above series and comment on its trend and seasonal variations.
- Compute seasonal indices and use them to seasonally adjust the series.
- Fit the appropriate trend model to the seasonally adjusted series.
- Use decomposition analysis to make monthly forecasts for the last three months of 2010.

47. **FILE Revenue\_Miles.** Use the data from exercise 46 to:

- Estimate a linear trend model with seasonal dummies.
- Estimate an exponential trend model with seasonal dummies.
- Use the *MSE* and *MAD* to compare these models.
- Use the appropriate model to make monthly forecasts for the last three months of 2010.

48. **FILE Lowe's\_Sales.** The following data represent a portion of quarterly net sales (in millions of dollars) of Lowe's Companies, Inc., over the past five years.

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2004	\$8,861	\$10,169	\$9,064	\$8,550
2005	9,913	11,929	10,592	10,808
⋮	⋮	⋮	⋮	⋮
2008	12,009	14,509	11,728	9,984

SOURCE: All data retrieved from Annual Reports for Lowe's Companies, Inc.

- Estimate a linear trend model with seasonal dummy variables and compute the *MSE* and *MAD*.
- Estimate an exponential trend model with seasonal dummy variables and compute the *MSE* and *MAD*.
- Which model is more appropriate? Use this model to forecast net sales for Lowe's Companies, Inc., for fiscal year 2009.

49. **FILE Genzyme.** The S&P 500 Index is a value-weighted index of prices of 500 large-cap common stocks actively traded in the United States. A research analyst at an investment firm is attempting to forecast the daily stock price of Genzyme Corporation, one of the world's leading biotech companies, using both the S&P 500 Index as well as Genzyme's past stock prices. The following table shows a portion of the daily adjusted closing prices of Genzyme (*y*) and the S&P 500 Index (*x*) from December 1, 2010, to December 22, 2010.

Date	<i>y</i>	<i>x</i>
12/1/2010	71.14	1206.07
12/2/2010	70.97	1221.53
⋮	⋮	⋮
12/22/2010	71.52	1258.84

SOURCE: www.finance.yahoo.

Estimate three models: (a)  $y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$ , (b)  $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ , and (c)  $y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 y_{t-1} + \varepsilon_t$ . Use the most suitable model to forecast the Genzyme daily stock price for December 23, 2010.

50. **FILE Income\_Consumption.** In August 2010, the Department of Commerce reported that economic weakness continues across the country, with consumer spending continuing to stagnate. The government is considering various tax benefits to stimulate consumer spending through increased disposable income. The consumption function is one of the key relationships in economics, where consumption

(y) depends on disposable income (x). Consider the following table, which presents a portion of quarterly data on disposable income and personal consumption expenditure for the U.S. Both variables are measured in billions of dollars and are seasonally adjusted.

Year	Quarter	Income (x)	Consumption (y)
2006	1	9705.2	9148.2
2006	2	9863.8	9266.6
⋮	⋮	⋮	⋮
2010	4	11514.7	10525.2

SOURCE: U.S. Department of Commerce.

- Plot the consumption series. Estimate the appropriate polynomial trend model to forecast consumption expenditure for the 1st quarter of 2011.
- Estimate  $y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$  to forecast consumption expenditure for the 1st quarter of 2011.
- Which of these two models is more appropriate for making forecasts? Explain.

## CASE STUDIES

**CASE STUDY 18.1** Fried dough is a popular North American food associated with outdoor food stands at carnivals, amusement parks, fairs, and festivals, etc. Usually dusted with powdered sugar and drenched in oil, it is not particularly good for you but it sure is tasty! Jose Sanchez owns a small stall at Boston Commons in Boston, Massachusetts, where he sells fried dough and soft drinks. Although business is good, he is apprehensive about the variation in sales for no apparent reason. He asks a friend to help him make a forecast for fried dough as well as soft drinks. The accompanying table shows a portion of data on the number of plates of fried dough and soft drinks that he sold over the last 20 days.

**Data for Case Study 18.1** Data on Fried Dough and Soft Drinks

Day	Fried Dough	Soft Drinks
1	70	150
2	69	145
⋮	⋮	⋮
20	61	153

**FILE**  
*Fried\_Dough*

In a report, use the sample information to:

- Construct the exponentially smoothed series for fried dough and soft drinks using  $\alpha = 0.30$  and  $\alpha = 0.70$ .
- Calculate *MSE* and *MAD* for each series.
- Forecast sales of fried dough and soft drinks for day 21 with the best-fitting series.

**CASE STUDY 18.2** Madelyn Davis is a research analyst for a large investment firm. She has been assigned the task of forecasting sales for Walmart Stores, Inc., for fiscal year 2011. She collects quarterly sales for Walmart Stores, Inc. (in millions \$) for the 10-year period 2001 through 2010, a portion of which is shown in the accompanying table.

**Data for Case Study 18.2** Walmart Quarterly Sales (in millions \$)

Year	Quarters Ended			
	April 30	July 31	October 31	January 31
2001	42,985	46,112	45,676	56,556
2002	48,052	52,799	52,738	64,210
⋮	⋮	⋮	⋮	⋮
2010	93,471	100,082	98,667	112,826

SOURCE: All data retrieved from Annual Reports for Walmart Stores, Inc.

**FILE**  
*Walmart\_Sales*

In a report, use the sample information to:

- Use a scatterplot to determine which model best depicts trend for Walmart's sales.
- Determine whether or not a seasonal component exists in the series, using the seasonal dummy variable approach.
- Given the conclusions on trend and the seasonal component, provide forecast values for the four quarters of 2011 as well as total projected sales for fiscal year 2011.



**CASE STUDY 18.3** Gary Martin is a research analyst at an investment firm in Chicago. He follows the oil industry and has developed a pretty sophisticated model that forecasts an oil company's stock price. However, given the recent strife in the Middle East, he wonders if simpler causal models might do a better job at predicting stock prices in the near future. He collects data on the daily adjusted closing price of ExxonMobil Corporation (XOM) as well as the Dow Jones Industrial Average (DJIA) for February 2011. A portion of the data is shown in the accompanying table.

**Data for Case Study 18.3** XOM and DJIA Adjusted Closing Prices, February 2011

$t$	DJIA	XOM
February 1	12,040.16	83.47
February 2	12,041.97	82.97
$\vdots$	$\vdots$	$\vdots$
February 28	12,226.34	85.53

SOURCE: www.finance.yahoo.

**FILE**  
XOM

In a report, use the sample information to:

1. Estimate three models: (a)  $XOM_t = \beta_0 + \beta_1 DJIA_{t-1} + \varepsilon_t$ , (b)  $XOM_t = \beta_0 + \beta_1 XOM_{t-1} + \varepsilon_t$ , and (c)  $XOM_t = \beta_0 + \beta_1 DJIA_{t-1} + \beta_2 XOM_{t-1} + \varepsilon_t$ .
2. Determine which model best fits the data.
3. Use the most appropriate model to forecast the daily stock price for March 1, 2011.

## APPENDIX 18.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for *R* can be found on McGraw-Hill's Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### Smoothing Techniques – Moving Average

- A. (Replicating Example 18.1) From the menu choose **Stat > Time Series > Moving Average**.
- B. After **Variable**, select Production, and enter 3 after **MA length**. Select **Generate forecasts**, and after **Number of forecasts**, enter 1. Choose **Storage** and select **Moving averages**, **Fits (one-period-ahead forecasts)**, **Residuals**, and **Forecasts**.

#### Smoothing Techniques – Exponential Smoothing

- A. (Replicating Example 18.2) From the menu choose **Stat > Time Series > Single Exp Smoothing**.
- B. After **Variable**, select Production. Under **Weight to use in smoothing**, select **Use** and enter 0.2. Select **Generate forecasts** and after **Number of forecasts**, enter 1. Choose **Options**, and after **Use average of first K observations K =** enter 1. Choose **Storage** and select **Smoothed Data**, **Fits (one-period-ahead forecasts)**, **Residuals**, and **Forecasts**.

#### Estimating a Lagged Regression Model

- A. (Replicating Example 18.7 – Model 3) In order to create the variable  $Housing_{t-1}$ , from the menu choose **Calc > Calculator**. After **Store result in variable**, enter  $lag(Housing)$ . After **Expression**, enter  $lag('Housing Units Sold', 1)$ . Repeat these steps in order to create  $GDP_{t-1}$ .
- B. Estimate the regression model using the standard commands.

**FILE**  
Gas\_Production

**FILE**  
Gas\_Production

**FILE**  
Housing\_Units



## SPSS

### Estimating a Lagged Regression Model

- A. (Replicating Example 18.7 – Model 3) In order to create the variable  $\text{Housing}_{t-1}$ , from the menu choose **Transform > Compute Variable**. Under **Target Variable**, enter `lagHousing`. Under **Function Group**, select **All**, and under **Functions and Special Variables**, double-click on **Lag(1)**. Under **Numeric Expression**, select `Housing`. Repeat these steps in order to create  $\text{GDP}_{t-1}$ .
- B. Estimate the regression model using the standard commands.

FILE

*Housing\_Units*

## JMP

### Smoothing Techniques – Moving Average

- A. (Replicating Example 18.1) From the menu choose **Analyze > Modeling > Time Series**.
- B. Under **Select Columns**, select `Production`, and then under **Cast Selected Columns Into Roles**, select **Y, Time Series**. Under **Select Columns**, select `Week`, and then under **Cast Selected Columns Into Roles**, select **X, Time ID**.
- C. Click the red triangle next to **Time Series Production**, then select **Smoothing Model > Simple Moving Average**.
- D. Input the value 3 after **Enter smoothing window width**. Deselect **Centered**.
- E. Click the red triangle next to **Simple Moving Average**, then select **Smoothing Model > Simple Moving Average**; then select **Save data to table**.

FILE

*Gas\_Production*

### Smoothing Techniques – Exponential Smoothing

- A. (Replicating Example 18.2) From the menu choose **Analyze > Modeling > Time Series**.
- B. Under **Select Columns**, select `Production`, and then under **Cast Selected Columns Into Roles**, select **Y, Time Series**. Under **Select Columns**, select `Week`, and then under **Cast Selected Columns Into Roles**, select **X, Time ID**.
- C. Click the red triangle next to **Time Series Production**, then select **Smoothing Model > Simple Exponential Smoothing**.
- D. When constructing an exponential smoothing series, JMP uses an algorithm to pick the value of  $w$ . If you want to specify  $w$ , you have to construct the series by defining a formula in a column and using the formula calculator. Here, we allow JMP to pick the value of  $w$ . Click **Estimate**.
- E. Click the red triangle next to **Model: Simple Exponential Smoothing (Zero to One)**, then select **Save Columns**.

FILE

*Gas\_Production*

### Estimating a Lagged Regression Model

- A. (Replicating Example 18.2) In order to create the variable  $\text{Housing}_{t-1}$ , right-click on a new column in the spreadsheet and label it `lag(Housing)`. Right-click on the column of `lag(Housing)`, and select **Formula**. Under **Functions (grouped)**, select **Row > Lag**. Select `Housing` in the **Lag** formula. Repeat these steps in order to create  $\text{GDP}_{t-1}$ .
- B. Estimate the regression model using the standard commands.

FILE

*Housing\_Units*

# 19

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 19.1 Define and compute investment returns.
- LO 19.2 Use the Fisher equation to convert nominal returns into real returns and vice versa.
- LO 19.3 Calculate and interpret a simple price index.
- LO 19.4 Calculate and interpret an unweighted aggregate price index.
- LO 19.5 Compare the Laspeyres and the Paasche methods for computing a weighted aggregate price index.
- LO 19.6 Use price indices to deflate economic time series and derive the inflation rate.

# Returns, Index Numbers, and Inflation

In Chapter 18, we derived seasonal indices to adjust time series for seasonal changes. Policy makers often analyze time series in this format, as they are not particularly interested in its seasonal variations. Other transformations of time series also facilitate interpretation and statistical analysis. For example, financial analysts are interested in the analysis of investment returns. The underlying data may consist of asset prices and income distributions, but these can easily be transformed into investment returns. Similarly, economists are often interested in measuring the magnitude of economic changes over time. They can create index numbers that transform the original data into figures representing percentage changes. Finally, many time series are reported both in nominal as well as real terms. While the nominal values represent dollar amounts, the corresponding real values incorporate inflation to represent the purchasing power of money. In this chapter, we will compute and interpret all such transformed time series.

## INTRODUCTORY CASE

### Analyzing Beer and Wine Price Changes

Jehanne-Marie Roche is the owner of a convenience store in Mt. Angel, a cozy little town in Oregon, nestled between foothills and farmland. Although Jehanne-Marie sells selected grocery and household items, the major source of revenue is from the sale of liquor. However, a significant decline in consumer demand for liquor has occurred, due to the economic crisis that began in the fall of 2008. Jehanne-Marie has been forced to offer numerous price discounts to sell beer and wine at the store. Recently, she asked her nephew to help her understand the price movement of liquor at her store during the 2007–2009 time period. She gives him the average price and quantity information for red wine, white wine, and beer listed in Table 19.1.

**TABLE 19.1** Average Price and Quantity of Wine and Beer

Year		Red Wine	White Wine	6-pack of Beer
2007	Price	\$12.30	\$11.90	\$8.10
	Quantity	1,560	1,410	2,240
2008	Price	\$12.10	\$11.05	\$8.25
	Quantity	1,490	1,390	2,310
2009	Price	\$9.95	\$10.60	\$7.95
	Quantity	1,280	1,010	2,190

Jehanne-Marie wants to use the above information to:

1. Determine the percentage price change of red wine, white wine, and beer from 2007 to 2009.
2. Derive and interpret the aggregate price index of liquor.

A synopsis of this case is provided at the end of Section 19.2.

Define and compute investment returns.

In earlier chapters, the focus of many examples was on the analysis of **investment returns**. Here we describe a simple method to compute them. The time period used for computing an investment return may be a day, a week, a month, a year, or multiple years, and the investment may be in assets such as stocks, bonds, currencies, Treasury bills, or real estate. The investment may be in an individual asset or a portfolio of assets (for example, a mutual fund). An investment return consists of two components. The income component is the direct cash payments from the underlying asset, such as dividends, interest, or rental income. The price change component is the capital gain or loss resulting from an increase or decrease in the value of the asset.

Consider a share of Microsoft Corporation stock that an investor purchased a year ago for \$25. If the price of this share jumps to \$28 in a year, then \$3 (\$28 – \$25) is the annual capital gain from this stock. In percentage terms, it is computed as  $(3/25) \times 100 = 12\%$ . If Microsoft has also paid a dividend of \$1 per share during the year, the income component, in percentage terms, is  $(1/25) \times 100 = 4\%$ . Therefore, the total annual return from investing in Microsoft is 16% ( $12\% + 4\%$ ).

### CALCULATING AN INVESTMENT RETURN

An **investment return**  $R_t$  at the end of time  $t$  is calculated as

$$R_t = \frac{P_t - P_{t-1} + I_t}{P_{t-1}},$$

where  $P_t$  and  $P_{t-1}$  are the price of the asset at times  $t$  (current) and  $t-1$  (prior), respectively, and  $I_t$  is the income distributed during the investment period. The ratios  $\frac{P_t - P_{t-1}}{P_{t-1}}$  and  $\frac{I_t}{P_{t-1}}$  are the **capital gains yield** and the **income yield** components, respectively.

The process for computing an investment return is the same for all assets. The income component is dividends for stocks, interest for bonds, and rental income for a real estate investment. For some assets, like Treasury bills, there is no income component and the investment return consists entirely of a capital gain or loss.

### EXAMPLE 19.1

Helen Watson purchased a corporate bond for \$950 a year ago. She received a coupon payment (interest) of \$60 during the year. The bond is currently selling for \$975. Compute Helen's (a) capital gains yield, (b) income yield, and (c) investment return.

#### SOLUTION:

- We calculate the capital gains yield as  $\frac{P_t - P_{t-1}}{P_{t-1}} = \frac{975 - 950}{950} = 0.0263$  or 2.63%.
- Given the interest payment of \$60, we calculate the income yield as  $\frac{I_t}{P_{t-1}} = \frac{60}{950} = 0.0632$  or 6.32%.
- The investment return is the sum of the capital gains yield and the income yield, that is,  $0.0263 + 0.0632 = 0.0895$  or 8.95%. We can also compute it directly as  $R_t = \frac{P_t - P_{t-1} + I_t}{P_{t-1}} = \frac{975 - 950 + 60}{950} = \frac{85}{950} = 0.0895$  or 8.95%.

### EXAMPLE 19.2

Last year, Jim Hamilton bought a stock for \$35 and recently received a dividend of \$1.25. The stock is now selling for \$31. Find Jim's (a) capital gains yield, (b) income yield, and (c) investment return.

#### SOLUTION:

a. The capital gains yield is  $\frac{P_t - P_{t-1}}{P_{t-1}} = \frac{31 - 35}{35} = -0.1143$ , or  $-11.43\%$ .

b. The income yield is  $\frac{I_t}{P_{t-1}} = \frac{1.25}{35} = 0.0357$  or  $3.57\%$ .

c. The investment return is  $-0.1143 + 0.0357 = -0.0786$ , or  $-7.86\%$ .

Equivalently, we can compute the investment return as

$$R_t = \frac{P_t - P_{t-1} + I_t}{P_{t-1}} = \frac{31 - 35 + 1.25}{35} = \frac{-2.75}{35} = -0.0786, \text{ or } -7.86\%.$$

Note that the investment return is unaffected by the decision to sell or hold assets. A common misconception is that if you do not sell an asset, there is no capital gain or loss involved, as a given price increase or decrease leads only to paper gain or loss. This misconception often leads an investor to hold a "loser" asset longer than necessary because of the reluctance to admit a bad investment decision. The nonrecognition of the loss is relevant for tax purposes, since only realized income must be reported in tax returns. However, whether or not you have liquidated an asset is irrelevant when measuring its performance.

## The Adjusted Closing Price

Historical returns are often used by investors, analysts, and other researchers to assess past performance of a stock. In Example 19.2, we saw that dividend payments also influence stock returns. Therefore, we need the dividend data along with the price data to compute historical returns. Similarly, we need information on stock splits and reverse stock splits in computing returns. Tabulating corporate decisions such as the announcement of dividends, stock splits, and reverse stock splits can be very cumbersome. For these reasons, most data sources for stock price information, such as <http://finance.yahoo.com>, also include data on the **adjusted closing price**. Here, price data are adjusted using appropriate dividend and split multipliers; we recommend an introductory finance book for further details.

Given that the adjustment has been made for all applicable splits and dividend distributions, we can compute the total investment return on the basis of the price appreciation or depreciation of the adjusted closing prices.

#### USING ADJUSTED CLOSING PRICES TO CALCULATE AN INVESTMENT RETURN

Let  $P_t^*$  and  $P_{t-1}^*$  represent the **adjusted closing price** of a stock at times  $t$  (current) and  $t-1$  (prior), respectively. The investment return  $R_t$  at the end of time  $t$  is calculated as

$$R_t = \frac{P_t^* - P_{t-1}^*}{P_{t-1}^*}.$$

### EXAMPLE 19.3

Consider the adjusted closing stock prices of Microsoft Corporation in Table 19.2. Find the monthly returns for November and December for 2010.

**TABLE 19.2** Monthly Stock Prices for Microsoft Corporation

Date	Adjusted Closing Price
December 2010	\$26.04
November 2010	\$25.11
October 2010	\$26.35

SOURCE: Data obtained from finance.yahoo.com on December 2010.

**SOLUTION:** We compute the monthly return for November 2010 as  $R_t = \frac{25.11 - 26.35}{26.35} = -0.0471$ , or  $-4.71\%$ . Similarly, the monthly return for December 2010 is  $R_t = \frac{26.04 - 25.11}{25.11} = 0.0370$ , or  $3.70\%$ .

### LO 19.2

Use the Fisher equation to convert nominal returns into real returns and vice versa.

## Nominal versus Real Rates of Return

So far we have focused on **nominal returns**, which make no allowance for inflation. Financial rates such as interest rates, discount rates, and rates of return are generally reported in nominal terms. However, the nominal return does not represent a true picture because it does not capture the erosion of the purchasing power of money due to inflation. Consider, for example, an investment of \$100 that becomes \$105 after one year. While the nominal return on this investment is 5%, the purchasing power of the money is likely to have increased by less than 5%. Once the effects of inflation have been factored in, investors can determine the real, or true, return on their investment. In sum, the **real return** captures the change in the purchasing power, whereas the nominal return simply reflects the change in the number of dollars.

The relationship between the nominal and the real return was developed by Irving Fisher (1867–1947), a prominent economist. The **Fisher equation** is a theoretical relationship between nominal returns, real returns, and the expected inflation rate.

### THE FISHER EQUATION

Let  $R$  be the nominal rate of return,  $r$  the real rate of return, and  $i$  the expected inflation rate. The **Fisher equation** is defined as

$$1 + r = \frac{1 + R}{1 + i}.$$

When the expected inflation rate is relatively low, a reasonable approximation to the Fisher equation is  $r = R - i$ ; we will not be using this approximation in this chapter.

### EXAMPLE 19.4

The quoted rate of return on a one-year U.S. Treasury bill in January 2010 is 0.45% (www.ustreas.gov). Compute and interpret the real rate of return that investors can earn if the inflation rate is expected to be 1.6%.

**SOLUTION:** Using the Fisher equation,  $1 + r = \frac{1 + R}{1 + i} = \frac{1.0045}{1.0160} = 0.9887$ ; we derive the real rate of return as  $r = 0.9887 - 1 = -0.0113$ , or  $-1.13\%$ . The negative real rate of return implies that investors were extremely cautious and were even willing to accept a small drop in their purchasing power during the financial crisis period.



## EXAMPLE 19.5

A bond produces a real rate of return of 5.30% for a time period when the inflation rate is expected to be 3%. What is the nominal rate of return on the bond?

**SOLUTION:** The Fisher equation can be rewritten as  $1 + R = (1 + r)(1 + i)$ . Therefore, given the real rate of return of 5.30% and the inflation rate of 3%, we can easily compute,  $1 + R = (1.053)(1.03) = 1.0846$  giving us the nominal return of  $R = 1.0846 - 1 = 0.0846$ , or 8.46%.

## EXERCISES 19.1

1. You borrowed \$2,000 to take a vacation in the Caribbean islands. At the end of the year, you had to pay back \$2,200. What is the annual interest that you paid on your loan?
2. You bought a corporate bond last year for \$980. You received a coupon payment (interest) of \$60 and the bond is currently selling for \$990. What is the (a) income yield, (b) capital gains yield, and (c) total return on the investment?
3. The price of a stock has gone up from \$24 to \$35 in one year. It also paid a year-end dividend of \$1.20. What is the stock's (a) income yield, (b) capital gains yield, and (c) total return?
4. The year-end price and dividend information on a stock is given in the following table.

Year	Price	Dividend
1	\$23.50	NA
2	24.80	\$0.18
3	22.90	0.12

Note: NA denotes not applicable.

- a. What is the nominal return of the stock in years 2 and 3?
  - b. What is the corresponding real return if the inflation rates for years 2 and 3 were 2.8% and 1.6%, respectively?
5. A portfolio manager invested \$1,500,000 in bonds in 2007. In one year the market value of the bonds dropped to \$1,485,000. The interest payments during the year totaled \$105,000.
    - a. What was the manager's total rate of return for the year?
    - b. What was the manager's real rate of return if the inflation rate during the year was 2.3%?
  6. Bill Anderson purchased 1,000 shares of Microsoft Corporation stock for \$17,100 at the beginning of 2009. At the end of the year, he sold all of his Microsoft shares at \$30.48 a share. He also earned a dividend of \$0.52 per share during the year.

- a. What is Bill's total return on the investment?
  - b. What is the dollar gain from the investment?
7. You would like to invest \$20,000 for a year in a risk-free investment. A conventional certificate of deposit (CD) offers a 4.6% annual rate of return. You are also considering an "Inflation-Plus" CD which offers a real rate of return of 2.2% regardless of the inflation rate.
    - a. What is the implied (expected) inflation rate?
    - b. You decide to invest \$10,000 in the conventional CD and \$10,000 in the "Inflation-Plus" CD. What is your expected dollar value at the end of the year?
    - c. Which of the two CDs is a better investment if the actual inflation rate for the year turns out to be 2.2%?
  8. Consider the following adjusted closing stock prices for Intel Corporation. Find the monthly returns for November and December of 2010.

Date	Adjusted Closing Price
December 2010	\$21.48
November 2010	\$20.98
October 2010	\$19.73

SOURCE: finance.yahoo.com.

9. Consider the following adjusted closing stock prices for Johnson and Johnson (J&J) and Caterpillar, Inc. Compute and compare the monthly returns for both companies.

Date	J&J	Caterpillar
March 2011	60.70	99.86
February 2011	61.44	102.93
January 2011	59.23	97.01
December 2010	61.30	93.22
November 2010	61.00	84.20
October 2010	62.63	78.23

SOURCE: finance.yahoo.com.

## 19.2 INDEX NUMBERS

An **index number** is an easy-to-interpret numerical value that reflects a percentage change in price or quantity from a base value. In this chapter, we focus on price indices. The base value for a price index is set equal to 100 for the selected base period, and values in other periods are adjusted in proportion to the base. Thus, if the price index for a given year is 125, it implies that the price has increased by 25% from the base year. Similarly, a price index of 90 implies that the price has decreased by 10% from the base year. By working in a manner similar to percentages, index numbers make changes over time easier to compare. Index numbers enable policy makers and analysts to focus on the movements in variables rather than on their raw absolute values.

### LO 19.3

Calculate and interpret a simple price index.

### Simple Price Indices

Consider the price of a hamburger that increases from \$3.25 in 1995 to \$4.75 in 2010. We can easily determine that the price of a hamburger has increased by  $\frac{4.75 - 3.25}{3.25} = 0.46$ , or 46%. Alternatively, if we use 1995 as the base year with an index value of 100, then the corresponding index value for 2010 is 146, implying a 46% increase in price. This is an example of a **simple price index**.

#### A SIMPLE PRICE INDEX

A **simple price index** for any item is the ratio of the price in period  $t$ ,  $p_t$ , and the price in the base period,  $p_0$ , expressed as a percentage. It is calculated as  $\frac{p_t}{p_0} \times 100$ .

### EXAMPLE 19.6

Consider the data presented in the introductory case of this chapter in Table 19.1. Use the base year of 2007 to compute and interpret the 2008 and 2009 simple price indices for:

- a. Red wine
- b. White wine
- c. A 6-pack of beer

**SOLUTION:** Since 2007 is the base year, we set the corresponding index value equal to 100. The index values for other years are computed below.

- a. For red wine, the simple price index for 2008 is

$$\frac{\text{Price in 2008}}{\text{Price in 2007}} \times 100 = \frac{12.10}{12.30} \times 100 = 98.37.$$

Similarly, for 2009, it is

$$\frac{\text{Price in 2009}}{\text{Price in 2007}} \times 100 = \frac{9.95}{12.30} \times 100 = 80.89.$$

Therefore, the average price of red wine in 2008 and 2009 was 98.37% and 80.89%, respectively, of what it was in 2007. In other words, as compared to 2007, the price of red wine dropped by 1.63% in 2008 and 19.11% in 2009.

- b. For white wine, the simple price index for 2008 is  $(11.05/11.90) \times 100 = 92.86$ , and  $(10.60/11.90) \times 100 = 89.08$  for 2009. Therefore, relative to 2007, the average price of white wine dropped by 7.14% in 2008 and 10.92% in 2009.
- c. The simple price index for a six-pack of beer for 2008 is  $(8.25/8.10) \times 100 = 101.85$ , and  $(7.95/8.10) \times 100 = 98.15$  for 2009. Interestingly, while the prices of both red and white wines experienced substantial declines, the price of beer stayed fairly stable. Relative to the base year of 2007, there was a 1.85% increase in the price of beer in 2008 and a 1.85% decline in 2009.

## EXAMPLE 19.7

Table 19.3 shows the average price and corresponding price index for gasoline from 2000 to 2008. Interpret the price indices for 2001 and 2008.

**TABLE 19.3** Price and Corresponding Price Index for Unleaded Gasoline in U.S., Base Year 2000

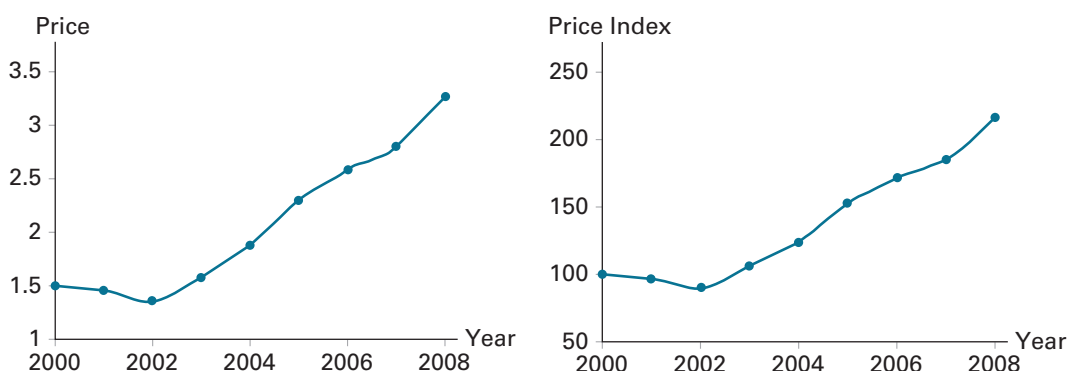
Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Price	1.51	1.46	1.36	1.59	1.88	2.30	2.59	2.80	3.27
Price Index (Base = 2000)	100	96.69	90.07	105.30	124.50	152.32	171.52	185.43	216.56

SOURCE: Bureau of Labor Statistics.

**SOLUTION:** Since 2000 is treated as the base year, the index number for 2000 is 100. The index number for 2001 is calculated as  $(1.46/1.51) \times 100 = 96.69$ . Thus, the gasoline price in 2001 was 96.69% of what it was in 2000, or 3.31% lower. Given a price index of 216.56 in 2008, the gasoline price in 2008 was 116.56% higher relative to 2000.

In Figure 19.1, we plot the raw price and price indices for gasoline from 2000 to 2008. Note that although the units of the gasoline price and index number graphs are different, the basic shape of the two graphs is similar. This shows that the main purpose of index numbers is to provide an easy interpretation of the changes of the series over time.

**FIGURE 19.1** Price of gasoline and the corresponding index numbers for 2000–2008



It is important to note that index numbers provide direct comparisons only with respect to the base year. Similar direct comparisons cannot be made between non-base years. For instance, based on the index numbers for 2005 and 2008, we cannot say that prices rose by 64.24% ( $216.56\% - 152.32\%$ ) from 2005 to 2008. The actual percentage change from 2005 and 2008 is  $\frac{216.56 - 152.32}{152.32} \times 100 = 42.17$ , implying that prices rose by 42.17% from 2005 to 2008.

Alternatively, we can use index numbers directly to compare prices between 2005 and 2008 by making 2005 the base year. It may be more meaningful to compare 2008 values with those in 2005 rather than the values in 2000. In fact, federal agencies routinely update the base year used in their calculations of statistical indices. For example, the reported imports and exports price indices in 2008 have been updated from the base year of 1995 to a revised base year of 2000.

It is fairly simple to revise the base period of an index. We basically transform the index of the newly chosen base period as 100 and values in other periods are adjusted by the same proportion.

### REVISING THE BASE PERIOD

A simple index can easily be updated with a revised base period as

$$\text{Updated Index} = \frac{\text{Old Index Value}}{\text{Old Index Value of New Base}} \times 100.$$

### EXAMPLE 19.8

Update the index numbers in Table 19.3 with a base year revised from 2000 to 2005.

**SOLUTION:** With a revised base of 2005, the index number for 2005 is updated from 152.32 to 100. Other indices are adjusted according to the revision rule. For instance, the index number for 2006 is updated as  $(171.52/152.32) \times 100 = 112.61$ . Table 19.4 contains index numbers that have been similarly updated.

**TABLE 19.4** Price Index for Gasoline Using Base Year of 2000 and 2005

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Price	1.51	1.46	1.36	1.59	1.88	2.30	2.59	2.80	3.27
Price Index (Base = 2000)	100	96.69	90.07	105.30	124.50	152.32	171.52	185.43	216.56
Price Index (Base = 2005)	65.65	63.48	59.13	69.13	81.74	100.00	112.61	121.74	142.17

With the revised base of 2005, we can directly deduce that the gasoline price in 2008 was 142.17% of what it was in 2005, or 42.17% higher.

### LO 19.4

Calculate and interpret an unweighted aggregate price index.

## Unweighted Aggregate Price Index

An **aggregate price index** is used to represent relative price movements for a group of items. Examples include the closely watched consumer price index (CPI) and the producer price index (PPI). An aggregate price index can be weighted or unweighted. An **unweighted aggregate price index** is based entirely on aggregate prices with no emphasis placed on quantity. In other words, it does not incorporate the information that consumers may not be consuming equal quantities over the years of the items comprising the index. Weighted methods, on the other hand, use quantity as weights in the calculations.

### CALCULATING AN UNWEIGHTED AGGREGATE PRICE INDEX

Let  $p_{it}$  represent the price of item  $i$  in period  $t$  and let  $p_{i0}$  be the corresponding price in the base period ( $t = 0$ ). The **unweighted aggregate price index** in period  $t$  is  $\frac{\sum p_{it}}{\sum p_{i0}} \times 100$ .

### EXAMPLE 19.9

A real estate firm based in Florida collects data on the average selling price of condominiums, single-family homes, and multifamily homes that it sold over the last three years. Table 19.5 shows the results. Compute the unweighted price index for the properties, using 2007 as the base year.

**TABLE 19.5** Average Price (in \$1,000s) of Properties Sold in Florida (Example 19.9)

Year	Condominiums	Single-family	Multifamily
2007	225	375	440
2008	148	250	390
2009	130	235	400

**SOLUTION:** In order to find the unweighted aggregate price index, we first aggregate prices for each year by adding up the prices of condominiums, single-family homes, and multifamily homes. For 2007, the aggregate price is computed as  $\Sigma p_{0i} = 225 + 375 + 440 = 1040$ . Similarly, the aggregate prices are  $\Sigma p_{1i} = 148 + 250 + 390 = 788$  for 2008 and  $\Sigma p_{2i} = 130 + 235 + 400 = 765$  for 2009. Then, using 2007 as the base year, the unweighted aggregate price indices are computed as

$$\text{Price Index for 2008} = \frac{788}{1040} \times 100 = 75.77, \text{ and}$$

$$\text{Price Index for 2009} = \frac{765}{1040} \times 100 = 73.56.$$

Thus, according to the unweighted aggregate price index, property values in 2008 were 75.77% of what they were in 2007, or equivalently, they were 24.23% lower. Similarly, relative to 2007, property values in 2009 were 73.56% lower. Although the unweighted aggregate price index captures the overall drop in property values in Florida, the drop seems slightly lower than what has been reported in the popular press. A possible explanation is that the unweighted index unfairly treats all property prices equally. The drop in property values would be greater if we take into account the fact that most properties in Florida consisted of condominiums and single-family homes, which witnessed a steeper price decline than multifamily homes.

## Weighted Aggregate Price Index

A **weighted aggregate price index** does not treat prices of different items equally. A higher weight is given to the items that are sold in higher quantities. However, there is no unique way to determine the weights, as they depend on the period in which the quantities are evaluated. One option is to evaluate the changing quantities over the years to derive the weighted average. However, in many applications, the quantity information is not readily available and we have to rely on its evaluation in a single time period. Two popular choices for weights are based on the quantities evaluated in the base period and in the current period. A **Laspeyres price index** uses the quantities evaluated in the base period to compute a weighted aggregate price index.

### LO 19.5

Compare the Laspeyres and the Paasche methods for computing a weighted aggregate price index.

#### CALCULATING A WEIGHTED AGGREGATE PRICE INDEX: THE LASPEYRES PRICE INDEX

Let  $p_{it}$  and  $q_{it}$  represent the price and quantity of item  $i$  in period  $t$  and let  $p_{i0}$  and  $q_{i0}$  be the corresponding values in the base period ( $t = 0$ ). Using only the base period quantities  $q_{i0}$ , the **Laspeyres price index** for period  $t$  is

$$\frac{\Sigma p_{it} q_{i0}}{\Sigma p_{i0} q_{i0}} \times 100.$$

### EXAMPLE 19.10

Table 19.6 shows the number of condominiums, single-family homes, and multi-family homes sold in Florida. Use these quantities, along with the price information from Table 19.5, to compute the Laspeyres price index for real estate, given a base year of 2007.

**TABLE 19.6** Number of Properties Sold in Florida

Year	Condominiums	Single-Family	Multifamily
2007	42	104	20
2008	28	76	16
2009	32	82	10

**SOLUTION:** Since the Laspeyres price index evaluates the quantities in the base period, we will only use the number of properties sold in 2007 in the calculation. Table 19.7 aids in the calculation of the Laspeyres index.

**TABLE 19.7** Calculations for Example 19.10

Year	Weighted Price = $\sum p_{it} q_{i0}$	The Laspeyres Index
2007	$225 \times 42 + 375 \times 104 + 440 \times 20 = 57250$	100
2008	$148 \times 42 + 250 \times 104 + 390 \times 20 = 40016$	$(40016/57250) \times 100 = 69.90$
2009	$130 \times 42 + 235 \times 104 + 400 \times 20 = 37900$	$(37900/57250) \times 100 = 66.20$

Based on the Laspeyres index, the real estate prices in 2008 were 69.90% of what they were in 2007, or equivalently they were 30.10% lower. Similarly, the real estate prices in 2009 were 66.20% of what they were in 2007, or equivalently they were 33.80% lower. Note that the computed drop in property values based on the Laspeyres price index is sharper than the one inferred from the unweighted aggregate price index.

As mentioned earlier, the choice of weights for a weighted aggregate price index depends on the quantity evaluated in a given period. Whereas a Laspeyres index uses the base period quantities as weights, a **Paasche index** uses the current period quantities in deriving the weights. Since the choice of weights for the two methods are different, the Laspeyres and Paasche indices differ for the period under evaluation.

#### CALCULATING A WEIGHTED AGGREGATE PRICE INDEX: THE PAASCHE PRICE INDEX

Let  $p_{it}$  and  $q_{it}$  represent the price and quantity of item  $i$  in period  $t$  and let  $p_{i0}$  and  $q_{i0}$  be the corresponding values in the base period ( $t = 0$ ). Using only the current period ( $t = n$ ) quantities  $q_{in}$ , the **Paasche price index** for period  $t$  is

$$\frac{\sum p_{it} q_{in}}{\sum p_{i0} q_{in}} \times 100.$$

#### EXAMPLE 19.11

Consider Tables 19.5 and 19.6, representing the price and quantity data for properties sold in Florida. Use this information to compute the Paasche price index for real estate, given a base year of 2007.

**SOLUTION:** Since the Paasche price index uses the quantities evaluated in the current period, we use only the numbers of properties sold in 2009 in the calculations. Table 19.8 aids in the calculation of the Paasche index.

**TABLE 19.8** Calculations for Example 19.11

Year	Weighted Price = $\sum p_{it} q_{in}$	The Paasche Index
2007	$225 \times 32 + 375 \times 82 + 440 \times 10 = 42350$	100
2008	$148 \times 32 + 250 \times 82 + 390 \times 10 = 29136$	$(29136/42350) \times 100 = 68.80$
2009	$130 \times 32 + 235 \times 82 + 400 \times 10 = 27430$	$(27430/42350) \times 100 = 64.77$



The Paasche index is calculated as 68.80 for 2008 and 64.77 for 2009. Therefore, according to the Paasche index with a base year of 2007, property values dropped by 31.20% in 2008 and 35.23% in 2009.

In general, the Laspeyres and Paasche indices provide similar results if the periods being compared are not too far apart. The two indices tend to differ when the length of time between the periods increases since the relative quantities of items (weights) adjust to the changes in consumer demand over time. Consumers tend to adjust their consumption patterns by decreasing (increasing) the quantity of items that undergo a larger relative price increase (decrease). For instance, a sharp increase in the price of an item is typically accompanied by a decrease in the quantity demanded, making its relative weight go down in value. Similarly, a sharp decrease in the price of an item will make its relative weight go up. Therefore, a Paasche index that uses the updated weights theoretically produces a lower estimate than a Laspeyres index when prices are increasing and a higher estimate when prices are decreasing. Our results regarding property values are consistent with this reasoning. During the period of price decline, the Laspeyres index suggests that relative to 2007, property values have dropped by 30.10% and 33.80% in 2008 and 2009, respectively. According to the Paasche index for the same period, property values had larger drops of 31.20% and 35.23%, respectively.

The Paasche index is attractive because it incorporates current expenditure patterns. However, its data requirements are more stringent than those of the Laspeyres index. The Paasche index requires that the weights be updated each year and the index numbers be recomputed for all of the previous years. The additional cost required to process current expenditure data, needed to revise the weights, can be substantial. It may not always be possible to produce a timely Paasche index. Therefore, the Laspeyres index is a more widely used weighted aggregate price index. The base period is changed periodically to ensure that the Laspeyres index does not become outdated. Here the base period revision involves updated calculations using quantity weights of the new base period.

### EXAMPLE 19.12

Let us revisit the introductory case with the data presented in Table 19.1. Using 2007 as the base year, compute and interpret the weighted aggregate price indices for liquor using

- a. The Laspeyres method
- b. The Paasche method

**SOLUTION:** Since 2007 is used as the base year, its value for both indices is set equal to 100.

- a. For the Laspeyres price index, the prices are weighted by the quantities evaluated in the base period of 2007. Therefore, the weighted price for 2007 is computed as

$$\sum p_{it} q_{i0} = 12.30 \times 1560 + 11.90 \times 1410 + 8.10 \times 2240 = \$54,111.$$

Similarly, the weighted price equals

$$12.10 \times 1560 + 11.05 \times 1410 + 8.25 \times 2240 = \$52,936.5 \text{ for 2008, and} \\ 9.95 \times 1560 + 10.60 \times 1410 + 7.95 \times 2240 = \$48,276 \text{ for 2009.}$$

The corresponding price index is  $(52936.5/54111) \times 100 = 97.83$  for 2008 and  $(48276/54111) \times 100 = 89.22$  for 2009. Therefore, based on the Laspeyres index, liquor prices are 97.83% in 2008 and 89.22% in 2009 of what they were in 2007. In other words, overall liquor prices dropped by 2.17% in 2008 and by 10.78% in 2009.

- b. For the Paasche price index, the prices are weighted by the quantities evaluated in the current period, which in our example is 2009. Therefore, the weighted price for 2007 is computed as

$$\Sigma p_{it} q_{in} = 12.30 \times 1280 + 11.90 \times 1010 + 8.10 \times 2190 = \$45,502.$$

Similarly, the weighted prices equal

$$12.10 \times 1280 + 11.05 \times 1010 + 8.25 \times 2190 = \$44,716 \text{ for 2008, and}$$

$$9.95 \times 1280 + 10.60 \times 1010 + 7.95 \times 2190 = \$40,852.5 \text{ for 2009.}$$

The corresponding price index is  $(44716/45502) \times 100 = 98.27$  for 2008 and  $(40852.5/45502) \times 100 = 89.78$  for 2009. Therefore, based on the Paasche index, liquor prices are 98.27% in 2008 and 89.78% in 2009 of what they were in 2007. In other words, relative to 2007, overall liquor prices dropped by 1.73% in 2008 and by 10.22% in 2009.

## SYNOPSIS OF INTRODUCTORY CASE



The global financial crisis that began in 2008 has had major consequences in all aspects of the American economy. The staggering number of layoffs highlights the effects of the financial crisis being passed on to the real economy. Jehanne-Marie, the owner of a small convenience store in Oregon, has not been spared the effects of the crisis. She has been forced to offer numerous price discounts to counter the plummeting demand for liquor. Interestingly, the cutbacks by consumers have not been uniform across red wine, white wine, and beer. While the price of red wine has dropped by 19.11% from 2007 to 2009, the corresponding drop in price has been 10.92% for white wine and only 1.85% for beer. In order to capture the overall price movement of liquor, two weighted aggregate price indices are also computed. These indices devote a higher weight to the price of items that are sold in higher quantities. The weights are defined by the base period quantities for the Laspeyres index and the current period quantities for the Paasche index. Both indices suggest that relative to 2007, Jehanne-Marie has experienced an overall price decline of about 2% in 2008 and a larger 10.50% in 2009. In sum, Jehanne-Marie is advised to focus more on beer sales, rather than wine, during harsh economic times. A comprehensive analysis that includes other grocery items like bread, cheese, and soda would better describe the full impact of the economic crisis on her total sales.

## EXERCISES 19.2

### Mechanics

10. Consider the following price data from 1994 to 2002.

Year	1994	1995	1996	1997	1998	1999	2000	2001	2002
Price	62	60	64	67	66	70	74	72	70

- a. Compute the simple price index using 1994 as the base year.

- b. Determine the percentage change in prices from 1994 to 1998.

11. Consider the following simple price index created with a base year of 2004.

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012
Price Index	100	102.2	106.3	110.8	109.4	107.2	108.9	110.5	114.7

- Update the index numbers using a revised base year of 2008.
- Determine the percentage change in price from 2004 to 2012.
- Determine the percentage change in price from 2008 to 2012.

12. Consider the following price and quantity data for three products from 2008 to 2010.

Year		Product 1	Product 2	Product 3
2008	Price	\$14.30	\$13.90	\$18.10
	Quantity	992	1110	800
2009	Price	\$14.90	\$13.70	\$18.50
	Quantity	980	1220	790
2010	Price	\$15.50	\$13.80	\$17.90
	Quantity	140	1290	810

- Compute the simple price index for each product, using 2008 as the base year.
  - Compare the relative price movements of the three products.
13. Use the price and quantity information in Exercise 12 to compute the following aggregate price indices, given a base year of 2008:
- The unweighted aggregate price index
  - The Laspeyres price index
  - The Paasche price index

## Applications

14. Consider the following average monthly prices for regular gasoline in California in 2008.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Price	3.25	3.18	3.56	3.82	3.97	4.48	4.46	4.16	3.79	3.39	2.46	1.82

SOURCE: [www.energyalmanac.ca.gov](http://www.energyalmanac.ca.gov).

- Construct a simple price index with January 2008 as the base.
  - Determine the percentage change in the average gasoline price in California from January to June.
15. The following table shows the monthly adjusted closing price per share of Microsoft Corporation for 2009.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Price	16.6	15.8	18.0	19.8	20.6	23.4	23.2	24.4	25.5	27.5	29.2	30.3

SOURCE: <http://finance.yahoo.com>.

- Construct a simple price index with January 2009 as the base.
- What is the percentage price change in July relative to January?
- What is the percentage price change in December relative to January?

16. The MIT Sloan School of Management is one of the leading business schools in the U.S. The following table contains the tuition data for the masters program in the Sloan School of Management.

Year	2004	2005	2006	2007	2008	2009
Tuition	\$36,850	\$39,844	\$42,634	\$44,556	\$46,784	\$48,650

SOURCE: <http://web.mit.edu/ir/financial/tuition.html>.

- Use 2004 as the base year to form a simple price index for tuition.
  - Use 2007 as the base year to form a simple price index for tuition.
  - Compare the percentage tuition increase from 2004 through 2007 and 2007 through 2009.
17. **FILE Returns 2009.** According to dollar cost averaging, a fixed amount of money is invested periodically in a portfolio. Consequently, more units of a financial asset are purchased when prices are low and fewer units are purchased when prices are high. Robert Dudek follows dollar cost averaging by making a monthly investment of \$500 toward retirement. His monthly investment is divided equally among two T. Rowe Price mutual funds: Equity Income and Short-term Bond funds. The following table represents the monthly adjusted closing price of the funds in 2009.

Month	EqInc	Bond	Month	EqInc	Bond
January	14.77	4.47	July	18.40	4.72
February	12.93	4.49	August	19.45	4.75
March	14.14	4.53	September	19.92	4.78
April	16.04	4.58	October	19.53	4.80
May	16.87	4.64	November	20.60	4.85
June	16.89	4.66	December	20.99	4.82

SOURCE: <http://finance.yahoo.com>.

- Compute and interpret the Laspeyres price index.
  - Compute and interpret the Paasche price index.
  - Why are the results of the two indices different?
18. JJ Diner is a small mom and pop restaurant in Lincoln, Nebraska. They offer three choices for breakfast: omelets, pancakes, or cereal. The average prices (in dollars) for these options for 2007, 2008, and 2009 are shown in the accompanying table.

Year	Omelet	Pancakes	Cereal
2007	4.75	3.50	3.50
2008	5.25	4.25	4.00
2009	5.00	4.50	4.25

- Compute and interpret the simple price index for each breakfast, using 2007 as the base year.
- Compute and interpret the unweighted aggregate price index for breakfast, using 2007 as the base year.

19. The following table shows the number (in 1,000s) of breakfasts sold at JJ Diner.

Year	Omelet	Pancakes	Cereal
2007	9.26	7.98	2.44
2008	11.82	9.20	2.62
2009	10.48	8.50	2.12

Use this information, along with the price data provided in Exercise 18, to

- Compute and interpret the Laspeyres price index.
  - Compute and interpret the Paasche price index.
20. With the collapse of house prices that started in 2006, the American Dream has become a nightmare for many of the 75 million Americans who own a home (*CBS Evening News*, February 2, 2010). However, the drop in house prices has not been uniform across the country. The accompanying table represents median home prices (in \$1,000s) by region for 2007, 2008, and 2009.

Region	2007	2008	2009
Northeast	288.1	271.5	240.7
Midwest	161.4	150.5	142.5
South	178.8	169.4	154.6
West	342.5	276.1	224.2

SOURCE: [www.realtor.org](http://www.realtor.org).

- Use 2007 as the base year to construct a simple price index for each region.
- Use the percentage decline in home values to discuss regional differences in price drops.

21. Consider the following table, which reports the sale (quantity in 1,000s) of homes by region for 2007, 2008, and 2009.

Region	2007	2008	2009
Northeast	1006	849	868
Midwest	1327	1129	1165
South	2235	1865	1913
West	1084	1070	1210

SOURCE: [www.realtor.org](http://www.realtor.org).

Use this information, along with the price data provided in Exercise 20, to

- Compute and interpret the Laspeyres aggregate home price index for the U.S.
- Compute and interpret the Paasche aggregate home price index for the U.S.
- Comment on the differences between the two indices.

## 19.3 USING PRICE INDICES TO DEFLATE A TIME SERIES

### LO 19.6

Use price indices to deflate economic time series and derive the inflation rate.

Most business and economic time series are generally reported in nominal terms, implying that they are measured in dollar amounts. Since inflation erodes the value of money over time, the dollar differences over time do not quite tell the whole story. For instance, we cannot directly compare the starting salary of a recent college graduate with that of a college graduate five years ago. Due to price increases, the purchasing power of recent graduates may be lower even if they make more money than their predecessors. Similarly, a hardware store may have doubled its revenue over 20 years, but the true increase in value may be much smaller once it has been adjusted for inflation.

An important function of the price indices, mentioned in the previous section, is to serve as deflators. A **deflated** series is obtained by adjusting the given time series for changes in prices, or inflation. We use the price indices to remove the effect of inflation so that we can evaluate business and economic time series in a more meaningful way.

### NOMINAL VERSUS REAL VALUES

A time series that has been deflated is said to be represented in **real terms**. The unadjusted time series is said to be represented in **nominal terms**. We use a price index to convert the nominal value of a time series into its real value as

$$\text{Real Value} = \frac{\text{Nominal Value}}{\text{Price Index}} \times 100.$$

Consider the following example. Lisa Redford has worked in a small marketing firm in Florida for the last three years. Due to the recent economic crisis, her salary has dropped from \$80,000 in 2007 to \$64,000 in 2009. While her salary has dropped by 20%, a larger drop in property values for the same time period may have made it easier for Lisa to own a home in Florida. In Example 19.10, we used the base year of 2007 to derive the Laspeyres price index of property values for 2009 as 66.20, implying that real estate prices were 33.80% lower in 2009 than in 2007. It is more meaningful to compare Lisa's salary of \$80,000 in 2007 (the base year) with the price-adjusted (real) salary of  $(\$64,000/66.20) \times 100 = \$96,677$  in 2009. Using the Laspeyres price index of property values for adjustment, Lisa is actually slightly better off in 2009 than she was in 2007, despite the salary cut. However, it is not reasonable to adjust Lisa's salary solely on the basis of the price index of property values in Florida. Since her expenditure is not limited to mortgage payments, a more comprehensive price index is needed to make the price adjustment to the salary. In fact, when we say that a series has been deflated, we imply that the series has been adjusted on the basis of the price of a comprehensive basket of goods and services.

The two most commonly used price indices used to deflate economic time series are the **Consumer Price Index, CPI**, and the **Producer Price Index, PPI**. While both the CPI and PPI measure the percentage price change over time for a fixed basket of goods and services, they differ in the composition of the basket and in the types of prices used in the analysis. The general process of computing the CPI and PPI is similar to the method outlined in the preceding section. However, we will not elaborate on their composition in this chapter.

The CPI is perhaps the best-known weighted aggregate price index. The U.S. Bureau of Labor Statistics computes a monthly CPI based on the prices paid by urban consumers for a representative basket of goods and services. As of 2010, the CPI uses 1982 as the base year. The prices of several hundred consumption items are included in the index. In addition, randomly selected consumers help determine the expenditure for the representative basket of goods and services. The corresponding quantities of items in the base year are used for computing the weights for the index.

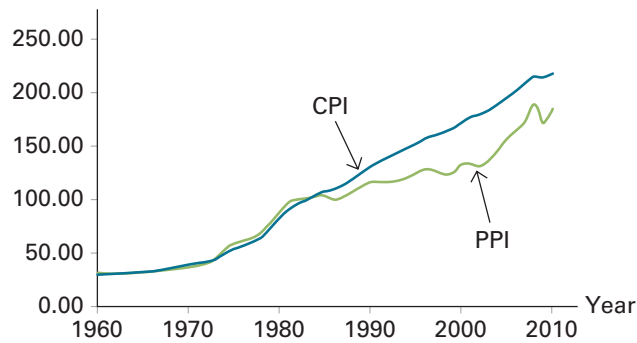
The PPI is a weighted aggregate price index of prices measured at the wholesale, or producer level. Prior to 1978, the PPI was called the Wholesale Price Index, WPI. The U.S. Bureau of Labor Statistics computes a monthly PPI based on the selling prices received by domestic producers for their entire marketed output. The target set includes purchases of goods and services by consumers—directly from the producer or indirectly from a retailer—and by other producers as inputs to their production or as capital investment.

Note that the CPI is based on out-of-pocket expenditures of an urban consumer and the PPI is based on the portion that is actually received by the producer. Therefore, although sales and excise taxes are included in the CPI, they are not included in the PPI because they do not represent revenue to the producer. The differences between the PPI and CPI are consistent with the way these indices are used for deflation. It is common to use the CPI to adjust wages for changes in the cost of living. The PPI, on the other hand, is useful to deflate revenue in order to obtain real growth in output.

It is often assumed that the direction and magnitude of a price change in the PPI for finished goods anticipates a similar change in the CPI for all items. This is not always the case. In Figure 19.2, we use the **CPI\_PPI** data to plot the annual CPI and PPI from 1960–2010, with a base of 1982–1984. Interestingly, the two indices moved in sync until the early 1980s. Beyond that, changes in prices that consumers paid far exceeded those received by producers, with the difference peaking in 2002. Also, noteworthy is the fact that while there was a significant dip in the PPI, the CPI showed a very slight decline during the peak of the financial crisis in 2009.



**FIGURE 19.2**  
CPI and PPI for  
1960–2010; base period  
1982–1984



Source: Bureau of Labor Statistics.

### EXAMPLE 19.13

Tom Denio has been a project manager in a small construction firm in Atlanta since 2000. He started with a salary of \$52,000, which grew to \$84,000 in 2008. The revenue of the construction firm also grew over the years, increasing from \$13 million in 2000 to \$18 million in 2008. According to the Bureau of Labor Statistics, the values of the consumer price index with a base of 1982–1984 for 2000 and 2008 are 172.20 and 215.30, respectively. The corresponding values of the producer price index are 132.70 and 189.60, respectively.

- Compute and analyze the nominal and real increase in Tom's salary.
- Compute and analyze the nominal and real revenue growth of the construction firm.

#### SOLUTION:

- Tom's nominal salary grew by  $\frac{84000 - 52000}{52000} = 0.62$ , or by 62% from 2000 to 2008. This nominal salary makes no cost of living adjustment. We use the CPI to compute his real salary as  $(52000/172.20) \times 100 = \$30,197$  in 2000 and  $(84000/215.30) \times 100 = \$39,015$  in 2008. These are Tom's real salaries based on 1982–1984 prices. Thus, while Tom's salary increased by 62% in dollar amounts, his purchasing power increased by only  $\frac{39015 - 30197}{30197} = 0.29$ , or by 29%.
- The nominal revenue of the construction firm grew by  $\frac{18 - 13}{13} = 0.38$ , or by 38% from 2000 to 2008. We use the producer price index to compute the revenue growth in real terms. The real revenue is  $(13/132.70) \times 100 = \$9.80$  million in 2000 and  $(18/189.60) \times 100 = \$9.49$  million in 2008. Therefore, the real growth in revenue for the construction firm has been  $\frac{9.49 - 9.80}{9.80} = -0.03$ , or -3%.

## Inflation Rate

The **inflation rate** is the percentage rate of change of a price index over time. We generally use the CPI to compute the inflation rate in the U.S. Also, although it is common to quote the inflation rate in annual terms, the CPI can be used to calculate the inflation rate for any time period.

### CALCULATING THE INFLATION RATE

The reported **inflation rate**  $i_t$  for a given period is generally based on the consumer price index, CPI. It is computed as  $i_t = \frac{CPI_t - CPI_{t-1}}{CPI_{t-1}}$ .



### EXAMPLE 19.14

The consumer price indices for the years 2006, 2007, and 2008 are reported as 201.59, 207.34, and 215.30, respectively (Source: Bureau of Labor Statistics). Use this information to compute the annual inflation rate for 2007 and 2008.

**SOLUTION:** The inflation rates for 2007 and 2008 are computed as:

$$i_{2007} = \frac{CPI_{2007} - CPI_{2006}}{CPI_{2006}} = \frac{207.34 - 201.59}{201.59} = 0.0285 \text{ or } 2.85\%.$$

$$i_{2008} = \frac{CPI_{2008} - CPI_{2007}}{CPI_{2007}} = \frac{215.30 - 207.34}{207.34} = 0.0384 \text{ or } 3.84\%.$$

Therefore, the inflation rate increased from 2.85% in 2007 to 3.84% in 2008.

### EXAMPLE 19.15

At the beginning of 2007, Joe Gonzales invested \$1,000 in a mutual fund, which grew to \$1,050 in a year. The consumer price index, with a base of 1982–1984, is 203.37 for January 2007 and 212.23 for January 2008. Compute the real annual rate of return for Joe.

**SOLUTION:** The real rate of return is based on the deflated investment values, which for the two years are computed as  $\frac{1000}{203.37} \times 100 = \$491.71$  and  $\frac{1050}{212.23} \times 100 = \$494.75$ , respectively. The resulting real return of investment is derived as  $\frac{494.75 - 491.71}{491.71} = 0.0062$ , or 0.62%.

In Section 19.1, we used the Fisher equation to convert the nominal return into the real return. The Fisher equation will give us this same value for the real return on the investment. The nominal return for Joe is  $\frac{1050 - 1000}{1000} = 0.05$  and the corresponding inflation rate is  $\frac{212.23 - 203.37}{203.37} = 0.0436$ . Therefore, using the Fisher equation, we can compute  $1 + r = \frac{1 + R}{1 + i} = \frac{1.05}{1.0436} = 1.0061$  to get the real rate of return  $r = 0.0061$ , which varies slightly from the previous calculation for the real rate of return due to rounding.

## EXERCISES 19.3

### Mechanics

22. The nominal values for four years are given by 32, 37, 39, and 42. Convert these values to real terms if the price index values for the corresponding years are given by 100, 102, 103, 108.
23. An item increases in value from 240 to 280 in one year. What is the percentage change in the value of this item? Compute the percentage change in real terms if overall prices have increased by 5% for the same period.
24. Let revenues increase by 10% from \$100,000 to \$110,000. Calculate the percentage change in real terms if the relevant price index increases by 4% from 100 to 104.

25. The following table represents the nominal values of an item and the corresponding price index for 2007 and 2008.

Year	Nominal Value	Price Index
2007	38	112
2008	40	120

- a. Compute the inflation rate for 2008.
  - b. Compute the annual percentage change of the item in real terms.
26. The following table represents the nominal values of an item and the corresponding price index from 2009 to 2011.

Year	Nominal Value	Price Index
2009	38	100
2010	40	103
2011	42	112

- Compare the percentage change in the nominal values with the corresponding real values from 2009 to 2010.
- Compare the percentage change in the nominal values with the corresponding real values from 2010 to 2011.
- Use the price data to compute the inflation rate for 2010 and 2011.

## Applications

27. **FILE Sales\_2009.** Economists often look at retail sales data to gauge the state of the economy. This is especially so in a recession year, when consumer spending has decreased. Consider the following table, which shows U.S. monthly nominal retail sales for 2009. Sales are measured in millions of dollars and have been seasonally adjusted. Also included in the table is the corresponding producer price index (PPI) for 2009.

Month	Sales	PPI	Month	Sales	PPI
January	340,439	171.2	July	342,489	171.6
February	342,356	170.9	August	350,800	174.1
March	339,228	169.6	September	343,687	173.3
April	338,344	170.6	October	347,641	174.0
May	339,873	170.6	November	354,467	176.6
June	342,912	173.7	December	353,817	177.3

SOURCE: Federal Reserve Bank of Dallas.

- How many times were nominal sales below that of the previous month?
  - Use the PPI to compute sales in real terms. How many times were real sales below that of the previous month?
  - Compute the total percentage increase in nominal as well as real retail sales in 2009.
  - Can economists feel optimistic about the economy based on the retail sales data?
28. Japan was the first Asian country to challenge the dominance of the U.S. in the 1980s. However, since then, its economy has been in a slow but relentless decline (*The New York Times*, October 16, 2010). This country has been trapped in low growth and a downward spiral of prices, known as deflation. Consider the following CPI for Japan for the years 2001 through 2009. Compute and interpret the annual inflation rates in Japan in the 2000s.

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009
CPI	120.1	119.0	118.7	118.7	118.3	118.7	118.7	120.3	118.7

SOURCE: Bureau of Labor Statistics.

29. **FILE Earnings\_2008.** Each month the Current Employment Statistics (CES) program surveys numerous businesses and government agencies in order to obtain detailed data on earnings of workers. Consider the following data on the national average of hourly earnings for 2008. Also included is the corresponding consumer price index for 2008.

Month	Earnings	CPI	Month	Earnings	CPI
January	21.25	173.3	July	21.66	183.7
February	21.29	173.9	August	21.74	181.9
March	21.43	175.8	September	21.80	182.0
April	21.43	176.5	October	21.84	177.3
May	21.52	178.8	November	21.93	172.3
June	21.60	181.5	December	21.96	169.4

SOURCE: Bureau of Labor Statistics.

- Use the CPI to deflate the national average of hourly earnings.
- Compute the percentage change in the nominal as well as real hourly earnings in 2008.
- Were consumers getting better off over 2008? Explain.

Use the following information on CPI and PPI for Exercises 30, 31, and 32.

Year	CPI (1982–84 = 100)	PPI (1982 = 100)
2006	201.59	164.80
2007	207.34	172.70
2008	215.30	189.60
2009	214.54	172.90

SOURCE: Bureau of Labor Statistics.

30. The total revenue for The Walt Disney Company was \$35,510,000 for 2007, \$37,843,000 for 2008, and \$36,149,000 for 2009 (Source: <http://finance.yahoo.com>).
- Deflate the total revenue with the relevant price index.
  - Discuss the revenue trend during the 2007–2009 period using nominal as well as real values.
31. According to the New Hampshire Department of Education, the average teacher salary in public school districts in New Hampshire was \$46,797 in 2006, \$48,310 in 2007, and \$46,797 in 2008. Comment on the percentage change in the dollar value (nominal) as well as the purchasing power (real) of salaries.
32. According to Fisher College of Business at the Ohio State University, the starting salary of their graduates in the MBA program in 2008 was \$89,156. What must be the starting salary of the MBAs in 2009 if the salary increase makes the exact cost of living adjustment?

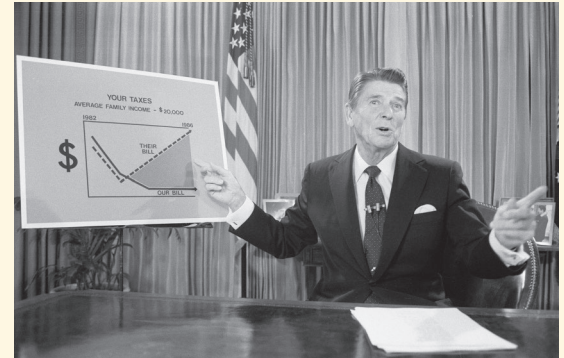
## WRITING WITH STATISTICS

Valerie Barnes is a graduate student in the department of political science at Michigan State University. She has been asked to write a brief report on the changes in the economic climate during the presidency of Ronald Reagan from 1981–1989. Valerie collects information on various economic indicators at the beginning and the end of President Reagan’s term, as shown in Table 19.9.

**TABLE 19.9** Select Economic Indicators during the Reagan Presidency

Economic Indicators	1981	1989
Federal Debt (\$ billions)	\$994.8	\$2,868.0
Median Household Income	\$19,074	\$28,906
Cost of a New Home	\$83,000	\$148,800
Dow Jones Industrial Average High	1,024	2,791
Cost of a Gallon of Regular Gasoline	\$1.38	\$1.12
Consumer Price Index (1982–1984 = 100)	90.9	124

SOURCE: [www.1980sflashback.com](http://www.1980sflashback.com).



Valerie would like to use the above information to:

1. Evaluate the change in prices over the Reagan Era, including the annual inflation rate.
2. Calculate and interpret corresponding deflated economic indicators for 1981 and 1989.
3. Comment on changes in select economic indicators during Reagan’s presidency.

Ronald Wilson Reagan became the 40th President of the United States in 1981 after serving eight years as governor of California. He took office at a time when the U.S. was experiencing economic stagnation and inflation. As president, Reagan advocated reduced business regulation and extensive tax cuts to boost economic growth. Arguably, the Reagan era signifies a period of significant growth as the economy recovered from the recession.

Crucial economic indicators were analyzed during Reagan’s presidency. The consumer price index (CPI) values imply that prices were 9.1% lower in 1981 and 24% higher in 1989 than during the base years of 1982–1984. The percentage price increase during Reagan’s term is calculated as 36.41%, resulting in an annualized inflation rate of  $(1 + 0.3641)^{1/8} - 1 = 3.96\%$ . The CPI is also used to deflate crucial economic indicators. For instance, while the median household income increased from \$19,074 to \$28,906, or by 51.55%, the corresponding deflated incomes increased from \$20,984 to \$23,311, or by 11.09%. Other similarly deflated economic indicators are presented in Table 19.A.

**TABLE 19.A** Deflated Economic Indicators

Economic Indicators	1981	1989
Federal Debt (\$ billions)	\$1,094.4	\$2,312.9
Median Household Income	\$20,984	\$23,311
Cost of a New Home	\$91,309	\$120,000
Dow Jones Industrial Average High	1,127	2,251
Cost of a Gallon of Regular Gasoline	\$1.52	\$0.90

The significant increase in the federal debt during the Reagan era is noteworthy. When Reagan took office, he used deficit spending through tax cuts to stimulate the economy. However, the debt continued to grow throughout the boom years. The resulting deflated

**Sample Report—  
Economic Indicators  
during Reagan’s  
Presidency**

federal debt rose sharply from \$1,094.4 billion in 1981 to \$2,312.9 billion in 1989, or by 111%. The deflated cost of a new home grew from \$91,309 to \$120,000, or by 31.42%. Therefore, despite the 11.09% growth in real income, a higher percentage increase in home values made owning a new home more difficult. Interestingly, the deflated Dow Jones Industrial Average High grew by a whopping 99.73% from 1,127 in 1981 to 2,251 in 1989. Finally, there was a steep decline of 40.79% in the deflated price of gasoline from \$1.52 per gallon to \$0.90 per gallon. Perhaps the price decline was the consequence of the falling demand as consumers reacted to the energy crisis of the 1970s.

President Reagan's policies reflected his personal belief in individual freedom. According to Reagan supporters, his policies resulted in the largest peacetime economic boom in American history. His critics, on the other hand, argue that the Reagan era is associated with a widening of inequality, where the rich got richer with little economic gains for most Americans. This argument is partly reflected by a meager 11.09% real increase in the median household income during the supposedly good years.

## CONCEPTUAL REVIEW

### LO 19.1 Define and compute investment returns.

The **investment return**  $R_t$  is calculated as  $R_t = \frac{P_t - P_{t-1} + I_t}{P_{t-1}}$ , where  $\frac{P_t - P_{t-1}}{P_{t-1}}$  and  $\frac{I_t}{P_{t-1}}$  are the **capital gains yield** and the **income yield** components, respectively.

The **adjusted closing price** makes appropriate adjustments for dividend distributions, stock splits and reverse stock splits. Let  $P_t^*$  and  $P_{t-1}^*$  represent the adjusted closing price of a stock at times  $t$  (current) and  $t-1$  (prior), respectively. Using adjusted closing prices, the investment return  $R_t$  at the end of time  $t$  is calculated as  $R_t = \frac{P_t^* - P_{t-1}^*}{P_{t-1}^*}$ .

### LO 19.2 Use the Fisher equation to convert nominal returns into real returns and vice versa.

The **Fisher equation**,  $1 + r = \frac{1 + R}{1 + i}$ , represents the relationship between the nominal return  $R$ , the real return  $r$ , and the expected inflation rate  $i$ .

### LO 19.3 Calculate and interpret a simple price index.

An **index number** is an easy-to-interpret numerical value that reflects a percentage change in price or quantity from a base value. A **simple price index** is a ratio of the price in period  $t$ ,  $p_t$ , and the price in the base period,  $p_0$ , expressed as a percentage. It is calculated as  $\frac{p_t}{p_0} \times 100$ . It is common to update the base period over time. We update a simple index, with a revised base period, as  $\text{Updated Index} = \frac{\text{Old Index Value}}{\text{Old Index Value of New Base}} \times 100$ .

### LO 19.4 Calculate and interpret an unweighted aggregate price index.

Let  $p_{it}$  represent the price of item  $i$  in period  $t$  and let  $p_{i0}$  be the corresponding price in the base period ( $t = 0$ ). An **unweighted aggregate price index** in period  $t$  is  $\frac{\sum p_{it}}{\sum p_{i0}} \times 100$ .

### LO 19.5 Compare the Laspeyres and the Paasche methods for computing a weighted aggregate price index.

Let  $p_{it}$  and  $q_{it}$  represent the price and quantity of item  $i$  in period  $t$  and let  $p_{i0}$  and  $q_{i0}$  be the corresponding values in the base period ( $t = 0$ ). Using only the base period quantities  $q_{i0}$ , the **Laspeyres price index** for period  $t$  is  $\frac{\sum p_{it}q_{i0}}{\sum p_{i0}q_{i0}} \times 100$ . Using only the current period ( $t = n$ ) quantities  $q_{in}$ , the **Paasche price index** for period  $t$  is  $\frac{\sum p_{it}q_{in}}{\sum p_{i0}q_{in}} \times 100$ .

**LO 19.6 Use price indices to deflate economic time series and derive the inflation rate.**

A **deflated** time series is obtained by adjusting it for changes in prices, or inflation. A time series that has been deflated is said to be represented in **real terms**. The unadjusted time series is said to be represented in **nominal terms**. We use a price index to convert the nominal value of a time series into its real value as  $\text{Real Value} = \frac{\text{Nominal Value}}{\text{Price Index}} \times 100$ .

Two commonly used price indices used to deflate economic time series are the **Consumer Price Index (CPI)** and the **Producer Price Index (PPI)**. It is common to use the CPI to adjust wages for changes in the cost of living. On the other hand, the PPI is useful to deflate revenue in order to obtain real growth in output. The reported **inflation rate**  $i_t$  for a given period is generally based on the CPI and is computed as  $i_t = \frac{CPI_t - CPI_{t-1}}{CPI_{t-1}}$ .

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

33. Kim Baek invested \$20,000 for a year in corporate bonds. Each bond sold for \$1,000 and earned a coupon payment of \$80 each during the year. The price of the bond at the end of the year has dropped to \$980.
- Calculate Kim's investment return.
  - Calculate Kim's total dollar gain or loss on his investment.
34. Toyota Motor Corp., once considered a company synonymous with reliability and customer satisfaction, has been engulfed in a perfect storm with millions of cars recalled (*BBC News*, March 19, 2010). The following table shows the monthly adjusted closing price per share of Toyota from October 2009 to March 2010.

Date	Adjusted Closing Price	Date	Adjusted Closing Price
October 2009	78.89	January 2010	77.00
November 2009	78.54	February 2010	74.83
December 2009	84.16	March 2010	79.56

SOURCE: <http://finance.yahoo.com>.

- Form a simple price index with October 2009 as the base.
  - Update the simple price index, using January 2010 as the base.
  - What is the percentage price change from October 2009 to December 2009?
  - What is the percentage price change from January 2010 to March 2010?
35. Consider the following price data from 2002 to 2010.

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010
Price	3.20	3.46	3.51	4.02	4.18	4.30	4.59	4.50	4.70

- Compute the simple price index using 2002 as the base year.
- Update the index numbers with a base year revised from 2002 to 2005.

- Plot the index numbers with a base year of 2002 and a base year of 2005. Compare the two plots.
36. Consider the following price data from 2009 to 2011.

Year	Product 1	Product 2	Product 3
2009	38	94	45
2010	40	92	48
2011	42	98	56

- Compute and interpret the simple price index for each product, using 2009 as the base year.
  - Compute and interpret the unweighted aggregate price index, using 2009 as the base year.
37. Let the quantities corresponding to the prices in Exercise 36 be given by the following table.

Year	Product 1	Product 2	Product 3
2009	90	32	48
2010	82	34	46
2011	76	30	36

- Compute the Laspeyres price index, using 2009 as the base year.
  - Compute the Paasche price index, using 2009 as the base year.
  - Comment on the differences between the two indices.
38. Lindsay Kelly bought 100 shares of Google, 300 shares of Microsoft, and 500 shares of Nokia in January 2005. The adjusted closing prices of these stocks over the next three years are shown in the accompanying table.

Year	Google	Microsoft	Nokia
2005	195.62	24.11	13.36
2006	432.66	26.14	16.54
2007	505.00	28.83	19.83

SOURCE: <http://finance.yahoo.com>.

- Compute and interpret the unweighted aggregate price index for Lindsay's portfolio, using 2005 as the base year.



- b. Compute and interpret the corresponding weighted price index using the Laspeyres approach.
- c. Why are the results from parts a and b so different?
39. Citigroup, Inc., is a major financial services company based in New York. It suffered huge losses during the global financial crisis and was rescued in November 2008 in a massive bailout by the U.S. government. Consider the following table, representing the net revenue and net income of Citigroup for 2006 to 2009. Both variables are measured in billions of dollars.

Year	Net Revenue	Net Income
2006	146.6	21.2
2007	159.2	3.6
2008	105.8	-27.7
2009	111.0	-1.6

SOURCE: <http://money.cnn.com>.

- a. Compute and interpret the simple price index for net revenue, using 2006 as the base year.
- b. Compute and interpret the simple price index for net income, using 2006 as the base year.
40. Consider the following consumer price index and producer price index for 2006–2009.

Year	CPI (1982–84 = 100)	PPI (1982 = 100)
2006	201.59	164.80
2007	207.34	172.70
2008	215.30	189.60
2009	214.54	172.90

SOURCE: Bureau of Labor Statistics.

- a. Use the relevant price index to deflate the data on net revenue of Citigroup, given in Exercise 39.
- b. Use the relevant price index to deflate the data on net income of Citigroup, given in Exercise 39.
41. An investor bought 1,000 shares of Citigroup in January 2009 for \$3.55 a share. She sold all of her shares in December 2009 for \$3.31 a share.
- a. What annual rate of return did the investor earn?
- b. Use the CPI information from Exercise 40 to compute the inflation rate for 2009.
- c. What is the investor's real rate of return?
42. The adjusted closing prices of Wendy's/Arby's Group, Inc., for the first three months of 2008 are presented in the following table. Also included in the table is the corresponding consumer price index (CPI).

Date	Adjusted Closing Price	CPI (Base 1982–1984)
January, 2008	8.94	212.225
February, 2008	8.29	212.703
March, 2008	6.01	213.543

SOURCE: <http://finance.yahoo.com>; and Bureau of Labor Statistics.

- a. Find the real rate of return for the three months by first using the CPI to deflate the adjusted closing price.
- b. Replicate the above result with Fisher's equation, based on the nominal rate of return and the inflation rate.

## CASE STUDIES

**CASE STUDY 19.1** The dot-com period, roughly between 1995–2000, was characterized by extreme investor optimism for Internet-based businesses. This period was also marked by young, bold managers, who made a good deal of money by reaching consumers only over the Internet. Arguably, the dot-com boom was a case of too much too fast and was consequently followed by a crash in March 2000. Jose Menges is a business student at a California State University. For his senior seminar course, he has been asked to compare the stock performance of Internet-based companies with non-Internet-based companies during the dot-com boom-bust period. He collects monthly data on the adjusted closing prices from 1999 to 2000 for four companies. Amazon (AMZN) and eBay (EBAY) are chosen to represent the Internet-based companies, whereas Coca-Cola (COKE) and Johnson and Johnson (JNJ) reflect non-Internet companies. A portion of the data is shown in the accompanying table.

### FILE Dotcom

**Data for Case Study 19.1** Monthly Adjusted Closing Prices for Four Firms, 1999–2000

Month	AMZN	EBAY	COKE	JNJ
January, 1999	58.47	11.57	44.00	33.93
February, 1999	64.06	13.92	43.80	34.13
⋮	⋮	⋮	⋮	⋮
December, 2000	15.56	8.25	30.84	41.86

SOURCE: <http://finance.yahoo.com>.



In a report, use the sample information to:

1. Compute monthly returns for all companies for 1999 and 2000.
2. Compare the stock performance for the Internet-based companies with non-Internet based companies in the dot-com boom-bust period.

**CASE STUDY 19.2** The U.S. is often blamed for triggering the 2008 global financial crisis because many of the excesses and bad practices originated in the U.S. The crisis has had consequences on all aspects of the global economy. According to a recent report by Brookings Institute, the U.S. economic crisis is linked to a huge drop in world trade. Since U.S. imports have been an important component of world demand, a drop in imports has had repercussions in its exports. Rami Horowitz is a young economist working for a trade policy institute. He wishes to analyze the changes in U.S. imports and exports over the 2007–2009 time period. Rami collects quarterly data on real exports and real imports, where the values are seasonally adjusted and measured in billions of 2005 dollars. A portion of the data is shown in the accompanying table.

**Data for Case Study 19.2** U.S. Real Exports and Imports

Period	Real Exports	Real Imports
2007: Quarter 1	1485.9	2190.8
2007: Quarter 2	1504.8	2188.1
⋮	⋮	⋮
2009: Quarter 4	1555.5	1902.7

SOURCE: Federal Reserve Bank of Dallas.

**FILE**  
*World\_Trade*

In a report, use the sample information to:

1. Create simple indices for real exports and real imports with Quarter 1, 2007, used as the base period.
2. Interpret the percentage changes in real exports and real imports over the three-year period.

**CASE STUDY 19.3** The Cheesecake Factory, Inc., is a popular restaurant chain in the U.S. Although it started as a small restaurant in 1978, it currently has over 140 branches all over the country. The restaurants are characterized by extensive menus, custom décor, and large portions of food. Jeff Watson works as the kitchen manager in one of their regional branches. He is responsible for managing the operations as well as food and labor costs. He constantly monitors market conditions and, in his annual reports, analyzes the changing retail cost of the ingredients used in cooking. In his current report, he decides to analyze meat prices. He collects data on monthly average retail prices of three varieties of ground beef. This information is important, as the restaurant purchases about 1,400 pounds of regular, 800 pounds of ground chuck, and 500 pounds of lean ground beef each month. A portion of the data is shown in the accompanying table.

**Data for Case Study 19.3** 2009 Monthly Retail Cost of Ground Beef (in \$ per pound)

Month	Regular Beef	Ground Chuck	Lean Ground Beef
January	2.357	2.961	3.426
February	2.436	3.019	3.440
⋮	⋮	⋮	⋮
December	2.186	2.828	3.391

SOURCE: United States Department of Agriculture.

**FILE**  
*Ground\_Beef*

In a report, use the sample information to:

1. Compute and interpret simple indices for each variety of meat, using January 2009 as the base period.
2. Compute and interpret the weighted aggregate price index, using January 2009 as the base period.
3. Compare the above indices.

# 20

## LEARNING OBJECTIVES

After reading this chapter  
you should be able to:

- LO 20.1 Distinguish between parametric and nonparametric tests.
- LO 20.2 Make inferences about a population median.
- LO 20.3 Make inferences about the population median difference based on matched-pairs sampling.
- LO 20.4 Make inferences about the difference between two population medians based on independent sampling.
- LO 20.5 Make inferences about the difference between three or more population medians.
- LO 20.6 Conduct a hypothesis test for the population Spearman rank correlation coefficient.
- LO 20.7 Make inferences about the difference between two populations of ordinal data based on matched-pairs sampling.
- LO 20.8 Determine whether the elements of a sequence appear in a random order.

# Nonparametric Tests

The hypothesis tests presented in earlier chapters make certain assumptions about the underlying population. We refer to these tests as parametric tests. A  $t$  or an  $F$  test, for example, requires that the observations come from a normal distribution. These tests are quite “robust,” in the sense that they are still useful when the assumptions are not exactly fulfilled, especially when the sample size is large. However, in situations when the underlying population is markedly nonnormal, we apply distribution-free alternative techniques called nonparametric methods. In addition to not needing to fulfill a given distribution requirement, another benefit of nonparametric methods is that they do not require a level of measurement as strong as that necessary for parametric tests. For instance, we cannot calculate means and variances with ordinal data (required calculations for parametric tests) because the numbers on an ordinal scale have no meaning except to indicate rank order. In this chapter we explore a variety of nonparametric tests that make fewer assumptions about the distribution of the underlying population and/or treat data of a weaker scale.



## INTRODUCTORY CASE

### Analyzing Mutual Fund Returns

In Chapter 3 we were introduced to Rebecca Johnson, an investment counselor at a large bank. One of her clients has narrowed his investment options to two top-performing mutual funds from the last decade: Vanguard's Precious Metals and Mining fund (henceforth, Metals) and Fidelity's Strategic Income fund (henceforth, Income). He has some final questions for Rebecca with respect to each fund's return. Rebecca explains that her analysis will use techniques that do not rely on stringent assumptions concerning the distribution of the underlying population, since return data often diverge from the normal distribution. Table 20.1 shows a portion of the annual return data for each fund and some relevant descriptive statistics over the last decade.

**TABLE 20.1** Annual Returns (in percent) for Metals and Income Funds, 2000–2009

**FILE**  
*Fund\_Returns*

Year	Metals	Income
2000	−7.34	4.07
2001	18.33	6.52
⋮	⋮	⋮
2009	76.46	31.77
	$\bar{x} = 24.65\%$ median = 33.83% $s = 37.13\%$	$\bar{x} = 8.51\%$ median = 7.34% $s = 11.07\%$

SOURCE: <http://finance.yahoo.com>.

Rebecca would like to use the above sample information to:

1. Determine whether the median return for the Metals fund is greater than 5%.
2. Determine whether the median difference between the two funds' returns differs from zero.
3. Determine whether the funds' returns are correlated.

A synopsis of this case is provided at the end of Section 20.4.

Distinguish between parametric and nonparametric tests.

The parametric tests presented in earlier chapters make certain assumptions about the underlying population. These conventional tests can be misleading if the underlying assumptions are not met. **Nonparametric tests**, also referred to as distribution-free tests, are attractive when the parametric assumptions seem unreasonable. Nonparametric tests use fewer and weaker assumptions than those associated with parametric tests. For instance, these tests do not assume that the sample data originate from a normal distribution. Nonparametric tests are especially useful when sample sizes are small. Finally, while parametric tests require data of interval or ratio scale, nonparametric tests can be performed on data of nominal or ordinal scale. (For a review of these data concepts, refer to Section 3 in Chapter 1.)

Nonparametric tests have disadvantages, too. If the parametric assumptions are valid yet we choose to use a nonparametric test, the nonparametric test is less powerful (more prone to Type II error) than its parametric counterpart. The reason for less power is that a nonparametric test uses the data less efficiently. As we will see shortly, nonparametric tests often focus on the rank of the data rather than the magnitude of the sample values, thus possibly ignoring useful information.

Table 20.2 summarizes some of the parametric tests that we examined in earlier chapters. The first column shows the parametric test of interest, the second column states the underlying assumptions of the test, and the third column lists where the test was covered in the text. Each one of these parametric tests has a nonparametric counterpart. At the end of Section 20.4, we will present a table that lists the corresponding nonparametric test for each parametric test.

**TABLE 20.2** Summary of Select Parametric Tests

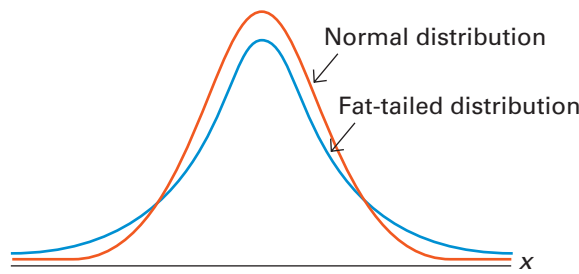
Parametric Test	Population Characteristics and Other Description	Reference Section
<i>t</i> test concerning the population mean	Sampling from a normal population or large sample; $\sigma$ unknown	9.3
<i>t</i> test to determine whether the population mean difference differs from some assumed value based on matched-pairs sampling	Sampling from a normal population or large sample	10.2
<i>t</i> test to determine whether two population means differ from some assumed value based on independent sampling	Sampling from normal populations or large samples; $\sigma_1$ and $\sigma_2$ unknown	10.1
<i>F</i> test to determine whether three or more population means differ	Sampling from normal populations or large samples; $\sigma_1, \sigma_2, \sigma_3, \dots$ unknown but assumed equal	13.1
<i>t</i> test to determine whether two variables are correlated	Sampling from a normal population or large sample	14.1

Make inferences about a population median.

## The Wilcoxon Signed-Rank Test for a Population Median

In Chapter 9, we used a *t* test to determine whether the population mean  $\mu$  ( $\sigma$  unknown) differs from some assumed value. However, as shown in Table 20.2, a *t* test requires that we sample from a normal distribution. If we cannot assume that the data are normally distributed and/or we want to test whether the population *median* differs from some hypothesized value, we can apply the **Wilcoxon signed-rank test**. The Wilcoxon signed-rank test makes no assumptions concerning the distribution of the population except that it is continuous and symmetric.

Let's revisit the introductory case. Here we learn that Rebecca's analysis of the fund return data will use nonparametric techniques. She chooses these methods because the distribution of return data often has "fatter tails" as compared to the normal



**FIGURE 20.1** Normal distribution versus “fat-tailed” distribution

distribution; that is, the likelihood of extreme returns (area under the tail) is higher for a fatter-tailed distribution than for a normal distribution. Figure 20.1 shows a normal distribution versus a distribution with fatter tails. If Rebecca were to rely on tests that incorrectly assume that the data are normally distributed, then there is a chance that she may make erroneous conclusions. She chooses to use the Wilcoxon signed-rank test for the population median.

Following the methodology outlined in earlier chapters, when conducting a hypothesis test for the population median  $m$ , we want to test whether  $m$  is not equal to, greater than, or less than  $m_0$ , the value of the population median postulated in the null hypothesis. The null and alternative hypotheses will assume one of the following forms:

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: m = m_0$	$H_0: m \leq m_0$	$H_0: m \geq m_0$
$H_A: m \neq m_0$	$H_A: m > m_0$	$H_A: m < m_0$

For the Metals fund return data in Table 20.1, we would like to determine whether the median return for this fund is greater than 5%. We formulate the one-tailed test as

$$H_0: m \leq 5$$

$$H_A: m > 5$$

To arrive at the sample value for the Wilcoxon signed-rank test statistic  $T$ , several calculations are necessary.

- A.** We first calculate the difference  $d_i$  between each observed value and the hypothesized median. In this case,  $d_i = x_i - 5$ , as shown in the second column of Table 20.3.

**TABLE 20.3** Calculations for the Wilcoxon Signed-Rank Test Statistic

Return, $x$ (1)	$d = x - m_0$ (2)	$ d $ (3)	Rank (4)	Ranks of Negative Differences (5)	Ranks of Positive Differences (6)
-7.34	$-7.34 - 5 = -12.34$	12.34	2	2	
18.33	13.33	13.33	3		3
33.35	28.35	28.35	4		4
59.45	54.45	54.45	8		8
8.09	3.09	3.09	1		1
43.79	38.79	38.79	7		7
34.30	29.30	29.30	5		5
36.13	31.13	31.13	6		6
-56.02	-61.02	61.02	9	9	
76.46	71.46	71.46	10		10
				$T^- = 11$	$T^+ = 44$

- B. We then take the absolute value of each difference,  $|d_i|$ . See the third column of Table 20.3. Any differences of zero are discarded from the sample; no zero-differences occur in this example. We calculate  $|d_i|$  because if the median is 5 (the null hypothesis is true), then positive or negative differences of a given magnitude are equally likely.
- C. Next we rank the absolute value of each difference, assigning 1 to the smallest  $|d|$  and  $n$  to the largest  $|d|$ . Note that  $n$  would be smaller than the original sample size if there were some zero-difference observations, which we discarded; here,  $n$  equals the original sample size of 10. Any ties in the ranks of differences are assigned the average of the tied ranks. For instance, if two observations have the rank of 5 (occupying the 5th and 6th positions), each is assigned the rank of  $(5 + 6)/2 = 5.5$ . Or, if three observations have a ranking of 1, each is assigned a rank of  $(1 + 2 + 3)/3 = 2$ . In the Metals fund return example, there are no ties. The rankings for the differences are shown in the fourth column of Table 20.3.
- D. We then sum the ranks of the negative differences (denoted  $T^-$ ) and sum the ranks of the positive differences (denoted  $T^+$ ). In this example we find two negative differences, whose rank sum is  $T^- = 11$ , and eight positive differences, whose rank sum is  $T^+ = 44$ . These calculations are shown in the fifth and sixth columns of Table 20.3.

The sum of  $T^-$  and  $T^+$  should equal  $n(n + 1)/2$ , which is the formula for the sum of consecutive integers from 1 to  $n$ . In our example,  $T^- + T^+ = 11 + 44 = 55$ . Also,  $n(n + 1)/2 = 10(10 + 1)/2 = 55$ . As we will see shortly, the value of  $T^-$  is not used in the actual analysis, but its calculation can help us avoid errors.

Values of  $T^-$  and  $T^+$  relatively close to one another indicate that the rank sums more or less offset one another and provide evidence in support of the null hypothesis. However, a small value of  $T^-$  relative to  $T^+$ , for example, implies larger positive deviations from the hypothesized median value of 5. This would suggest that the median is greater than 5. We can view a small value of  $T^+$  relative to  $T^-$  in a like manner.

#### THE TEST STATISTIC $T$ FOR THE WILCOXON SIGNED-RANK TEST

The **test statistic  $T$**  for the Wilcoxon signed-rank test is defined as  $T = T^+$ , where  $T^+$  denotes the sum of the ranks of the positive differences from the hypothesized median  $m_0$ .

There are two scenarios when conducting the Wilcoxon signed-rank test:

1. If  $n \leq 10$ , then the test statistic  $T$  has a distribution whose critical values are shown in Table 6 of Appendix A.
2. If  $n \geq 10$ ,  $T$  can be approximated by the normal distribution with mean  $\mu_T = \frac{n(n + 1)}{4}$  and standard deviation  $\sigma_T = \sqrt{\frac{n(n + 1)(2n + 1)}{24}}$ , and hence the value of the resulting test statistic is computed as  $z = \frac{T - \mu_T}{\sigma_T}$ .

For ease of exposition, we do not make a distinction between the random variable and the particular outcomes of the random variable in this chapter. For example, we use the test statistic  $T$  to represent a random variable as well as its sample value. We adopt this same practice for the test statistics  $W$ ,  $H$ ,  $r_s$ , and  $R$  that we introduce in later sections.

Table 20.4 shows a portion of lower  $T_L$  and upper  $T_U$  critical values for the Wilcoxon signed-rank test. A more complete version is found in Table 6 of Appendix A. As usual,  $T_L$  is used for a left-tailed test,  $T_U$  is used for a right-tailed test, and both  $T_L$  and  $T_U$  are used for a two-tailed test.



**TABLE 20.4** Portion of the Lower  $T_L$  and Upper  $T_U$  Critical Values for the Wilcoxon Signed-Rank Test

Two-Tailed Test: One-Tailed Test:	$\alpha = 0.10$ $\alpha = 0.05$	$\alpha = 0.05$ $\alpha = 0.025$	$\alpha = 0.02$ $\alpha = 0.01$	$\alpha = 0.01$ $\alpha = 0.005$
$n = 8$	5, 31	3, 33	1, 35	0, 36
9	8, 37	5, 40	3, 42	1, 44
10	10, <b>45</b>	8, 47	5, 50	3, 52

**EXAMPLE 20.1**

Given the return data in Table 20.1, determine whether the median return for the Metals fund is significantly greater than 5% with  $\alpha = 0.05$ .

**SOLUTION:** As discussed earlier, the competing hypotheses for the test are  $H_0: m \leq 5$  versus  $H_A: m > 5$ . In this example, since the sample size equals exactly ten, we can implement the Wilcoxon test with or without the normal distribution approximation. Here, we implement the test without the normal approximation; we will use the normal approximation in Example 20.3. For a right-tailed test with  $\alpha = 0.05$  and  $n = 10$ , the decision rule is to reject the null hypothesis if the value of the test statistic  $T$  is *greater than or equal to*  $T_U = 45$  (see the boldface value in Table 20.4). In Table 20.3, we calculated the value of the test statistic as  $T = T^+ = 44$ . Since  $T = 44 < 45 = T_U$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that the median return is greater than 5% for the Metals fund.

**EXAMPLE 20.2**

Given the return data in Table 20.1, determine whether the median return for the Income fund is significantly greater than 5% with  $\alpha = 0.05$ .

**SOLUTION:** We specify the same competing hypotheses as in Example 20.1; that is,  $H_0: m \leq 5$  versus  $H_A: m > 5$ . Table 20.5 reports a portion of the Minitab output when testing whether the median return for the Income fund is greater than 5%. Minitab reports the value of the Wilcoxon signed-rank test statistic as  $T = T^+ = 40$  with a corresponding  $p$ -value of 0.111. At the 5% significance level, we do not reject the null hypothesis; we cannot conclude that the median return for the Income fund is significantly greater than 5%.

**TABLE 20.5** Minitab Output for Example 20.2

Test of median = 5.000 versus median > 5.000					
	N	N for Test	Wilcoxon Statistic	P	Estimated Median
Income	10	10	40.0	0.111	7.410

**Using a Normal Distribution Approximation for  $T$** 

The sampling distribution of  $T$  can be approximated by the normal distribution if  $n$  has at least 10 observations.<sup>1</sup> We can then easily implement a  $z$  test with this approximation.

<sup>1</sup>Since the normality assumption for parametric tests becomes less stringent in large samples, the main appeal of rank-based tests tends to be with relatively small samples.

### EXAMPLE 20.3

Redo the test specified in Example 20.1 assuming that  $T$  is normally distributed.

**SOLUTION:** Again we use the competing hypotheses,  $H_0: m \leq 5$  versus  $H_A: m > 5$ , and the value of the test statistic,  $T = T^+ = 44$ . Since there are 10 years of return data ( $n = 10$ ), we calculate the mean and the standard deviation of the sampling distribution of  $T$  as

$$\mu_T = \frac{n(n+1)}{4} = \frac{10(10+1)}{4} = 27.5 \text{ and}$$
$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{10(10+1)(2 \times 10 + 1)}{24}} = 9.81.$$

The corresponding value of the test statistic is

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{44 - 27.5}{9.81} = 1.68.$$

Therefore, with the normal distribution approximation, we find the corresponding  $p$ -value as  $P(Z \geq 1.68) = 0.0465$ . Since the  $p$ -value is slightly lower than  $\alpha = 0.05$ , we reject  $H_0$ . This conclusion differs from that made in Example 20.1. We note that the test implemented without the normal approximation in Example 20.1 is more reliable; the reliability of the test with the normal approximation improves with the sample size.

## EXERCISES 20.1

### Mechanics

1. Consider the following competing hypotheses and sample data.

$$H_0: m = 20 \quad n = 7 \quad T^- = 2 \quad T^+ = 26$$
$$H_A: m \neq 20$$

- Specify the decision rule at the 5% significance level.
  - What is the conclusion? Explain.
2. Consider the following competing hypotheses and sample data.

$$H_0: m \geq 150 \quad n = 9 \quad T^- = 42 \quad T^+ = 3$$
$$H_A: m < 150$$

- Specify the decision rule at the 1% significance level.
  - What is the conclusion? Explain.
3. Consider the following competing hypotheses and sample data.

$$H_0: m \leq 150 \quad n = 30 \quad T^- = 200 \quad T^+ = 265$$
$$H_A: m > 150$$

- Assuming that the sampling distribution of  $T$  is normally distributed, determine the value of the test statistic.
- Calculate the  $p$ -value.
- At the 5% significance level, what is the conclusion? Explain.

4. Consider the following sample data.

8	5	11	7	6	5
---	---	----	---	---	---

- Specify the competing hypotheses to determine whether the median is less than 10.
- With  $\alpha = 0.05$ , what is the decision rule?
- Calculate the value of the test statistic  $T$ .
- Is the median less than 10? Explain.

5. Consider the following sample data.

105	90	110	80	85	85	103	70	115	75
-----	----	-----	----	----	----	-----	----	-----	----

Assume the normal approximation for  $T$ .

- Specify the competing hypotheses to determine whether the median differs from 100.
- Determine the critical value(s) of the test at  $\alpha = 0.10$ .
- Calculate the value of the test statistic.
- Is the median different from 100? Explain.

### Applications

6. A random sample of eight drugstores shows the following prices (in \$) for a popular pain reliever:

5.00	4.25	3.75	5.50	5.75	6.25	5.25	4.25
------	------	------	------	------	------	------	------

- Specify the competing hypotheses to determine whether the median price is less than \$4.50.
  - At the 5% significance level, what is the decision rule?
  - Calculate the value of the test statistic.
  - At the 5% significance level, can you conclude that the median price is less than \$4.50?
7. The following table lists the annual returns (in percent) over a 10-year period for ING Russia, a top-performing mutual fund.

Year	Return	Year	Return
2000	-17.80	2005	70.94
2001	80.32	2006	67.52
2002	24.72	2007	30.69
2003	75.88	2008	-71.51
2004	5.91	2009	129.97

Source: www.finance.yahoo.com

- Specify the competing hypotheses to determine whether the median return differs from 8%.
  - At the 10% significance level, what is the decision rule? Do not assume the normal approximation for  $T$ .
  - Calculate the value of the test statistic.
  - Does the median return differ from 8%?
8. During the fourth quarter of 2009, rents declined in almost all major cities in the United States. The largest fall was in New York, where average rents fell nearly 20% to \$44.69 per square foot annually (*The Wall Street Journal*, January 8, 2010). The following table lists the average rent per square foot for 10 cities during the fourth quarter of 2009.

City	Rent	City	Rent
New York	\$45	Miami	\$24
Washington, D.C.	42	Seattle	24
San Francisco	30	Chicago	21
Boston	30	Houston	20
Los Angeles	27	Philadelphia	20

- Specify the competing hypotheses to determine whether the median rent is greater than \$25 per square foot.
  - Calculate the value of the test statistic. Assume the normal approximation for  $T$ .
  - Calculate the  $p$ -value.
  - At the 1% significance level, can you conclude that the median rent exceeds \$25 per square foot?
9. **FILE Wage.** An economist wants to test whether the median hourly wage is less than \$22.
- Specify the competing hypotheses for the test.
  - At the 5% significance level, what is the decision rule? Assume the normal approximation for  $T$ .
  - Using the **Wage** data, calculate the value of the test statistic.
  - At the 5% significance level, can you conclude that the median hourly wage is less than \$22?
10. **FILE MV\_Houses.** A realtor in Mission Viejo, California, believes that the median price of a house is more than \$500,000.
- Specify the competing hypotheses for the test.
  - Using the **MV\_Houses** data, calculate the value of the test statistic. Assume the normal approximation for  $T$ .
  - Calculate the  $p$ -value.
  - At the 5% significance level, is the realtor's claim supported by the data?
11. **FILE Convenience.** An entrepreneur examines monthly sales (in \$1,000s) for 40 convenience stores in Rhode Island.
- Specify the competing hypotheses to determine whether median sales differ from \$130,000.
  - At the 5% significance level, what is the decision rule? Assume the normal approximation for  $T$ .
  - Using the **Convenience** data, calculate the value of the test statistic.
  - At the 5% significance level, do median sales differ from \$130,000?

## 20.2 TESTING TWO POPULATION MEDIANS

In Chapter 10, we presented  $t$  tests to determine whether significant differences existed between population means from matched-pairs and independent samples. When using a  $t$  test, we assume that we are sampling from normal populations. If we wish to compare central tendencies from nonnormal populations, then the **Wilcoxon signed-rank test** serves as the nonparametric counterpart to the matched-pairs  $t$  test. The **Wilcoxon rank-sum test**, also referred to as the **Mann-Whitney test**, is used for independent samples. We again note that if the normality assumption is not unreasonable, then these tests are less powerful than the standard  $t$  tests. We begin this section by examining the Wilcoxon signed-rank test for a matched-pairs experiment followed by the Wilcoxon rank-sum test for independent samples.

## LO 20.3

Make inferences about the population median difference based on matched-pairs sampling.

# The Wilcoxon Signed-Rank Test for a Matched-Pairs Sample

In this application of matched-pairs sampling, the parameter of interest is referred to as the median difference  $m_D$  where  $D = X - Y$ , and the random variables  $X$  and  $Y$  are matched in a pair. We refer you to Chapter 10 for details on matched-pairs sampling. When we wish to test whether  $m_D$  is not equal to, greater than, or less than 0, we set up the competing hypotheses as follows.

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: m_D = 0$	$H_0: m_D \leq 0$	$H_0: m_D \geq 0$
$H_A: m_D \neq 0$	$H_A: m_D > 0$	$H_A: m_D < 0$

Since applying the Wilcoxon signed-rank test to a matched-pairs sample is nearly identical to its use for a single sample, we describe its use with an example and computer output.

## EXAMPLE 20.4

We again use the *Fund\_Returns* data from the introductory case. At the 5% significance level, determine whether the median difference between the funds' returns differs from zero.

**SOLUTION:** We must first recognize that these samples are dependent, in that each return observation is blocked by year. We apply the Wilcoxon signed-rank test to determine whether significant differences exist between the median difference of the returns and formulate the two-tailed test as

$$H_0: m_D = 0$$

$$H_A: m_D \neq 0$$

Table 20.6 summarizes the method for calculating the value of the  $T$  statistic; that is, we first calculate differences between the returns (column 4), find absolute differences (column 5), and determine rankings (column 6). Then we compute the sum of the ranks of negative differences (column 7,  $T^- = 12$ ), and the sum of the ranks of positive differences (column 8,  $T^+ = 43$ ). The value of the test statistic  $T$  is  $T = T^+ = 43$ . Given the critical values in Table 20.4, the decision rule is to reject  $H_0$  if  $T \leq 8$  or  $T \geq 47$ . Since  $8 < T = 43 < 47$ , we do not reject  $H_0$ ; at the 5% significance level we cannot conclude that the median difference between the returns differs from zero.

**TABLE 20.6** Calculations for Wilcoxon Signed-Rank Test

Year (1)	Metals $x$ (2)	Income $y$ (3)	$d = x - y$ (4)	$ d $ (5)	Rank (6)	Ranks of Negative Differences (7)	Ranks of Positive Differences (8)
2000	-7.34	4.07	$-7.34 - 4.07 = -11.41$	11.41	2	2	
2001	18.33	6.52	11.81	11.81	3		3
2002	33.35	9.38	23.97	23.97	4		4
2003	59.45	18.62	40.83	40.83	8		8
2004	8.09	9.44	-1.35	1.35	1	1	
2005	43.79	3.12	40.67	40.67	7		7
2006	34.30	8.15	26.15	26.15	5		5
2007	36.13	5.44	30.69	30.69	6		6
2008	-56.02	-11.37	-44.65	44.65	9	9	
2009	76.46	31.77	44.69	44.69	10		10
						$T^- = 12$	$T^+ = 43$

## Using the Computer for the Wilcoxon Signed-Rank Test

Table 20.7 reports a portion of the Minitab output when testing whether the median difference between the Metals fund and the Income fund differs from zero. Minitab reports a Wilcoxon Statistic of 43, which is the same value that we calculated manually. The  $p$ -value corresponding to this test statistic is 0.126. At the 5% significance level, we do not reject the null hypothesis; we cannot conclude that the median difference between funds' returns differs from zero.

**TABLE 20.7** Minitab Output for Example 20.4

Test of median = 0.000 versus median not = 0.000				
	N	N for Test	Wilcoxon Statistic	P
Metals-Income	10	10	43.0	0.126

## The Wilcoxon Rank-Sum Test for Independent Samples

Now we discuss whether significant differences exist between two population medians when the underlying populations are nonnormal and the samples are independent. In this situation, we use the Wilcoxon rank-sum test. The parameter of interest is the difference between two population medians  $m_1 - m_2$ . When we wish to test whether  $m_1 - m_2$  is not equal to, greater than, or less than 0, we set up the competing hypotheses as follows.

### LO 20.4

Make inferences about the difference between two population medians based on independent sampling.

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: m_1 - m_2 = 0$	$H_0: m_1 - m_2 \leq 0$	$H_0: m_1 - m_2 \geq 0$
$H_A: m_1 - m_2 \neq 0$	$H_A: m_1 - m_2 > 0$	$H_A: m_1 - m_2 < 0$

Consider the next example.

An undergraduate at a local university has narrowed her choice of major to computer science or finance. She wonders whether her choice will significantly influence her salary upon graduation. She gathers salary data (in \$1,000s) on 10 recent graduates who majored in computer science and 10 recent graduates who majored in finance. The data are shown in Table 20.8.

**TABLE 20.8** Salary Information on Computer Science and Finance Graduates (in \$1,000s)

Computer Science		Finance	
66	59	61	55
60	67	52	52
58	64	54	52
65	68	50	47
70	69	62	46

**FILE**  
Undergrad\_Salaries

In order to determine whether salaries are significantly different depending on major, we apply the Wilcoxon rank-sum test. Let  $m_1$  and  $m_2$  denote the population median salary for computer science and finance majors, respectively. We formulate the two-tailed test as

$$H_0: m_1 - m_2 = 0$$

$$H_A: m_1 - m_2 \neq 0$$

To arrive at the value of the test statistic for the Wilcoxon rank-sum test  $W$ , several steps are necessary.

- A. We first pool the data from sample 1 (Computer Science) and sample 2 (Finance), with  $n_1$  and  $n_2$  observations, and arrange **all** the data in ascending order of magnitude. That is, we treat the independent samples as if they are one large sample of size  $n_1 + n_2 = n$ . See column 1 of Table 20.9.
- B. We then rank the observations from lowest to highest, assigning the numbers 1 to  $n$ . Since we have a multiple tie at ranks 4, 5, and 6, we assign to each of the tied observations the mean of the ranks which they jointly occupy, or  $(4 + 5 + 6)/3 = 5$ . See columns 2 and 3 of Table 20.9. We note that the finance salaries occupy the lower ranks, whereas the computer science salaries occupy the higher ranks.
- C. We then sum the ranks of the computer science salaries (denoted  $W_1$ ) and sum the ranks of the finance salaries (denoted  $W_2$ ). Here we find that  $W_1 = 149$  and  $W_2 = 61$ ; see columns 4 and 5 of Table 20.9. To check that we have performed the calculations properly, we confirm that the sum of the rank sums,  $W_1 + W_2$ , equals  $\frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$ , which is equivalent to the sum of the integers from 1 to  $n_1 + n_2$ . We first find that  $W_1 + W_2 = 149 + 61 = 210$ . Since  $n_1 = 10$  and  $n_2 = 10$ , we then find that  $\frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} = \frac{(10 + 10)(10 + 10 + 1)}{2} = 210$ .

**TABLE 20.9** Calculations for Wilcoxon Rank-Sum Test

Salary (1)	Sample of Origin (2)	Rank (3)	Computer Science Ranks (4)	Finance Ranks (5)
46	Finance	1		1
47	Finance	2		2
50	Finance	3		3
52	Finance	5		5
52	Finance	5		5
52	Finance	5		5
54	Finance	7		7
55	Finance	8		8
58	Computer Science	9	9	
59	Computer Science	10	10	
60	Computer Science	11	11	
61	Finance	12		12
62	Finance	13		13
64	Computer Science	14	14	
65	Computer Science	15	15	
66	Computer Science	16	16	
67	Computer Science	17	17	
68	Computer Science	18	18	
69	Computer Science	19	19	
70	Computer Science	20	20	
			$W_1 = 149$	$W_2 = 61$

If the median salary of computer science majors is equal to the median salary of finance majors, then we would expect each major to produce about as many low ranks as high ranks, so that  $W_1$  is relatively close to  $W_2$ . However, if the median salaries are significantly different, then most of the higher ranks will be occupied by one major and most of the lower ranks will be occupied by the other major, so that  $W_1$  will significantly differ from  $W_2$ . We determine whether  $W_1$  is close to or far from  $W_2$  by comparing one of these values to the appropriate critical value.



## THE TEST STATISTIC $W$ FOR THE WILCOXON RANK-SUM TEST

The **test statistic  $W$**  for the Wilcoxon rank-sum test is defined as:

- $W = W_1$  if  $n_1 \leq n_2$ , or
- $W = W_2$  if  $n_1 > n_2$ ,

where  $W_1$  and  $W_2$  denote the sums of the ranks of the values in samples 1 and 2.

There are two scenarios when conducting the Wilcoxon rank-sum test:

1. If  $n_1 \leq 10$  and  $n_2 \leq 10$ , then the test statistic  $W$  has a distribution whose critical values are shown in Table 7 of Appendix A.
2. If  $n_1 \geq 10$  and  $n_2 \geq 10$ , then  $W$  can be approximated by the normal distribution with mean  $\mu_W = \frac{(n_1 + n_2 + 1) \times \min(n_1, n_2)}{2}$  and standard deviation  $\sigma_W = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$ , and hence the value of the resulting test statistic is computed as  $z = \frac{W - \mu_W}{\sigma_W}$ .

In the above example, since  $n_1 = n_2 = 10$ , we select the value of the test statistic as the rank sum of sample 1, so  $W = W_1 = 149$ .

Table 20.10 shows a portion of lower  $W_L$  and upper  $W_U$  critical values for the Wilcoxon rank-sum test with  $n_1$  and  $n_2$  as the number of observations in the respective samples. (A more complete version is found in Table 7 of Appendix A.) Based on the specification of the hypothesis test, the rejection region will be in either one or both sides of the distribution.

**TABLE 20.10** Lower  $W_L$  and Upper  $W_U$  Critical Values for the Wilcoxon Rank-Sum Test

$\alpha = 0.025$ for one-tailed test and $\alpha = 0.05$ for a two-tailed test								
	$n_1$ : 3	4	5	6	7	8	9	10
$n_2$ : 8	8, 28	14, 38	21, 49	29, 61	39, 73	49, 87	51, 93	54, 98
9	8, 31	15, 41	22, 53	31, 65	41, 78	51, 93	63, 108	66, 114
10	9, 33	16, 44	24, 56	32, 70	43, 83	54, 98	66, 114	<b>79, 131</b>

### EXAMPLE 20.5

Use the salary data in Table 20.8 to determine whether the median computer science salary differs from the median finance salary at the 5% significance level.

**SOLUTION:** As discussed earlier, the competing hypotheses for the test are  $H_0: m_1 - m_2 = 0$  versus  $H_A: m_1 - m_2 \neq 0$ . For a two-tailed test with  $\alpha = 0.05$  and  $n_1 = n_2 = 10$ , the decision rule is to reject the null hypothesis if  $W \leq 79$  or  $W \geq 131$  (see the boldface values in Table 20.10).

Since the value of the test statistic  $W = W_1 = 149$  is greater than 131, we reject  $H_0$ . At the 5% significance level, we conclude that the median computer science salary differs from the median finance salary.

## Using a Normal Distribution Approximation for $W$

When  $n_1$  and  $n_2$  both have at least 10 observations, we can use the normal distribution approximation to implement a  $z$  test.

### EXAMPLE 20.6

Assuming that the distribution of  $W$  is approximately normal, let's again determine whether the median computer science salary differs from the median finance salary at the 5% significance level.

**SOLUTION:** We specify the same competing hypotheses,  $H_0: m_1 - m_2 = 0$  versus  $H_A: m_1 - m_2 \neq 0$ , and compute the value of the test statistic as  $W = W_1 = 149$ . We now compute the mean and the standard deviation as

$$\mu_W = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{10(10 + 10 + 1)}{2} = 105, \text{ and}$$
$$\sigma_W = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(10 \times 10)(10 + 10 + 1)}{12}} = 13.23.$$

The value of the test statistic is calculated as

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{149 - 105}{13.23} = 3.33.$$

Using the  $z$  table, we find the  $p$ -value as  $2 \times P(Z \geq 3.33) = 0.0009$ . Since the  $p$ -value is less than the significance level of  $\alpha = 0.05$ , we reject  $H_0$  and conclude, as before, that the median computer science salary differs from the median finance salary.

### Using the Computer for the Wilcoxon Rank-Sum Test

Table 20.11 shows a portion of the Minitab output for Example 20.6. (Minitab references the test as the Mann-Whitney Test.) The value of the test statistic  $W$  matches the value that we calculated by hand. Minitab refers to  $m_1$  and  $m_2$  as  $\eta_1$  and  $\eta_2$ , respectively, and provides a  $p$ -value of 0.0010. At the 5% significance level, we again conclude that the median salaries between the two majors differ.

**TABLE 20.11** Minitab Output for Example 20.6

$W = 149.0$
Test of $\eta_1 = \eta_2$ vs $\eta_1 \neq \eta_2$ is significant at 0.0010.
The test is significant at 0.0010 (adjusted for ties).

## EXERCISES 20.2

### Mechanics

12. Consider the following competing hypotheses and accompanying sample data drawn from a matched-pairs sample.

$$\begin{array}{lll} H_0: m_D \geq 0 & n = 9 & T^- = 40 \quad T^+ = 5 \\ H_A: m_D < 0 & & \end{array}$$

- a. Specify the decision rule at the 5% significance level.  
b. What is the conclusion?
13. Consider the following competing hypotheses and accompanying sample data drawn from a matched-pairs sample.

$$\begin{array}{lll} H_0: m_D = 0 & n = 50 & T^- = 400 \quad T^+ = 875 \\ H_A: m_D \neq 0 & & \end{array}$$

- a. Determine the value of the test statistic using a normal approximation for  $T$ .

- b. Calculate the  $p$ -value.  
c. At the 5% significance level, what is the conclusion?

14. **FILE Exercise 20.14.** The following table contains information on a matched-pairs sample.

Number	Sample 1	Sample 2
1	18	21
2	12	11
3	21	23
4	22	20
5	16	20
6	14	17
7	17	17
8	18	22

- Specify the competing hypotheses that determine whether the median difference between Population 1 and Population 2 is less than zero.
- With  $\alpha = 0.05$ , what is the decision rule?
- Calculate the value of the test statistic  $T$ .
- Is the median difference between Population 1 and Population 2 less than zero? Explain.

15. **FILE Exercise 20.15.** The following data are derived from a matched-pairs sample.

Observation	Sample 1	Sample 2
1	120	125
2	156	160
3	143	140
4	100	90
5	115	132
6	140	124
7	111	112
8	142	144
9	175	200
10	190	182

- Specify the competing hypotheses that determine whether the population median difference differs from zero.
  - Assuming that  $T$  is normally distributed, determine the value of the test statistic.
  - Calculate the  $p$ -value.
  - At the 1% significance level, what is the conclusion?
16. Consider the following competing hypotheses and accompanying data drawn from two independent populations.
- $$H_0: m_1 - m_2 = 0 \quad W_1 = 80 \quad W_2 = 40$$
- $$H_A: m_1 - m_2 \neq 0 \quad n_1 = 7 \quad n_2 = 8$$
- Specify the decision rule at the 5% significance level.
  - Determine the value of the test statistic.
  - What is the conclusion?
17. Consider the following competing hypotheses and accompanying data drawn from two independent populations.
- $$H_0: m_1 - m_2 \geq 0 \quad W_1 = 20 \quad W_2 = 35$$
- $$H_A: m_1 - m_2 < 0 \quad n_1 = 5 \quad n_2 = 5$$
- Specify the decision rule at the 5% significance level.
  - Determine the value of the test statistic.
  - What is the conclusion?
18. The following data were randomly drawn from two independent populations.

Sample 1	15	23	19	34	30	
Sample 2	28	25	34	35	37	40

- Specify the competing hypotheses to determine whether the median of Population 1 is less than the median of Population 2.
- With  $\alpha = 0.05$ , what is the decision rule?
- Calculate the value of the test statistic  $W$ .
- Is the median of Population 1 less than the median of Population 2? Explain.

19. The following data were randomly drawn from two independent populations.

Sample 1	-2	0	4	-5	2	
Sample 2	-3	-8	-1	0	-10	3

- Specify the competing hypotheses to determine whether the median of Population 1 is greater than the median of Population 2.
  - With  $\alpha = 0.05$ , what is the decision rule?
  - Calculate the value of the test statistic  $W$ .
  - Is the median of Population 1 greater than the median of Population 2? Explain.
20. The following data are provided for two samples drawn from independent populations:  $W = 700$ ,  $n_1 = 25$ , and  $n_2 = 20$ . Suppose the distribution of  $W$  is approximately normal.
- Calculate the mean and the standard deviation of the distribution of  $W$ .
  - Specify the competing hypotheses to determine whether the median of Population 1 is greater than the median of Population 2.
  - What is the decision rule with  $\alpha = 0.05$ ?
  - Calculate the value of the test statistic  $Z$ .
  - What is the conclusion?
21. The following data are provided for two samples drawn from independent populations:  $W = 545$ ,  $n_1 = 25$ , and  $n_2 = 25$ . Suppose the distribution of  $W$  is approximately normal.
- Calculate the mean and the standard deviation of the distribution of  $W$ .
  - Specify the competing hypotheses to determine whether the median of Population 1 differs from the median of Population 2.
  - Calculate the value of the test statistic  $Z$ .
  - Calculate the  $p$ -value.
  - At the 10% significance level, what is the conclusion?

## Applications

22. **FILE Mock SAT.** Suppose eight college-bound students take a mock SAT, complete a three-month test-prep course, and then take the real SAT. A portion of the data is shown in the accompanying table.

Student	Score on Mock SAT	Score on Real SAT
1	1830	1840
2	1760	1800
:	:	:
8	1710	1780

- Specify the competing hypotheses to determine whether the median score on the real SAT is greater than the median score on the mock SAT.
- At the 5% significance level, what is the decision rule?
- Calculate the value of the test statistic  $T$ .
- Is there sufficient evidence to conclude that the median score on the real SAT is greater than the median score on the mock SAT?

23. A diet center claims that it has the most effective weight loss program in the region. Its advertisement says "Participants in our program really lose weight." Five clients of this program are weighed on the first day of the diet and then three months later.

Client	Weight on First Day of Diet	Weight Three Months Later
1	155	151
2	205	203
3	167	168
4	186	183
5	194	195

- Specify the null and alternative hypotheses to test whether the median difference for weight loss supports the diet center's claim.
- At the 5% significance level, what is the decision rule?
- Calculate the value of the test statistic  $T$ .
- Do the data support the diet center's claim? Explain.

24. **FILE Appraisals.** A bank employs two appraisers. When approving borrowers for mortgages, it is imperative that the appraisers value the same types of properties consistently. To make sure that this is the case, the bank asks the appraisers to value 10 different properties. A portion of the data is shown in the following table.

Property	Value from Appraiser 1	Value from Appraiser 2
1	\$235,000	\$239,000
2	195,000	190,000
⋮	⋮	⋮
10	575,000	583,000

- Specify the competing hypotheses to determine whether the median difference between the values from appraiser 1 and appraiser 2 differs from zero.
- Calculate the value of the test statistic  $T$ . Assume the normal approximation for  $T$ .
- Calculate the  $p$ -value.
- At the 5% significance level, is there sufficient evidence to conclude that the appraisers are not consistent in their appraisals? Explain.

25. A professor teaches two sections of an introductory statistics course. He gives each section the same final and wonders if any significant differences exist between the medians of these sections. He randomly draws a sample of seven scores from Section A and six scores from Section B.

Section A	75	62	87	93	74	77	65
Section B	64	95	72	78	85	80	

- Set up the hypotheses to test the claim that the median test score in Section A differs from the median test score in Section B.
- With  $\alpha = 0.05$ , what is the decision rule?
- Calculate the value of the test statistic  $W$ .
- Do the median test scores differ? Explain.

26. A recent analysis of census data suggests that married men have a higher median income than unmarried men (*The Boston Globe*, January 19, 2010). Suppose the incomes (in \$1,000s) of six married men and seven unmarried men produce the following representative results:

Married	84	75	83	67	70	76	
Unmarried	67	63	62	66	71	64	68

- Set up the hypotheses to test the claim that the median income of married men is greater than the median income of unmarried men.
- With  $\alpha = 0.05$ , what is the decision rule?
- Calculate the value of the test statistic  $W$ .
- Is the claim supported by the data? Explain.

27. **FILE South Koreans.** According to the Organization of Economic Cooperation and Development, South Koreans spend more hours per year on the job than people in any other developed country (*The Wall Street Journal*, March 1, 2010). Suppose 10 workers in South Korea and 10 workers in the United States are asked to report the number of hours worked in the last year. The results are shown in the accompanying table.

South Korea	2624	1560	2698	2730	2879	3215	1753	2457	2669	2259
United States	2132	1432	1718	1456	2323	1795	2861	1600	1104	1041

- Set up the hypotheses to test the claim that the median annual hours worked in South Korea are greater than the median annual hours worked in the United States.
- Calculate the value of the test statistic  $W$ .
- Assume the normal approximation for  $W$ . With  $\alpha = 0.05$ , is the claim supported by the data? Explain.

28. **FILE Spending Gender.** Researchers at the Wharton School of Business have found that men and women shop for different reasons. While women enjoy the shopping experience, men are on a mission to get the job done. Men do not shop as frequently, but when they

do, they make big purchases like expensive electronics. The accompanying table shows a portion of the amount spent over the weekend by 40 men and 60 women at a local mall.

Spending by Men (\$)	Spending by Women (\$)
85	90
102	79
⋮	⋮

- Specify the competing hypotheses to determine whether the median amount spent by men is more than the amount spent by women.
- Calculate the value of the test statistic  $W$ . Assume the normal approximation for  $W$ .
- Calculate the  $p$ -value.
- At the 5% significance level, is there sufficient evidence to conclude that the median amount spent by men is greater than the median amount spent by women? Explain.

## 20.3 TESTING THREE OR MORE POPULATION MEDIANS

LO 20.5

In Chapter 13, we applied the one-way ANOVA  $F$  test to compare three or more population means. In order to implement this test, we assumed that for each population the variable of interest was normally distributed with the same variance. The **Kruskal-Wallis test** is a nonparametric alternative to the one-way ANOVA test that can be used when the assumptions of normality and/or equal population variances cannot be validated. It is based on ranks and is used for testing the equality of three or more population medians. Since the Kruskal-Wallis test is essentially an extension of the Wilcoxon rank-sum test, we discuss its application through an example.

Make inferences about the difference between three or more population medians.

### The Kruskal-Wallis Test

An undergraduate admissions officer would like to examine whether SAT scores differ by ethnic background. She collects a representative sample of SAT scores from Blacks, Hispanics, Whites, and Asians. Her results are shown in Table 20.12. She decides not to pursue the one-way ANOVA  $F$  test because she does not believe that the population variances are equal. Instead she chooses to apply the Kruskal-Wallis test.

**TABLE 20.12** SAT Scores by Ethnic Background

Blacks	1246	1148	1300	1404	1396	1450	
Hispanics	1267	1228	1450	1351	1280		
Whites	1581	1649	981	1877	1629	1800	1423
Asians	1623	1550	1936	1800	1750		

FILE  
KW\_SAT

Let  $m_1$ ,  $m_2$ ,  $m_3$ , and  $m_4$  denote the median SAT scores for Blacks, Hispanics, Whites, and Asians, respectively. We formulate the competing hypotheses as

$$H_0: m_1 = m_2 = m_3 = m_4$$

$H_A$ : Not all population medians are equal.

As in the Wilcoxon rank-sum test, we follow several steps to arrive at the value for the Kruskal-Wallis test statistic  $H$ .

- First, we pool the  $k$  independent samples (here,  $k = 4$ ) and then rank the observations from 1 to  $n$ . Since the total number of observations is 23, we rank the scores from 1 to 23. As before, if there are any ties, then we assign to each of the tied observations the mean of the ranks which they jointly occupy. In this sample, two individuals score 1450 and each is assigned the rank of 12.5, since the values jointly occupy the 12th and 13th ranks. Also, two individuals score 1800 and each is assigned the rank of 20.5. Table 20.13 shows the rank for each SAT score.

- B. We then calculate a ranked sum, denoted  $R_i$ , for each of the  $k$  samples. For instance, the ranked sum for Blacks is calculated as  $4 + 2 + 7 + 10 + 9 + 12.5 = 44.5$ . These sums are shown in the second-to-last row of Table 20.13.

**TABLE 20.13** Calculations for the Kruskal-Wallis Test

Blacks	Rank	Hispanics	Rank	Whites	Rank	Asians	Rank
1246	4	1267	5	1581	15	1623	16
1148	2	1228	3	1649	18	1550	14
1300	7	1450	12.5	981	1	1936	23
1404	10	1351	8	1877	22	1800	20.5
1396	9	1280	6	1629	17	1750	19
1450	12.5			1800	20.5		
				1423	11		
$R_1 = 44.5$		$R_2 = 34.5$		$R_3 = 104.5$		$R_4 = 92.5$	
$\frac{R_1^2}{n_1} = 330.0$		$\frac{R_2^2}{n_2} = 238.1$		$\frac{R_3^2}{n_3} = 1560.0$		$\frac{R_4^2}{n_4} = 1711.3$	

If median SAT scores across ethnic groups are the same, we expect the ranked sums to be relatively close to one another. However, if some sums deviate substantially from others, then this is evidence that not all population medians are the same. We determine whether the variability of some ranked sums differs significantly from others by first calculating the value of the test statistic  $H$ .

#### THE TEST STATISTIC $H$ FOR THE KRUSKAL-WALLIS TEST

The **test statistic  $H$**  for the Kruskal-Wallis test is defined as

$$H = \left( \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} \right) - 3(n+1),$$

where  $R_i$  and  $n_i$  are the rank sum and the size of the  $i$ th sample,  $n = \sum_{i=1}^k n_i$ , and  $k$  is the number of populations (independent samples). If  $n_i \geq 5$  for  $i = 1, 2, \dots, k$ , then  $H$  can be approximated by the  $\chi^2$  distribution with  $k - 1$  degrees of freedom.

For small sample values ( $n_i < 5$ ), the test may be based on special tables; however, we will not pursue that case. The Kruskal-Wallis test is always a right-tailed test.

#### EXAMPLE 20.7

Use the data in Table 20.13 to determine whether some median SAT scores differ by ethnic background at the 5% significance level.

**SOLUTION:** As discussed earlier, the appropriate hypotheses for the test are

$$H_0: m_1 = m_2 = m_3 = m_4$$

$$H_A: \text{Not all population medians are equal.}$$

With  $\alpha = 0.05$  and  $k = 4$ , so that degrees of freedom equal  $df = k - 1 = 3$ , we reference the chi-square table and find a critical value of  $\chi_{0.05,3}^2 = 7.815$ . The decision rule is to reject  $H_0$  if  $H > 7.815$ . We compute the value of the test statistic  $H$  as

$$\begin{aligned} H &= \left( \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} \right) - 3(n+1) \\ &= \left( \frac{12}{23(23+1)} (330.0 + 238.1 + 1560.0 + 1711.3) \right) - 3(23+1) \\ &= 83.5 - 72 = 11.5 \end{aligned}$$



Since the value of the test statistic,  $H = 11.5$ , is greater than the critical value of 7.815, we reject  $H_0$ . At the 5% significance level, not all median SAT scores across ethnic groups are the same.

## Using the Computer for the Kruskal-Wallis Test

Table 20.14 shows a portion of the Minitab output as applied to the SAT example. The value of the test statistic computed by Minitab is the same value as the one that we calculated by hand ( $H = 11.48$ ). In addition, Minitab reports the  $p$ -value as 0.009; thus, given that 0.009 is less than the significance level of 0.05, we again reject  $H_0$  and conclude that not all median scores are the same.

**TABLE 20.14** Minitab Output for Example 20.7

Kruskal-Wallis Test on SAT		
$H = 11.48$	$DF = 3$	$P = 0.009$ (adjusted for ties)

## EXAMPLE 20.8

A sociologist suspects differences in median incomes in three major eastern cities. He randomly samples 100 workers from each city. The top of Table 20.15 shows a portion of the data. Unsure that the median income is the same in each city, he uses Minitab to conduct a Kruskal-Wallis test and produces the results shown at the bottom of Table 20.15. At the 5% significance level, can he conclude that some differences exist in the median incomes in these three cities?

**TABLE 20.15** Data and Minitab Output for Example 20.8

City 1	City 2	City 3
86.5	88.1	102.4
76.7	73.3	127.9
$\vdots$	$\vdots$	$\vdots$
85	63.3	80.1
Kruskal-Wallis Test on Income		
$H = 85.50$ $DF = 2$ $P = 0.000$ (adjusted for ties)		

**FILE**  
City\_Income

**SOLUTION:** We let  $m_1$ ,  $m_2$ , and  $m_3$  denote the population median incomes for City 1, City 2, and City 3, respectively, and formulate the competing hypotheses as

$$H_0: m_1 = m_2 = m_3$$

$H_A$ : Not all population median incomes are equal.

The value of the test statistic computed by Minitab is  $H = 85.50$  with an associated  $p$ -value  $\approx 0$ . Since the  $p$ -value is less than the significance level of 0.05, we reject  $H_0$  and conclude that not all median incomes across the three cities are the same.

It is important to note that if we reject the null hypothesis, we can only conclude that not all population medians are equal. The Kruskal-Wallis test does not allow us to infer which medians differ. Further analysis of the difference between population medians is beyond the scope of this text.

## EXERCISES 20.3

### Mechanics

29. Consider the following sample information:  $k = 3$  and  $H = 4.5$ .
- Specify the competing hypotheses to test whether some differences exist between the medians.
  - At the 10% significance level, specify the critical value and the decision rule.
  - Do some medians differ? Explain.
30. Consider the following sample information:  $k = 5$  and  $H = 12.4$ .
- Specify the competing hypotheses to test whether some differences exist between the medians.
  - Approximate the  $p$ -value.
  - At the 5% significance level, do some medians differ? Explain.
31. **FILE Exercise 20.31.** Random samples were drawn from three independent populations. The results are shown in the accompanying table.

Sample 1	120	95	115	110	90	
Sample 2	100	85	105	80	75	90
Sample 3	72	65	100	76	66	55

- Specify the competing hypotheses to test whether some differences exist between the medians.
  - Calculate the value of the test statistic  $H$ .
  - At the 10% significance level, what is the critical value?
  - Do some medians differ? Explain.
32. **FILE Exercise 20.32.** Random samples were drawn from four independent populations. The results are shown in the accompanying table.

Sample 1	-10	-15	0	5	10
Sample 2	-2	-3	-4	-5	-6
Sample 3	2	4	6	8	10
Sample 4	5	7	9	11	13

- Specify the competing hypotheses to test whether some differences exist between the medians.
- Calculate the value of the test statistic  $H$ .
- Approximate the  $p$ -value.
- At the 1% significance level, do some medians differ? Explain.

### Applications

33. **FILE Unemployment.** A research analyst wants to test whether the median unemployment rate differs from one region of the country to another. She collects the

unemployment rate (in percent) of similar-sized cities in three regions of the United States. The results are shown in the accompanying table.

Region A	12.5	13.0	8.5	10.7	9.3
Region B	9.0	9.5	7.0	6.7	8.2
Region C	8.2	7.4	10.9	11.1	10.4

- Specify the competing hypotheses to test whether some differences exist in the median unemployment rates between the three regions.
  - Calculate the value of the test statistic  $H$ .
  - At the 10% significance level, what is the critical value?
  - Do some median unemployment rates differ by region? Explain.
34. **FILE Bulb.** A quality-control manager wants to test whether there is any difference in the median length of life of light bulbs between three different brands. Random samples were drawn from each brand where the duration of each light bulb (in hours) was measured. The results are shown in the accompanying table.

Brand 1	Brand 2	Brand 3
375	280	350
400	290	415
425	300	425
410	325	380
420	350	405

- Specify the competing hypotheses to test whether some differences exist in the median length of life of light bulbs between the three brands.
  - Calculate the value of the test statistic  $H$ .
  - Approximate the  $p$ -value.
  - At the 10% significance level, do some differences exist between the median length of life of light bulbs by brand? Explain.
35. **FILE Industry Returns.** A research analyst examines annual returns (in percent) for Industry A, Industry B, and Industry C, as shown in the accompanying table.

Industry A	Industry B	Industry C
16.86	15.41	13.53
5.11	10.87	9.58
12.45	4.43	18.75
-32.44	-18.45	-28.77
32.11	20.96	32.17

- Specify the competing hypotheses to test whether some differences exist in the median returns by industry.

- b. Calculate the value of the test statistic  $H$ .
- c. At the 10% significance level, what is the critical value?
- d. Do some differences exist between the median returns by industry? Explain.

36. **FILE Detergents.** A well-known conglomerate claims that its detergent “whitens and brightens better than all the rest.” In order to compare the cleansing action of the top three detergents, 15 swatches of white cloth were soiled with red wine and grass stains and then washed in front-loading machines with the respective detergents. The following whiteness readings are shown in the accompanying table.

Detergent 1	Detergent 2	Detergent 3
84	78	87
79	74	80
87	81	91
85	86	77
94	86	78

- a. Specify the competing hypotheses to test whether some differences exist in the median cleansing action of the three detergents.
  - b. Calculate the value of the test statistic  $H$ .
  - c. Approximate the  $p$ -value.
  - d. At the 1% significance level, do some differences exist between the median cleansing action by detergent? Explain.
37. **FILE Exam Scores.** A statistics instructor wonders whether significant differences exist in her students’

median exam scores in her three different sections. She randomly selects scores from 10 different students in each section. A portion of the data is shown in the accompanying table.

Section 1	Section 2	Section 3
85	91	74
68	84	69
:	:	:
74	75	73

Do these data provide enough evidence at the 5% significance level to indicate that there are some differences in median scores in the three sections?

38. **FILE Job Satisfaction.** A human resource specialist wants to determine whether the median job satisfaction score (on a scale of 0 to 100) differs depending on a person’s field of employment. She collects scores from 30 employees in three different fields. A portion of the data is shown in the accompanying table.

Field 1	Field 2	Field 3
80	76	81
76	73	77
:	:	:
79	67	80

At the 10% significance level, can we conclude that there are some differences in job satisfaction depending on field of employment?

## 20.4 TESTING THE CORRELATION BETWEEN TWO VARIABLES

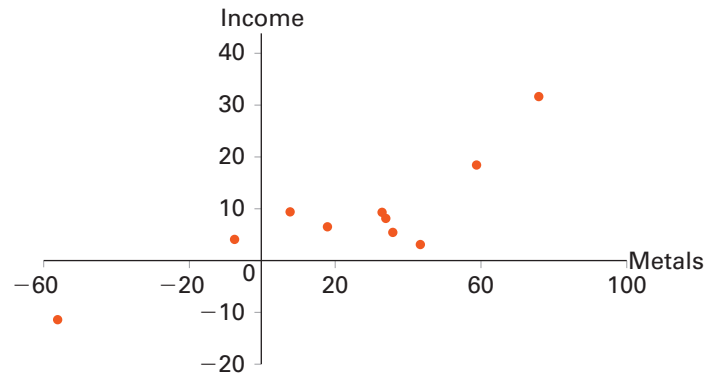
LO 20.6

In earlier chapters, we used the correlation coefficient, also referred to as the Pearson correlation coefficient, to measure the strength and direction of the linear relationship between two random variables. Recall that the value of this correlation coefficient falls between  $-1$  and  $+1$ ; as its absolute value approaches one, the linear relationship becomes stronger. We used a  $t$  test to determine whether the population correlation coefficient differs from zero, which assumes that we sample from normal populations. Since this assumption breaks down in some situations, we need a nonparametric alternative. The **Spearman rank correlation test** serves as this option. The Spearman rank correlation coefficient measures the correlation between two random variables based on rank orderings. Its value also falls between  $-1$  and  $+1$  and is interpreted in the same way as the Pearson correlation coefficient.

Figure 20.2 shows a scatterplot of the return data for the Metals and Income funds from the introductory case. Each point in the scatterplot represents a pairing of each fund’s return for a given year. It appears that the two funds are positively related.

Conduct a hypothesis test for the population Spearman rank correlation coefficient.

**FIGURE 20.2**  
Scatterplot of return  
data for the Metals  
and Income funds



Suppose we want to determine whether the observed relationship is real or due to chance. As we noted before, return data often do not follow the normal distribution. Therefore, using the  $t$  test to analyze the Pearson correlation coefficient is not appropriate. Instead, we let  $\rho_s$  denote the population Spearman rank correlation coefficient, and we formulate a two-tailed test as

$$H_0: \rho_s = 0$$

$$H_A: \rho_s \neq 0$$

To conduct the test, we first calculate the sample Spearman rank correlation coefficient  $r_s$  using the following steps.

- A.** We rank the observations from the Metals fund from smallest to largest. In the case of ties, we assign to each tied observation the average of the ranks that they jointly occupy. We perform the same procedure for the Income fund. Columns 2 and 3 of Table 20.16 show the original return data, and columns 4 and 5 show the funds' ranked values.

**TABLE 20.16** Calculations for the Spearman Rank Correlation Coefficient

Year (1)	Metals $x$ (2)	Income $y$ (3)	Rank for Metals (4)	Rank for Income (5)	Difference $d$ (6)	Difference Squared $d^2$ (7)
2000	-7.34	4.07	2	3	-1	1
2001	18.33	6.52	4	5	-1	1
2002	33.35	9.38	5	7	-2	4
2003	59.45	18.62	9	9	0	0
2004	8.09	9.44	3	8	-5	25
2005	43.79	3.12	8	2	6	36
2006	34.30	8.15	6	6	0	0
2007	36.13	5.44	7	4	3	9
2008	-56.02	-11.37	1	1	0	0
2009	76.46	31.77	10	10	0	0
					$\Sigma d_i = 0$	$\Sigma d_i^2 = 76$

- B.** We calculate the difference  $d_i$  between the ranks of each pair of observations. See column 6 of Table 20.16. As a check, when we sum the differences,  $\Sigma d_i$ , we should obtain zero.
- C.** We then sum the squared differences. The resulting value is shown in the last cell of column 7 in Table 20.16; that is,  $\Sigma d_i^2 = 76$ .

### THE SPEARMAN RANK CORRELATION COEFFICIENT

The sample **Spearman rank correlation coefficient**  $r_s$  between two variables  $x$  and  $y$  is defined as

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where  $d_i$  is the difference between the ranks of observations  $x_i$  and  $y_i$ .

There are two scenarios when conducting the Spearman rank correlation test:

1. If  $n \leq 10$ ,  $|r_s|$  is compared against a positive critical value(s) as shown in Table 8 of Appendix A.
2. If  $n \geq 10$ ,  $r_s$  can be approximated by the normal distribution with zero mean and standard deviation of  $\sqrt{\frac{1}{n-1}}$ , and hence the value of the resulting test statistic is computed as  $z = r_s \sqrt{n-1}$ .

For the sample Spearman correlation coefficient between the Metals and Income funds, we calculate  $r_s$  as

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 76}{10 \times (10^2 - 1)} = 1 - 0.46 = 0.54.$$

A value of  $r_s = 0.54$  implies that the Metals and Income funds have a positive, rather moderate, relationship.

For  $n \leq 10$ , the rank correlation test is based on special tables determined from the exact distribution of  $r_s$ . Table 20.17 shows a portion of upper critical values for one- and two-tailed tests concerning  $\rho_s$  at various  $\alpha$  values. (A more complete version is found in Table 8 of Appendix A.) For a two-tailed test with  $\alpha = 0.05$  and  $n = 10$ , the decision rule is to reject the null hypothesis if  $|r_s| > 0.648$ . See the relevant critical value in boldface in Table 20.17.

**TABLE 20.17** Upper Critical Values for Testing the Spearman Rank Correlation Coefficient

Two-Tailed Test: One-Tailed Test:	$\alpha = 0.10$ $\alpha = 0.05$	$\alpha = 0.05$ $\alpha = 0.025$	$\alpha = 0.02$ $\alpha = 0.01$	$\alpha = 0.01$ $\alpha = 0.005$
$n = 8$	0.643	0.738	0.833	0.881
9	0.600	0.683	0.783	0.833
10	0.564	<b>0.648</b>	0.745	0.794

Since the value of the test statistic,  $r_s = 0.54$ , is not greater than 0.648, we cannot reject  $H_0$ . At the 5% significance level, we cannot conclude that the rank correlation between the Metals fund and the Income fund differs from zero.

### Using a Normal Distribution Approximation for $r_s$

As mentioned above, when  $n \geq 10$ , we can also use the normal distribution approximation to implement a  $z$  test.

#### EXAMPLE 20.9

A sports statistician would like to analyze the relationship between a quarterback's salary (in \$ millions) and his age based on the Spearman rank correlation coefficient. He collects data on 32 quarterbacks and uses Minitab to conduct his analysis. The top half of Table 20.18 shows a portion of the data and the lower half shows the Minitab output. Can he conclude that salary and age are correlated at the 5% significance level?

**TABLE 20.18** Data and JMP Output for Example 20.9

Salary	Age
25.5566	27
22.0441	26
⋮	⋮
0.6260	29
Spearman's rho = 0.375	

**SOLUTION:** To determine whether salary and age are correlated, we formulate the competing hypotheses as

$$H_0: \rho_s = 0$$

$$H_A: \rho_s \neq 0$$

Referencing Table 20.18, we find that  $r_s = 0.375$ . Under normality, the value of the corresponding test statistic is calculated as  $z = 0.375 \sqrt{32 - 1} = 2.09$ , which yields a  $p$ -value of  $2 \times P(Z \geq 2.09) = 0.037$ . Since the  $p$ -value is less than the significance level of 5%, we reject  $H_0$  and conclude that salary and age are correlated.

### Summary of Parametric and Nonparametric Tests

Table 20.19 summarizes the select parametric tests referenced in Section 20.1 and their nonparametric counterparts. Nonparametric tests use fewer and weaker assumptions than those associated with parametric tests and are especially attractive when the underlying population is markedly nonnormal. However, a nonparametric test ignores useful information since it focuses on the rank rather than the magnitude of sample values. Therefore, in situations when the parametric assumptions are valid, the nonparametric test is less powerful than its parametric counterpart. In general, when the assumptions for a parametric test are met, it is preferable to use a parametric test rather than a nonparametric test. Since the normality assumption for parametric tests is less stringent in large samples, the main appeal of nonparametric tests tends to be with relatively small samples.

**TABLE 20.19** Parametric Test versus Nonparametric Alternative

Parametric Test	Nonparametric Alternative
$t$ test concerning the population mean	Wilcoxon signed-rank test concerning the population median
$t$ test to determine whether the population mean difference differs from zero based on matched-pairs sampling	Wilcoxon signed-rank test to determine whether the population median difference differs from zero based on matched-pairs sampling
$t$ test to determine whether two population means differ based on independent sampling	Wilcoxon rank-sum test to determine whether two population medians differ based on independent sampling
$F$ test to determine whether three or more population means differ	Kruskal-Wallis test to determine whether three or more population medians differ
$t$ test to determine whether two variables are correlated	Spearman rank correlation test to determine whether two variables are correlated



## SYNOPSIS OF INTRODUCTORY CASE

Vanguard's Precious Metals and Mining fund (referred to as Metals) and Fidelity's Strategic Income fund (referred to as Income) were two top-performing mutual funds from 2000–2009. Annual return data for the funds were obtained from <http://finance.yahoo.com>. Given that return data often have “fatter tails” than the normal distribution, the analysis focuses on nonparametric techniques. These techniques do not rely on the normality assumption concerning the underlying population. When applying the Wilcoxon signed-rank test at the 5% significance level, it is found that the median return for the Metals fund is not significantly greater than 5%. The same test is also used to conclude that the median difference between the Metals and Income returns does not differ from zero at the 5% significance level. Finally, the sample Spearman rank correlation coefficient is calculated as 0.54, implying a moderate, positive relationship between the returns of the two funds. However, a test conducted at the 5% significance level finds that the population Spearman rank correlation coefficient is not significantly different from zero. Interestingly, the Metals and the Income fund are not related, despite the fact that they are both influenced by underlying market conditions.



## EXERCISES 20.4

### Mechanics

39. Consider the following competing hypotheses and accompanying sample data.

$$\begin{aligned} H_0: \rho_S &= 0 & r_S &= 0.92 \text{ and } n = 8 \\ H_A: \rho_S &\neq 0 \end{aligned}$$

- Specify the decision rule at the 1% significance level.
- What is the value of the test statistic?
- What is the conclusion?

40. Consider the following competing hypotheses and accompanying sample data.

$$\begin{aligned} H_0: \rho_S &\geq 0 & r_S &= 0.64 \text{ and } n = 9 \\ H_A: \rho_S &< 0 \end{aligned}$$

- Specify the decision rule at the 5% significance level.
- What is the value of the test statistic?
- What is the conclusion?

41. Consider the following sample data:

x	12	18	20	22	25	15
y	15	20	25	22	27	19

- Calculate and interpret  $r_s$ .
- Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient differs from zero.

- At the 5% significance level, what is the decision rule?
- What is the conclusion?

42. Consider the following sample data:

x	-2	0	3	-1	4	7
y	-4	-3	-8	-5	-9	-10

- Calculate and interpret  $r_s$ .
- Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient is less than zero.
- At the 1% significance level, what is the decision rule?
- What is the conclusion?

43. Consider the following competing hypotheses and accompanying sample data.

$$\begin{aligned} H_0: \rho_S &= 0 & r_S &= 0.85 \text{ and } n = 65 \\ H_A: \rho_S &\neq 0 \end{aligned}$$

- What is the value of the test statistic and its associated  $p$ -value? Assume the normal approximation for  $r_s$ .
- At the 10% significance level, what is the conclusion?

44. Consider the following competing hypotheses and accompanying sample data.

$$H_0: \rho_s \leq 0 \quad r_s = 0.64 \text{ and } n = 50$$

$$H_A: \rho_s > 0$$

- What is the value of the test statistic and its associated  $p$ -value? Assume the normal approximation for  $r_s$ .
- At the 1% significance level, what is the conclusion?

## Applications

45. The following table shows the ranks given by two judges to the performance of six finalists in a men's figure skating competition:

Skater	A	B	C	D	E	F
Judge 1	6	5	4	3	2	1
Judge 2	5	6	4	3	1	2

- Calculate and interpret the Spearman rank correlation coefficient  $r_s$ .
  - Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient is different from zero.
  - At the 10% significance level, specify the critical value and the decision rule.
  - What is the conclusion to the test? Are the ranks significantly related?
46. **FILE WB Ranking.** The following table shows the World Bank's 2008 ranking of the richest countries, as measured by per capita GNP. In addition, it gives each country's respective rank with respect to infant mortality according to the Central Intelligence Agency. A higher rank indicates a lower mortality rate.

Country	Per Capita GNP Rank	Infant Mortality Rank
Luxembourg	1	7
Norway	2	4
Singapore	3	1
United States	4	10
Ireland	5	9
Switzerland	6	5
Netherlands	7	8
Austria	8	6
Sweden	9	2
Iceland	10	3

- Calculate and interpret the Spearman rank correlation coefficient  $r_s$ .
- Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient is different from zero.
- At the 5% significance level, specify the critical value and the decision rule.
- Are GNP and the infant mortality rate correlated? Explain.

47. You are interested in whether the returns on Asset A are negatively correlated with the returns on Asset B. Consider the following annual return data on the two assets:

	Asset A	Asset B
Year 1	-20%	8%
Year 2	-5	5
Year 3	18	-1
Year 4	15	-2
Year 5	4	3
Year 6	-12	2

- Calculate and interpret the Spearman rank correlation coefficient  $r_s$ .
  - Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient is less than zero.
  - At the 1% significance level, specify the critical value.
  - What is the conclusion to the test? Are the returns negatively correlated?
48. **FILE Price\_Days.** In an attempt to determine whether a relationship exists between the price of a home and the number of days it takes to sell the home, a real estate agent collected the following data from recent sales in his city.

Price (in \$1,000s)	Days to Sell Home
265	136
225	125
160	120
325	140
430	145
515	121
180	122
423	145

- Calculate and interpret the Spearman rank correlation coefficient  $r_s$ .
  - Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient differs from zero.
  - At the 5% significance level, specify the critical value and the decision rule.
  - Can you conclude that the price of a home and the number of days it takes to sell the home are correlated? Explain.
49. **FILE GRE\_GPA.** The director of graduate admissions at a local university is analyzing the relationship between scores on the Graduate Record Examination (GRE) and subsequent performance in graduate school,

as measured by a student's grade point average (GPA). She uses a sample of 7 students who graduated within the past five years, as shown in the accompanying table.

Student	1	2	3	4	5	6	7
GRE	1500	1400	1000	1050	1100	1250	800
GPA	3.4	3.5	3.0	2.9	3.0	3.3	2.7

- Calculate and interpret the Spearman rank correlation coefficient  $r_s$ .
  - Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient is greater than zero.
  - At the 5% significance level, specify the critical value and the decision rule.
  - Are GRE and GPA positively correlated? Explain.
50. A social scientist analyzes the relationship between educational attainment and salary. For 65 individuals he collects data on each individual's educational attainment (in years) and his/her salary (in \$1,000s). He then calculates a Spearman rank correlation coefficient of 0.85.
- Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient differs from zero.
  - Assume that the distribution of  $r_s$  is approximately normal. Calculate the value of the test statistic and the  $p$ -value of the test.
  - At the 5% significance level, are educational attainment and salary correlated?
51. An engineer examines the relationship between the weight of a car and its average miles per gallon (MPG). For a sample of 100 cars, he calculates a Spearman rank correlation coefficient of  $-0.60$ .
- Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient is less than zero.

- Assume that the distribution of  $r_s$  is approximately normal. Calculate the value of the test statistic and the  $p$ -value.
- At the 5% significance level, does a negative relationship exist between a car's weight and its average MPG?

52. **FILE Happiness.** Many attempts have been made to relate happiness with various factors. One such study relates happiness with age and finds that, holding everything else constant, people are least happy when they are in their mid-40s (*The Economist*, December 16, 2010). The accompanying table shows a portion of data on a respondent's age and his/her perception of well-being on a scale from 0 to 100.

Age	Happiness
49	62
51	66
:	:
69	72

Using the Spearman rank correlation coefficient, determine whether age and happiness are positively correlated at the 5% significance level.

53. **FILE Gambling.** The accompanying table shows a portion of the number of cases of crime related to gambling and offenses against the family and children for the 50 states in the United States during 2010.

State	Gambling	Family Abuse
Alabama	47	1,022
Alaska	10	315
:	:	:
Wyoming	0	194

Using the Spearman rank correlation coefficient, determine whether gambling and family abuse are correlated at the 5% significance level.

## 20.5 THE SIGN TEST

### LO 20.7

In some applications, a matched-pairs sample originates from ordinal data rather than from interval- or ratio-scaled data. Let's review the definition of ordinal data first introduced in Chapter 1. With ordinal data we are able to categorize and rank the data with respect to some characteristic or trait. The weakness with ordinal-scaled data is that we cannot interpret the difference between the ranked values because the actual numbers used are arbitrary. For example, suppose you are asked to classify the service at a particular hotel as excellent, good, fair, or poor. A standard way to record the ratings is

Excellent	4	Fair	2
Good	3	Poor	1

Make inferences about the difference between two populations of ordinal data based on matched-pairs sampling.

Here the value attached to excellent (4) is higher than the value attached to good (3), indicating that the response of excellent is preferred to good. However, another representation of the ratings might be

Excellent	100	Fair	70
Good	80	Poor	40

Excellent still receives a higher value than good, but now the difference between the two categories is 20 (100–80), as compared to a difference of 1 (4–3) when we use the first classification. In other words, differences between categories are meaningless with ordinal data.

If we have a matched-pairs sample of ordinal data, we can use the **sign test** to determine whether there are significant differences between the populations. When applying the sign test, we are only interested in whether the difference between two values in a pair is greater than, equal to, or less than zero. The difference between each pairing is replaced by a plus sign (+) if the difference is positive (that is, the first value exceeds the second value) or by a minus sign (–) if the difference between the pair is negative. If the difference between the pair is zero, we discard that particular observation from the sample.

If significant differences do not exist between the two populations, then we expect just as many plus signs as minus signs. Equivalently, we should observe plus signs 50% of the time and minus signs 50% of the time. Suppose we let  $p$  denote the population proportion of plus signs. (We could just as easily allow  $p$  to represent the population proportion of minus signs without loss of generality.) The competing hypotheses for the sign test take one of the following forms.

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: p = 0.50$	$H_0: p \leq 0.50$	$H_0: p \geq 0.50$
$H_A: p \neq 0.50$	$H_A: p > 0.50$	$H_A: p < 0.50$

A two-tailed test allows us to determine whether the proportion of plus signs differs from the proportion of minus signs. A right-tailed (left-tailed) test allows us to determine whether the proportion of plus signs is greater than (less than) the proportion of minus signs.

Let  $\bar{P} = X/n$  be the estimator of the population proportion of plus signs. As discussed in Chapter 7, if  $np$  and  $n(1 - p)$  are both 5 or more, then the distribution of  $\bar{P}$  is approximately normal, with mean  $E(\bar{P}) = p$  and standard error  $se(\bar{P}) = \sqrt{p(1 - p)/n}$ . Assuming a probability of success  $p = 0.50$ ,  $se(\bar{P}) = 0.5/\sqrt{n}$  and the normal distribution approximation is satisfactory as long as  $n \geq 10$ . When  $n < 10$ , we rely on the binomial distribution to conduct the sign test; we will not consider such cases.

#### THE TEST STATISTIC FOR THE SIGN TEST

The value of the test statistic for the sign test is computed as  $z = \frac{\bar{p} - 0.5}{0.5/\sqrt{n}}$ , where  $\bar{p} = x/n$  is the sample proportion of plus signs. The test is valid when  $n \geq 10$ .

#### EXAMPLE 20.10

In December 2009, Domino's Pizza, Inc. released untraditional ads citing that its old recipe for pizza produced crust that tasted like cardboard and sauce that tasted like ketchup. Domino's Pizza claims that its reformulated pizza is a vast improvement over the old recipe; for instance, garlic and parsley are now baked into the crust and a new sweeter, bolder tomato sauce is used. Suppose 20 customers are asked to sample the old recipe and then sample the new recipe. Each person is asked to rate the pizzas

on a 5-point scale, where 1 = inedible and 5 = very tasty. The ratings are shown in Table 20.20. Do these data provide sufficient evidence to allow us to conclude that the new recipe is preferred to the old recipe? Use  $\alpha = 0.05$ .

**TABLE 20.20** Calculations for Sign Test in Example 20.10

Customer	Old Recipe	New Recipe	Sign	Customer	Old Recipe	New Recipe	Sign
1	3	4	–	11	3	4	–
2	3	2	+	12	4	5	–
3	2	5	–	13	1	2	–
4	4	4	0	14	3	3	0
5	2	5	–	15	5	3	+
6	1	3	–	16	3	4	–
7	3	2	+	17	1	5	–
8	1	2	–	18	4	2	+
9	2	4	–	19	3	4	–
10	4	5	–	20	2	5	–

**SOLUTION:** If customers feel that there is no difference between the old recipe and the new recipe, then we expect 50% of the customers to prefer the old recipe and 50% to prefer the new recipe. Let  $p$  denote the population proportion of consumers who prefer the old recipe. We want to specify the competing hypotheses such that rejection of the null hypothesis provides evidence that customers prefer the new recipe (implying that  $p$  is significantly less than 0.50). We set up the competing hypotheses as

$$H_0: p \geq 0.50$$

$$H_A: p < 0.50$$

Table 20.20 shows the signs for each customer. For example, customer 1 ranks the old recipe with the value 3 and the new recipe with the value 4, which yields a minus sign when the difference between the old recipe and the new recipe is calculated:  $3 - 4 = -1$ . This difference implies that this customer prefers the new recipe. We find 4 positive signs, 14 negative signs, and 2 ties (ranks of zero). We then let  $n$  denote the number of matched-paired observations such that the sign between the rankings is nonzero; thus,  $n$  equals 18. We denote  $\bar{p}$  as the sample proportion of plus signs. Given that there are four plus signs, the sample proportion is calculated as  $\bar{p} = 4/18 = 0.22$ . (Note that if we had calculated the sample proportion of minus signs, then  $\bar{p} = 0.78$ ; the resulting value of the test statistic only differs in its sign. In this instance, we would conduct a right-tailed hypothesis test.) We calculate the value of the test statistic as

$$z = \frac{\bar{p} - 0.5}{0.5/\sqrt{n}} = \frac{0.22 - 0.5}{0.5/\sqrt{18}} = \frac{-0.28}{0.118} = -2.37.$$

Using the  $z$  table, we find the  $p$ -value for a left-tailed test as  $P(Z \leq -2.37) = 0.0089$ . Since the  $p$ -value is less than the significance level of  $\alpha = 0.05$ , we reject  $H_0$  and conclude that consumers prefer the reformulated version as compared to the old recipe at the 5% significance level.

The sign test can be used with quantitative as well as ordinal data. However, since the sign test ignores the magnitude in the difference between two observations, it is advisable to use the Wilcoxon signed-rank test if quantitative data are available.

## EXERCISES 20.5

### Mechanics

54. Consider the following competing hypotheses and sample data.

$$H_0: p = 0.50 \quad n = 40 \quad \bar{p} = 0.30$$

$$H_A: p \neq 0.50$$

- Calculate the value of the test statistic for the sign test.
  - Calculate the  $p$ -value.
  - At the 5% significance level, what is the conclusion? Explain.
55. Consider the following competing hypotheses and sample data.

$$H_0: p \leq 0.50 \quad n = 25 \quad \bar{p} = 0.64$$

$$H_A: p > 0.50$$

- At the 1% significance level, what is the decision rule?
  - Calculate the value of the test statistic for the sign test.
  - What is the conclusion? Explain.
56. Consider the following sign data, produced from a matched-pairs sample of ordinal data.

+ + + - + + - + + + + - + + - - + + + +

- Specify the competing hypotheses to determine whether the proportion of negative signs differs from the proportion of positive signs.
  - Calculate the value of the test statistic.
  - Calculate the  $p$ -value.
  - At the 5% significance level, what is the conclusion? Explain.
57. Consider the following sign data, produced from a matched-pairs sample of ordinal data.

+ - - + - - + - + - + - - - + -

- Specify the competing hypotheses to determine whether the proportion of negative signs is significantly greater than the proportion of positive signs.
- At the 1% significance level, what is the decision rule?
- Calculate the value of the test statistic.
- What is the conclusion? Explain.

### Applications

58. **FILE Water.** Concerned with the increase of plastic water bottles in landfills, a leading environmentalist wants to determine whether there is any difference in taste between the local tap water and the leading bottled water. She randomly selects 14 consumers and conducts a blind taste test. She asks the consumers to rank the taste on a scale of one to five (with "five" indicating excellent taste). The sample results are shown in the accompanying table.

| Consumer | Tap Water | Bottled Water | Consumer | Tap Water | Bottled Water |
|----------|-----------|---------------|----------|-----------|---------------|
| 1        | 4         | 5             | 8        | 5         | 2             |
| 2        | 3         | 2             | 9        | 3         | 4             |
| 3        | 5         | 4             | 10       | 2         | 4             |
| 4        | 4         | 3             | 11       | 5         | 4             |
| 5        | 3         | 5             | 12       | 4         | 3             |
| 6        | 5         | 3             | 13       | 5         | 2             |
| 7        | 2         | 1             | 14       | 3         | 4             |

- Using the sign test, specify the competing hypotheses to determine whether there are significant differences in preferences between tap water and bottled water.
  - Calculate the value of the test statistic.
  - Calculate the  $p$ -value.
  - At the 5% significance level, what is the conclusion? Do the results indicate that significant differences exist in preferences?
59. In March 2009, 100 registered voters were asked to rate the "effectiveness" of President Obama. In March 2010, these same people were again asked to make the same assessment. Seventy percent of the second ratings were lower than the first ratings and 30% were higher.
- Using the sign test, specify the competing hypotheses to determine whether the President's rating has significantly declined.
  - Calculate the value of the test statistic.
  - Calculate the  $p$ -value.
  - At the 5% significance level, do the data suggest that the President's rating has significantly declined?
60. **FILE PhD\_Rating.** For scholarship purposes, two graduate faculty members rate 12 applicants to the PhD program on a scale of 1 to 10 (with 10 indicating an excellent candidate). These ratings are shown in the following table.

| Candidate | Faculty A's Rating | Faculty B's Rating | Candidate | Faculty A's Rating | Faculty B's Rating |
|-----------|--------------------|--------------------|-----------|--------------------|--------------------|
| 1         | 5                  | 6                  | 7         | 2                  | 2                  |
| 2         | 7                  | 8                  | 8         | 8                  | 9                  |
| 3         | 8                  | 5                  | 9         | 9                  | 10                 |
| 4         | 7                  | 7                  | 10        | 6                  | 4                  |
| 5         | 9                  | 10                 | 11        | 8                  | 9                  |
| 6         | 4                  | 3                  | 12        | 6                  | 8                  |

- Using the sign test, specify the competing hypotheses to determine whether the ratings differ between the two faculty members.
- Determine the critical value(s) at the 10% significance level.



- c. Calculate the value of the test statistic.
  - d. Do the data suggest that faculty ratings differ? Explain.
61. A new diet and exercise program claims that it significantly lowers a participant's cholesterol level. In order to test this claim, a sample of 60 participants is taken. Their cholesterol levels are measured before and after the three-month program. Forty of the participants recorded lower cholesterol levels at the end of the program, 18 participants recorded higher cholesterol levels, and 2 participants recorded no change.
- a. Using the sign test, specify the competing hypotheses to test the program's claim.
  - b. Calculate the value of the test statistic.
  - c. Calculate the  $p$ -value.
  - d. At the 5% significance level, do the data support the program's claim? Explain.

## 20.6 TESTS BASED ON RUNS

LO 20.8

In many applications, we wish to determine whether some observed values occur in a truly random fashion or whether some form of a nonrandom pattern exists. In other words, we want to test if the elements of the sequence are mutually independent. The **Wald-Wolfowitz runs test** is a procedure used to examine whether the elements in a sequence appear in a random order. It can be applied to either quantitative or qualitative data so long as we can separate the sample data into two categories.

Determine whether the elements of a sequence appear in a random order.

Suppose we observe a machine filling 16-ounce cereal boxes. Since a machine is unlikely to dispense exactly 16 ounces in each box, we expect the weight of each box to deviate from 16 ounces. We might conjecture that a machine is operating properly if the deviations from 16 ounces occur in a random order. Let's sample 30 cereal boxes and denote those boxes that are overfilled with the letter O and those that are underfilled with the letter U. The following sequence of Os and Us is produced:

**Sequence:** OOOOUUUOOOOOUOOOUUUUOOOOOUOOOOO

One possible way to test whether or not a machine is operating properly is to determine if the elements of a particular sequence of Os and Us occur randomly. If we observe a long series of consecutive Os (or Us), then the machine may be significantly overfilling (or underfilling) the cereal boxes. Adjustment of the machine is likely necessary if this is the case. Given the observed sequence, can we conclude that the machine is operating properly in the sense that the series of Os and Us occur randomly?

In general, when applying the runs test, we specify the competing hypotheses as

- $H_0$ : The elements occur randomly.
- $H_A$ : The elements do not occur randomly.

In this particular application, the null hypothesis implies that the machine properly fills the boxes and the alternative hypothesis implies that it does not. Before deriving the test statistic, it is first necessary to introduce some terminology. We define a **run** as an uninterrupted sequence of one letter, symbol, or attribute, such as O or U. We rewrite the observed sequence, but now include single horizontal lines below the letter O. The five single lines indicate that we observe five runs of O, or  $R_O = 5$ . Similarly, the double horizontal lines below the letter U show that we have four runs of U, or  $R_U = 4$ . Thus, the total number of runs  $R$  is equal to nine:  $R = R_O + R_U = 5 + 4 = 9$ . Also, note that we have a total of 30 observations, of which 20 are Os and 10 are Us, or  $n = n_O + n_U = 20 + 10 = 30$ .

**Sequence:** OOOO UUU OOOO U OOO UUUU OOOO UU OOOOO

We then ask: "Are nine runs consisting of 30 observations too few or too many compared with the number of runs expected in a strictly random sequence of 30 observations?"

In general, the runs test is a two-tailed test; that is, too many runs are deemed just as unlikely as too few runs. For example, consider the following two sequences:

Sequence A: OOOOOOOOOOOOO UUUUUU OOOOOOOOOOOOO

Sequence B: O U O U O U O U O U O U O U O U O U O U O U O U O U O U O U O U

If the null hypothesis of randomness is true, Sequence A seems unlikely in the sense that there appear to be too few runs given a sample of 30 observations. Sequence B also seems unlikely since O and U alternate systematically, or equivalently, there appear to be too many runs. It is more readily apparent in the machine-filling application that a sequence that produces too few runs indicates a machine that is not operating properly; that is, the machine has a pattern of consistently overfilling and/or underfilling the cereal boxes. However, a machine that exhibits a perfect regularity of overfilling, underfilling, overfilling, underfilling, etc. (too many runs) may be just as problematic. If there are too many runs, then this may indicate some sort of repeated alternating pattern.

Let  $n_1$  and  $n_2$  denote the numbers of Os and Us in an  $n$ -element sequence. In general, the sampling distribution of  $R$  (the distribution for the runs test) is quite complex and its critical values are provided in specially constructed tables. However, if  $n_1$  and  $n_2$  are at least 10, then the distribution of  $R$  is approximately normal.

#### THE TEST STATISTIC $R$ FOR THE WALD-WOLFOWITZ RUNS TEST

The value of the test statistic for the Wald-Wolfowitz runs test is computed as  $z = \frac{R - \mu_R}{\sigma_R}$ , where  $R$  represents the number of runs with mean  $\mu_R = \frac{2n_1n_2}{n} + 1$  and standard deviation  $\sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n - 1)}}$ ;  $n_1$  and  $n_2$  are the number of elements in a sequence possessing and not possessing a certain attribute and  $n = n_1 + n_2$ . The test is valid when  $n_1 \geq 10$  and  $n_2 \geq 10$ .

For the machine example, we found that  $R = 9$ ; in addition, we have  $n_1 = n_O = 20$  and  $n_2 = n_U = 10$ . We calculate the mean and the standard deviation of the distribution of  $R$  as

$$\mu_R = \frac{2n_1n_2}{n} + 1 = \frac{2(20)(10)}{30} + 1 = 14.33 \text{ and}$$

$$\sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n - 1)}} = \sqrt{\frac{(2 \times 20 \times 10)(2 \times 20 \times 10 - 30)}{30^2(30 - 1)}} = \sqrt{\frac{148000}{26100}} = 2.38.$$

Thus, the expected number of runs in a sample with 30 observations is 14.33 and the standard deviation is 2.38. We calculate the value of the test statistic as  $z = \frac{R - \mu_R}{\sigma_R} = \frac{9 - 14.33}{2.38} = -2.24$ . We find the  $p$ -value for a two-tailed test as  $2 \times P(Z \leq -2.24) = 0.0250$ . Since the  $p$ -value is less than  $\alpha = 0.05$ , we reject  $H_0$  and conclude that the machine does not properly fill the boxes. At the 5% significance level, adjustment of the machine is necessary.

## The Method of Runs Above and Below the Median

As mentioned earlier, the runs test can be applied to both qualitative and quantitative data. Any sample with numerical values can be treated similarly by using letters, say  $A$  and  $B$ , to denote values falling above and below the median of the sample, respectively. The resulting  $A$ s and  $B$ s can be tested for randomness by applying the **method of runs above and below the median**. This test is especially useful in detecting trends and cyclical patterns in economic data. A finding of too few runs is suggestive of a trend; that is, we first observe mostly  $A$ s and later mostly  $B$ s (or vice versa). In computing the value of the test statistic, we omit values that are equal to the median. A systematic alternation of  $A$ s and  $B$ s—that is, too many runs—implies a cyclical pattern, whereas too few runs implies a trend. Consider the following example.

## EXAMPLE 20.11

Table 20.21 shows the growth rate in the gross domestic product (GDP) for the United States from 1980 through 2009. Use the method of runs above and below the median with a significance level of 10% to test the null hypothesis of randomness against the alternative that a trend or cyclical pattern occurs.

**TABLE 20.21** GDP Growth Rates (in percent) for the United States, 1980–2009

| Year | GDP   | Year | GDP   | Year | GDP   |
|------|-------|------|-------|------|-------|
| 1980 | −0.24 | 1990 | 1.86  | 2000 | 3.69  |
| 1981 | 2.52  | 1991 | −0.19 | 2001 | 0.76  |
| 1982 | −1.97 | 1992 | 3.34  | 2002 | 1.61  |
| 1983 | 4.52  | 1993 | 2.69  | 2003 | 2.52  |
| 1984 | 7.20  | 1994 | 4.06  | 2004 | 3.65  |
| 1985 | 4.10  | 1995 | 2.54  | 2005 | 3.08  |
| 1986 | 3.43  | 1996 | 3.75  | 2006 | 2.87  |
| 1987 | 3.34  | 1997 | 4.55  | 2007 | 2.00  |
| 1988 | 4.12  | 1998 | 4.22  | 2008 | 1.10  |
| 1989 | 3.53  | 1999 | 4.49  | 2009 | −2.40 |

SOURCE: <http://data.worldbank.org/indicator>

**FILE**  
US\_GDP

**SOLUTION:** Since we are testing the null hypothesis of randomness against the alternative that there is a trend or cyclical pattern, we formulate the competing hypotheses as

$H_0$ : The GDP growth rate is random.

$H_A$ : The GDP growth rate is not random.

We first calculate the median GDP growth rate as 3.21%. Letting  $A$  and  $B$  denote an observation that falls above the median and below the median, respectively, we rewrite the data using the following sequence of  $A$ s and  $B$ s:

**Sequence:** BBB AAAAAAA BB A B A B AAAAA BBB A BBBBB

We see that the number of runs below the median  $R_B$  is 6, while the number of runs above the median  $R_A$  is 5, so the total number of runs  $R$  is 11. Also, since no values were discarded (no value was equal to the median), the total number of observations are  $n = 30$ , where the number of observations below the median and the number of observations above the median are  $n_B = 15$  and  $n_A = 15$ , respectively. Using this information, we compute the mean and the standard deviation of the distribution of  $R$  as

$$\begin{aligned}\mu_R &= \frac{2n_A n_B}{n} + 1 = \frac{2(15)(15)}{30} + 1 = 16 \text{ and} \\ \sigma_R &= \sqrt{\frac{2n_A n_B (2n_A n_B - n)}{n^2(n-1)}} = \sqrt{\frac{(2 \times 15 \times 15)(2 \times 15 \times 15 - 30)}{30^2(30-1)}} = \sqrt{\frac{189000}{26100}} \\ &= 2.69.\end{aligned}$$

Thus, the value of the test statistic is  $z = \frac{R - \mu_R}{\sigma_R} = \frac{11 - 16}{2.69} = -1.86$ . Using the  $z$  table, we find the  $p$ -value for this two-tailed test as  $2 \times P(Z \leq -1.86) = 0.0628$ . Since the  $p$ -value is less than  $\alpha = 0.10$ , we reject  $H_0$  and conclude that the sample is not

random. In fact, since the observed number of runs ( $R = 11$ ) is significantly less than the expected number of runs ( $\mu_R = 16$ ), there is evidence of a trend. The test results do not enable us to determine whether there is an upward or downward trend in the data.

## Using the Computer for the Runs Test

Table 20.22 shows a portion of Minitab output for the GDP example. The letter *K* denotes the median growth rate of 3.21%. Note that the observed number of runs and the expected number of runs of 11 and 16, respectively, match those that we calculated by hand. In addition, the *p*-value also matches our hand-calculated value, allowing us to reject the null hypothesis of randomness at the 10% significance level.

**TABLE 20.22** Minitab Output for Example 20.11

| Runs Test for GDP                    |
|--------------------------------------|
| Runs above and below $K = 3.21$      |
| The observed number of runs = 11     |
| The expected number of runs = 16     |
| 15 observations above $K$ , 15 below |
| P-value = 0.063                      |

## EXERCISES 20.6

## Mechanics

62. Consider the following information:  $n_1 = 24$ ,  $n_2 = 28$  and  $R = 18$ , where  $R$  is the number of runs,  $n_1$  and  $n_2$  are the number of elements in a sequence possessing and not possessing a certain attribute, and  $n_1 + n_2 = n$ .
  - a. Specify the competing hypotheses to test for nonrandomness.
  - b. Calculate the value of the test statistic.
  - c. Calculate the  $p$ -value.
  - d. At the 5% significance level, are the observations nonrandom?
63. Consider the following information:  $n_1 = 10$ ,  $n_2 = 13$  and  $R = 8$ , where  $R$  is the number of runs,  $n_1$  and  $n_2$  are the number of elements in a sequence possessing and not possessing a certain attribute, and  $n_1 + n_2 = n$ .
  - a. Specify the competing hypotheses to test for nonrandomness.
  - b. Calculate the value of the test statistic.
  - c. Calculate the  $p$ -value.
  - d. At the 5% significance level, are the observations nonrandom?
64. Let A and B be two possible outcomes of a single experiment. The sequence of the outcomes is as follows:

BBAABAABBABABBBABBAAABABBABBABA

At the 5% significance level, conduct a hypothesis test to determine if the outcomes are nonrandom.

65. Let  $D$  denote a desirable outcome and  $U$  denote an undesirable outcome. The sequence of the outcomes is as follows:

DDDUUDUUUUUDDDUUDUUUDDDUUUUDDDD

At the 1% significance level, conduct a hypothesis test to determine if the outcomes are nonrandom.

## Applications

66. Given the digits zero through nine, a computer program is supposed to generate even and odd numbers randomly. The computer produced the following sequence of numbers:

5346802977168315243392

- Specify the competing hypotheses to test for nonrandomness.
- At the 1% significance level, what is the decision rule?
- What is the value of the test statistic?
- Is the program operating properly? Explain.

67. A gambler suspects that a coin may be weighted more heavily toward the outcome of tails (T) over heads (H). He flips the coin 25 times and notes the following sequence:

TTHTTTHHTHTTHTTTTHHTHTHTTH

- Specify the competing hypotheses to test the gambler's belief on nonrandomness.
- What is the value of the test statistic?
- Calculate the  $p$ -value.
- At the 5% significance level, is the gambler's belief supported by the data?

68. **FILE India\_GDP.** The following table shows a portion of the growth rate in the gross domestic product (GDP) for India from 1980 through 2008. Use the method of runs above and below the median with a significance level of 5% to test the null hypothesis of randomness against the alternative that there is a trend or cyclical pattern.

| Year | GDP  |
|------|------|
| 1980 | 6.74 |
| 1981 | 6.00 |
| ⋮    | ⋮    |
| 2008 | 6.07 |

SOURCE: <http://data.worldbank.org/indicator>.

69. **FILE Absenteeism.** The superintendent of a large suburban high school must decide whether to close the school for at least two days due to the spread of flu. If she can confirm a trend in absenteeism, then she will close the high school. The following are the number of students absent from the high school on 25 consecutive school days.

44, 56, 55, 40, 42, 51, 50, 59, 58, 45, 44, 52, 52,  
43, 48, 58, 57, 42, 60, 65, 69, 75, 70, 72, 72

Use the method of runs above and below the median and  $\alpha = 0.05$  to test the null hypothesis of randomness against the alternative that there is a trend.

70. **FILE Amgen.** A research analyst follows the biotechnology industry and examines the daily stock price of Amgen, Inc. over the past year. The table below shows a portion of the daily stock price of Amgen for the 252 trading days in 2010. The research analyst wants to test the random-walk hypothesis that suggests that stock prices move randomly over time with no discernible pattern.

| Date       | Adjusted Stock Price |
|------------|----------------------|
| 1/4/2010   | \$57.72              |
| 1/5/2010   | 57.22                |
| ⋮          | ⋮                    |
| 12/31/2010 | 54.9                 |

- Use the method of runs above and below the median to test the null hypothesis of randomness against the alternative that there is a trend at the 5% significance level.
- Can the research analyst conclude that the movement of Amgen's stock price is consistent with the random-walk hypothesis?

## WRITING WITH STATISTICS

Meg Suzuki manages a trendy sushi restaurant in Chicago, Illinois. She is planning an aggressive advertising campaign to offset the loss of business due to competition from other restaurants. She knows advertising costs increase overall costs, but she hopes this effort will positively affect sales, as it has done in the past under her tenure. She collects monthly data on sales (in \$1,000s) and advertising costs (in \$) over the past two years and produces the following regression equation:

$$\begin{aligned}\text{Estimated Sales} &= 17.77 + 0.03\text{Advertising Costs} \\ t\text{-statistics} &= (17.77) \quad (21.07)\end{aligned}$$

At the 5% significance level, Meg initially concludes that advertising is significant in explaining sales. However, to estimate this regression model, she had to make certain assumptions that might not be valid. Specifically, with a time series analysis, the assumption maintaining the independence of the error terms often breaks down. In other words, the regression model often suffers from serial correlation. Table 20.23 shows a portion of the values of the residuals.



**TABLE 20.23** Values of Residuals

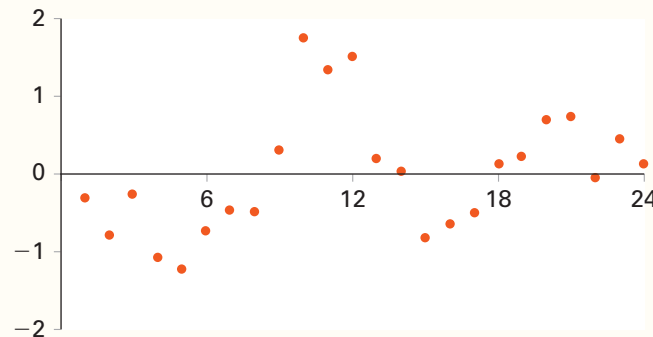
| Observation | Residual |
|-------------|----------|
| 1           | -0.31    |
| 2           | -0.80    |
| ⋮           | ⋮        |
| 24          | 0.12     |

Meg would like to use the runs test to determine whether the positive and negative residuals occur randomly at the 5% significance level.

## Sample Report— Testing the Independence of Residuals

One of the underlying assumptions of a linear regression model is that the error term is uncorrelated across observations. In a regression model relating sales to advertising costs, there is reason to believe that correlated observations may be a problem because the data are time series. Figure 20.A is a scatterplot of the residuals against time. If the residuals show no pattern around the horizontal axis, then the observations are not likely correlated. Given the wavelike movement in the residuals over time (clustering below the horizontal axis, then above the horizontal axis, etc.), the observations are likely correlated.

**FIGURE 20.A** Scatterplot of Residuals against Time



The above graphical analysis is supplemented with a runs test to determine if the residuals fail to follow a random pattern. A residual is given a + symbol if the residual is positive and a - symbol if the residual is negative. There are 12 positive residuals and 12 negative residuals, or  $n_+ = 12$  and  $n_- = 12$ , respectively. A run is then defined as an uninterrupted sequence of a + or a - sign. The sample data exhibits three positive runs,  $R_+ = 3$ , and three negative runs,  $R_- = 3$ , for a total number of runs equal to six,  $R = 6$ .

Are six runs consisting of 24 observations too few or too many compared with the number of runs expected in a strictly random sequence of 24 observations? To answer this question, the mean and the standard deviation for the distribution of  $R$  are calculated. The mean number of runs in a sample of 24 observations is 13 with a standard deviation of 2.4. Table 20.A provides summary data to conduct the runs test.

**TABLE 20.A** Data for Runs Test,  $n = 24$

- Mean number of runs,  $\mu_R = 13$ , versus actual number of runs,  $R = 6$ .
- Standard deviation of the sampling distribution of  $R$ :  $\sigma_R = 2.4$ .
- z-statistic = -2.92; the  $p$ -value (two-tailed) = 0.0036.

The sample value of the test statistic is  $z = -2.92$  with an associated  $p$ -value of 0.0036. The null hypothesis of the randomness of the residuals is rejected at the 5% level; the pattern of the residuals is nonrandom. Corrective measures should be taken before statistical inference is conducted on the estimated model.



## CONCEPTUAL REVIEW

### LO 20.1 Distinguish between parametric and nonparametric tests.

**Nonparametric tests**, also referred to as distribution-free tests, do not require the stringent assumptions of parametric tests and are especially attractive when the underlying population is markedly nonnormal. Also, while parametric tests require data of interval or ratio scale, nonparametric tests can be performed on data of nominal or ordinal scale.

However, a nonparametric test ignores useful information since it often focuses on the rank rather than the magnitude of sample values. Therefore, in situations when the parametric assumptions are valid, the nonparametric test is less powerful (more prone to Type II error) than its parametric counterpart. In general, when the assumptions for a parametric test are met, it is preferable to use a parametric test rather than a nonparametric test. Since the normality assumption for parametric tests is less stringent in large samples, the main appeal of rank-based nonparametric tests tends to be with relatively small samples.

### LO 20.2 Make inferences about a population median.

If we cannot assume that the data are normally distributed and/or we want to test the population median, we apply the **Wilcoxon signed-rank test**. The value of the test statistic  $T$  for the Wilcoxon signed-rank test is  $T = T^+$ , where  $T^+$  denotes the sum of the ranks of the positive differences from the hypothesized median  $m_0$ .

If the sample size  $n \leq 10$ , we use a special table to derive the decision rule for the hypothesis test. The sampling distribution of  $T$  can also be approximated by the normal distribution if  $n \geq 10$ . With the normal approximation, the value of the test statistic is calculated as  $z = \frac{T - \mu_T}{\sigma_T}$ , where  $\mu_T = \frac{n(n+1)}{4}$  and  $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$ .

### LO 20.3 Make inferences about the population median difference based on matched-pairs sampling.

We can also apply the **Wilcoxon signed-rank test** as the nonparametric counterpart to the  $t$  test that was used to determine whether the population mean difference differs from zero based on matched-pairs sampling. The measurement of interest is the difference between paired observations, or  $d_i = x_i - y_i$ . We conduct the test by following analogous steps to those applied for a one-sample Wilcoxon signed-rank test.

### LO 20.4 Make inferences about the difference between two population medians based on independent sampling.

We use the **Wilcoxon rank-sum test** to determine whether two populations have different medians based on independent sampling. We pool the data and calculate the rank sum of sample 1,  $W_1$ , and the rank sum of sample 2,  $W_2$ . If the two sample sizes satisfy  $n_1 \leq n_2$ , then the test statistic  $W$  is  $W = W_1$ . Otherwise, we set  $W = W_2$ .

If either sample is less than or equal to 10, we use a special table to obtain the critical value(s) for the hypothesis test. The sampling distribution of  $W$  can be approximated by the normal distribution if both sample sizes are greater than or equal to 10. The value of the test statistic is calculated as  $z = \frac{W - \mu_W}{\sigma_W}$ , where  $\mu_W = \frac{(n_1 + n_2 + 1) \times \min(n_1, n_2)}{2}$  and  $\sigma_W = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$ .

### LO 20.5 Make inferences about the difference between three or more population medians.

We employ the **Kruskal-Wallis test** as the nonparametric alternative to the one-way ANOVA  $F$  test. It is based on ranks and is used for testing the differences between the

medians of  $k$  populations. The value of the test statistic for the Kruskal-Wallis test is  $H = \left( \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} \right) - 3(n+1)$ , where  $R_i$  and  $n_i$  are the rank sum and the size of the  $i$ th sample,  $n = \sum_{i=1}^k n_i$ , and  $k$  is the number of populations (independent samples). So long as  $n_i \geq 5$ , the test statistic  $H$  follows the  $\chi^2$  distribution with  $k - 1$  degrees of freedom.

**LO 20.6 Conduct a hypothesis test for the population Spearman rank correlation coefficient.**

The **Spearman rank correlation coefficient**  $r_s$  measures the sample correlation between two random variables. We compute it as  $r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$ , where  $d_i$  is the difference between the ranks assigned to the variables. When the sample size is small ( $n \leq 10$ ), we use a special table to determine the significance of the population Spearman rank correlation coefficient  $\rho_s$ . When  $n \geq 10$ , it is reasonable to assume that the distribution of  $r_s$  is approximately normal. The resulting value of the test statistic is calculated as  $z = r_s \sqrt{n - 1}$ .

**LO 20.7 Make inferences about the difference between two populations of ordinal data based on matched-pairs sampling.**

We use the **sign test** to determine whether significant differences exist between two matched-pairs populations of ordinal data. The value of the test statistic is computed as  $z = \frac{\bar{p} - 0.5}{0.5/\sqrt{n}}$ , where  $\bar{p}$  is the sample proportion of positive signs. The test is valid when  $n \geq 10$ .

**LO 20.8 Determine whether the elements of a sequence appear in a random order.**

We apply the **Wald-Wolfowitz runs test** to examine whether or not the attributes in a sequence appear in a random order. The sampling distribution of  $R$ , representing the number of runs, can be approximated by the normal distribution if  $n_1 \geq 10$  and  $n_2 \geq 10$ . The value of the test statistic is computed as  $z = \frac{R - \mu_R}{\sigma_R}$ , where  $\mu_R = \frac{2n_1n_2}{n} + 1$  and  $\sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n - 1)}}$ . The test is valid when  $n_1 \geq 10$  and  $n_2 \geq 10$ . We can use the runs test with quantitative data to investigate whether the values randomly fall above and below the sample's median. This test is especially useful in detecting trends and cyclical patterns in economic data.

## ADDITIONAL EXERCISES AND CASE STUDIES

### Exercises

71. The following are the closing stock prices for a pharmaceutical firm over the past two weeks.

| Day        | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Price (\$) | 61.22 | 60.99 | 61.91 | 61.59 | 61.76 | 61.91 | 61.30 | 61.37 | 61.95 | 62.82 |

- Specify the competing hypotheses to determine whether the median stock price is greater than \$61.25.
- At the 5% significance level, what is the decision rule? Do not assume that the sampling distribution of  $T$  is normally distributed.
- Calculate the value of the Wilcoxon signed-rank test statistic  $T$ .
- Is the median stock price greater than \$61.25? Explain.

72. A farmer is concerned that a change in fertilizer to an organic variant might change his crop yield. He subdivides 6 lots and uses the old fertilizer on one half of each lot and the new fertilizer on the other half. The following table shows the results.

| Lot | Crop Yield Using Old Fertilizer | Crop Yield Using New Fertilizer |
|-----|---------------------------------|---------------------------------|
| 1   | 10                              | 12                              |
| 2   | 11                              | 10                              |
| 3   | 10                              | 13                              |
| 4   | 12                              | 9                               |
| 5   | 12                              | 11                              |
| 6   | 11                              | 12                              |

- Specify the competing hypotheses to determine whether the median difference between the crop yields differs from zero.
- At the 5% significance level, what is the decision rule?
- Calculate the value of the Wilcoxon signed-rank test statistic  $T$ .
- Is there sufficient evidence to conclude that the median difference between the crop yields differs from zero? Should the farmer be concerned? Explain.

73. **FILE Comparison.** The table below shows a portion of the returns for Fidelity's Equity Income mutual fund and Vanguard's Equity Income mutual fund from 2000 through 2010.

| Year | Fidelity | Vanguard |
|------|----------|----------|
| 2000 | 3.88     | 13.57    |
| 2001 | -5.02    | -2.34    |
| ⋮    | ⋮        | ⋮        |
| 2010 | 15.13    | 14.88    |

Source: <http://finance.yahoo.com>.

- Specify the competing hypotheses to determine whether the median difference between the returns differs from zero.
  - Calculate the value of the Wilcoxon signed-rank test statistic  $T$ . Assume normality of  $T$ .
  - At the 5% significance level, what is the decision rule?
  - Does the median difference between the returns differ from zero? Explain.
74. **FILE Refrigerator.** A consumer advocate researches the length of life between two brands of refrigerators, Brand A and Brand B. He collects data on the longevity of 40 refrigerators for Brand A and repeats the sampling for Brand B. A portion of the data is shown in the accompanying table.

| Brand A | Brand B |
|---------|---------|
| 16      | 16      |
| 14      | 20      |
| ⋮       | ⋮       |
| 18      | 17      |

- Specify the competing hypotheses to test whether the median length of life differs between the two brands.
- Calculate the value of the Wilcoxon rank-sum statistic  $W$ .
- Calculate the value of the test statistic, assuming that the sampling distribution of  $W$  is approximately normal.

- With  $\alpha = 0.05$ , does median longevity differ between the two brands? Explain.

75. **FILE Test Centers.** A psychiatrist believes that the location of a test center may influence a test taker's performance. To test his claim, he collects SAT scores from four different locations.

| Location 1 | Location 2 | Location 3 | Location 4 |
|------------|------------|------------|------------|
| 1350       | 1300       | 1000       | 1450       |
| 1275       | 1320       | 1350       | 1700       |
| 1200       | 1260       | 1100       | 1600       |
| 1450       | 1400       | 1050       | 1325       |
| 1150       | 1425       | 1025       | 1200       |

- Specify the competing hypotheses to test whether some median test scores differ by location.
- At the 5% significance level, what is the critical value?
- Calculate the value of the test statistic  $H$ .
- Do the data support the psychiatrist's belief? Explain.

76. **FILE PE Ratio.** An economist wants to determine whether the Price/Earnings (P/E) ratio is the same for firms in three industries. Five firms were randomly selected from each industry. Their P/E ratios are shown in the accompanying table.

| Industry A | 12.19 | 12.44 | 7.28  | 9.96  | 10.51 |
|------------|-------|-------|-------|-------|-------|
| Industry B | 14.34 | 17.80 | 9.32  | 14.90 | 9.41  |
| Industry C | 26.38 | 24.75 | 16.88 | 16.87 | 16.70 |

- Specify the competing hypotheses to test whether some median P/E ratios differ by industry.
- Calculate the value of the test statistic  $H$ .
- At the 5% significance level, what is the critical value?
- Do some P/E ratios differ by industry? Explain.

77. **FILE Electronics Utilities.** The following table shows a portion of the annual returns (in percent) for two of Fidelity's mutual funds: the Fidelity Advisor's Electronic Fund and the Fidelity Advisor's Utilities Fund.

| Year | Electronics | Utilities |
|------|-------------|-----------|
| 2001 | 47.41       | -15.28    |
| 2002 | 20.74       | -29.48    |
| ⋮    | ⋮           | ⋮         |
| 2010 | 16.84       | 11.08     |

Source: <http://finance.yahoo.com>.

- Calculate and interpret the Spearman rank correlation coefficient  $r_s$ .

- b. Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient differs from zero.
- c. At the 5% significance level, specify the critical value and the decision rule. Assume that the sampling distribution of  $r_s$  is approximately normal.
- d. What is the conclusion to the test? Are the returns correlated?

78. A research analyst believes that a positive relationship exists between a firm's advertising expenditures and its sales. For 65 firms, she collects data on each firm's yearly advertising expenditures and subsequent sales. She calculates a Spearman rank correlation coefficient of 0.45.

- a. Specify the competing hypotheses to determine whether the Spearman rank correlation coefficient differs from zero.
- b. Assume that the sampling distribution of  $r_s$  is approximately normal. Calculate the value of the test statistic and the  $p$ -value.
- c. At the 5% significance level, are advertising and sales correlated? Explain.

79. **FILE Inspectors.** In order to ensure the public's health and safety, state health inspectors are required to rate the cleanliness and quality of all restaurants in the state. Restaurants that consistently score below a certain level often lose their licenses to operate. From a sample of 10 restaurants, two health inspectors give the ratings shown in the accompanying table, where a score of 10 denotes excellence in cleanliness and quality.

| Restaurant | Inspector A's Rating | Inspector B's Rating |
|------------|----------------------|----------------------|
| 1          | 9                    | 8                    |
| 2          | 5                    | 6                    |
| 3          | 10                   | 8                    |
| 4          | 5                    | 4                    |
| 5          | 2                    | 3                    |
| 6          | 7                    | 5                    |
| 7          | 8                    | 6                    |
| 8          | 4                    | 1                    |
| 9          | 3                    | 2                    |
| 10         | 8                    | 9                    |

- a. Using the sign test, specify the competing hypotheses to determine whether the ratings differ between the two health inspectors.
- b. Calculate the value of the test statistic.
- c. Calculate the  $p$ -value.

- d. At the 5% significance level, do the data suggest that the ratings differ?

80. **FILE China GDP.** The following table shows a portion of the growth rate in the gross domestic product (GDP) for China from 1980 through 2008. Use the method of runs above and below the median with a significance level of 5% to test the null hypothesis of randomness against the alternative that there is a trend or cyclical pattern.

| Year | GDP |
|------|-----|
| 1980 | 7.8 |
| 1981 | 5.2 |
| ⋮    | ⋮   |
| 2008 | 9.0 |

SOURCE: <http://data.worldbank.org/indicator>.

81. **FILE Dow Jones.** A research analyst follows the monthly price data for the Dow Jones Industrial Average for the years 2008–2010. The accompanying table shows a portion of the price data. The analyst wants to test the random-walk hypothesis that suggests that prices move randomly over time with no discernible pattern.

| Date      | Adjusted Closing Price |
|-----------|------------------------|
| 1/2/2008  | 12,650.36              |
| 2/1/2008  | 12,266.39              |
| ⋮         | ⋮                      |
| 12/1/2010 | 11,577.51              |

SOURCE: <http://finance.yahoo.com>.

- a. Use the method-of-runs above and below the median to test the null hypothesis of randomness against the alternative that there is a trend or cyclical pattern at the 5% significance level.
- b. Can the research analyst conclude that the movement of the Dow Jones Industrial Average is consistent with the random-walk hypothesis?

82. **FILE US CPI.** The following table shows a portion of the percent change in the consumer price index (CPI) for the United States from 1980 through 2008. Use the method of runs above and below the median with a significance level of 5% to test the null hypothesis of randomness against the alternative that there is a trend or cyclical pattern.

| Year | CPI  |
|------|------|
| 1980 | 12.5 |
| 1981 | 8.9  |
| ⋮    | ⋮    |
| 2008 | 0.1  |

SOURCE: <http://data.worldbank.org/indicator>.

## CASE STUDIES

**CASE STUDY 20.1** The economic recovery in California has become increasingly divided between coastal and inland areas (*The Wall Street Journal*, February 2, 2010). For instance, the median home price in Southern California increased 7.5% in December 2009 from a year earlier to \$360,000; however, the median home price declined by 10% to \$180,000 over the same time period in the Inland Empire counties of San Bernardino and Riverside. An economist gathers 10 recent home sales (in \$1,000s) in Southern California and 10 recent home sales (in \$1,000s) in the Inland Empire counties. A portion of the results are shown in the accompanying table.

**Data for Case Study 20.1** California Home Prices

| Home | Southern California | Inland Empire |
|------|---------------------|---------------|
| 1    | 418                 | 167           |
| 2    | 491                 | 186           |
| ⋮    | ⋮                   | ⋮             |
| 10   | 885                 | 262           |

**FILE**  
California

In a report, use the sample information to:

1. Calculate and interpret relevant summary measures for California home prices in these two regions.
2. Explain why the  $t$  test for comparing means from independent samples might be inappropriate in this case.
3. Use the Wilcoxon rank-sum method to determine whether the median home price in Southern California is greater than the median home price in the Inland Empire.

**CASE STUDY 20.2** There has been a lot of discussion lately surrounding the levels and structure of executive compensation. It is well documented that in general, compensation received by senior executives has risen steeply in recent years. The accompanying table lists a portion of total compensation for the top 10 CEOs in four industry classifications: Manufacturing (technology); Manufacturing (other); Services (financial); Services (other). Total compensation for 2006 is measured in \$ millions.

**Data for Case Study 20.2** Top Executive Compensation (in \$ millions), 2006

| Manufacturing (Technology) | Manufacturing (Other) | Services (Financial) | Services (Other) |
|----------------------------|-----------------------|----------------------|------------------|
| 39.82                      | 64.63                 | 91.38                | 24.02            |
| 32.85                      | 60.73                 | 48.13                | 21.51            |
| ⋮                          | ⋮                     | ⋮                    | ⋮                |
| 24.08                      | 30.80                 | 25.75                | 13.04            |

**FILE**  
Compensation

SOURCE: Compustat.

In a report, use the sample information to:

1. Calculate and interpret relevant summary measures for executive compensation in these four industries.

2. Explain why the one-way ANOVA  $F$  test for comparing three or more population means may be inappropriate in this case.
3. At the 5% significance level, use the Kruskal-Wallis test to determine whether some median compensations vary across classifications.

**CASE STUDY 20.3** The consumption function, developed by John Maynard Keynes, captures one of the key relationships in economics. It expresses consumption as a function of disposable income, where disposable income is defined as income after taxes. The accompanying table shows a portion of average U.S. annual consumption and disposable income for the years 1985–2006.

**Data for Case Study 20.3** Consumption and Disposable Income, 1985–2006

| Year | Consumption | Disposable Income |
|------|-------------|-------------------|
| 1985 | \$23,490    | \$22,887          |
| 1986 | 23,866      | 23,172            |
| ⋮    | ⋮           | ⋮                 |
| 2006 | 48,398      | 58,101            |

SOURCE: *The Statistical Abstract of the United States*.

**FILE**

*Consumption\_Function*

In a report, use the sample information to:

1. Estimate and interpret the model:  $\text{Consumption} = \beta_0 + \beta_1 \text{Disposable Income} + \varepsilon$ .
2. Indicate which assumption might be violated, given that the analysis uses time series data.
3. Use the runs test to determine whether the pattern of the residuals is nonrandom at the 5% significance level.

## APPENDIX 20.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, and JMP. More detailed instructions for these packages and for  $R$  can be found on McGraw-Hill's Connect or through your instructor. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands.

### Minitab

#### The Wilcoxon Signed-Rank Test

(Replicating Example 20.2) From the menu choose **Stat > Nonparametrics > 1-Sample Wilcoxon**. After **Variables**, select **Income**. Select **Test median** and enter the value 5. After **Alternative**, select “greater than.”

#### The Wilcoxon Rank-Sum Test

(Replicating Example 20.6) From the menu choose **Stat > Nonparametrics > Mann-Whitney**. Select **Computer Science** for the **First Sample**, and then select **Finance** for the **Second Sample**. After **Alternative**, select “not equal.”

**FILE**

*Fund\_Returns*

**FILE**

*Undergrad\_Salaries*



## The Kruskal-Wallis Test

- A. (Replicating Example 20.7) Pool the data into one column and label SAT. In an adjacent column, labeled Category, denote all Black observations with the value 1, all Hispanic observations with the value 2, all White observations with the value 3, and all Asian observations with the value 4.
- B. From the menu choose **Stat**, > **Nonparametrics** > **Kruskal-Wallis**.
- C. After **Response**, select SAT, and after **Factor**, select Category.

**FILE**  
*KW\_SAT*

## The Spearman Rank Correlation Test

- A. (Replicating Example 20.9) Choose **Data** > **Rank** to rank the Salary and Age data.
- B. Choose **Stat** > **Basic Statistics** > **Correlation** to compute the Pearson's correlation coefficient on the columns of the ranked data. (Uncheck **Display p-values** because the *p*-value that Minitab provides is not accurate for the Spearman Rank Correlation test.)

**FILE**  
*Quarterback\_Salary*

## The Runs Test

- A. (Replicating Example 20.11) From the menu choose **Stat** > **Nonparametrics** > **Runs Test**. Under **Variables**, select GDP. Select **Above and below** and enter the value 3.21.

**FILE**  
*US\_GDP*

## SPSS

### The Wilcoxon Signed-Rank Test

- A. (Replicating Example 20.2) From the menu select **Analyze** > **Nonparametric Tests** > **One-Sample**.
- B. Choose the **Objective** tab and after “What is your objective?” select **Custom Analysis**. Choose the **Fields** tab and after **Test Fields** select Income. Choose the **Settings** tab, select **Customize tests**, and select **Compare median to hypothesized (Wilcoxon signed-rank test)**. Enter 5 after **Hypothesized Mean**.

**FILE**  
*Fund\_Returns*

### The Wilcoxon Rank-Sum Test

- A. (Replicating Example 20.6) Pool the data into one column and label Salary. In an adjacent column, labeled Major, input the corresponding major with each salary.
- B. In the menu select **Analyze** > **Nonparametric Tests** > **Independent Samples**.
- C. Choose the **Objective** tab and after “What is your objective?” select **Custom Analysis**. Choose the **Field** tab, after **Test Field** select Salary, and after **Groups** select Major. Choose the **Settings** tab, select **Customize settings** and select **Mann-Whitney U (2 samples)**.

**FILE**  
*Undergrad\_Salaries*

## The Kruskal-Wallis Test

- A. (Replicating Example 20.7) To arrange the data, follow step A under the Minitab section for the Kruskal-Wallis Test.
- B. In the menu select **Analyze** > **Nonparametric Tests** > **Independent Samples**.
- C. Choose the **Objective** tab and after “What is your objective?” select **Custom Analysis**. Choose the **Field** tab, after **Test Field** select SAT, and after **Groups** select Category. Choose the **Settings** tab, select **Customize settings** and select **Kruskal-Wallis 1-Way Anova (k samples)**.

**FILE**  
*KW\_SAT*

**FILE***Quarterback\_Salary*

### The Spearman Rank Correlation Test

- A. (Replicating Example 20.9) From the menu choose **Analyze > Correlate > Bivariate**.
- B. Under **Variables**, select Salary and Age. Under **Correlation Coefficients**, select Spearman.

**FILE***US\_GDP*

### The Runs Test

- A. (Replicating Example 20.11) From the menu select **Analyze > Nonparametric Tests > Legacy Dialogs > Runs**.
- B. Under **Test Variable List**, select GDP. Under **Cut Point**, select **Custom** and enter 3.21.

## JMP

**FILE***Fund\_Returns*

### The Wilcoxon Signed-Rank Test

- A. (Replicating Example 20.4) Create a column of differences between the returns (Metals – Income) and label the column Difference.
- B. From the menu choose **Analyze > Distribution**.
- C. Under **Select Columns**, select Difference, then under **Cast Selected Columns into Roles**, select **Y, Columns**.
- D. Click on the red triangle next to **Difference**, and select **Test Mean**. After **Specify Hypothesized Mean**, enter 0, and check the box before **Wilcoxon Signed Rank**.

**FILE***Undergrad\_Salaries*

### The Wilcoxon Rank-Sum Test

- A. (Replicating Example 20.6) To arrange the data, follow step A under the SPSS section for the Wilcoxon Rank-Sum Test.
- B. From the menu select **Analyze > Fit Y by X**.
- C. Under **Select Columns**, select Salary, then under **Cast Selected Columns into Roles**, select **Y, Columns**. Under **Select Columns**, select Major, and then under **Cast Selected Columns into Roles**, select **X, Factor**.
- D. Click on the red triangle next to **Oneway Analysis of Salary By Major** and select **Nonparametric > Wilcoxon Test**.

**FILE***KW\_SAT*

### The Kruskal-Wallis Test

- A. (Replicating Example 20.7) To arrange the data, follow step A under the Minitab section for the Kruskal-Wallis Test.
- B. From the menu select **Analyze > Fit Y by X**.
- C. Under **Select Columns**, select SAT, then under **Cast Selected Columns into Roles**, select **Y, Columns**. Under **Select Columns**, select Category, and then under **Cast Selected Columns into Roles**, select **X, Factor**.
- D. Click on the red triangle next to **Oneway Analysis of All By Category** and select **Nonparametric > Wilcoxon Test**.

**FILE***Quarterback\_Salary*

### The Spearman Rank Correlation Test

- A. (Replicating Example 20.9) From the menu choose **Analyze > Multivariate Methods > Multivariate**.

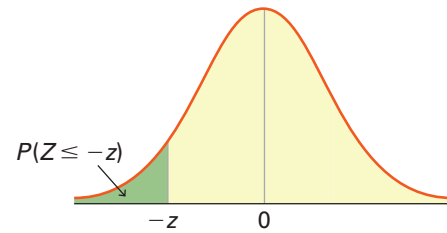
- B.** Under **Select Columns**, select Salary and Age, and under **Cast Selected Columns into Roles**, select **Y, Columns**.
- C.** Click the red triangle beside **Multivariate**. Select **Nonparametric Correlations > Spearman's  $\rho$** .

# APPENDIX A

## Tables

**TABLE 1** Standard Normal Curve Areas

Entries in this table provide cumulative probabilities, that is, the area under the curve to the left of  $-z$ . For example,  $P(Z \leq -1.52) = 0.0643$ .

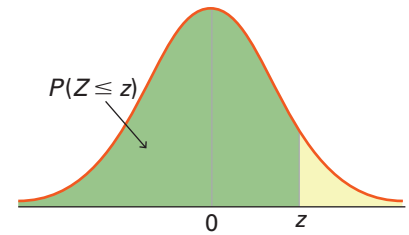


| $z$  | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| -3.9 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| -3.8 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| -3.7 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| -3.6 | 0.0002 | 0.0002 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| -3.5 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 |
| -3.4 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0002 |
| -3.3 | 0.0005 | 0.0005 | 0.0005 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0003 |
| -3.2 | 0.0007 | 0.0007 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0005 | 0.0005 | 0.0005 |
| -3.1 | 0.0010 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0007 | 0.0007 |
| -3.0 | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 | 0.0011 | 0.0011 | 0.0010 | 0.0010 |
| -2.9 | 0.0019 | 0.0018 | 0.0018 | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |
| -2.8 | 0.0026 | 0.0025 | 0.0024 | 0.0023 | 0.0023 | 0.0022 | 0.0021 | 0.0021 | 0.0020 | 0.0019 |
| -2.7 | 0.0035 | 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0027 | 0.0026 |
| -2.6 | 0.0047 | 0.0045 | 0.0044 | 0.0043 | 0.0041 | 0.0040 | 0.0039 | 0.0038 | 0.0037 | 0.0036 |
| -2.5 | 0.0062 | 0.0060 | 0.0059 | 0.0057 | 0.0055 | 0.0054 | 0.0052 | 0.0051 | 0.0049 | 0.0048 |
| -2.4 | 0.0082 | 0.0080 | 0.0078 | 0.0075 | 0.0073 | 0.0071 | 0.0069 | 0.0068 | 0.0066 | 0.0064 |
| -2.3 | 0.0107 | 0.0104 | 0.0102 | 0.0099 | 0.0096 | 0.0094 | 0.0091 | 0.0089 | 0.0087 | 0.0084 |
| -2.2 | 0.0139 | 0.0136 | 0.0132 | 0.0129 | 0.0125 | 0.0122 | 0.0119 | 0.0116 | 0.0113 | 0.0110 |
| -2.1 | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 | 0.0150 | 0.0146 | 0.0143 |
| -2.0 | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 | 0.0192 | 0.0188 | 0.0183 |
| -1.9 | 0.0287 | 0.0281 | 0.0274 | 0.0268 | 0.0262 | 0.0256 | 0.0250 | 0.0244 | 0.0239 | 0.0233 |
| -1.8 | 0.0359 | 0.0351 | 0.0344 | 0.0336 | 0.0329 | 0.0322 | 0.0314 | 0.0307 | 0.0301 | 0.0294 |
| -1.7 | 0.0446 | 0.0436 | 0.0427 | 0.0418 | 0.0409 | 0.0401 | 0.0392 | 0.0384 | 0.0375 | 0.0367 |
| -1.6 | 0.0548 | 0.0537 | 0.0526 | 0.0516 | 0.0505 | 0.0495 | 0.0485 | 0.0475 | 0.0465 | 0.0455 |
| -1.5 | 0.0668 | 0.0655 | 0.0643 | 0.0630 | 0.0618 | 0.0606 | 0.0594 | 0.0582 | 0.0571 | 0.0559 |
| -1.4 | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 | 0.0721 | 0.0708 | 0.0694 | 0.0681 |
| -1.3 | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 | 0.0869 | 0.0853 | 0.0838 | 0.0823 |
| -1.2 | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 | 0.1038 | 0.1020 | 0.1003 | 0.0985 |
| -1.1 | 0.1357 | 0.1335 | 0.1314 | 0.1292 | 0.1271 | 0.1251 | 0.1230 | 0.1210 | 0.1190 | 0.1170 |
| -1.0 | 0.1587 | 0.1562 | 0.1539 | 0.1515 | 0.1492 | 0.1469 | 0.1446 | 0.1423 | 0.1401 | 0.1379 |
| -0.9 | 0.1841 | 0.1814 | 0.1788 | 0.1762 | 0.1736 | 0.1711 | 0.1685 | 0.1660 | 0.1635 | 0.1611 |
| -0.8 | 0.2119 | 0.2090 | 0.2061 | 0.2033 | 0.2005 | 0.1977 | 0.1949 | 0.1922 | 0.1894 | 0.1867 |
| -0.7 | 0.2420 | 0.2389 | 0.2358 | 0.2327 | 0.2296 | 0.2266 | 0.2236 | 0.2206 | 0.2177 | 0.2148 |
| -0.6 | 0.2743 | 0.2709 | 0.2676 | 0.2643 | 0.2611 | 0.2578 | 0.2546 | 0.2514 | 0.2483 | 0.2451 |
| -0.5 | 0.3085 | 0.3050 | 0.3015 | 0.2981 | 0.2946 | 0.2912 | 0.2877 | 0.2843 | 0.2810 | 0.2776 |
| -0.4 | 0.3446 | 0.3409 | 0.3372 | 0.3336 | 0.3300 | 0.3264 | 0.3228 | 0.3192 | 0.3156 | 0.3121 |
| -0.3 | 0.3821 | 0.3783 | 0.3745 | 0.3707 | 0.3669 | 0.3632 | 0.3594 | 0.3557 | 0.3520 | 0.3483 |
| -0.2 | 0.4207 | 0.4168 | 0.4129 | 0.4090 | 0.4052 | 0.4013 | 0.3974 | 0.3936 | 0.3897 | 0.3859 |
| -0.1 | 0.4602 | 0.4562 | 0.4522 | 0.4483 | 0.4443 | 0.4404 | 0.4364 | 0.4325 | 0.4286 | 0.4247 |
| -0.0 | 0.5000 | 0.4960 | 0.4920 | 0.4880 | 0.4840 | 0.4801 | 0.4761 | 0.4721 | 0.4681 | 0.4641 |

Source: Probabilities calculated with Excel.

**TABLE 1** (Continued)

Entries in this table provide cumulative probabilities, that is, the area under the curve to the left of  $z$ . For example,  $P(Z \leq 1.52) = 0.9357$ .

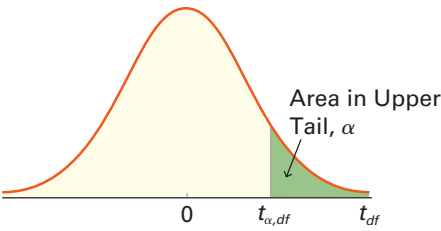


| <b>z</b> | <b>0.00</b> | <b>0.01</b> | <b>0.02</b> | <b>0.03</b> | <b>0.04</b> | <b>0.05</b> | <b>0.06</b> | <b>0.07</b> | <b>0.08</b> | <b>0.09</b> |
|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 0.0      | 0.5000      | 0.5040      | 0.5080      | 0.5120      | 0.5160      | 0.5199      | 0.5239      | 0.5279      | 0.5319      | 0.5359      |
| 0.1      | 0.5398      | 0.5438      | 0.5478      | 0.5517      | 0.5557      | 0.5596      | 0.5636      | 0.5675      | 0.5714      | 0.5753      |
| 0.2      | 0.5793      | 0.5832      | 0.5871      | 0.5910      | 0.5948      | 0.5987      | 0.6026      | 0.6064      | 0.6103      | 0.6141      |
| 0.3      | 0.6179      | 0.6217      | 0.6255      | 0.6293      | 0.6331      | 0.6368      | 0.6406      | 0.6443      | 0.6480      | 0.6517      |
| 0.4      | 0.6554      | 0.6591      | 0.6628      | 0.6664      | 0.6700      | 0.6736      | 0.6772      | 0.6808      | 0.6844      | 0.6879      |
| 0.5      | 0.6915      | 0.6950      | 0.6985      | 0.7019      | 0.7054      | 0.7088      | 0.7123      | 0.7157      | 0.7190      | 0.7224      |
| 0.6      | 0.7257      | 0.7291      | 0.7324      | 0.7357      | 0.7389      | 0.7422      | 0.7454      | 0.7486      | 0.7517      | 0.7549      |
| 0.7      | 0.7580      | 0.7611      | 0.7642      | 0.7673      | 0.7704      | 0.7734      | 0.7764      | 0.7794      | 0.7823      | 0.7852      |
| 0.8      | 0.7881      | 0.7910      | 0.7939      | 0.7967      | 0.7995      | 0.8023      | 0.8051      | 0.8078      | 0.8106      | 0.8133      |
| 0.9      | 0.8159      | 0.8186      | 0.8212      | 0.8238      | 0.8264      | 0.8289      | 0.8315      | 0.8340      | 0.8365      | 0.8389      |
| 1.0      | 0.8413      | 0.8438      | 0.8461      | 0.8485      | 0.8508      | 0.8531      | 0.8554      | 0.8577      | 0.8599      | 0.8621      |
| 1.1      | 0.8643      | 0.8665      | 0.8686      | 0.8708      | 0.8729      | 0.8749      | 0.8770      | 0.8790      | 0.8810      | 0.8830      |
| 1.2      | 0.8849      | 0.8869      | 0.8888      | 0.8907      | 0.8925      | 0.8944      | 0.8962      | 0.8980      | 0.8997      | 0.9015      |
| 1.3      | 0.9032      | 0.9049      | 0.9066      | 0.9082      | 0.9099      | 0.9115      | 0.9131      | 0.9147      | 0.9162      | 0.9177      |
| 1.4      | 0.9192      | 0.9207      | 0.9222      | 0.9236      | 0.9251      | 0.9265      | 0.9279      | 0.9292      | 0.9306      | 0.9319      |
| 1.5      | 0.9332      | 0.9345      | 0.9357      | 0.9370      | 0.9382      | 0.9394      | 0.9406      | 0.9418      | 0.9429      | 0.9441      |
| 1.6      | 0.9452      | 0.9463      | 0.9474      | 0.9484      | 0.9495      | 0.9505      | 0.9515      | 0.9525      | 0.9535      | 0.9545      |
| 1.7      | 0.9554      | 0.9564      | 0.9573      | 0.9582      | 0.9591      | 0.9599      | 0.9608      | 0.9616      | 0.9625      | 0.9633      |
| 1.8      | 0.9641      | 0.9649      | 0.9656      | 0.9664      | 0.9671      | 0.9678      | 0.9686      | 0.9693      | 0.9699      | 0.9706      |
| 1.9      | 0.9713      | 0.9719      | 0.9726      | 0.9732      | 0.9738      | 0.9744      | 0.9750      | 0.9756      | 0.9761      | 0.9767      |
| 2.0      | 0.9772      | 0.9778      | 0.9783      | 0.9788      | 0.9793      | 0.9798      | 0.9803      | 0.9808      | 0.9812      | 0.9817      |
| 2.1      | 0.9821      | 0.9826      | 0.9830      | 0.9834      | 0.9838      | 0.9842      | 0.9846      | 0.9850      | 0.9854      | 0.9857      |
| 2.2      | 0.9861      | 0.9864      | 0.9868      | 0.9871      | 0.9875      | 0.9878      | 0.9881      | 0.9884      | 0.9887      | 0.9890      |
| 2.3      | 0.9893      | 0.9896      | 0.9898      | 0.9901      | 0.9904      | 0.9906      | 0.9909      | 0.9911      | 0.9913      | 0.9916      |
| 2.4      | 0.9918      | 0.9920      | 0.9922      | 0.9925      | 0.9927      | 0.9929      | 0.9931      | 0.9932      | 0.9934      | 0.9936      |
| 2.5      | 0.9938      | 0.9940      | 0.9941      | 0.9943      | 0.9945      | 0.9946      | 0.9948      | 0.9949      | 0.9951      | 0.9952      |
| 2.6      | 0.9953      | 0.9955      | 0.9956      | 0.9957      | 0.9959      | 0.9960      | 0.9961      | 0.9962      | 0.9963      | 0.9964      |
| 2.7      | 0.9965      | 0.9966      | 0.9967      | 0.9968      | 0.9969      | 0.9970      | 0.9971      | 0.9972      | 0.9973      | 0.9974      |
| 2.8      | 0.9974      | 0.9975      | 0.9976      | 0.9977      | 0.9977      | 0.9978      | 0.9979      | 0.9979      | 0.9980      | 0.9981      |
| 2.9      | 0.9981      | 0.9982      | 0.9982      | 0.9983      | 0.9984      | 0.9984      | 0.9985      | 0.9985      | 0.9986      | 0.9986      |
| 3.0      | 0.9987      | 0.9987      | 0.9987      | 0.9988      | 0.9988      | 0.9989      | 0.9989      | 0.9989      | 0.9990      | 0.9990      |
| 3.1      | 0.9990      | 0.9991      | 0.9991      | 0.9991      | 0.9992      | 0.9992      | 0.9992      | 0.9992      | 0.9993      | 0.9993      |
| 3.2      | 0.9993      | 0.9993      | 0.9994      | 0.9994      | 0.9994      | 0.9994      | 0.9994      | 0.9995      | 0.9995      | 0.9995      |
| 3.3      | 0.9995      | 0.9995      | 0.9995      | 0.9996      | 0.9996      | 0.9996      | 0.9996      | 0.9996      | 0.9996      | 0.9997      |
| 3.4      | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9998      |
| 3.5      | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      |
| 3.6      | 0.9998      | 0.9998      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      |
| 3.7      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      |
| 3.8      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      |
| 3.9      | 1.0000      | 1.0000      | 1.0000      | 1.0000      | 1.0000      | 1.0000      | 1.0000      | 1.0000      | 1.0000      | 1.0000      |

Source: Probabilities calculated with Excel.

**TABLE 2** Student's *t* Distribution

Entries in this table provide the values of  $t_{\alpha,df}$  that correspond to a given upper-tail area  $\alpha$  and a specified number of degrees of freedom  $df$ . For example, for  $\alpha = 0.05$  and  $df = 10$ ,  $P(T_{10} \geq 1.812) = 0.05$ .



| df | $\alpha$ |       |       |        |        |        |
|----|----------|-------|-------|--------|--------|--------|
|    | 0.20     | 0.10  | 0.05  | 0.025  | 0.01   | 0.005  |
| 1  | 1.376    | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2  | 1.061    | 1.886 | 2.920 | 4.303  | 6.965  | 9.925  |
| 3  | 0.978    | 1.638 | 2.353 | 3.182  | 4.541  | 5.841  |
| 4  | 0.941    | 1.533 | 2.132 | 2.776  | 3.747  | 4.604  |
| 5  | 0.920    | 1.476 | 2.015 | 2.571  | 3.365  | 4.032  |
| 6  | 0.906    | 1.440 | 1.943 | 2.447  | 3.143  | 3.707  |
| 7  | 0.896    | 1.415 | 1.895 | 2.365  | 2.998  | 3.499  |
| 8  | 0.889    | 1.397 | 1.860 | 2.306  | 2.896  | 3.355  |
| 9  | 0.883    | 1.383 | 1.833 | 2.262  | 2.821  | 3.250  |
| 10 | 0.879    | 1.372 | 1.812 | 2.228  | 2.764  | 3.169  |
| 11 | 0.876    | 1.363 | 1.796 | 2.201  | 2.718  | 3.106  |
| 12 | 0.873    | 1.356 | 1.782 | 2.179  | 2.681  | 3.055  |
| 13 | 0.870    | 1.350 | 1.771 | 2.160  | 2.650  | 3.012  |
| 14 | 0.868    | 1.345 | 1.761 | 2.145  | 2.624  | 2.977  |
| 15 | 0.866    | 1.341 | 1.753 | 2.131  | 2.602  | 2.947  |
| 16 | 0.865    | 1.337 | 1.746 | 2.120  | 2.583  | 2.921  |
| 17 | 0.863    | 1.333 | 1.740 | 2.110  | 2.567  | 2.898  |
| 18 | 0.862    | 1.330 | 1.734 | 2.101  | 2.552  | 2.878  |
| 19 | 0.861    | 1.328 | 1.729 | 2.093  | 2.539  | 2.861  |
| 20 | 0.860    | 1.325 | 1.725 | 2.086  | 2.528  | 2.845  |
| 21 | 0.859    | 1.323 | 1.721 | 2.080  | 2.518  | 2.831  |
| 22 | 0.858    | 1.321 | 1.717 | 2.074  | 2.508  | 2.819  |
| 23 | 0.858    | 1.319 | 1.714 | 2.069  | 2.500  | 2.807  |
| 24 | 0.857    | 1.318 | 1.711 | 2.064  | 2.492  | 2.797  |
| 25 | 0.856    | 1.316 | 1.708 | 2.060  | 2.485  | 2.787  |
| 26 | 0.856    | 1.315 | 1.706 | 2.056  | 2.479  | 2.779  |
| 27 | 0.855    | 1.314 | 1.703 | 2.052  | 2.473  | 2.771  |
| 28 | 0.855    | 1.313 | 1.701 | 2.048  | 2.467  | 2.763  |
| 29 | 0.854    | 1.311 | 1.699 | 2.045  | 2.462  | 2.756  |
| 30 | 0.854    | 1.310 | 1.697 | 2.042  | 2.457  | 2.750  |



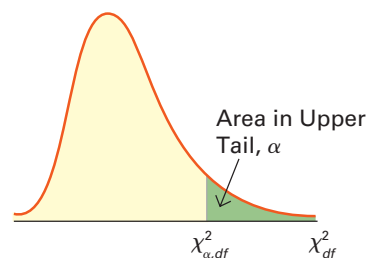
**TABLE 2** (Continued)

| df       | $\alpha$ |       |       |       |       |       |
|----------|----------|-------|-------|-------|-------|-------|
|          | 0.20     | 0.10  | 0.05  | 0.025 | 0.01  | 0.005 |
| 31       | 0.853    | 1.309 | 1.696 | 2.040 | 2.453 | 2.744 |
| 32       | 0.853    | 1.309 | 1.694 | 2.037 | 2.449 | 2.738 |
| 33       | 0.853    | 1.308 | 1.692 | 2.035 | 2.445 | 2.733 |
| 34       | 0.852    | 1.307 | 1.691 | 2.032 | 2.441 | 2.728 |
| 35       | 0.852    | 1.306 | 1.690 | 2.030 | 2.438 | 2.724 |
| 36       | 0.852    | 1.306 | 1.688 | 2.028 | 2.434 | 2.719 |
| 37       | 0.851    | 1.305 | 1.687 | 2.026 | 2.431 | 2.715 |
| 38       | 0.851    | 1.304 | 1.686 | 2.024 | 2.429 | 2.712 |
| 39       | 0.851    | 1.304 | 1.685 | 2.023 | 2.426 | 2.708 |
| 40       | 0.851    | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 |
|          |          |       |       |       |       |       |
| 41       | 0.850    | 1.303 | 1.683 | 2.020 | 2.421 | 2.701 |
| 42       | 0.850    | 1.302 | 1.682 | 2.018 | 2.418 | 2.698 |
| 43       | 0.850    | 1.302 | 1.681 | 2.017 | 2.416 | 2.695 |
| 44       | 0.850    | 1.301 | 1.680 | 2.015 | 2.414 | 2.692 |
| 45       | 0.850    | 1.301 | 1.679 | 2.014 | 2.412 | 2.690 |
| 46       | 0.850    | 1.300 | 1.679 | 2.013 | 2.410 | 2.687 |
| 47       | 0.849    | 1.300 | 1.678 | 2.012 | 2.408 | 2.685 |
| 48       | 0.849    | 1.299 | 1.677 | 2.011 | 2.407 | 2.682 |
| 49       | 0.849    | 1.299 | 1.677 | 2.010 | 2.405 | 2.680 |
| 50       | 0.849    | 1.299 | 1.676 | 2.009 | 2.403 | 2.678 |
|          |          |       |       |       |       |       |
| 51       | 0.849    | 1.298 | 1.675 | 2.008 | 2.402 | 2.676 |
| 52       | 0.849    | 1.298 | 1.675 | 2.007 | 2.400 | 2.674 |
| 53       | 0.848    | 1.298 | 1.674 | 2.006 | 2.399 | 2.672 |
| 54       | 0.848    | 1.297 | 1.674 | 2.005 | 2.397 | 2.670 |
| 55       | 0.848    | 1.297 | 1.673 | 2.004 | 2.396 | 2.668 |
| 56       | 0.848    | 1.297 | 1.673 | 2.003 | 2.395 | 2.667 |
| 57       | 0.848    | 1.297 | 1.672 | 2.002 | 2.394 | 2.665 |
| 58       | 0.848    | 1.296 | 1.672 | 2.002 | 2.392 | 2.663 |
| 59       | 0.848    | 1.296 | 1.671 | 2.001 | 2.391 | 2.662 |
| 60       | 0.848    | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 |
|          |          |       |       |       |       |       |
| 80       | 0.846    | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 |
| 100      | 0.845    | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 |
| 150      | 0.844    | 1.287 | 1.655 | 1.976 | 2.351 | 2.609 |
| 200      | 0.843    | 1.286 | 1.653 | 1.972 | 2.345 | 2.601 |
| 500      | 0.842    | 1.283 | 1.648 | 1.965 | 2.334 | 2.586 |
| 1000     | 0.842    | 1.282 | 1.646 | 1.962 | 2.330 | 2.581 |
| $\infty$ | 0.842    | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |

SOURCE: t values calculated with Excel.

**TABLE 3**  $\chi^2$  (Chi-Square) Distribution

Entries in this table provide the values of  $\chi^2_{\alpha,df}$  that correspond to a given upper-tail area  $\alpha$  and a specified number of degrees of freedom  $df$ . For example, for  $\alpha = 0.05$  and  $df = 10$ ,  $P(\chi^2_{10} \geq 18.307) = 0.05$ .



| df | $\alpha$ |        |        |        |        |        |        |        |        |        |
|----|----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
|    | 0.995    | 0.990  | 0.975  | 0.950  | 0.900  | 0.100  | 0.050  | 0.025  | 0.010  | 0.005  |
| 1  | 0.000    | 0.000  | 0.001  | 0.004  | 0.016  | 2.706  | 3.841  | 5.024  | 6.635  | 7.879  |
| 2  | 0.010    | 0.020  | 0.051  | 0.103  | 0.211  | 4.605  | 5.991  | 7.378  | 9.210  | 10.597 |
| 3  | 0.072    | 0.115  | 0.216  | 0.352  | 0.584  | 6.251  | 7.815  | 9.348  | 11.345 | 12.838 |
| 4  | 0.207    | 0.297  | 0.484  | 0.711  | 1.064  | 7.779  | 9.488  | 11.143 | 13.277 | 14.860 |
| 5  | 0.412    | 0.554  | 0.831  | 1.145  | 1.610  | 9.236  | 11.070 | 12.833 | 15.086 | 16.750 |
| 6  | 0.676    | 0.872  | 1.237  | 1.635  | 2.204  | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 |
| 7  | 0.989    | 1.239  | 1.690  | 2.167  | 2.833  | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 |
| 8  | 1.344    | 1.646  | 2.180  | 2.733  | 3.490  | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 |
| 9  | 1.735    | 2.088  | 2.700  | 3.325  | 4.168  | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 |
| 10 | 2.156    | 2.558  | 3.247  | 3.940  | 4.865  | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |
| 11 | 2.603    | 3.053  | 3.816  | 4.575  | 5.578  | 17.275 | 19.675 | 21.920 | 24.725 | 26.757 |
| 12 | 3.074    | 3.571  | 4.404  | 5.226  | 6.304  | 18.549 | 21.026 | 23.337 | 26.217 | 28.300 |
| 13 | 3.565    | 4.107  | 5.009  | 5.892  | 7.042  | 19.812 | 22.362 | 24.736 | 27.688 | 29.819 |
| 14 | 4.075    | 4.660  | 5.629  | 6.571  | 7.790  | 21.064 | 23.685 | 26.119 | 29.141 | 31.319 |
| 15 | 4.601    | 5.229  | 6.262  | 7.261  | 8.547  | 22.307 | 24.996 | 27.488 | 30.578 | 32.801 |
| 16 | 5.142    | 5.812  | 6.908  | 7.962  | 9.312  | 23.542 | 26.296 | 28.845 | 32.000 | 34.267 |
| 17 | 5.697    | 6.408  | 7.564  | 8.672  | 10.085 | 24.769 | 27.587 | 30.191 | 33.409 | 35.718 |
| 18 | 6.265    | 7.015  | 8.231  | 9.390  | 10.865 | 25.989 | 28.869 | 31.526 | 34.805 | 37.156 |
| 19 | 6.844    | 7.633  | 8.907  | 10.117 | 11.651 | 27.204 | 30.144 | 32.852 | 36.191 | 38.582 |
| 20 | 7.434    | 8.260  | 9.591  | 10.851 | 12.443 | 28.412 | 31.410 | 34.170 | 37.566 | 39.997 |
| 21 | 8.034    | 8.897  | 10.283 | 11.591 | 13.240 | 29.615 | 32.671 | 35.479 | 38.932 | 41.401 |
| 22 | 8.643    | 9.542  | 10.982 | 12.338 | 14.041 | 30.813 | 33.924 | 36.781 | 40.289 | 42.796 |
| 23 | 9.260    | 10.196 | 11.689 | 13.091 | 14.848 | 32.007 | 35.172 | 38.076 | 41.638 | 44.181 |
| 24 | 9.886    | 10.856 | 12.401 | 13.848 | 15.659 | 33.196 | 36.415 | 39.364 | 42.980 | 45.559 |
| 25 | 10.520   | 11.524 | 13.120 | 14.611 | 16.473 | 34.382 | 37.652 | 40.646 | 44.314 | 46.928 |
| 26 | 11.160   | 12.198 | 13.844 | 15.379 | 17.292 | 35.563 | 38.885 | 41.923 | 45.642 | 48.290 |
| 27 | 11.808   | 12.879 | 14.573 | 16.151 | 18.114 | 36.741 | 40.113 | 43.195 | 46.963 | 49.645 |
| 28 | 12.461   | 13.565 | 15.308 | 16.928 | 18.939 | 37.916 | 41.337 | 44.461 | 48.278 | 50.993 |
| 29 | 13.121   | 14.256 | 16.047 | 17.708 | 19.768 | 39.087 | 42.557 | 45.722 | 49.588 | 52.336 |
| 30 | 13.787   | 14.953 | 16.791 | 18.493 | 20.599 | 40.256 | 43.773 | 46.979 | 50.892 | 53.672 |

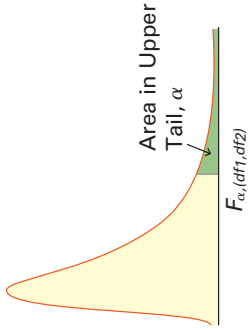
TABLE 3 (Continued)

| df  | $\alpha$ |        |        |        |        |         |         |         |         |         |
|-----|----------|--------|--------|--------|--------|---------|---------|---------|---------|---------|
|     | 0.995    | 0.990  | 0.975  | 0.950  | 0.900  | 0.100   | 0.050   | 0.025   | 0.010   | 0.005   |
| 31  | 14.458   | 15.655 | 17.539 | 19.281 | 21.434 | 41.422  | 44.985  | 48.232  | 52.191  | 55.003  |
| 32  | 15.134   | 16.362 | 18.291 | 20.072 | 22.271 | 42.585  | 46.194  | 49.480  | 53.486  | 56.328  |
| 33  | 15.815   | 17.074 | 19.047 | 20.867 | 23.110 | 43.745  | 47.400  | 50.725  | 54.776  | 57.648  |
| 34  | 16.501   | 17.789 | 19.806 | 21.664 | 23.952 | 44.903  | 48.602  | 51.966  | 56.061  | 58.964  |
| 35  | 17.192   | 18.509 | 20.569 | 22.465 | 24.797 | 46.059  | 49.802  | 53.203  | 57.342  | 60.275  |
| 36  | 17.887   | 19.233 | 21.336 | 23.269 | 25.643 | 47.212  | 50.998  | 54.437  | 58.619  | 61.581  |
| 37  | 18.586   | 19.960 | 22.106 | 24.075 | 26.492 | 48.363  | 52.192  | 55.668  | 59.893  | 62.883  |
| 38  | 19.289   | 20.691 | 22.878 | 24.884 | 27.343 | 49.513  | 53.384  | 56.896  | 61.162  | 64.181  |
| 39  | 19.996   | 21.426 | 23.654 | 25.695 | 28.196 | 50.660  | 54.572  | 58.120  | 62.428  | 65.476  |
| 40  | 20.707   | 22.164 | 24.433 | 26.509 | 29.051 | 51.805  | 55.758  | 59.342  | 63.691  | 66.766  |
| 41  | 21.421   | 22.906 | 25.215 | 27.326 | 29.907 | 52.949  | 56.942  | 60.561  | 64.950  | 68.053  |
| 42  | 22.138   | 23.650 | 25.999 | 28.144 | 30.765 | 54.090  | 58.124  | 61.777  | 66.206  | 69.336  |
| 43  | 22.859   | 24.398 | 26.785 | 28.965 | 31.625 | 55.230  | 59.304  | 62.990  | 67.459  | 70.616  |
| 44  | 23.584   | 25.148 | 27.575 | 29.787 | 32.487 | 56.369  | 60.481  | 64.201  | 68.710  | 71.893  |
| 45  | 24.311   | 25.901 | 28.366 | 30.612 | 33.350 | 57.505  | 61.656  | 65.410  | 69.957  | 73.166  |
| 46  | 25.041   | 26.657 | 29.160 | 31.439 | 34.215 | 58.641  | 62.830  | 66.617  | 71.201  | 74.437  |
| 47  | 25.775   | 27.416 | 29.956 | 32.268 | 35.081 | 59.774  | 64.001  | 67.821  | 72.443  | 75.704  |
| 48  | 26.511   | 28.177 | 30.755 | 33.098 | 35.949 | 60.907  | 65.171  | 69.023  | 73.683  | 76.969  |
| 49  | 27.249   | 28.941 | 31.555 | 33.930 | 36.818 | 62.038  | 66.339  | 70.222  | 74.919  | 78.231  |
| 50  | 27.991   | 29.707 | 32.357 | 34.764 | 37.689 | 63.167  | 67.505  | 71.420  | 76.154  | 79.490  |
| 55  | 31.735   | 33.570 | 36.398 | 38.958 | 42.060 | 68.796  | 73.311  | 77.380  | 82.292  | 85.749  |
| 60  | 35.534   | 37.485 | 40.482 | 43.188 | 46.459 | 74.397  | 79.082  | 83.298  | 88.379  | 91.952  |
| 65  | 39.383   | 41.444 | 44.603 | 47.450 | 50.883 | 79.973  | 84.821  | 89.177  | 94.422  | 98.105  |
| 70  | 43.275   | 45.442 | 48.758 | 51.739 | 55.329 | 85.527  | 90.531  | 95.023  | 100.425 | 104.215 |
| 75  | 47.206   | 49.475 | 52.942 | 56.054 | 59.795 | 91.061  | 96.217  | 100.839 | 106.393 | 110.286 |
| 80  | 51.172   | 53.540 | 57.153 | 60.391 | 64.278 | 96.578  | 101.879 | 106.629 | 112.329 | 116.321 |
| 85  | 55.170   | 57.634 | 61.389 | 64.749 | 68.777 | 102.079 | 107.522 | 112.393 | 118.236 | 122.325 |
| 90  | 59.196   | 61.754 | 65.647 | 69.126 | 73.291 | 107.565 | 113.145 | 118.136 | 124.116 | 128.299 |
| 95  | 63.250   | 65.898 | 69.925 | 73.520 | 77.818 | 113.038 | 118.752 | 123.858 | 129.973 | 134.247 |
| 100 | 67.328   | 70.065 | 74.222 | 77.929 | 82.358 | 118.498 | 124.342 | 129.561 | 135.807 | 140.169 |

SOURCE:  $\chi^2$  values calculated with Excel.

**TABLE 4** *F* Distribution

Entries in this table provide the values of  $F_{\alpha(df_1, df_2)}$  that correspond to a given upper-tail area  $\alpha$  and a specified number of degrees of freedom in the numerator  $df_1$  and degrees of freedom in the denominator  $df_2$ . For example, for  $\alpha = 0.05$ ,  $df_1 = 8$ , and  $df_2 = 6$ ,  $P(F_{(8,6)} \geq 4.15) = 0.05$ .



| $df_2$ | $\alpha$ | $df_1$  |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
|--------|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
|        |          | 1       | 2       | 3       | 4       | 5       | 6       | 7       | 8       | 9       | 10      | 15      | 25      | 50      | 100     | 500     |
| 1      | 0.10     | 39.86   | 49.50   | 53.59   | 55.83   | 57.24   | 58.2    | 58.91   | 59.44   | 59.86   | 60.19   | 61.22   | 62.05   | 62.69   | 63.01   | 63.26   |
|        | 0.05     | 161.45  | 199.50  | 215.71  | 224.58  | 230.16  | 233.99  | 236.77  | 238.88  | 240.54  | 241.88  | 245.95  | 249.26  | 251.77  | 253.04  | 254.06  |
|        | 0.025    | 647.79  | 799.50  | 864.16  | 899.58  | 921.85  | 937.11  | 948.22  | 956.66  | 963.28  | 968.63  | 984.87  | 998.08  | 1008.12 | 1013.17 | 1017.24 |
|        | 0.01     | 4052.18 | 4999.50 | 5403.35 | 5624.58 | 5763.65 | 5858.99 | 5928.36 | 5981.07 | 6022.47 | 6055.85 | 6157.28 | 6239.83 | 6302.52 | 6334.11 | 6359.50 |
| 2      | 0.10     | 8.53    | 9.00    | 9.16    | 9.24    | 9.29    | 9.33    | 9.35    | 9.37    | 9.38    | 9.39    | 9.42    | 9.45    | 9.47    | 9.48    | 9.49    |
|        | 0.05     | 18.51   | 19.00   | 19.16   | 19.25   | 19.30   | 19.33   | 19.35   | 19.37   | 19.38   | 19.40   | 19.43   | 19.46   | 19.48   | 19.49   | 19.49   |
|        | 0.025    | 38.51   | 39.00   | 39.17   | 39.25   | 39.30   | 39.33   | 39.36   | 39.37   | 39.39   | 39.40   | 39.43   | 39.46   | 39.48   | 39.49   | 39.50   |
|        | 0.01     | 98.50   | 99.00   | 99.17   | 99.25   | 99.30   | 99.33   | 99.36   | 99.37   | 99.39   | 99.40   | 99.43   | 99.46   | 99.48   | 99.49   | 99.50   |
| 3      | 0.10     | 5.54    | 5.46    | 5.39    | 5.34    | 5.31    | 5.28    | 5.27    | 5.25    | 5.24    | 5.23    | 5.20    | 5.17    | 5.15    | 5.14    | 5.14    |
|        | 0.05     | 10.13   | 9.55    | 9.28    | 9.12    | 9.01    | 8.94    | 8.89    | 8.85    | 8.81    | 8.79    | 8.70    | 8.63    | 8.58    | 8.55    | 8.53    |
|        | 0.025    | 17.44   | 16.04   | 15.44   | 15.10   | 14.88   | 14.73   | 14.62   | 14.54   | 14.47   | 14.42   | 14.25   | 14.12   | 14.01   | 13.96   | 13.91   |
|        | 0.01     | 34.12   | 30.82   | 29.46   | 28.71   | 28.24   | 27.91   | 27.67   | 27.49   | 27.35   | 27.23   | 26.87   | 26.58   | 26.35   | 26.24   | 26.15   |
| 4      | 0.10     | 4.54    | 4.32    | 4.19    | 4.11    | 4.05    | 4.01    | 3.98    | 3.95    | 3.94    | 3.92    | 3.87    | 3.83    | 3.80    | 3.78    | 3.76    |
|        | 0.05     | 7.71    | 6.94    | 6.59    | 6.39    | 6.26    | 6.16    | 6.09    | 6.04    | 6.00    | 5.96    | 5.86    | 5.77    | 5.70    | 5.66    | 5.64    |
|        | 0.025    | 12.22   | 10.65   | 9.98    | 9.60    | 9.36    | 9.20    | 9.07    | 8.98    | 8.90    | 8.84    | 8.66    | 8.50    | 8.38    | 8.32    | 8.27    |
|        | 0.01     | 21.20   | 18.00   | 16.69   | 15.98   | 15.52   | 15.21   | 14.98   | 14.80   | 14.66   | 14.55   | 14.20   | 13.91   | 13.69   | 13.58   | 13.49   |
| 5      | 0.10     | 4.06    | 3.78    | 3.62    | 3.52    | 3.45    | 3.40    | 3.37    | 3.34    | 3.32    | 3.30    | 3.24    | 3.19    | 3.15    | 3.13    | 3.11    |
|        | 0.05     | 6.61    | 5.79    | 5.41    | 5.19    | 5.05    | 4.95    | 4.88    | 4.82    | 4.77    | 4.74    | 4.62    | 4.52    | 4.44    | 4.41    | 4.37    |
|        | 0.025    | 10.01   | 8.43    | 7.76    | 7.39    | 7.15    | 6.98    | 6.85    | 6.76    | 6.68    | 6.62    | 6.43    | 6.27    | 6.14    | 6.08    | 6.03    |
|        | 0.01     | 16.26   | 13.27   | 12.06   | 11.39   | 10.97   | 10.67   | 10.46   | 10.29   | 10.16   | 10.05   | 9.72    | 9.45    | 9.24    | 9.13    | 9.04    |
| 6      | 0.10     | 3.78    | 3.46    | 3.29    | 3.18    | 3.11    | 3.05    | 3.01    | 2.98    | 2.96    | 2.94    | 2.87    | 2.81    | 2.77    | 2.75    | 2.73    |
|        | 0.05     | 5.99    | 5.14    | 4.76    | 4.53    | 4.39    | 4.28    | 4.21    | 4.15    | 4.10    | 4.06    | 3.94    | 3.83    | 3.75    | 3.71    | 3.68    |
|        | 0.025    | 8.81    | 7.26    | 6.60    | 6.23    | 5.99    | 5.82    | 5.70    | 5.60    | 5.52    | 5.46    | 5.27    | 5.11    | 4.98    | 4.92    | 4.86    |
|        | 0.01     | 13.75   | 10.92   | 9.78    | 9.15    | 8.75    | 8.47    | 8.26    | 8.10    | 7.98    | 7.87    | 7.56    | 7.30    | 7.09    | 6.99    | 6.90    |
| 7      | 0.10     | 3.59    | 3.26    | 3.07    | 2.96    | 2.88    | 2.83    | 2.78    | 2.75    | 2.72    | 2.70    | 2.63    | 2.57    | 2.52    | 2.50    | 2.48    |
|        | 0.05     | 5.59    | 4.74    | 4.35    | 4.12    | 3.97    | 3.87    | 3.79    | 3.73    | 3.68    | 3.64    | 3.51    | 3.40    | 3.32    | 3.27    | 3.24    |
|        | 0.025    | 8.07    | 6.54    | 5.89    | 5.52    | 5.29    | 5.12    | 4.99    | 4.90    | 4.82    | 4.76    | 4.57    | 4.40    | 4.28    | 4.21    | 4.16    |
|        | 0.01     | 12.25   | 9.55    | 8.45    | 7.85    | 7.46    | 7.19    | 6.99    | 6.84    | 6.72    | 6.62    | 6.31    | 6.06    | 5.86    | 5.75    | 5.67    |

|                 |       | df <sub>1</sub> |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|-----------------|-------|-----------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| df <sub>2</sub> | α     | 1               | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | 15   | 25   | 50   | 100  | 500  |
| 8               | 0.10  | 3.46            | 3.11 | 2.92 | 2.81 | 2.73 | 2.67 | 2.62 | 2.59 | 2.56 | 2.54 | 2.46 | 2.40 | 2.35 | 2.32 | 2.30 |
|                 | 0.05  | 5.32            | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.22 | 3.11 | 3.02 | 2.97 | 2.94 |
|                 | 0.025 | 7.57            | 6.06 | 5.42 | 5.05 | 4.82 | 4.65 | 4.53 | 4.43 | 4.36 | 4.30 | 4.10 | 3.94 | 3.81 | 3.74 | 3.68 |
|                 | 0.01  | 11.26           | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.91 | 5.81 | 5.52 | 5.26 | 5.07 | 4.96 | 4.88 |
| 9               | 0.10  | 3.36            | 3.01 | 2.81 | 2.69 | 2.61 | 2.55 | 2.51 | 2.47 | 2.44 | 2.42 | 2.34 | 2.27 | 2.22 | 2.19 | 2.17 |
|                 | 0.05  | 5.12            | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 3.01 | 2.89 | 2.80 | 2.76 | 2.72 |
|                 | 0.025 | 7.21            | 5.71 | 5.08 | 4.72 | 4.48 | 4.32 | 4.20 | 4.10 | 4.03 | 3.96 | 3.77 | 3.60 | 3.47 | 3.40 | 3.35 |
|                 | 0.01  | 10.56           | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.35 | 5.26 | 4.96 | 4.71 | 4.52 | 4.41 | 4.33 |
| 10              | 0.10  | 3.29            | 2.92 | 2.73 | 2.61 | 2.52 | 2.46 | 2.41 | 2.38 | 2.35 | 2.32 | 2.24 | 2.17 | 2.12 | 2.09 | 2.06 |
|                 | 0.05  | 4.96            | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.85 | 2.73 | 2.64 | 2.59 | 2.55 |
|                 | 0.025 | 6.94            | 5.46 | 4.83 | 4.47 | 4.24 | 4.07 | 3.95 | 3.85 | 3.78 | 3.72 | 3.52 | 3.35 | 3.22 | 3.15 | 3.09 |
|                 | 0.01  | 10.04           | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.94 | 4.85 | 4.56 | 4.31 | 4.12 | 4.01 | 3.93 |
| 11              | 0.10  | 3.23            | 2.86 | 2.66 | 2.54 | 2.45 | 2.39 | 2.34 | 2.30 | 2.27 | 2.25 | 2.17 | 2.10 | 2.04 | 2.01 | 1.98 |
|                 | 0.05  | 4.84            | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.72 | 2.60 | 2.51 | 2.46 | 2.42 |
|                 | 0.025 | 6.72            | 5.26 | 4.63 | 4.28 | 4.04 | 3.88 | 3.76 | 3.66 | 3.59 | 3.53 | 3.33 | 3.16 | 3.03 | 2.96 | 2.90 |
|                 | 0.01  | 9.65            | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.74 | 4.63 | 4.54 | 4.25 | 4.01 | 3.81 | 3.71 | 3.62 |
| 12              | 0.10  | 3.18            | 2.81 | 2.61 | 2.48 | 2.39 | 2.33 | 2.28 | 2.24 | 2.21 | 2.19 | 2.10 | 2.03 | 1.97 | 1.94 | 1.91 |
|                 | 0.05  | 4.75            | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.62 | 2.50 | 2.40 | 2.35 | 2.31 |
|                 | 0.025 | 6.55            | 5.10 | 4.47 | 4.12 | 3.89 | 3.73 | 3.61 | 3.51 | 3.44 | 3.37 | 3.18 | 3.01 | 2.87 | 2.80 | 2.74 |
|                 | 0.01  | 9.33            | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.64 | 4.50 | 4.39 | 4.30 | 4.01 | 3.76 | 3.57 | 3.47 | 3.38 |
| 13              | 0.10  | 3.14            | 2.76 | 2.56 | 2.43 | 2.35 | 2.28 | 2.23 | 2.20 | 2.16 | 2.14 | 2.05 | 1.98 | 1.92 | 1.88 | 1.85 |
|                 | 0.05  | 4.67            | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 | 2.53 | 2.41 | 2.31 | 2.26 | 2.22 |
|                 | 0.025 | 6.41            | 4.97 | 4.35 | 4.00 | 3.77 | 3.60 | 3.48 | 3.39 | 3.31 | 3.25 | 3.05 | 2.88 | 2.74 | 2.67 | 2.61 |
|                 | 0.01  | 9.07            | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 | 4.10 | 3.82 | 3.57 | 3.38 | 3.27 | 3.19 |
| 14              | 0.10  | 3.10            | 2.73 | 2.52 | 2.39 | 2.31 | 2.24 | 2.19 | 2.15 | 2.12 | 2.10 | 2.01 | 1.93 | 1.87 | 1.83 | 1.80 |
|                 | 0.05  | 4.60            | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 | 2.46 | 2.34 | 2.24 | 2.19 | 2.14 |
|                 | 0.025 | 6.30            | 4.86 | 4.24 | 3.89 | 3.66 | 3.50 | 3.38 | 3.29 | 3.21 | 3.15 | 2.95 | 2.78 | 2.64 | 2.56 | 2.50 |
|                 | 0.01  | 8.86            | 6.51 | 5.56 | 5.04 | 4.69 | 4.46 | 4.28 | 4.14 | 4.03 | 3.94 | 3.66 | 3.41 | 3.22 | 3.11 | 3.03 |
| 15              | 0.10  | 3.07            | 2.70 | 2.49 | 2.36 | 2.27 | 2.21 | 2.16 | 2.12 | 2.09 | 2.06 | 1.97 | 1.89 | 1.83 | 1.79 | 1.76 |
|                 | 0.05  | 4.54            | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 | 2.40 | 2.28 | 2.18 | 2.12 | 2.08 |
|                 | 0.025 | 6.20            | 4.77 | 4.15 | 3.80 | 3.58 | 3.41 | 3.29 | 3.20 | 3.12 | 3.06 | 2.86 | 2.69 | 2.55 | 2.47 | 2.41 |
|                 | 0.01  | 8.68            | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 | 3.80 | 3.52 | 3.28 | 3.08 | 2.98 | 2.89 |
| 16              | 0.10  | 3.05            | 2.67 | 2.46 | 2.33 | 2.24 | 2.18 | 2.13 | 2.09 | 2.06 | 2.03 | 1.94 | 1.86 | 1.79 | 1.76 | 1.73 |
|                 | 0.05  | 4.49            | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 | 2.35 | 2.23 | 2.12 | 2.07 | 2.02 |
|                 | 0.025 | 6.12            | 4.69 | 4.08 | 3.73 | 3.50 | 3.34 | 3.22 | 3.12 | 3.05 | 2.99 | 2.79 | 2.61 | 2.47 | 2.40 | 2.33 |
|                 | 0.01  | 8.53            | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.78 | 3.69 | 3.41 | 3.16 | 2.97 | 2.86 | 2.78 |

TABLE 4 (Continued)

|                 |       | df <sub>1</sub> |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|-----------------|-------|-----------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| df <sub>2</sub> | α     | 1               | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | 15   | 25   | 50   | 100  | 500  |
| 17              | 0.10  | 3.03            | 2.64 | 2.44 | 2.31 | 2.22 | 2.15 | 2.10 | 2.06 | 2.03 | 2.00 | 1.91 | 1.83 | 1.76 | 1.73 | 1.69 |
|                 | 0.05  | 4.45            | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 | 2.31 | 2.18 | 2.08 | 2.02 | 1.97 |
|                 | 0.025 | 6.04            | 4.62 | 4.01 | 3.66 | 3.44 | 3.28 | 3.16 | 3.06 | 2.98 | 2.92 | 2.72 | 2.55 | 2.41 | 2.33 | 2.26 |
|                 | 0.01  | 8.40            | 6.11 | 5.18 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.68 | 3.59 | 3.31 | 3.07 | 2.87 | 2.76 | 2.68 |
| 18              | 0.10  | 3.01            | 2.62 | 2.42 | 2.29 | 2.20 | 2.13 | 2.08 | 2.04 | 2.00 | 1.98 | 1.89 | 1.80 | 1.74 | 1.70 | 1.67 |
|                 | 0.05  | 4.41            | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 | 2.27 | 2.14 | 2.04 | 1.98 | 1.93 |
|                 | 0.025 | 5.98            | 4.56 | 3.95 | 3.61 | 3.38 | 3.22 | 3.10 | 3.01 | 2.93 | 2.87 | 2.67 | 2.49 | 2.35 | 2.27 | 2.20 |
|                 | 0.01  | 8.29            | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.60 | 3.51 | 3.23 | 2.98 | 2.78 | 2.68 | 2.59 |
| 19              | 0.10  | 2.99            | 2.61 | 2.40 | 2.27 | 2.18 | 2.11 | 2.06 | 2.02 | 1.98 | 1.96 | 1.86 | 1.78 | 1.71 | 1.67 | 1.64 |
|                 | 0.05  | 4.38            | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 | 2.23 | 2.11 | 2.00 | 1.94 | 1.89 |
|                 | 0.025 | 5.92            | 4.51 | 3.90 | 3.56 | 3.33 | 3.17 | 3.05 | 2.96 | 2.88 | 2.82 | 2.62 | 2.44 | 2.30 | 2.22 | 2.15 |
|                 | 0.01  | 8.18            | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.52 | 3.43 | 3.15 | 2.91 | 2.71 | 2.60 | 2.51 |
| 20              | 0.10  | 2.97            | 2.59 | 2.38 | 2.25 | 2.16 | 2.09 | 2.04 | 2.00 | 1.96 | 1.94 | 1.84 | 1.76 | 1.69 | 1.65 | 1.62 |
|                 | 0.05  | 4.35            | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 | 2.20 | 2.07 | 1.97 | 1.91 | 1.86 |
|                 | 0.025 | 5.87            | 4.46 | 3.86 | 3.51 | 3.29 | 3.13 | 3.01 | 2.91 | 2.84 | 2.77 | 2.57 | 2.40 | 2.25 | 2.17 | 2.10 |
|                 | 0.01  | 8.10            | 5.85 | 4.94 | 4.43 | 4.10 | 3.87 | 3.70 | 3.56 | 3.46 | 3.37 | 3.09 | 2.84 | 2.64 | 2.54 | 2.44 |
| 21              | 0.10  | 2.96            | 2.57 | 2.36 | 2.23 | 2.14 | 2.08 | 2.02 | 1.98 | 1.95 | 1.92 | 1.83 | 1.74 | 1.67 | 1.63 | 1.60 |
|                 | 0.05  | 4.32            | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 | 2.18 | 2.05 | 1.94 | 1.88 | 1.83 |
|                 | 0.025 | 5.83            | 4.42 | 3.82 | 3.48 | 3.25 | 3.09 | 2.97 | 2.87 | 2.80 | 2.73 | 2.53 | 2.36 | 2.21 | 2.13 | 2.06 |
|                 | 0.01  | 8.02            | 5.78 | 4.87 | 4.37 | 4.04 | 3.81 | 3.64 | 3.51 | 3.40 | 3.31 | 3.03 | 2.79 | 2.58 | 2.48 | 2.38 |
| 22              | 0.10  | 2.95            | 2.56 | 2.35 | 2.22 | 2.13 | 2.06 | 2.01 | 1.97 | 1.93 | 1.90 | 1.81 | 1.73 | 1.65 | 1.61 | 1.58 |
|                 | 0.05  | 4.30            | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 | 2.15 | 2.02 | 1.91 | 1.85 | 1.80 |
|                 | 0.025 | 5.79            | 4.38 | 3.78 | 3.44 | 3.22 | 3.05 | 2.93 | 2.84 | 2.76 | 2.70 | 2.50 | 2.32 | 2.17 | 2.09 | 2.02 |
|                 | 0.01  | 7.95            | 5.72 | 4.82 | 4.31 | 3.99 | 3.76 | 3.59 | 3.45 | 3.35 | 3.26 | 2.98 | 2.73 | 2.53 | 2.42 | 2.33 |
| 23              | 0.10  | 2.94            | 2.55 | 2.34 | 2.21 | 2.11 | 2.05 | 1.99 | 1.95 | 1.92 | 1.89 | 1.80 | 1.71 | 1.64 | 1.59 | 1.56 |
|                 | 0.05  | 4.28            | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 | 2.27 | 2.13 | 2.00 | 1.88 | 1.82 | 1.77 |
|                 | 0.025 | 5.75            | 4.35 | 3.75 | 3.41 | 3.18 | 3.02 | 2.90 | 2.81 | 2.73 | 2.67 | 2.47 | 2.29 | 2.14 | 2.06 | 1.99 |
|                 | 0.01  | 7.88            | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.30 | 3.21 | 2.93 | 2.69 | 2.48 | 2.37 | 2.28 |
| 24              | 0.10  | 2.93            | 2.54 | 2.33 | 2.19 | 2.10 | 2.04 | 1.98 | 1.94 | 1.91 | 1.88 | 1.78 | 1.70 | 1.62 | 1.58 | 1.54 |
|                 | 0.05  | 4.26            | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 | 2.25 | 2.11 | 1.97 | 1.86 | 1.80 | 1.75 |
|                 | 0.025 | 5.72            | 4.32 | 3.72 | 3.38 | 3.15 | 2.99 | 2.87 | 2.78 | 2.70 | 2.64 | 2.44 | 2.26 | 2.11 | 2.02 | 1.95 |
|                 | 0.01  | 7.82            | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.26 | 3.17 | 2.89 | 2.64 | 2.44 | 2.33 | 2.24 |



|                 |       | df <sub>1</sub> |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|-----------------|-------|-----------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| df <sub>2</sub> | α     | 1               | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | 15   | 25   | 50   | 100  | 500  |
| 25              | 0.10  | 2.92            | 2.53 | 2.32 | 2.18 | 2.09 | 2.02 | 1.97 | 1.93 | 1.89 | 1.87 | 1.77 | 1.68 | 1.61 | 1.56 | 1.53 |
|                 | 0.05  | 4.24            | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 | 2.09 | 1.96 | 1.84 | 1.78 | 1.73 |
|                 | 0.025 | 5.69            | 4.29 | 3.69 | 3.35 | 3.13 | 2.97 | 2.85 | 2.75 | 2.68 | 2.61 | 2.41 | 2.23 | 2.08 | 2.00 | 1.92 |
|                 | 0.01  | 7.77            | 5.57 | 4.68 | 4.18 | 3.85 | 3.63 | 3.46 | 3.32 | 3.22 | 3.13 | 2.85 | 2.60 | 2.40 | 2.29 | 2.19 |
| 26              | 0.10  | 2.91            | 2.52 | 2.31 | 2.17 | 2.08 | 2.01 | 1.96 | 1.92 | 1.88 | 1.86 | 1.76 | 1.67 | 1.59 | 1.55 | 1.51 |
|                 | 0.05  | 4.23            | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 | 2.07 | 1.94 | 1.82 | 1.76 | 1.71 |
|                 | 0.025 | 5.66            | 4.27 | 3.67 | 3.33 | 3.10 | 2.94 | 2.82 | 2.73 | 2.65 | 2.59 | 2.39 | 2.21 | 2.05 | 1.97 | 1.90 |
|                 | 0.01  | 7.72            | 5.53 | 4.64 | 4.14 | 3.82 | 3.59 | 3.42 | 3.29 | 3.18 | 3.09 | 2.81 | 2.57 | 2.36 | 2.25 | 2.16 |
| 27              | 0.10  | 2.90            | 2.51 | 2.30 | 2.17 | 2.07 | 2.00 | 1.95 | 1.91 | 1.87 | 1.85 | 1.75 | 1.66 | 1.58 | 1.54 | 1.50 |
|                 | 0.05  | 4.21            | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 | 2.20 | 2.06 | 1.92 | 1.81 | 1.74 | 1.69 |
|                 | 0.025 | 5.63            | 4.24 | 3.65 | 3.31 | 3.08 | 2.92 | 2.80 | 2.71 | 2.63 | 2.57 | 2.36 | 2.18 | 2.03 | 1.94 | 1.87 |
|                 | 0.01  | 7.68            | 5.49 | 4.60 | 4.11 | 3.78 | 3.56 | 3.39 | 3.26 | 3.15 | 3.06 | 2.78 | 2.54 | 2.33 | 2.22 | 2.12 |
| 28              | 0.10  | 2.89            | 2.50 | 2.29 | 2.16 | 2.06 | 2.00 | 1.94 | 1.90 | 1.87 | 1.84 | 1.74 | 1.65 | 1.57 | 1.53 | 1.49 |
|                 | 0.05  | 4.20            | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 | 2.19 | 2.04 | 1.91 | 1.79 | 1.73 | 1.67 |
|                 | 0.025 | 5.61            | 4.22 | 3.63 | 3.29 | 3.06 | 2.90 | 2.78 | 2.69 | 2.61 | 2.55 | 2.34 | 2.16 | 2.01 | 1.92 | 1.85 |
|                 | 0.01  | 7.64            | 5.45 | 4.57 | 4.07 | 3.75 | 3.53 | 3.36 | 3.23 | 3.12 | 3.03 | 2.75 | 2.51 | 2.30 | 2.19 | 2.09 |
| 29              | 0.10  | 2.89            | 2.50 | 2.28 | 2.15 | 2.06 | 1.99 | 1.93 | 1.89 | 1.86 | 1.83 | 1.73 | 1.64 | 1.56 | 1.52 | 1.48 |
|                 | 0.05  | 4.18            | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 | 2.18 | 2.03 | 1.89 | 1.77 | 1.71 | 1.65 |
|                 | 0.025 | 5.59            | 4.20 | 3.61 | 3.27 | 3.04 | 2.88 | 2.76 | 2.67 | 2.59 | 2.53 | 2.32 | 2.14 | 1.99 | 1.90 | 1.83 |
|                 | 0.01  | 7.60            | 5.42 | 4.54 | 4.04 | 3.73 | 3.50 | 3.33 | 3.20 | 3.09 | 3.00 | 2.73 | 2.48 | 2.27 | 2.16 | 2.06 |
| 30              | 0.10  | 2.88            | 2.49 | 2.28 | 2.14 | 2.05 | 1.98 | 1.93 | 1.88 | 1.85 | 1.82 | 1.72 | 1.63 | 1.55 | 1.51 | 1.47 |
|                 | 0.05  | 4.17            | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.16 | 2.01 | 1.88 | 1.76 | 1.70 | 1.64 |
|                 | 0.025 | 5.57            | 4.18 | 3.59 | 3.25 | 3.03 | 2.87 | 2.75 | 2.65 | 2.57 | 2.51 | 2.31 | 2.12 | 1.97 | 1.88 | 1.81 |
|                 | 0.01  | 7.56            | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.07 | 2.98 | 2.70 | 2.45 | 2.25 | 2.13 | 2.03 |
| 50              | 0.10  | 2.81            | 2.41 | 2.20 | 2.06 | 1.97 | 1.90 | 1.84 | 1.80 | 1.76 | 1.73 | 1.63 | 1.53 | 1.44 | 1.39 | 1.34 |
|                 | 0.05  | 4.03            | 3.18 | 2.79 | 2.56 | 2.40 | 2.29 | 2.20 | 2.13 | 2.07 | 2.03 | 1.87 | 1.73 | 1.60 | 1.52 | 1.46 |
|                 | 0.025 | 5.34            | 3.97 | 3.39 | 3.05 | 2.83 | 2.67 | 2.55 | 2.46 | 2.38 | 2.32 | 2.11 | 1.92 | 1.75 | 1.66 | 1.57 |
|                 | 0.01  | 7.17            | 5.06 | 4.20 | 3.72 | 3.41 | 3.19 | 3.02 | 2.89 | 2.78 | 2.70 | 2.42 | 2.17 | 1.95 | 1.82 | 1.71 |
| 100             | 0.10  | 2.76            | 2.36 | 2.14 | 2.00 | 1.91 | 1.83 | 1.78 | 1.73 | 1.69 | 1.66 | 1.56 | 1.45 | 1.35 | 1.29 | 1.23 |
|                 | 0.05  | 3.94            | 3.09 | 2.70 | 2.46 | 2.31 | 2.19 | 2.10 | 2.03 | 1.97 | 1.93 | 1.77 | 1.62 | 1.48 | 1.39 | 1.31 |
|                 | 0.025 | 5.18            | 3.83 | 3.25 | 2.92 | 2.70 | 2.54 | 2.42 | 2.32 | 2.24 | 2.18 | 1.97 | 1.77 | 1.59 | 1.48 | 1.38 |
|                 | 0.01  | 6.90            | 4.82 | 3.98 | 3.51 | 3.21 | 2.99 | 2.82 | 2.69 | 2.59 | 2.50 | 2.22 | 1.97 | 1.74 | 1.60 | 1.47 |
| 500             | 0.10  | 2.72            | 2.31 | 2.09 | 1.96 | 1.86 | 1.79 | 1.73 | 1.68 | 1.64 | 1.61 | 1.50 | 1.39 | 1.28 | 1.21 | 1.12 |
|                 | 0.05  | 3.86            | 3.01 | 2.62 | 2.39 | 2.23 | 2.12 | 2.03 | 1.96 | 1.90 | 1.85 | 1.69 | 1.53 | 1.38 | 1.28 | 1.16 |
|                 | 0.025 | 5.05            | 3.72 | 3.14 | 2.81 | 2.59 | 2.43 | 2.31 | 2.22 | 2.14 | 2.07 | 1.86 | 1.65 | 1.46 | 1.34 | 1.19 |
|                 | 0.01  | 6.69            | 4.65 | 3.82 | 3.36 | 3.05 | 2.84 | 2.68 | 2.55 | 2.44 | 2.36 | 2.07 | 1.81 | 1.57 | 1.41 | 1.23 |

Source: F-values calculated with Excel.

**TABLE 5** Studentized Range Values  $q_{\alpha, c, n_T - c}$  for Tukey's HSD Method

| $n_T - c$ | $\alpha$ | The number of means, $c$ |      |      |      |      |      |      |      |       |       |       |
|-----------|----------|--------------------------|------|------|------|------|------|------|------|-------|-------|-------|
|           |          | 2                        | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10    | 11    | 12    |
| 5         | 0.05     | 3.64                     | 4.60 | 5.22 | 5.67 | 6.03 | 6.33 | 6.58 | 6.80 | 6.99  | 7.17  | 7.32  |
|           | 0.01     | 5.70                     | 6.98 | 7.80 | 8.42 | 8.91 | 9.32 | 9.67 | 9.97 | 10.24 | 10.48 | 10.70 |
| 6         | 0.05     | 3.46                     | 4.34 | 4.90 | 5.30 | 5.63 | 5.90 | 6.12 | 6.32 | 6.49  | 6.65  | 6.79  |
|           | 0.01     | 5.24                     | 6.33 | 7.03 | 7.56 | 7.97 | 8.32 | 8.61 | 8.87 | 9.10  | 9.30  | 9.48  |
| 7         | 0.05     | 3.34                     | 4.16 | 4.68 | 5.06 | 5.36 | 5.61 | 5.82 | 6.00 | 6.16  | 6.30  | 6.43  |
|           | 0.01     | 4.95                     | 5.92 | 6.54 | 7.01 | 7.37 | 7.68 | 7.94 | 8.17 | 8.37  | 8.55  | 8.71  |
| 8         | 0.05     | 3.26                     | 4.04 | 4.53 | 4.89 | 5.17 | 5.40 | 5.60 | 5.77 | 5.92  | 6.05  | 6.18  |
|           | 0.01     | 4.75                     | 5.64 | 6.20 | 6.62 | 6.96 | 7.24 | 7.47 | 7.68 | 7.86  | 8.03  | 8.18  |
| 9         | 0.05     | 3.20                     | 3.95 | 4.41 | 4.76 | 5.02 | 5.24 | 5.43 | 5.59 | 5.74  | 5.87  | 5.98  |
|           | 0.01     | 4.60                     | 5.43 | 5.96 | 6.35 | 6.66 | 6.91 | 7.13 | 7.33 | 7.49  | 7.65  | 7.78  |
| 10        | 0.05     | 3.15                     | 3.88 | 4.33 | 4.65 | 4.91 | 5.12 | 5.30 | 5.46 | 5.60  | 5.72  | 5.83  |
|           | 0.01     | 4.48                     | 5.27 | 5.77 | 6.14 | 6.43 | 6.67 | 6.87 | 7.05 | 7.21  | 7.36  | 7.49  |
| 11        | 0.05     | 3.11                     | 3.82 | 4.26 | 4.57 | 4.82 | 5.03 | 5.20 | 5.35 | 5.49  | 5.61  | 5.71  |
|           | 0.01     | 4.39                     | 5.15 | 5.62 | 5.97 | 6.25 | 6.48 | 6.67 | 6.84 | 6.99  | 7.13  | 7.25  |
| 12        | 0.05     | 3.08                     | 3.77 | 4.20 | 4.51 | 4.75 | 4.95 | 5.12 | 5.27 | 5.39  | 5.51  | 5.61  |
|           | 0.01     | 4.32                     | 5.05 | 5.50 | 5.84 | 6.10 | 6.32 | 6.51 | 6.67 | 6.81  | 6.94  | 7.06  |
| 13        | 0.05     | 3.06                     | 3.73 | 4.15 | 4.45 | 4.69 | 4.88 | 5.05 | 5.19 | 5.32  | 5.43  | 5.53  |
|           | 0.01     | 4.26                     | 4.96 | 5.40 | 5.73 | 5.98 | 6.19 | 6.37 | 6.53 | 6.67  | 6.79  | 6.90  |
| 14        | 0.05     | 3.03                     | 3.70 | 4.11 | 4.41 | 4.64 | 4.83 | 4.99 | 5.13 | 5.25  | 5.36  | 5.46  |
|           | 0.01     | 4.21                     | 4.89 | 5.32 | 5.63 | 5.88 | 6.08 | 6.26 | 6.41 | 6.54  | 6.66  | 6.77  |
| 15        | 0.05     | 3.01                     | 3.67 | 4.08 | 4.37 | 4.59 | 4.78 | 4.94 | 5.08 | 5.20  | 5.31  | 5.40  |
|           | 0.01     | 4.17                     | 4.84 | 5.25 | 5.56 | 5.80 | 5.99 | 6.16 | 6.31 | 6.44  | 6.55  | 6.66  |
| 16        | 0.05     | 3.00                     | 3.65 | 4.05 | 4.33 | 4.56 | 4.74 | 4.90 | 5.03 | 5.15  | 5.26  | 5.35  |
|           | 0.01     | 4.13                     | 4.79 | 5.19 | 5.49 | 5.72 | 5.92 | 6.08 | 6.22 | 6.35  | 6.46  | 6.56  |
| 17        | 0.05     | 2.98                     | 3.63 | 4.02 | 4.30 | 4.52 | 4.70 | 4.86 | 4.99 | 5.11  | 5.21  | 5.31  |
|           | 0.01     | 4.10                     | 4.74 | 5.14 | 5.43 | 5.66 | 5.85 | 6.01 | 6.15 | 6.27  | 6.38  | 6.48  |
| 18        | 0.05     | 2.97                     | 3.61 | 4.00 | 4.28 | 4.49 | 4.67 | 4.82 | 4.96 | 5.07  | 5.17  | 5.27  |
|           | 0.01     | 4.07                     | 4.70 | 5.09 | 5.38 | 5.60 | 5.79 | 5.94 | 6.08 | 6.20  | 6.31  | 6.41  |
| 19        | 0.05     | 2.96                     | 3.59 | 3.98 | 4.25 | 4.47 | 4.65 | 4.79 | 4.92 | 5.04  | 5.14  | 5.23  |
|           | 0.01     | 4.05                     | 4.67 | 5.05 | 5.33 | 5.55 | 5.73 | 5.89 | 6.02 | 6.14  | 6.25  | 6.34  |
| 20        | 0.05     | 2.95                     | 3.58 | 3.96 | 4.23 | 4.45 | 4.62 | 4.77 | 4.90 | 5.01  | 5.11  | 5.20  |
|           | 0.01     | 4.02                     | 4.64 | 5.02 | 5.29 | 5.51 | 5.69 | 5.84 | 5.97 | 6.09  | 6.19  | 6.28  |

**TABLE 5** (Continued)

| $n_T - c$ | $\alpha$ | The number of means, $c$ |      |      |      |      |      |      |      |      |      |      |
|-----------|----------|--------------------------|------|------|------|------|------|------|------|------|------|------|
|           |          | 2                        | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | 11   | 12   |
| 24        | 0.05     | 2.92                     | 3.53 | 3.90 | 4.17 | 4.37 | 4.54 | 4.68 | 4.81 | 4.92 | 5.01 | 5.10 |
|           | 0.01     | 3.96                     | 4.55 | 4.91 | 5.17 | 5.37 | 5.54 | 5.69 | 5.81 | 5.92 | 6.02 | 6.11 |
| 30        | 0.05     | 2.89                     | 3.49 | 3.85 | 4.10 | 4.30 | 4.46 | 4.60 | 4.72 | 4.82 | 4.92 | 5.00 |
|           | 0.01     | 3.89                     | 4.45 | 4.80 | 5.05 | 5.24 | 5.40 | 5.54 | 5.65 | 5.76 | 5.85 | 5.93 |
| 40        | 0.05     | 2.86                     | 3.44 | 3.79 | 4.04 | 4.23 | 4.39 | 4.52 | 4.63 | 4.73 | 4.82 | 4.90 |
|           | 0.01     | 3.82                     | 4.37 | 4.70 | 4.93 | 5.11 | 5.26 | 5.39 | 5.50 | 5.60 | 5.69 | 5.76 |
| 60        | 0.05     | 2.83                     | 3.40 | 3.74 | 3.98 | 4.16 | 4.31 | 4.44 | 4.55 | 4.65 | 4.73 | 4.81 |
|           | 0.01     | 3.76                     | 4.28 | 4.59 | 4.82 | 4.99 | 5.13 | 5.25 | 5.36 | 5.45 | 5.53 | 5.60 |
| 120       | 0.05     | 2.80                     | 3.36 | 3.68 | 3.92 | 4.10 | 4.24 | 4.36 | 4.47 | 4.56 | 4.64 | 4.71 |
|           | 0.01     | 3.70                     | 4.20 | 4.50 | 4.71 | 4.87 | 5.01 | 5.12 | 5.21 | 5.30 | 5.37 | 5.44 |
| $\infty$  | 0.05     | 2.77                     | 3.31 | 3.63 | 3.86 | 4.03 | 4.17 | 4.29 | 4.39 | 4.47 | 4.55 | 4.62 |
|           | 0.01     | 3.64                     | 4.12 | 4.40 | 4.60 | 4.76 | 4.88 | 4.99 | 5.08 | 5.16 | 5.23 | 5.29 |

SOURCE: E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, vol. 1 (Cambridge: Cambridge University Press, 1966).

**TABLE 6** Lower ( $T_L$ ) and Upper ( $T_U$ ) Critical Values for the Wilcoxon Signed-Rank Test

| Two-Tail Test: | $\alpha = 0.10$ | $\alpha = 0.05$  | $\alpha = 0.02$ | $\alpha = 0.01$  |
|----------------|-----------------|------------------|-----------------|------------------|
| One-Tail Test: | $\alpha = 0.05$ | $\alpha = 0.025$ | $\alpha = 0.01$ | $\alpha = 0.005$ |
| $n = 5$        | 0, 15           | —, —             | —, —            | —, —             |
| 6              | 2, 19           | 0, 21            | —, —            | —, —             |
| 7              | 3, 25           | 2, 26            | 0, 28           | —, —             |
| 8              | 5, 31           | 3, 33            | 1, 35           | 0, 36            |
| 9              | 8, 37           | 5, 40            | 3, 42           | 1, 44            |
| 10             | 10, 45          | 8, 47            | 5, 50           | 3, 52            |

SOURCE: Adapted from "Extended Tables of the Wilcoxon Matched Pairs Signed Rank Statistics," *Journal of the American Statistical Association* 60 (1965), 864–71.

**TABLE 7** Lower ( $W_L$ ) and Upper ( $W_U$ ) Critical Values for the Wilcoxon Rank-Sum Test

| $\alpha = 0.025$ for one-tailed test and $\alpha = 0.05$ for a two-tailed test |           |        |        |        |        |        |         |         |
|--|-----------|--------|--------|--------|--------|--------|---------|---------|
|  | $n_1$ : 3 | 4      | 5      | 6      | 7      | 8      | 9       | 10      |
| $n_2$ : 3  | 5, 16     | 6, 18  | 6, 21  | 7, 23  | 7, 26  | 8, 28  | 8, 31   | 9, 33   |
| 4  | 6, 18     | 11, 25 | 12, 28 | 12, 32 | 13, 35 | 14, 38 | 15, 41  | 16, 44  |
| 5  | 6, 21     | 12, 28 | 18, 37 | 19, 41 | 20, 45 | 21, 49 | 22, 53  | 24, 56  |
| 6  | 7, 23     | 12, 32 | 19, 41 | 26, 52 | 28, 56 | 29, 61 | 31, 65  | 32, 70  |
| 7  | 7, 26     | 13, 35 | 20, 45 | 28, 56 | 37, 68 | 39, 73 | 41, 78  | 43, 83  |
| 8  | 8, 28     | 14, 38 | 21, 49 | 29, 61 | 39, 73 | 49, 87 | 51, 93  | 54, 98  |
| 9  | 8, 31     | 15, 41 | 22, 53 | 31, 65 | 41, 78 | 51, 93 | 63, 108 | 66, 114 |
| 10   | 9, 33     | 16, 44 | 24, 56 | 32, 70 | 43, 83 | 54, 98 | 66, 114 | 79, 131 |
| $\alpha = 0.05$ for one-tailed test and $\alpha = 0.10$ for a two-tailed test  |           |        |        |        |        |        |         |         |
|  | $n_1$ : 3 | 4      | 5      | 6      | 7      | 8      | 9       | 10      |
| $n_2$ : 3  | 6, 15     | 7, 17  | 7, 20  | 8, 22  | 9, 24  | 9, 27  | 10, 29  | 11, 31  |
| 4  | 7, 17     | 12, 24 | 13, 27 | 14, 30 | 15, 33 | 16, 36 | 17, 39  | 18, 42  |
| 5  | 7, 20     | 13, 27 | 19, 36 | 20, 40 | 22, 43 | 24, 46 | 25, 50  | 26, 54  |
| 6  | 8, 22     | 14, 30 | 20, 40 | 28, 50 | 30, 54 | 32, 58 | 33, 63  | 35, 67  |
| 7  | 9, 24     | 15, 33 | 22, 43 | 30, 54 | 39, 66 | 41, 71 | 43, 76  | 46, 80  |
| 8  | 9, 27     | 16, 36 | 24, 46 | 32, 58 | 41, 71 | 52, 84 | 54, 90  | 57, 95  |
| 9  | 10, 29    | 17, 39 | 25, 50 | 33, 63 | 43, 76 | 54, 90 | 66, 105 | 69, 111 |
| 10   | 11, 31    | 18, 42 | 26, 54 | 35, 67 | 46, 80 | 57, 95 | 69, 111 | 83, 127 |

SOURCE: F. Wilcoxon and R. A. Wilcox, *Some Rapid Approximate Statistical Procedures* (New York: American Cyanamid Company, 1964).**TABLE 8** Upper Critical Values for the Spearman Rank-Correlation Coefficient

| Two-Tail Test: | $\alpha = 0.10$ | $\alpha = 0.05$  | $\alpha = 0.02$ | $\alpha = 0.01$  |
|----------------|-----------------|------------------|-----------------|------------------|
| One-Tail Test: | $\alpha = 0.05$ | $\alpha = 0.025$ | $\alpha = 0.01$ | $\alpha = 0.005$ |
| $n = 5$        | 0.900           | —                | —               | —                |
| 6              | 0.829           | 0.886            | 0.943           | —                |
| 7              | 0.714           | 0.786            | 0.893           | —                |
| 8              | 0.643           | 0.738            | 0.833           | 0.881            |
| 9              | 0.600           | 0.683            | 0.783           | 0.833            |
| 10             | 0.564           | 0.648            | 0.745           | 0.794            |

SOURCE: E. G. Olds, "Distribution of Sums of Squares of Rank Differences for Small Samples," *Annals of Mathematical Statistics* 9 (1938).

## Answers to Selected Exercises

### Chapter 1

- 1.2 35 is likely the estimated average age. It would be rather impossible to reach all video game players.
- 1.4 a. The population is all marketing managers.  
b. No, the average salary was likely computed from a sample in order to save time and money.
- 1.6 Answers will vary depending on when data are retrieved. The numbers represent time series data.
- 1.8 Answers will vary depending on when data are retrieved. The numbers represent cross-sectional data.
- 1.10 Answers will vary depending on when data are retrieved. The numbers represent cross-sectional data.
- 1.12 a. Qualitative  
b. Quantitative, continuous  
c. Quantitative, discrete
- 1.14 a. Ratio  
b. Ordinal  
c. Nominal
- 1.16 a. Nominal

| Major      | Number of Students |
|------------|--------------------|
| Accounting | 5                  |
| Economics  | 7                  |
| Finance    | 5                  |
| Marketing  | 3                  |
| Management | 6                  |
| Undecided  | 4                  |

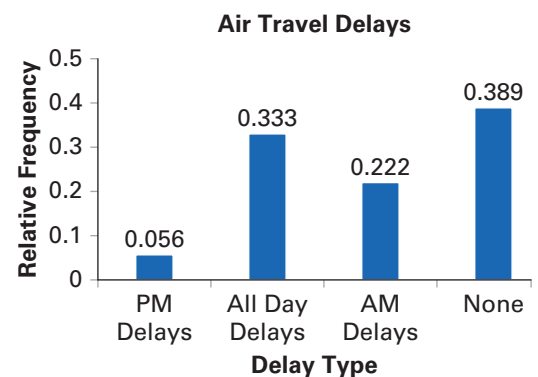
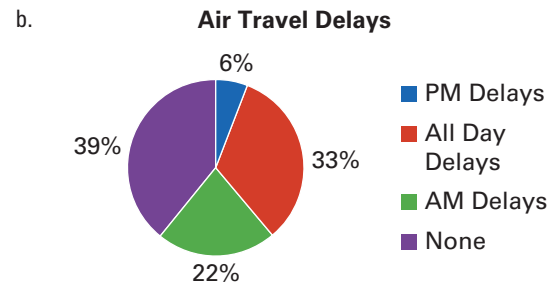
- c. Economics (Marketing) has the highest (lowest) number of students.

### Chapter 2

- 2.2 a.
- | Rating    | Frequency | Relative Frequency |
|-----------|-----------|--------------------|
| Excellent | 5         | 0.208              |
| Good      | 12        | 0.500              |
| Fair      | 4         | 0.167              |
| Poor      | 3         | 0.125              |
- b. The most common response is Good. Over 70 percent of the patients reveal that they are in either good or excellent health, suggesting overall health of first-time patients is strong.

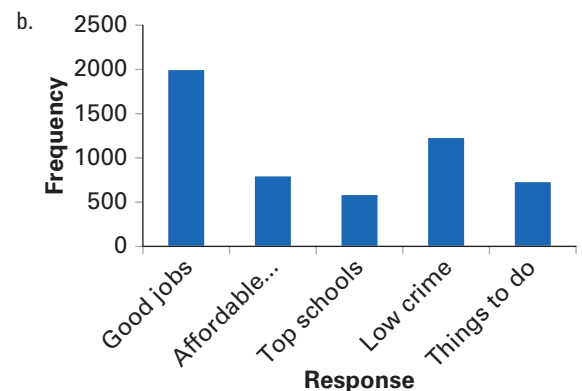
2.4 a.

| Delays         | Frequency | Relative Frequency |
|----------------|-----------|--------------------|
| PM Delays      | 1         | 0.056              |
| All Day Delays | 6         | 0.333              |
| AM Delays      | 4         | 0.222              |
| None           | 7         | 0.389              |



2.6 a.

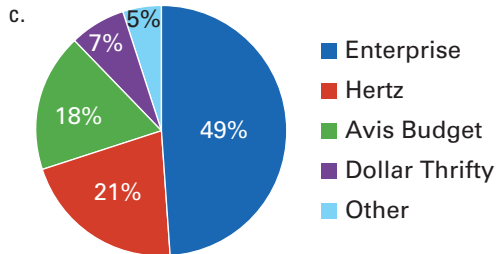
| Response         | Frequency |
|------------------|-----------|
| Good jobs        | 1970      |
| Affordable homes | 799       |
| Top schools      | 586       |
| Low crime        | 1225      |
| Things to do     | 745       |



2.8 a.

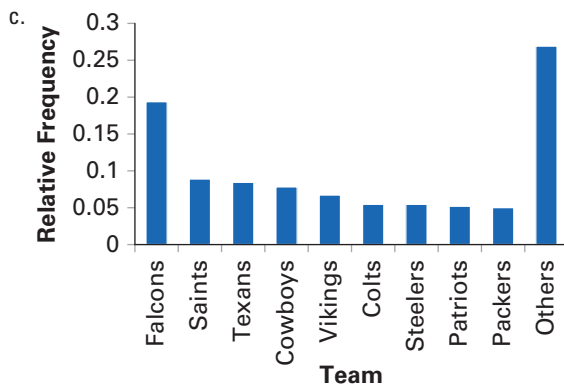
| Company        | Relative Frequency |
|----------------|--------------------|
| Enterprise     | 0.489              |
| Hertz          | 0.215              |
| Avis Budget    | 0.183              |
| Dollar Thrifty | 0.068              |
| Other          | 0.046              |

b. Hertz accounted for 21.5% of sales.



2.10 a. 5584

b. 0.052



2.14 This graph does not correctly depict what has happened to sales over the most recent five-year period. The vertical axis has been stretched so that the increase in sales appears more pronounced than warranted.

2.16 a.

| Classes     | Frequency  |
|-------------|------------|
| -10 up to 0 | 9          |
| 0 up to 10  | 31         |
| 10 up to 20 | 19         |
| 20 up to 30 | 8          |
| 30 up to 40 | 3          |
|             | Total = 70 |

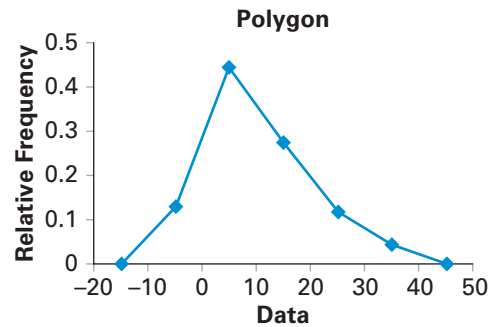
19 observations are at least 10 but less than 20.

b.

| Classes     | Relative Frequency    | Cumulative Relative Frequency                   |
|-------------|-----------------------|---|
| -10 up to 0 | $9/70 = 0.129$        | 0.129   |
| 0 up to 10  | $31/70 = 0.443$       | $0.129 + 0.443 = 0.572$                         |
| 10 up to 20 | $19/70 = 0.271$       | $0.129 + 0.443 + 0.271 = 0.843$                 |
| 20 up to 30 | $8/70 = 0.114$        | $0.129 + 0.443 + 0.271 + 0.114 = 0.957$         |
| 30 up to 40 | $3/70 = 0.043$        | $0.129 + 0.443 + 0.271 + 0.114 + 0.043 = 1.000$ |
|             | Total $\approx 1.000$ |   |

27.1% of the observations are at least 10 but less than 20; 84.3% are less than 20.

c.



The distribution is not symmetric. It is positively skewed.

2.18 a.

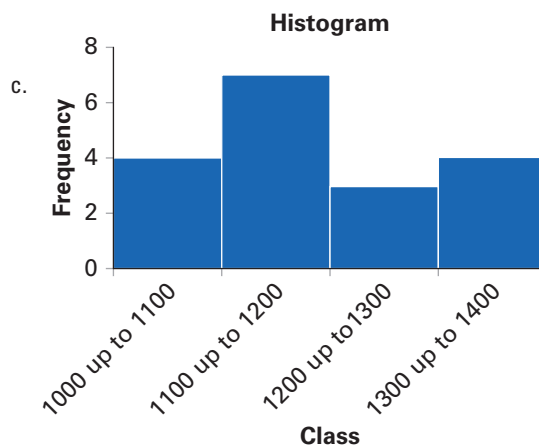
| Class             | Relative Frequency |
|-------------------|--------------------|
| 1,000 up to 1,100 | $2/16 = 0.1250$    |
| 1,100 up to 1,200 | $7/16 = 0.4375$    |
| 1,200 up to 1,300 | $3/16 = 0.1875$    |
| 1,300 up to 1,400 | $4/16 = 0.2500$    |
|                   | Total = 1.0000     |

43.75% of the observations are at least 1,100 but less than 1,200.

b.

| Class             | Cumulative Frequency | Cumulative Relative Frequency |
|-------------------|----------------------|-------------------------------|
| 1,000 up to 1,100 | 2                    | $2/16 = 0.125$                |
| 1,100 up to 1,200 | $2 + 7 = 9$          | $9/16 = 0.562$                |
| 1,200 up to 1,300 | $2 + 7 + 3 = 12$     | $12/16 = 0.750$               |
| 1,300 up to 1,400 | $2 + 7 + 3 + 4 = 16$ | $16/16 = 1.000$               |

12 observations are less than 1300.



c.

2.20 a.

| Class         | Frequency             |
|---------------|-----------------------|
| -20 up to -10 | $0.04 \times 50 = 2$  |
| -10 up to 0   | $0.28 \times 50 = 14$ |
| 0 up to 10    | $0.26 \times 50 = 13$ |
| 10 up to 20   | $0.22 \times 50 = 11$ |
| 20 up to 30   | $0.20 \times 50 = 10$ |
|               | Total = 50            |

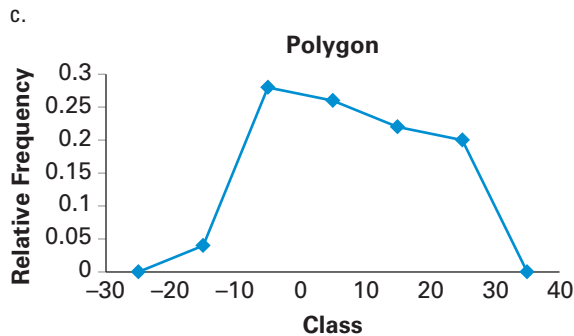
14 observations are at least -10 but less than 0.



b.

| Class         | Cumulative Frequency |
|---------------|----------------------|
| -20 up to -10 | 2                    |
| -10 up to 0   | 2 + 14 = 16          |
| 0 up to 10    | 16 + 13 = 29         |
| 10 up to 20   | 29 + 11 = 40         |
| 20 up to 30   | 40 + 10 = 50         |

40 observations are less than 20.



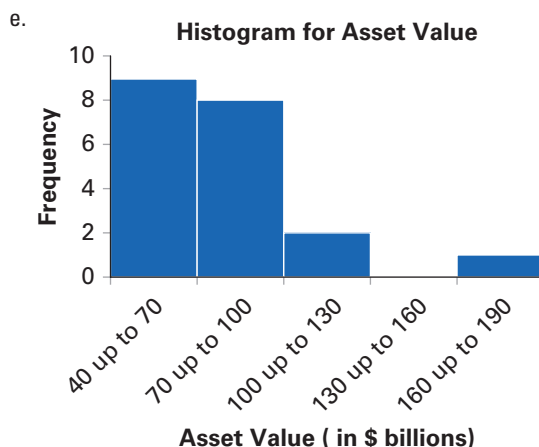
2.22 a.

| Assets (in billions) | Frequency  |
|----------------------|------------|
| 40 up to 70          | 9          |
| 70 up to 100         | 8          |
| 100 up to 130        | 2          |
| 130 up to 160        | 0          |
| 160 up to 190        | 1          |
|                      | Total = 20 |

b.

| Assets (in billions) | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
|----------------------|--------------------|----------------------|-------------------------------|
| 40 up to 70          | 9/20 = 0.45        | 9                    | 9/20 = 0.45                   |
| 70 up to 100         | 8/20 = 0.40        | 9 + 8 = 17           | 17/20 = 0.85                  |
| 100 up to 130        | 2/20 = 0.10        | 17 + 2 = 19          | 19/20 = 0.95                  |
| 130 up to 160        | 0/20 = 0           | 19 + 0 = 19          | 19/20 = 0.95                  |
| 160 up to 190        | 1/20 = 0.05        | 19 + 1 = 20          | 20/20 = 1                     |

- c. Two funds had assets of at least 100 but less than 130 (in \$ billions); 19 funds had assets less than \$160 billion.
- d. 40% of the funds had assets of at least \$70 but less than \$100 (in billions); 95% of the funds had assets less than \$130 billion.



The distribution is positively skewed.

Note: The histogram could have also been made with relative frequencies. It would have had the same positive skewness.

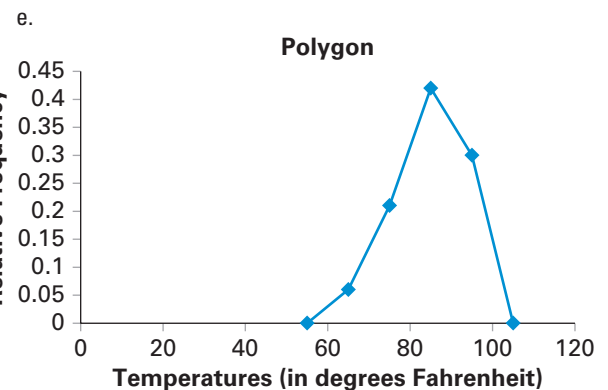
2.24 a.

| Temperature  | Frequency  |
|--------------|------------|
| 60 up to 70  | 2          |
| 70 up to 80  | 7          |
| 80 up to 90  | 14         |
| 90 up to 100 | 10         |
|              | Total = 33 |

b.

| Temperature  | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
|--------------|--------------------|----------------------|-------------------------------|
| 60 up to 70  | 2/33 = 0.061       | 2                    | 2/33 = 0.061                  |
| 70 up to 80  | 7/33 = 0.212       | 2 + 7 = 9            | 9/33 = 0.273                  |
| 80 up to 90  | 14/33 = 0.424      | 9 + 14 = 23          | 23/33 = 0.697                 |
| 90 up to 100 | 10/33 = 0.303      | 23 + 10 = 33         | 33/33 = 1.000                 |
|              | Total = 1.000      |                      |                               |

- c. 9 cities had temperatures less than 80°.
- d. 42.4% of the cities recorded temperatures of at least 80° but less than 90°; 69.7% of the cities had temperatures less than 90°.

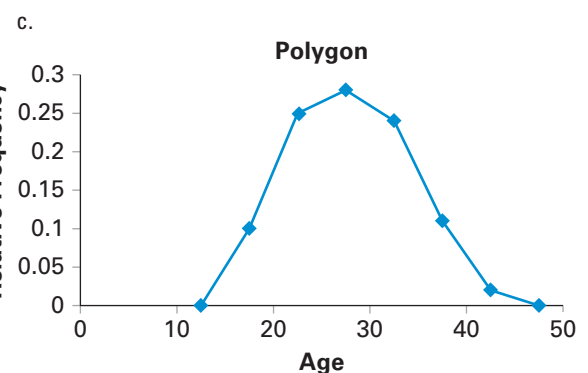


The distribution is slightly negatively skewed.

2.26 a.

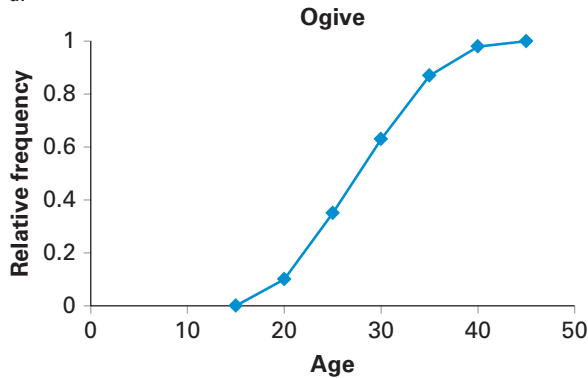
| Age         | Frequency        | Cumulative Frequency | Cumulative Relative Frequency |
|-------------|------------------|----------------------|-------------------------------|
| 15 up to 20 | 0.10(2000) = 200 | 200                  | 0.1                           |
| 20 up to 25 | 0.25(2000) = 500 | 200 + 500 = 700      | 0.10 + 0.25 = 0.35            |
| 25 up to 30 | 0.28(2000) = 560 | 700 + 560 = 1,260    | 0.35 + 0.28 = 0.63            |
| 30 up to 35 | 0.24(2000) = 480 | 1,260 + 480 = 1,740  | 0.63 + 0.24 = 0.87            |
| 35 up to 40 | 0.11(2000) = 220 | 1,740 + 220 = 1,960  | 0.87 + 0.11 = 0.98            |
| 40 up to 45 | 0.02(2000) = 40  | 1,960 + 40 = 2,000   | 0.98 + 0.02 = 1.00            |
|             | Total = 2000     |                      |                               |

- b. 28% of the women were at least 25 but less than 30 years old; 87% were less than 35 years old.



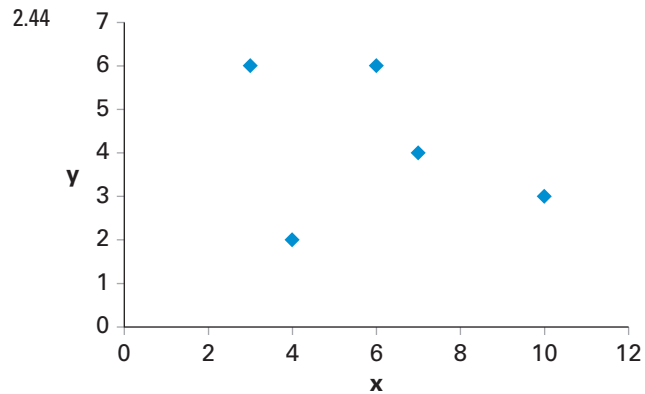
The distribution appears to be relatively symmetric with possibly a slight positive skew.

d.



If we draw a horizontal line that corresponds to the 0.5 value on the vertical axis, it will intersect the ogive at the age of approximately 28 years old.

Spain has a relatively younger team compared to Netherlands. Spain's ages range from 21 to 32, while Netherlands' ages range from 22 to 39. The majority of players in both teams are in their 20s. However, Netherlands has a couple of more players in their 30s than Spain.



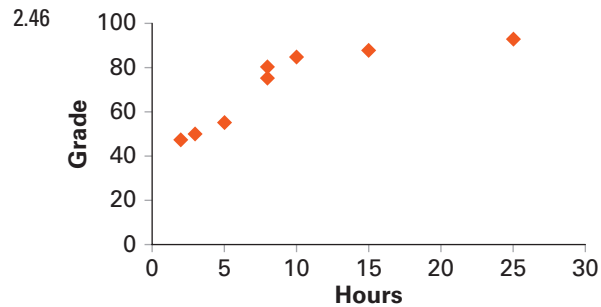
There is no evident relationship between  $x$  and  $y$ .

- 2.28 a. No. The distribution is not symmetric. It is positively skewed.  
 b. Forty-four percent of the states had median household income between \$45,000 and \$55,000.  
 c. Sixty-six percent of the states had median household income between \$35,000 and \$55,000.

2.38

| Stem | Leaf            |
|------|-----------------|
| -8   | 7 5 5 3 2 0 0 0 |
| -7   | 9 7 5 3 3 2 1   |
| -6   | 5 5 4           |
| -5   | 2 0             |

(Keep in mind that these values are negative.) The distribution is not symmetric; it is positively skewed. Most of the numbers are in the lower stems of -8 and -7.

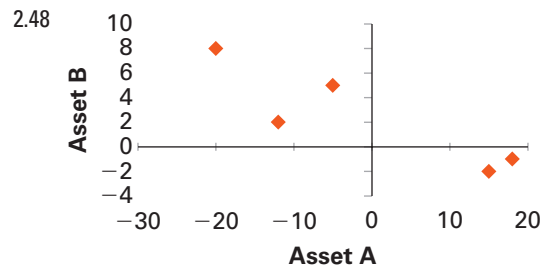


There is a positive relationship between number of hours spent studying and grades. As the number of hours spent studying increases, grades tend to increase.

2.40

| Stem | Leaf  |
|------|---|
| 7    | 3 4 6 7 8 8                                 |
| 8    | 0 1 2 3 4 4 4 4 7 8                         |
| 9    | 0 0 0 1 1 2 2 2 3 3 4 4 4 4 4 5 6 6 6 8 8 9 |
| 10   | 6 7   |

Temperatures ranged from a low of 73 to a high of 107. The distribution is not symmetric; it has negative skew. Temperatures in 90s were the most frequent.



There is a slightly negative relationship between the two assets. Therefore, it would be wise for the investor to include them in her portfolio.

2.42 Spain

| Stem | Leaf                              |
|------|-----------------------------------|
| 2    | 1 1 1 2 3 3 4 4 5 5 5 6 7 8 9 9 9 |
| 3    | 0 0 2                             |

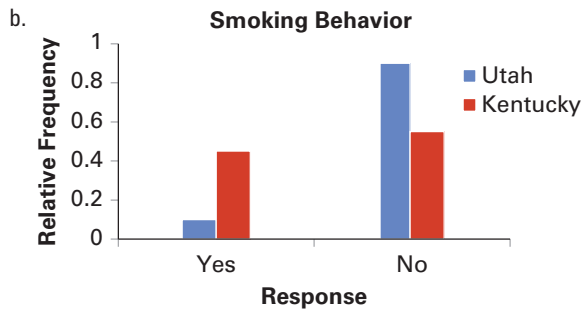
Netherlands

| Stem | Leaf                          |
|------|-------------------------------|
| 2    | 2 3 3 4 5 5 5 6 6 6 7 7 7 7 9 |
| 3    | 0 3 5 5 9                     |

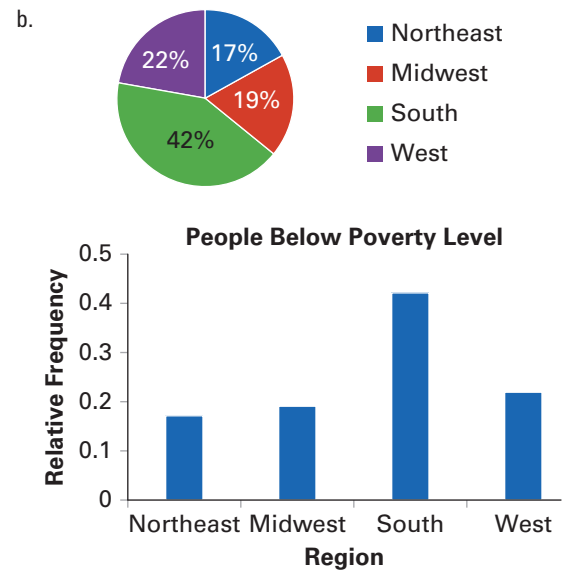
2.50 a.

| Responses | Utah Relative Frequency | Kentucky Relative Frequency |
|-----------|-------------------------|-----------------------------|
| Yes       | $2/20 = 0.10$           | $9/20 = 0.45$               |
| No        | $18/20 = 0.90$          | $11/20 = 0.55$              |
|           | Total = 1.00            | Total = 1.00                |

The sample responses show the difference regarding smoking behavior in the two states. Notice that 45% of the households in Kentucky allow smoking at home whereas only 10% do so in Utah.

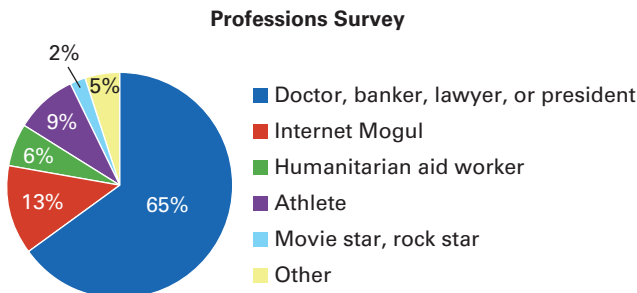
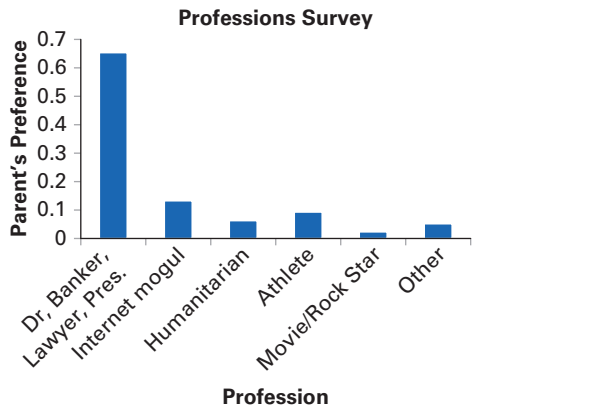


The bar chart shows that smoking at home is much more common in Kentucky than in Utah.



These charts show that the highest percentage of people who live below the poverty level live in the South, and the lowest percentage live in the Northeast.

2.52 a.



The charts reveal parent preferences. Sixty-five percent of parents want their children to have a profession such as a doctor, lawyer, banker or president. Less preferable are other professions such humanitarian-aid worker or a movie star.

- b. Since 9% of parents want their children to become an athlete, we find  $550 \times 0.09 \approx 50$ . Therefore, among 550 parents approximately 50 parents want their kids to become an athlete.

2.54 a.

| Region        | Relative Frequency      |
|---------------|-------------------------|
| Northeast     | $6,166/37,276 = 0.165$  |
| Midwest       | $7,237/37,276 = 0.194$  |
| South         | $15,501/37,276 = 0.416$ |
| West          | $8,372/37,276 = 0.225$  |
| Total = 1.000 |                         |

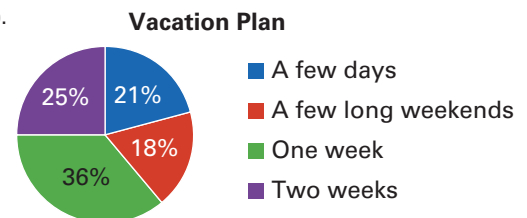
19.4% of people living below the poverty level live in the Midwest region.

2.56 a.

| Response            | Frequency           |
|---------------------|---------------------|
| A few days          | $0.21(3057) = 642$  |
| A few long weekends | $0.18(3057) = 550$  |
| One week            | $0.36(3057) = 1101$ |
| Two weeks           | $0.25(3057) = 764$  |
| Total = 3057        |                     |

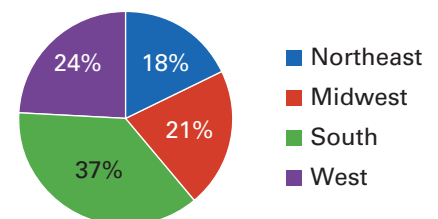
Approximately 1101 people are going to take a one week vacation.

b.



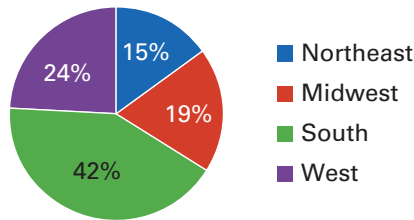
2.58 a. The pie chart is below.

#### Percentage of People in Each Region



The chart shows the highest percentage of people live in the South and the lowest percentage live in the Northeast.

b. **Below Poverty Level**



The chart shows the highest percentage of people living in poverty are in the South and the lowest percentage of people living in poverty are living in the Northeast. The percentage of people living in poverty in the South is higher than the percentage of people that live in South, and the percentage of people living in poverty in the Northeast is less than the percentage of people that live in the Northeast.

c.

| Stem | Leaf              |
|------|-------------------|
| 3    | 6 6               |
| 4    | 4 7               |
| 5    | 3 3 4 6           |
| 6    | 0 1 5 5 6 7 7 9   |
| 7    | 0 1 3 3 3 7 8 9 9 |

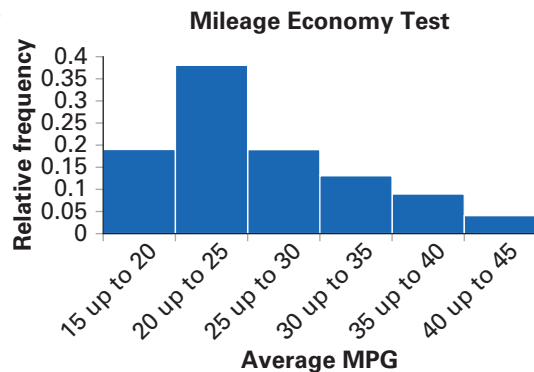
The distribution is not symmetric; it is negatively skewed. The majority of ages range from the 60s to 70s. Table 2.16 shows the majority of ages to be in the 50s and 60s. Further, this diagram shows ages ranging from 36 to 79, whereas Table 2.16 has ages ranging from 36 to 90.

2.60 a.

| Average MPG | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
|-------------|--------------------|----------------------|-------------------------------|
| 15 up to 20 | $15/80 = 0.1875$   | 15                   | 0.1875                        |
| 20 up to 25 | $30/80 = 0.3750$   | $15 + 30 = 45$       | $45/80 = 0.5625$              |
| 25 up to 30 | $15/80 = 0.1875$   | $45 + 15 = 60$       | $60/80 = 0.7500$              |
| 30 up to 35 | $10/80 = 0.1250$   | $60 + 10 = 70$       | $70/80 = 0.8750$              |
| 35 up to 40 | $7/80 = 0.0875$    | $70 + 7 = 77$        | $77/80 = 0.9625$              |
| 40 up to 45 | $3/80 = 0.0375$    | $77 + 3 = 80$        | $80/80 = 1.0000$              |
|             | Total = 1.0000     |                      |                               |

- b. 60 cars got less than 30 mpg; 37.5% of the cars got at least 20 but less than 25 mpg; 87.5% of the cars got less than 35 mpg; Since 87.5% got less than 35 mpg, 12.5% of the cars got 35 mpg or more.

c.

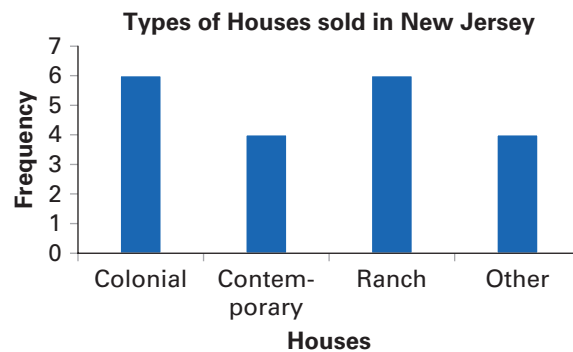
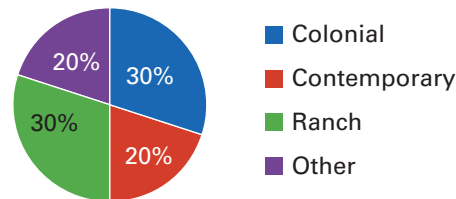


The distribution is not symmetric; it is positively skewed.

- 2.62 a. There were 4 people out of 25 with a net worth greater than \$20 billion. Since  $4/25 = 0.16$ , 16% of the wealthiest people had net worth greater than \$20 billion.
- b. Two people had a net worth less than \$10 billion, which is  $2/25 = 0.08$ , or 8%. From the previous question, we know that 16% had a net worth greater than \$20 billion. Therefore,  $16\% + 8\% = 24\%$  *did not* have a net worth between \$10 and \$20 billion. Consequently, 76% had net worth between \$10 billion and \$20 billion.

2.64 a.

**Types of Houses sold in New Jersey**

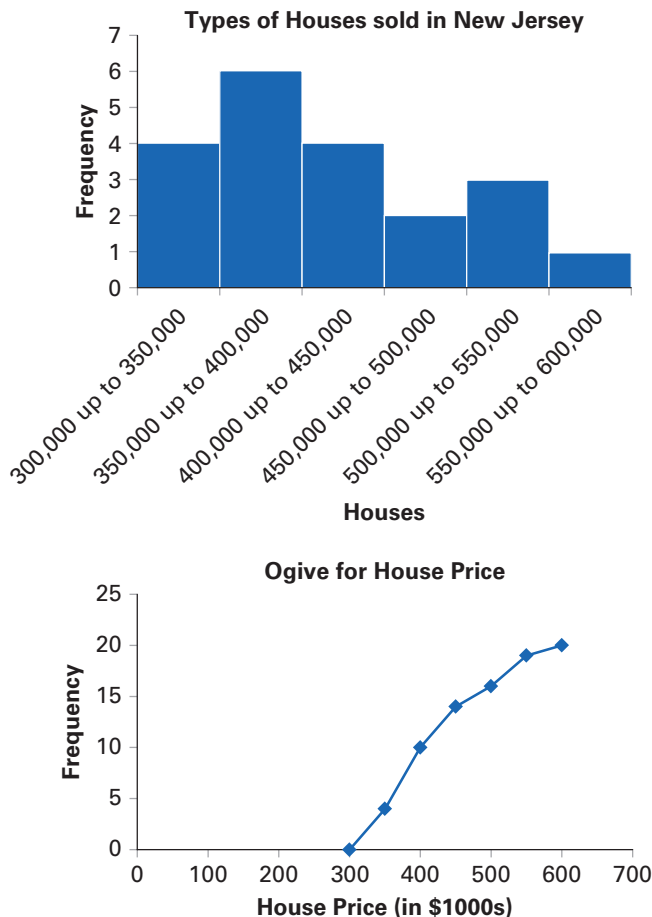


These charts show that the majority (60%) of houses were either Ranch or Colonial, but also 40% were either Contemporary or some other type.

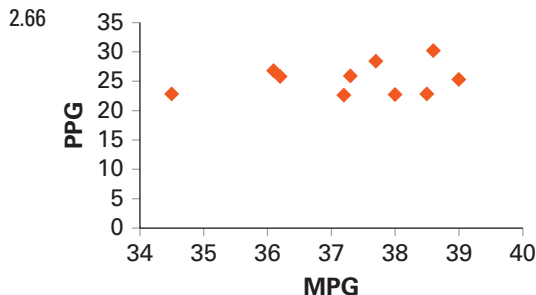
- b. To figure out how wide to make the classes, find the highest price and subtract the lowest price to get the range. That is  $\$568,000 - \$300,000 = \$268,000$ . Then, since we want 7 classes, divide the range by 7;  $268,000/7 = \$38,386$ . However, for ease of interpretation, round to the most sensible number: \$50,000. Therefore, our classes will have a width of \$50,000, with a lower bound of the first class of \$300,000.

| Classes               | Frequency  |
|-----------------------|------------|
| 300,000 up to 350,000 | 4          |
| 350,000 up to 400,000 | 6          |
| 400,000 up to 450,000 | 4          |
| 450,000 up to 500,000 | 2          |
| 500,000 up to 550,000 | 3          |
| 550,000 up to 600,000 | 1          |
|                       | Total = 20 |

c.



The histogram shows that the most frequent house price is in the \$350,000 up to \$400,000 range. The ogive shows that the middle price (with a frequency of 10/20 or 50%) is about \$400,000.



The scatterplot reveals no clear relationship between PPG and MPG.

### Chapter 3

3.2 Mean = -2.67; Median = -3.5; Mode = -4

3.4 Mean = 18.33; Median = 20; Mode = 15, 20

3.6 a. Sample mean:  $\bar{x} = \frac{3+4+3+3+5+2+4+2+5+6}{10} = 3.7$

The median is the average of the values at the 5<sup>th</sup> and 6<sup>th</sup> positions;

$$\text{Median} = \frac{3+4}{2} = 3.5.$$

The mode is 3.

Since we are describing the number of bedrooms in a home, the best measure of central location in this case is the mode. (A home cannot have 3.7 bedrooms or 3.5 bedrooms.)

3.10 a. Average price per share

$$= \frac{10.34 \times 100 + 13.98 \times 60 + 14.02 \times 100}{40 + 60 + 100} = 12.17$$

b. Average price per share

$$= \frac{10.34 \times 40 + 13.98 \times 60 + 14.02 \times 100}{40 + 60 + 100} = 13.27$$

3.12 a. For market capitalization: Mean = 164.10; Median = 167.50.

b. For total return: Mean = 40.71%; Median = 6.05%.

c. The mean and the median of the market capitalization are close in value. Since the values are close, the distribution is nearly symmetric. The mean and the median of the total return differ significantly. In this case, the mean is affected by outliers (extreme values), so the median is a better measure of central location.

3.16  $L_{20} = (7 + 1) \frac{20}{100} = 1.6$ . Thus the 20<sup>th</sup> percentile is located 60% of the distance between the first observation and the second observation, and it is  $120 + 0.6(187 - 120) = 160.20$ .

$L_{50} = (7 + 1) \frac{50}{100} = 5$ . Thus the 50<sup>th</sup> percentile (also the median) is located at 5<sup>th</sup> position, and it is 215.

$L_{80} = (7 + 1) \frac{80}{100} = 6.4$ . Thus the 80<sup>th</sup> percentile is located 40% of the distance between the 6<sup>th</sup> observation and the 7<sup>th</sup> observation, and it is  $312 + 0.4(343 - 312) = 324.4$ .

3.20 a.  $L_{25} = (7 + 1) \frac{25}{100} = 2$ . The 25<sup>th</sup> percentile is -0.05.

Approximately 25 percent of the observations have values less than -0.05.

$L_{50} = (7 + 1) \frac{50}{100} = 4$ . The 50<sup>th</sup> percentile is 0.04.

Approximately 50 percent of the observations have values less than 0.04.

$L_{75} = (7 + 1) \frac{75}{100} = 6$ . The 75<sup>th</sup> percentile is 0.10.

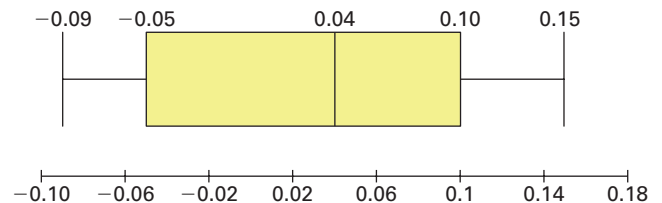
Approximately 75 percent of the observations have values less than 0.10.

b.  $IQR = Q_3 - Q_1 = 0.10 - (-0.05) = 0.15$ .

Lower limit =  $Q_1 - 1.5 \times IQR = -0.05 - 0.225 = -0.275$ .

Upper limit =  $Q_3 + 1.5 \times IQR = 0.10 + 0.225 = 0.325$ .

Therefore, there are no outliers.



3.26  $G_8 = \sqrt[8]{(1 + R_1) \times (1 + R_2) \times \dots \times (1 + R_8)} - 1$   
 $= \sqrt[8]{(1 + 0.04)(1 + 0.08)(1 - 0.05)(1 + 0.06)} - 1 = 0.0313$ ,  
 or 3.13%.

3.28  $G_g = \sqrt[3]{(1 + 0.10)(1 + 0.05)(1 - 0.15)} - 1 = -0.006$ , that is -0.6%.

3.30  $G_g = \sqrt[1.25]{(1 + 0.05)(1 + 0.03)} - 1 = 1.0647 - 1 = 0.0647$ , that is, 6.47%.

3.32 a.

|              | Year 1 – Year 2 | Year 2 – Year 3 | Year 3 – Year 4 |
|--------------|-----------------|-----------------|-----------------|
| Growth Rates | 0.0667          | 0.0781          | 0.1014          |

b.  $G_g = \sqrt[3]{(1 + 0.0667)(1 + 0.0781)(1 + 0.1014)} - 1 = 0.082$ , that is, 8.2%.

3.34 a. The arithmetic mean is given by

$$\bar{x} = \frac{33.42 + 13.82 - 44.73 + 31.91}{4} = 8.605$$

The average annual return between 2006 and 2009 is 8.61%.

- b. The geometric mean is given by

$$G_g = \sqrt[4]{(1 + 0.3342)(1 + 0.1382)(1 - 0.4473)(1 + 0.3191)} - 1 = 0.0258$$

The geometric mean return between 2006 and 2009 is 2.58%.

- c. By the end of 2009, the total accumulation would be  $(1,000)(1 + 0.0258)^4 = \$1,107.26$ .

- 3.36 a. The growth rates for each retailer are below.

|           | Home Depot | Lowe's  |
|-----------|------------|---------|
| 2008–2009 | −0.0784    | −0.001  |
| 2009–2010 | −0.0717    | −0.0209 |

- b. The average growth rates for each retailer are  $G_{HD} = -0.075$ ;  $G_{Lowe's} = -0.011$ .

- 3.38 a.  $G_g = \sqrt[4]{(1 + 0.0884)(1 + 0.0917)(1 + 0.1409)(1 + 0.0295)} - 1 = 0.0869$ , or  $\approx 8.7\%$ .

- b.  $G_g = \sqrt[4]{\frac{19,176}{13,470}} - 1 = 0.0869$  or 8.7%. As expected, both ways yield the same result.

- 3.40 a. Range = Max – Min = 10 – (−8) = 18

b.  $\mu = \frac{\sum x_i}{N} = \frac{0}{5} = 0$ ;  $MAD = \frac{\sum |x_i - \mu|}{N} = \frac{24}{5} = 4.8$

c.  $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{184.00}{5} = 36.80$

d.  $\sigma = \sqrt{\sigma^2} = 6.07$

- 3.42 a. Range = Max – Min = 12 – (−10) = 22

b.  $\bar{x} = \frac{\sum x_i}{n} = \frac{-6}{6} = -1$ ;  $MAD = \frac{\sum |x_i - \bar{x}|}{n} = \frac{44}{6} = 7.33$

c.  $s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{406}{5} = 81.2$ ;  $s = \sqrt{s^2} = 9.01$

- 3.44 a. For Starbucks,  $\bar{x} = \frac{\sum x_i}{n} = \frac{145}{6} = 24.17$ ,

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{12.8334}{5} = 2.57$$

For Panera Bread Co,  $\bar{x} = \frac{\sum x_i}{n} = \frac{454}{6} = 75.67$ ,

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{63.3334}{5} = 12.67$$

- b. Relative to Panera Bread's stock price, Starbucks' stock price has a lower variability as indicated by a lower standard deviation.

c. For Starbucks,  $CV = \frac{s}{\bar{x}} = \frac{1.6}{24.17} = 0.07$

For Panera,  $CV = \frac{s}{\bar{x}} = \frac{3.56}{75.67} = 0.05$

Starbucks has a greater relative dispersion since its sample coefficient of variation is higher.

- 3.46 a.  $\bar{x} = 164.10$ ;  $s^2 = 5,476.99$ ;  $s_{MKT} = \sqrt{s^2} = 74.01$ ;

$$CV_{MKT} = \frac{s}{\bar{x}} = \frac{74.01}{164.10} = 0.45$$

- b.  $\bar{x} = 40.71$ ;  $s^2 = 11,040.45$ ;  $s_{Return} = \sqrt{s^2} = 105.07$ ;

$$CV_{Return} = \frac{s}{\bar{x}} = \frac{105.07}{40.71} = 2.58$$

- c. Total return has a higher coefficient of variation which translates in a greater relative dispersion.

- 3.48 a. Investment B provides a higher return. Investment A provides the least risk since it has a smaller standard deviation.

b.  $Sharpe_A = \frac{\bar{x} - \bar{R}_f}{s_i} = \frac{10 - 1.4}{5} = 1.72$ ;  $Sharpe_B = \frac{\bar{x} - \bar{R}_f}{s_i} = \frac{15 - 1.4}{10} = 1.36$ .

The Sharpe Ratio is higher for investment A; hence it provides a higher reward per unit of risk.

- 3.50 a.  $\bar{x} = 3$ ,  $\bar{x}_2 = 5$ . The 2<sup>nd</sup> Investment provides a higher return since it has a higher mean.

- b.  $s_1 = 5.29$ ,  $s_2 = 9.02$ . The 1<sup>st</sup> investment provides the least risk because it has a lower standard deviation.

- c.  $Sharpe_1 = \frac{\bar{x} - \bar{R}_f}{s_1} = \frac{3 - 1.2}{5.29} = 0.34$ ;  $Sharpe_2 = \frac{\bar{x} - \bar{R}_f}{s_1} = \frac{5 - 1.2}{9} = 0.42$ . The 2<sup>nd</sup> investment performs better because it offers more reward per risk.

- 3.54 a. The values 70 and 90 are two standard deviations below the mean and above the mean, respectively. Using Chebyshev's Theorem and  $k = 2$ , we have  $1 - 1/2^2 = 0.75$ . In other words, Chebyshev's Theorem asserts that at least 75% of the scores fall within 70 and 90.

- b. The values 65 and 95 are three standard deviations below the mean and above the mean, respectively. Using Chebyshev's Theorem and  $k = 3$ , we have  $1 - 1/3^2 = 0.89$ . In other words, Chebyshev's Theorem asserts that at least 89% of the scores fall within 65 and 95.

- 3.56 a. We know that at least 75% of the observations fall within two standard deviations of the mean. We are given the mean and standard deviation of 500 and 25, respectively. Therefore, at least 75% of the observations fall within 450 and 550.

- b. We know that at least 89% of the observations fall within three standard deviations of the mean. We are given the mean and standard deviation of 500 and 25, respectively. Therefore, at least 89% of the observations fall within 425 and 575.

- 3.58 a. According to the empirical rule about 68% of the observations are between 700 and 800. Hence, half of the remaining 32%, that is 16%, of the observations are less than 700.

- b. 16% of 500 is 80 observations.

- 3.60 a.  $[\bar{x} - 2s, \bar{x} + 2s] = [5 - 2 \times 2.5, 5 + 2 \times 2.5] = [0, 10]$ . Given that 95% of the data falls within two standard deviations of the mean,  $95\% + 2.5\% = 97.5\%$  of the observations are positive.

- b.  $100\% - 97.5\% = 2.5\%$  of the observations are not positive.

- 3.64 a. The salaries \$66,000 and \$78,000 are two standard deviations below the mean and above the mean, respectively. Using Chebyshev's theorem and  $k = 2$ , we have  $1 - 1/2^2 = 0.75$ . In other words, Chebyshev's theorem asserts that at least 75% of the faculty earns at least \$66,000 but no more than \$78,000.

- b. The salaries \$63,000 and \$81,000 are three standard deviations below the mean and above the mean, respectively. Using Chebyshev's theorem and  $k = 3$ , we have  $1 - 1/3^2 = 0.89$ . In other words, Chebyshev's theorem asserts that at least 89% of the faculty earns at least \$63,000 but no more than \$81,000.

- 3.66 a. About 68% of the scores are in the interval [84, 116].

- b. About 95% of the scores are in the interval [68, 132]. About 2.5% of the scores are less than 68.

- c. Using a., about 16% of the scores are more than 116.

- 3.68 a. Talk times 2.4 and 5.6 hours are two standard deviations below the mean and above the mean, respectively. Using Chebyshev's theorem and  $k = 2$ , we have  $1 - 1/2^2 = 0.75$ . Therefore, at least 75% of cell phones will have talk time between 2.4 and 5.6 hours.

- b. Using the empirical rule we know that approximately 95% of cell phones will have talk time between 2.4 and 5.6 hours.



- 3.72 a.  $\bar{x} = \frac{\sum m_i f_i}{n} = \frac{2,305}{35} = 65.86$   
b.  $s^2 = \frac{\sum (m_i - \bar{x})^2 f_i}{n-1} = \frac{3,024.29}{34} = 88.95$ ;  $s = \sqrt{s^2} = 9.43$
- 3.74 a.  $\bar{x} = \frac{\sum m_i f_i}{n} = \frac{168}{50} = 3.36$   
b.  $s^2 = \frac{\sum (m_i - \bar{x})^2 f_i}{n-1} = \frac{189.52}{49} = 3.87$ ;  $s = \sqrt{s^2} = 1.97$
- 3.76 a.  $\bar{x} = \frac{\sum m_i f_i}{n} = \frac{538}{70} = 7.69$   
b.  $s^2 = \frac{\sum (m_i - \bar{x})^2 f_i}{n-1} = \frac{231.01}{69} = 3.35$ ;  $s = \sqrt{s^2} = 1.83$
- 3.78 a.  $\bar{x} = \frac{\sum m_i f_i}{n} = \frac{3,575}{88} = 40.63$   
b.  $s^2 = \frac{\sum (m_i - \bar{x})^2 f_i}{n-1} = \frac{6,708.63}{87} = 77.11$ ;  $s = \sqrt{s^2} = 8.78$
- 3.80 a.  $s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{-49.21}{4} = -12.3$   
b.  $r_{xy} = \frac{s_{xy}}{s_x s_y} = -0.96$ . There is a strong negative relationship between  $x$  and  $y$ .
- 3.82 a.  $s_{\text{Price, Days}} = \frac{4,419.75}{7} = 631.39$ . The covariance suggests that there is a positive linear relationship between the two variables.  
b.  $r_{\text{Price, Days}} = \frac{s_{\text{Price, Days}}}{s_{\text{Price}} s_{\text{Days}}} = 0.45$ . The correlation coefficient indicates that there is a moderate, positive relationship between the price of a home and the number of days it takes to sell it.
- 3.84 a.  $s_{\text{Educ, Sal}} = \frac{245}{7} = 35$ . The relationship between education and salary is positive.  
b.  $r_{\text{Educ, Sal}} = 0.95$ . The correlation coefficient indicates that the relationship between Education and Salary is positive and strong. Education seems to be a good indicator of Salary.
- 3.88 Mean = 809.14; Median = 366. Mode = *not available*. The median best reflects the typical sales as the value 3,300 is clearly an outlier that pulls the mean up.
- 3.92 a.  $G_g(\text{Gap}) = \sqrt{\frac{14.20}{15.73}} - 1 = -0.050$ , or  $-5\%$ ;  $G_g(\text{AE}) = \sqrt{\frac{2.99}{3.06}} - 1 = -0.012$ , or  $-1.2\%$   
b. American Eagle growth is less negative as compared to The Gap over this period.
- 3.94 a.  $\bar{x} = \frac{\sum m_i f_i}{n} = \frac{3262}{200} = 16.31$   
b.  $s^2 = \frac{\sum (m_i - \bar{x})^2 f_i}{n-1} = \frac{12,111.78}{199} = 60.86$ ,  $s = \sqrt{s^2} = 7.80$
- 3.96 a.  $s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{685.9425}{4} = 171.49$   
b.  $r_{xy} = \frac{s_{xy}}{s_x s_y} = 0.95$ . The correlation coefficient is close to 1. There is a strong positive relationship between the funds' annual returns.
- 4.10 a. Events  $A$  and  $B$  are not exhaustive because you may not have got an offer from either firm.  
b. Events  $A$  and  $B$  are not mutually exclusive because you may get an offer from both firms.
- 4.12 a. In 1971,  $P(\text{Age} \leq 40) = 0.26 + 0.45 = 0.71$   
In 2006,  $P(\text{Age} \leq 40) = 0.01 + 0.12 = 0.13$   
b. In 1971,  $P(\text{Age} \geq 51) = 0.05 + 0.03 = 0.08$   
In 2006,  $P(\text{Age} \geq 51) = 0.48 + 0.11 = 0.59$   
c. The data indicate that relative to 1971, younger workers are less attracted to municipal positions in 2006.
- 4.14 a. An odd of 15:8 means that an individual who, prior to the final, bet \$15 on Spain winning the World Cup would have won \$8 in gains if Spain did win. Therefore, net gain for the \$1000 bet is:  $\$1,000 \times \left(\frac{8}{15}\right) = \$533.33$   
b.  $P(\{\text{Spain wins}\}) = \frac{15}{(15+8)} = 0.652$
- 4.16  $P(A) = 0.65$ ;  $P(B) = 0.30$ ;  $P(A|B) = 0.45$   
a.  $P(A \cap B) = P(A|B)P(B) = (0.45)(0.30) = 0.135$   
b.  $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.65 + 0.30 - 0.135 = 0.815$   
c.  $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.135}{0.65} = 0.208$
- 4.18  $P(A) = 0.25$ ;  $P(B) = 0.30$   
a.  $P(A \cap B) = 0$ , by definition of mutually exclusive events  
b.  $P(A \cup B) = P(A) + P(B) = 0.25 + 0.30 = 0.55$   
c.  $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0}{0.30} = 0$  (Note that if  $B$  occurred then  $A$  cannot occur since the two are mutually exclusive.)
- 4.20  $P(A) = 0.65$ ,  $P(B) = 0.30$ , and  $P(A|B) = 0.45$   
a. No, because  $P(A|B) = 0.45 \neq 0.65 = P(A)$   
b.  $P(A \cap B) = P(A|B)P(B) = 0.45(0.30) = 0.135 \neq 0$ .  $A$  and  $B$  are not mutually exclusive events.  
c.  $P(A \cup B)^c = 1 - P(A \cup B)$ ;  $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.65 + 0.30 - 0.135 = 0.815$ ; Therefore,  $P(A \cup B)^c = 1 - 0.815 = 0.185$
- 4.22  $P(A) = 0.25$ ,  $P(B^c) = 0.40$ , and  $P(A \cap B) = 0.08$   
a.  $P(B) = 1 - P(B^c) = 1 - 0.40 = 0.60$   
b.  $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.08}{0.60} = 0.133$   
c.  $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.08}{0.25} = 0.32$
- 4.24  $P(A) = 0.40$ ,  $P(B) = 0.50$ , and  $P(A^c \cap B^c) = 0.24$   
a.  $P(A^c|B^c) = \frac{P(A^c \cap B^c)}{P(B^c)} = \frac{0.24}{1-0.50} = 0.48$   
b.  $P(A^c \cup B^c) = P(A^c) + P(B^c) - P(A^c \cap B^c) = (1 - 0.40) + (1 - 0.50) - 0.24 = 0.86$   
c.  $P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 0.24$ . Therefore,  $P(A \cup B) = 1 - P(A^c \cap B^c) = 1 - 0.24 = 0.76$
- 4.26 Let event  $O$  correspond to "students who ever go to their professor during office hours", and events  $MI$  and  $MA$  to "minor clarification" and "major clarification", respectively.  
 $P(O) = 0.20$ ,  $P(MI|O) = 0.3$ ,  $P(MA|O) = 0.7$   
a.  $P(MI \cap O) = P(MI|O)P(O) = (0.3)(0.2) = 0.06$   
b.  $P(MA \cap O) = P(MA|O)P(O) = (0.7)(0.2) = 0.14$
- 4.28 Let event  $R$  correspond to "Reduction in unemployment in the US", and event  $E$  to "Recession in Europe".  $P(R) = 0.18$ ,  $P(R|E) = 0.06$   
a.  $P(R^c) = 1 - P(R) = 1 - 0.18 = 0.82$   
b.  $P(R^c \cap E) = P(R^c|E)P(E) = (1 - 0.06)0.08 = (0.94)0.08 = 0.0752$
- 4.30 Let  $F$  correspond to "Foreign student" and  $S$  to "Smoke". We have  $P(F \cap S) = 0.05$  and  $P(S|F) = 0.50$ .

## Chapter 4

- 4.2 a. 1 to 1  
b. 9 to 1  
c. 0.67 to 1
- 4.4 a.  $A \cup B = \{1, 2, 3, 5, 6\} \neq S$ . Thus  $A$  and  $B$  are not exhaustive.  $A \cap B = A \neq \emptyset$ . Thus  $A$  and  $B$  are not mutually exclusive.  
b.  $A \cup C = \{1, 2, 3, 4, 6\} \neq S$ . Thus  $A$  and  $C$  are not exhaustive.  $A \cap C = \emptyset$ . Thus  $A$  and  $C$  are mutually exclusive.  
c.  $A \cup D = \{1, 2, 3, 4, 5, 6\} = S$ . Thus  $A$  and  $D$  are exhaustive.  $A \cap D = \emptyset$ . Thus  $A$  and  $D$  are mutually exclusive.  
d.  $B \cup C = \{1, 2, 3, 4, 5, 6\} = S$ . Thus  $B$  and  $C$  are exhaustive.  $B \cap C = \{6\} \neq \emptyset$ . Thus  $B$  and  $C$  are not mutually exclusive.

From  $P(S|F) = \frac{P(F \cap S)}{P(F)}$ , we get  $P(F) = \frac{P(F \cap S)}{P(S|F)} = \frac{0.05}{0.50} = 0.10$ .

Therefore, 10% of the student body at the university is foreign.

4.32 a.  $P(A) = \frac{3+5}{20} = 0.40$ ,  $P(A^c) = 1 - P(A) = 1 - 0.40 = 0.60$ ,  
 $P(B) = \frac{10}{20} = 0.50$

b. If 10 shirts in size L are of mixed colors, events  $A$  and  $B$  would be mutually exclusive. In this case,  $P(A \cap B) = 0$ . If any of the 10 shirts in size L are in white or blue, then events  $A$  and  $B$  would not be mutually exclusive. In this case,  $P(A \cap B) \neq 0$ . Events  $A$  and  $B$  are not exhaustive since  $P(A \cup B) < 1$ .

c. Mike's preference is described by the event  $A \cap B$ . It is the intersection of the events of getting both "a white or blue" and "size L" shirt.

4.34 Let event  $S_1$  correspond to "the first part is suitable" and  $S_2$  to "the second part is suitable".

a.  $P(S_1) = \frac{15}{20} = 0.75$

b.  $P(S_2 | S_1) = \frac{14}{19} = 0.7368$ .

c.  $P(S_2^c | S_1) = \frac{5}{19} = 0.2632$ .

4.36 For  $i = 1, 2$ , let event  $A_i$  be "the  $i$ -th selected member is in favor of the bonus".

a.  $P(A_1 \cap A_2) = \frac{10}{15} \times \frac{9}{14} = 0.4286$ .

b.  $P(A_1^c \cap A_2^c) = 5/15 \times 4/14 = 0.0952$ .

4.38 Let event  $A$  correspond to "Asians",  $B$  to "black",  $W$  to "white",  $H$  to "Hispanic", and  $T$  to "both parents at home". We have  $P(T|A) = 0.85$ ,  $P(T|W) = 0.78$ ,  $P(T|H) = 0.70$ , and  $P(T|B) = 0.38$ .

a.  $P(A \cap B) = 0$ . Events  $A$  and  $B$  are mutually exclusive.

$P(A) + P(B) < 1$ . Events  $A$  and  $B$  are not exhaustive.

b.  $P(W^c) = 1 - P(W) = 1 - 280/500 = 0.44$ .

c.  $P(W \cap T) = P(T|W)P(W) = 0.78(0.56) = 0.4368$ .

d.  $P(A \cap T) = P(T|A)P(A) = 0.85(0.10) = 0.085$ .

4.40 Let event  $H$  be "Woman faces sexual harassment", and event  $T$  be "Woman uses public transportation". We have  $P(H) = 2/3 \approx 0.67$ ,  $P(H|T) = 0.82$ , and  $P(T) = 0.28$ .

a.  $P(H \cap T) = P(H|T)P(T) = 0.82(0.28) \approx 0.23$

b.  $P(T|H) = \frac{P(H \cap T)}{P(H)} \approx \frac{0.23}{0.67} = 0.34$

4.42 Let event  $F$  correspond to "Foreclosed", event  $H$  to "centered in Arizona, California, Florida, and Nevada". Since  $P(F) = 0.0079$  and  $P(H|F) = 0.62$ , we obtain  $P(F \cap H) = P(H|F)P(F) = 0.0079(0.62) = 0.0049$  or 0.49%

4.48 a. Joint probability table:

| Global Warming | Political Affiliation |                |       |
|----------------|-----------------------|----------------|-------|
|                | Democrat (D)          | Republican (R) | Total |
| Yes (Y)        | 280                   | 120            | 400   |
| No (N)         | 120                   | 280            | 400   |
| Total          | 400                   | 400            | 800   |

b.  $P(R \cap Y) = \frac{120}{800} = 0.15$ .

c.  $P(N) = \frac{400}{800} = 0.50$ .

d.  $P(D|Y) = \frac{280}{400} = 0.70$ .

4.52 a.  $P(B^c) = 1 - P(B) = 1 - 0.60 = 0.40$

b.  $P(A \cap B) = P(A|B)P(B) = 0.80(0.60) = 0.48$

$P(A \cap B^c) = P(A|B^c)P(B^c) = 0.10(0.40) = 0.04$

c.  $P(A) = P(A \cap B) + P(A \cap B^c) = 0.48 + 0.04 = 0.52$

d.  $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.48}{0.52} = 0.9231$

4.56

| Prior Probabilities | Conditional Probabilities | Joint Probabilities                 | Posterior Probabilities  |
|---------------------|---------------------------|-------------------------------------|--|
| $P(B_1) = 0.10$     | $P(A B_1) = 0.40$         | $P(A \cap B_1) = (0.1)(0.4) = 0.04$ | $P(B_1 A) = \frac{P(A \cap B_1)}{P(A)} = \frac{0.04}{0.64} = 0.06$ |
| $P(B_2) = 0.60$     | $P(A B_2) = 0.60$         | $P(A \cap B_2) = (0.6)(0.6) = 0.36$ | $P(B_2 A) = \frac{P(A \cap B_2)}{P(A)} = \frac{0.36}{0.64} = 0.56$ |
| $P(B_3) = 0.30$     | $P(A B_3) = 0.80$         | $P(A \cap B_3) = (0.3)(0.8) = 0.24$ | $P(B_3 A) = \frac{P(A \cap B_3)}{P(A)} = \frac{0.24}{0.64} = 0.38$ |
| Total = 1.00        |                           | $P(A) = 0.04 + 0.36 + 0.24 = 0.64$  | Total = 1.00   |

4.58 Let event  $D$  be "Experience a decline", and event  $N$  be "Ratio is negative". We have  $P(D) = 0.20$ ,  $P(N|D) = 0.70$ , and  $P(N|D^c) = 0.15$ . In order to find  $P(D|N)$ , we first compute the probability that the ratio will be negative:

$$P(N) = P(N \cap D) + P(N \cap D^c) = P(N|D)P(D) + P(N|D^c)P(D^c) = 0.70(0.20) + 0.15(0.8) = 0.14 + 0.12 = 0.26$$

$$\text{Therefore, } P(D|N) = \frac{P(N \cap D)}{P(N)} = 0.14/0.26 = 0.54$$

4.62 Let  $F$  = "Player is fully fit to play",  $S$  = "Player is somewhat fit to play",  $N$  = "Player is not able to play", and  $W$  = "The Lakers win the game".

a. Consider the following table:

|               |                 |                                    |
|---------------|-----------------|------------------------------------|
| $P(F) = 0.40$ | $P(W F) = 0.80$ | $P(W \cap F) = 0.40(0.80) = 0.32$  |
| $P(S) = 0.30$ | $P(W S) = 0.60$ | $P(W \cap S) = 0.60(0.30) = 0.18$  |
| $P(N) = 0.30$ | $P(W N) = 0.40$ | $P(W \cap N) = 0.30(0.40) = 0.12$  |
| Total = 1.00  |                 | $P(W) = 0.32 + 0.18 + 0.12 = 0.62$ |

The Lakers have a 62% chance of winning the game.

b.  $P(F|W) = \frac{P(W \cap F)}{P(W)} = \frac{0.32}{0.62} = 0.52$

4.64 a.  $8! = 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 40,320$

$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$

b.  $\frac{8!}{(8-6)!} = \frac{8 \times 7}{2 \times 1} = 28$

c.  $\frac{8!}{(8-6)} = 8 \times 7 \times 6 \times 5 \times 4 \times 3 = 20,160$

4.66  $8! = 8 \times 7 \times 6 \times \dots \times 1 = 40,320$

4.68 a. We use the combination formula since the order does not matter;

$$\frac{10!}{(10-5)!5!} = \frac{10 \times 9 \times 8 \times 7 \times 6}{5 \times 4 \times 3 \times 2 \times 1} = 252$$

b. Here we use the permutation formula since the order does matter;  $\frac{10!}{(10-5)!} = 10 \times 9 \times 8 \times 7 \times 6 = 30,240$

4.86

|                     | Survived for Discharge (S) | Did not Survive for Discharge (S <sup>c</sup> ) | Total  |
|---------------------|----------------------------|---|--------|
| Day or Evening (D)  | 0.1338                     | 0.5417  | 0.6755 |
| Graveyard Shift (G) | 0.0477                     | 0.2768  | 0.3245 |
| Total               | 0.1815                     | 0.8185  | 1.00   |

a.  $P(G) = 0.3245$

b.  $P(S) = 0.1815$

c.  $P(S|G) = \frac{P(G \cap S)}{P(G)} = \frac{0.0477}{0.3245} = 0.1470$

d.  $P(G|S) = \frac{P(G \cap S)}{P(S)} = \frac{0.0477}{0.1815} = 0.2628$

e.  $P(S|G) = 0.1470 \neq 0.1815 = P(S)$ . (Also,  $P(G|S) = 0.2628 \neq 0.3245 = P(G)$ .)

Therefore, whether or not a patient survives is not independent of the timing of a heart attack. These results

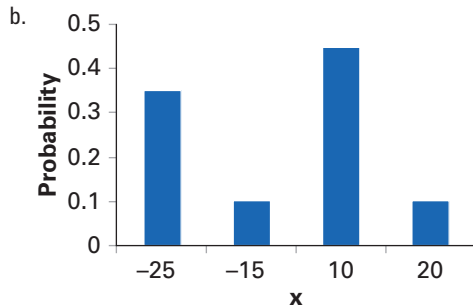
suggest that hospitals should have an equally adequate qualified professionals and resources available to patients at all times.

- 4.88 Let  $A$ : "US economy performs well" and  $B$ : "Asian countries perform well". We have  $P(A) = 0.40$ ,  $P(B|A) = 0.80$ , and  $P(B|A^c) = 0.30$ .

a.  $P(A \cap B) = P(B|A)P(A) = 0.80(0.40) = 0.32$   
 b.  $P(B) = P(B \cap A) + P(B \cap A^c) = P(B|A)P(A) + P(B|A^c)P(A^c) = 0.32 + 0.30(1 - 0.40) = 0.32 + 0.18 = 0.50$   
 c.  $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.32}{0.50} = 0.64$

## Chapter 5

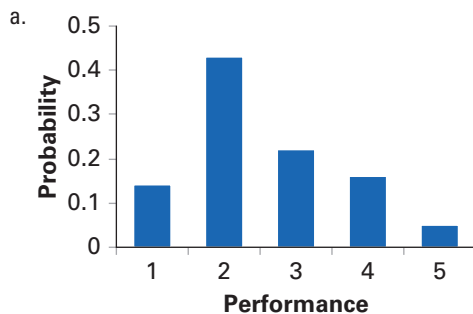
- 5.2 a.  $P(X = 10) = 0.45$



The distribution is not symmetric.

- c.  $P(X < 0) = 0.45$   
 d.  $P(X > -20) = 0.65$   
 e.  $P(X < 20) = 0.90$   
 5.4 a.  $P(X \leq 0) = 0.5$   
 b.  $P(X = 50) = 0.25$   
 c. Yes. The distribution has a finite number of values, each with an equal probability of occurring.

- 5.8 Let the random variable  $X$  represent performance.



The analyst has a somewhat pessimistic view based on the positively skewed distribution. There is only a 21% chance that the performance will be good (4) or very good (5).

- b.

| $x$           | $P(X \leq x)$ |
|---------------|---------------|
| 1 (Very poor) | 0.14          |
| 2 (Poor)      | 0.57          |
| 3 (Neutral)   | 0.79          |
| 4 (Good)      | 0.95          |
| 5 (Very good) | 1             |

- c.  $P(X \geq 4) = P(X = 4) + P(X = 5) = 0.16 + 0.05 = 0.21$ .

- 5.10 Let the random variable  $X$  represent confidence score.

- a. The probability that the confidence score is 2 (the consumer confidence will stay between 62 and 65) is  $1 - 0.35 - 0.25 = 0.40$ .

- b. The probability that the confidence score is not 1 (the consumer confidence will not fall below 62) is  $1 - 0.35 = 0.65$ .

- 5.14  $\mu = 5 \times 0.35 + 10 \times 0.30 + 15 \times 0.20 + 20 \times 0.15 = 10.75$   
 $\sigma^2 = (5 - 10.75)^2 \times 0.35 + (10 - 10.75)^2 \times 0.30 + (15 - 10.75)^2 \times 0.20 + (20 - 10.75)^2 \times 0.15 = 28.19$   $\sigma = \sqrt{28.19} = 5.31$   
 5.18 a.  $E(X) = -5 \times 0.30 + 0 \times 0.45 + 10 \times 0.25 = 1$   
 b.  $\sigma^2 = (-5 - 1)^2 \times 0.30 + (0 - 1)^2 \times 0.45 + (10 - 1)^2 \times 0.25 = 31.5$   
 $\sigma = \sqrt{31.5} = 5.61$   
 5.20  $E(X) = 4 \times 0.10 + 3 \times 0.30 + 2 \times 0.40 + 1 \times 0.10 + 0 \times 0.10 = 2.2$   
 5.24 a.  $E(X) = 1,000 \times 0.25 + 2,000 \times 0.45 + 5,000 \times 0.20 + 10,000 \times 0.10 = 3,150$   
 b. If Victor is risk neutral, he should not purchase the extended warranty because his expected cost (\$3,125) without the extended warranty is less than the cost of the warranty (\$3,400). The decision is not clear cut if he is risk averse; the decision will depend on his degree of risk aversion.  
 5.26 a.  $E(R) = 20 \times 0.20 + 10 \times 0.50 - 10 \times 0.30 = 6$   
 $\sigma = \sqrt{(20 - 6)^2 \times 0.20 + (10 - 6)^2 \times 0.50 + (-10 - 6)^2 \times 0.30}$   
 $= \sqrt{124} = 11.14$ , that is, 11.14%  
 b.  $E(R) = 40 \times 0.20 + 20 \times 0.50 - 40 \times 0.30 = 6$   
 $\sigma = \sqrt{(40 - 6)^2 \times 0.20 + (20 - 6)^2 \times 0.50 + (-40 - 6)^2 \times 0.30}$   
 $= \sqrt{964} = 31.05$   
 c. You would choose Fund 1 because it is less risky (smaller standard deviation) than Fund 2, with the same expected return of 6%.  
 5.28 The portfolio consists of  $\$20 \times 100 = \$2,000$  invested in Stock  $X$  and  $\$12 \times 200 = \$2,400$  invested in Stock  $Y$ . Total investment =  $\$2,000 + \$2,400 = \$4,400$ .  
 $w_X = \frac{2000}{4400} = 0.45$  and  $w_Y = \frac{2400}{4400} = 0.55$   
 5.30 a.  $w_X = \frac{200,000}{500,000} = 0.40$   
 $w_Y = \frac{300,000}{500,000} = 0.60$   
 b.  $E(R_p) = 0.40 \times 8 + 0.60 \times 12 = 10.4\%$   
 c.  $SD(R_p) = \sqrt{(0.40)^2(12)^2 + (0.60)^2(20)^2 + 2(0.40)(0.60)(0.40)(12)(20)}$   
 $= \sqrt{213.12} = 14.60$   
 5.32 a.  $E(R_p) = 0.60 \times 14 + 0.45 \times 8 = 11.6\%$   
 $\sigma^2 = (0.60)^2(26)^2 + (0.40)^2(14)^2 + 2(0.60)(0.40)(0.20)(26)(14)$   
 $= 309.66$   
 b.  $E(R_p) = 0.60 \times 14 + 0.40 \times 4 = 10\%$   
 $\sigma^2 = (0.60)^2(26)^2 + (0.40)^2(0)^2 + 2(0.60)(0.40)(0.20)(26)(0)$   
 $= 243.36$   
 c. The portfolios in parts a. and b. offer better expected returns than the bond alone, but have slightly higher variances than the bond fund with variance =  $(14)^2 = 196$ .  
 5.34 a.  $P(X = 0) = \frac{5!}{0!(5!)} (0.35)^0 (0.65)^5 = 0.1160$   
 b.  $P(X = 1) = \frac{5!}{1!(4!)} (0.35)^1 (0.65)^4 = 0.3124$   
 c.  $P(X \leq 1) = P(X = 0) + P(X = 1) = 0.1160 + 0.3124 = 0.4284$   
 5.36 a.  $P(3 < X < 5) = P(X = 4) = \frac{8!}{4!(4!)} (0.32)^4 (0.68)^4 = 0.1569$   
 b.  $P(3 < X \leq 5) = P(X = 4) + P(X = 5) = 0.1569 +$   
 $\left[ \frac{8!}{5!(3!)} (0.32)^5 (0.68)^3 \right] = 0.1569 + 0.0591 = 0.2160$   
 c.  $P(3 \leq X \leq 5) = P(X = 3) + P(X = 4) + P(X = 5) =$   
 $\left[ \frac{8!}{3!(5!)} (0.32)^3 (0.68)^5 \right] + 0.1569 + 0.0591 = 0.2668 + 0.1569$   
 $+ 0.0591 = 0.4828$

- 5.38 a.  $P(X \leq 50) = 0.2776$ ; Excel command: '=BINOM.DIST(50, 150, 0.36, 1)'
- b.  $P(X = 40) = 0.0038$ ; Excel command: '=BINOM.DIST(40, 150, 0.36, 0)'
- c.  $P(X > 60) = 1 - P(X \leq 60) = 1 - 0.8652 = 0.1348$ ; Excel command for  $P(X \leq 60)$ : '=BINOM.DIST(60, 150, 0.36, 1)'
- d.  $P(X \geq 55) = 1 - P(X \leq 54) = 1 - 0.5370 = 0.4630$ ; Excel command for  $P(X \leq 54)$ : '=BINOM.DIST(54, 150, 0.36, 1)'
- 5.40 a.  $P(X = 0) = \frac{8!}{0!(8!)} (0.20)^0 (0.80)^8 = 0.1678$
- b.  $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \left[ \frac{8!}{0!(8!)} (0.20)^0 (0.80)^8 \right] + \left[ \frac{8!}{1!(7!)} (0.20)^1 (0.80)^7 \right] + \left[ \frac{8!}{2!(6!)} (0.20)^2 (0.80)^6 \right] = 0.1678 + 0.3355 + 0.2936 = 0.7969$
- c.  $P(X \geq 7) = P(X = 7) + P(X = 8) = \left[ \frac{8!}{7!(1!)} (0.20)^7 (0.80)^1 \right] + \left[ \frac{8!}{8!(0!)} (0.20)^8 (0.80)^0 \right] = 0.0001 + 0.0000 = 0.0001$
- d.  $E(X) = 8(0.20) = 1.6$ , that is, 1.6 individuals
- e.  $\sigma^2 = 8(0.20)(0.80) = 1.28$   
 $\sigma = \sqrt{1.28} = 1.1314$
- 5.42 a.  $P(X < 5) = P(X \leq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = \left[ \frac{10!}{0!(10!)} (0.23)^0 (0.77)^{10} \right] + \left[ \frac{10!}{1!(9!)} (0.23)^1 (0.77)^9 \right] + (\text{etc.}) = 0.0733 + 0.2188 + 0.2942 + 0.2343 + 0.1225 = 0.9431$
- b.  $P(X < 5) = P(X \leq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = \left[ \frac{10!}{0!(10!)} (0.41)^0 (0.59)^{10} \right] + \left[ \frac{10!}{1!(9!)} (0.41)^1 (0.59)^9 \right] + (\text{etc.}) = 0.0051 + 0.0355 + 0.1111 + 0.2058 + 0.2503 = 0.6078$
- 5.44 a.  $P(X = 1) = \frac{6!}{1!(5!)} (0.76)^1 (0.24)^5 = 0.0036$
- b.  $P(X \geq 5) = P(X = 5) + P(X = 6) = \left[ \frac{6!}{5!(1!)} (0.76)^5 (0.24)^1 \right] + \left[ \frac{6!}{6!(0!)} (0.76)^6 (0.24)^0 \right] = 0.3651 + 0.1927 = 0.5578$
- c.  $P(X < 2) = P(X = 0) + P(X = 1) = \left[ \frac{6!}{0!(6!)} (0.76)^0 (0.24)^6 \right] + 0.0036 = 0.0002 + 0.0036 = 0.0038$
- d.  $E(X) = 6(0.76) = 4.56$   
 $P(X > 4.56) = P(X = 5) + P(X = 6) = 0.5578$  (from b)
- 5.46 a.  $P(X > 2) = P(X = 3) + P(X = 4) = \left[ \frac{4!}{3!(1!)} (0.50)^3 (0.50)^1 \right] + \left[ \frac{4!}{4!(0!)} (0.50)^4 (0.50)^0 \right] = 0.25 + 0.0625 = 0.3125$
- b.  $P(X > 2) = P(X = 3) + P(X = 4) = \left[ \frac{4!}{3!(1!)} (0.63)^3 (0.37)^1 \right] + \left[ \frac{4!}{4!(0!)} (0.63)^4 (0.37)^0 \right] = 0.3701 + 0.1575 = 0.5276$
- c.  $P(X > 2) = P(X = 3) + P(X = 4) = \left[ \frac{4!}{3!(1!)} (0.36)^3 (0.64)^1 \right] + \left[ \frac{4!}{4!(0!)} (0.36)^4 (0.64)^0 \right] = 0.1194 + 0.0168 = 0.1362$
- 5.50 Let  $X$  be the number of the students who specialize in finance.
- a.  $P(X = 10) = 0.1171$ ; Excel command: '=BINOM.DIST(10, 20, 0.40, 0)'
- b.  $P(X \leq 10) = 0.8725$ ; Excel command: '=BINOM.DIST(10, 20, 0.40, 1)'
- c.  $P(X \geq 15) = 1 - P(X \leq 14) = 1 - 0.9984 = 0.0016$ ; Excel command for  $P(X \leq 14)$  is '=BINOM.DIST(14, 20, 0.40, 1)'
- 5.52 a.  $P(X = 1) = \frac{e^{-1.5} 1.5^1}{1!} = 0.3347$
- b.  $P(X = 2) = \frac{e^{-1.5} 1.5^2}{2!} = 0.2510$
- c.  $P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - \left[ \frac{e^{-1.5} 1.5^0}{0!} + 0.3347 \right] = 1 - (0.2231 + 0.3347) = 1 - 0.5578 = 0.4422$
- 5.54 a.  $\mu_{0.5} = \frac{8}{2} = 4$
- b.  $P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - \left[ \frac{e^{-4} 4^0}{0!} + \frac{e^{-4} 4^1}{1!} \right] = 1 - (0.0183 + 0.0733) = 1 - 0.0916 = 0.9084$
- c.  $\mu_2 = 8(2) = 16$
- d.  $P(X = 10) = \frac{e^{-16} 16^{10}}{10!} = 0.0341$
- 5.56 a.  $P(X < 14) = P(X \leq 13) = 0.0661$ ; Excel command: '=POISSON.DIST(13, 20, 1)'
- b.  $P(X \geq 20) = 1 - P(X \leq 19) = 1 - 0.4703 = 0.5297$ ; Excel command for  $P(X \leq 19)$ : '=POISSON.DIST(19, 20, 1)'
- c.  $P(X = 25) = 0.0446$ ; Excel command: '=POISSON.DIST(25, 20, 0)'
- d.  $P(18 \leq X \leq 23) = P(X \leq 23) - P(X \leq 17) = 0.7875 - 0.2970 = 0.4905$   
 Excel commands: '=POISSON.DIST(23, 20, 1)' and: '=POISSON.DIST(17, 20, 1)'
- 5.60 a.  $\mu_{60} = 360$  cars over a 60-minute period thus  $\mu_1 = 360/60 = 6$  cars over a 1-minute period; so  $P(X = 2) = \frac{e^{-6} 6^2}{2!} = 0.0446$
- b.  $P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - \left[ \frac{e^{-6} 6^0}{0!} + \frac{e^{-6} 6^1}{1!} \right] = 1 - (0.0025 + 0.0149) = 1 - 0.0174 = 0.9826$
- c.  $\mu_{10} = 60$  for a 10-minute period ( $6 \times 10$ ), and  $P(X = 40) = \frac{e^{-60} 60^{40}}{40!} = 0.001$
- 5.62 a.  $P(X > 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)] = 1 - \left[ \frac{e^{-2} 2^0}{0!} + \frac{e^{-2} 2^1}{1!} + \frac{e^{-2} 2^2}{2!} \right] = 1 - (0.1353 + 0.2707 + 0.2707) = 1 - 0.6767 = 0.3233$
- b.  $\mu_5 = 2(5) = 10$ , that is, 10 cars per 5 minute period  $P(X = 6) = \frac{e^{-10} 10^6}{6!} = 0.0631$
- c.  $\mu_{180} = 2(3)(60) = 360$
- 5.64 a.  $P(X \leq 425) = 0.8980$ ; Excel command: '=POISSON.DIST(425, 400, 1)'
- b.  $P(X \geq 375) = 1 - P(X \leq 374) = 1 - 0.1002 = 0.8998$ ; Excel command for  $P(X \leq 374)$  is: '=POISSON.DIST(374, 400, 1)'
- 5.68 a.  $P(X = 0) = \frac{3!}{0!(3-0)!} \times \frac{(25-3)!}{(4-0)!(25-3-4+0)!} = \frac{(1)(7315)}{12,650} = 0.5783$
- b.  $P(X = 1) = \frac{3!}{1!(3-1)!} \times \frac{(25-3)!}{(4-1)!(25-3-4+1)!} = \frac{(3)(1540)}{12,650} = 0.3652$
- c.  $P(X \leq 1) = P(X = 0) + P(X = 1) = 0.5783 + 0.3652 = 0.9435$
- 5.70  $P(X = 0) = \frac{2!}{0!(2-0)!} \times \frac{(12-2)!}{(3-0)!(12-2-3+0)!} = \frac{(1)(120)}{220} = 0.5455$
- $\mu = 3 \left( \frac{2}{12} \right) = 0.50$   $\sigma = \sqrt{3 \left( \frac{2}{12} \right) \left( 1 - \frac{2}{12} \right) \left( \frac{12-3}{12-1} \right)} = \sqrt{0.3409} = 0.5839$
- 5.72  $P(X \geq 8) = 1 - P(X \leq 7) = 1 - 0.9223 = 0.0777$   
 Excel command for  $P(X \leq 7)$ : '=HYPGEOM.DIST(7, 20, 25, 100, 1)'
- $\mu = 20 \left( \frac{25}{100} \right) = 5$   
 $\sigma = \sqrt{20 \left( \frac{25}{100} \right) \left( 1 - \frac{25}{100} \right) \left( \frac{100-20}{100-1} \right)} = \sqrt{3.0303} = 1.7408$

- 5.76 a.  $P(X=3) = \frac{12!}{3!(12-3)!} \times \frac{(18-12)!}{(3-3)!(18-12-3+3)!} = \frac{(220)(1)}{816} = 0.2696$   
 b.  $P(X \geq 2) = 1 - [P(X=0) + P(X=1)] =$   

$$1 - \left[ \frac{12!}{0!(12-0)!} \times \frac{(18-12)!}{(3-0)!(18-12-3+0)!} + \frac{12!}{1!(12-1)!} \times \frac{(18-12)!}{(3-1)!(18-12-3+1)!} \right]$$
  

$$= 1 - \frac{(1)(20)}{816} + \frac{(12)(15)}{816} = 1 - 0.0245 + 0.2206 = 1 - 0.2451 = 0.7549$$
- 5.78  $P(X=2) = \frac{4!}{2!(4-2)!} \times \frac{(20-4)!}{(2-2)!(20-4-2+2)!} = \frac{(6)(1)}{190} = 0.0316$
- 5.80 a.  $P(X=2) = 0.0495$ ; Excel command: '=HYPGEOM.DIST (2, 5, 5, 59, 0)'  
 b.  $P(X=5) = 0.0000002$ ; Excel command: '=HYPGEOM.DIST (5, 5, 5, 59, 0)'  
 c.  $P(X=1) = 0.0256$ ; Excel command: '=HYPGEOM.DIST (1, 1, 1, 39, 0)'  
 d. Since the two stages are independent, we multiply the above probabilities to derive the probability of winning the jackpot:  $0.0000002 \times 0.0256 = 0.00000000512$
- 5.82 a.  $E(R) = (-15)(0.25) + (5)(0.35) + (10)(0.40) = 2$   
 b.  $\sigma^2 = (-15 - 2)^2(0.25) + (5 - 2)^2(0.35) + (10 - 2)^2(0.40) = 101$   
 $\sigma^2 = \sqrt{101} = 10.05$
- 5.84 a.  $E(R_v) = 0.35 \times (10) + 0.65 \times (5) = 6.75$   
 b.  $\sigma^2 = (0.35)^2(98) + (0.65)^2(26) + (2)(0.35)(0.65)(22) = 33$   
 $\sigma = \sqrt{33} = 5.74$
- 5.92 a.  $P(X=10) = 0.0272$ ; Excel command: '=BINOM.DIST (10, 30, 0.19, 0)'  
 b.  $P(10 \leq X \leq 20) = P(X \leq 20) - P(X \leq 9) = 1 - 0.9549 = 0.0451$   
 Excel commands: '=BINOM.DIST(20, 30, 0.19, 1)' and: '=BINOM.DIST(9, 30, 0.19, 1)'  
 c.  $P(X \leq 8) = 0.8996$ ; Excel command: '=BINOM.DIST (8, 30, 0.19, 1)'
- 5.94 a.  $P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.4405 = 0.5595$ ; Excel command for  $P(X \leq 4)$ : '=POISSON.DIST (4, 5, 1)'  
 b.  $P(X < 5) = P(X \leq 4) = 0.4405$
- 5.98 a.  $P(X=6) = 0.0115$ ; Excel command: '=HYPGEOM.DIST (6, 10, 20, 80, 0)'  
 b.  $P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.9353 = 0.0647$ ; Excel command for  $P(X \leq 4)$ : '=HYPGEOM.DIST (4, 10, 20, 80, 1)'  
 c.  $P(X \leq 2) = 0.5206$ ; Excel command: '=HYPGEOM.DIST (2, 10, 20, 80, 1)'  
 d.  $E(X) = 10\left(\frac{20}{80}\right) = 2.5$

## Chapter 6

- 6.2 a. 0.30  
 b. 0.16  
 c. 0.70
- 6.4 a.  $f(x) = 0.0333$   
 b.  $\mu = 20$ ;  $\sigma^2 = 75$ ;  $\sigma = 8.66$   
 c.  $P(X > 10) = 0.8325$
- 6.6 a.  $\mu = 20$ ;  $\sigma^2 = 33.33$ ;  $\sigma = 5.77$   
 b.  $f(x) = 0.05$ ;  $P(X > 22) = 0.4$   
 c.  $P(15 \leq X \leq 23) = 0.4$
- 6.8 a.  $\mu = 16$   
 b.  $f(x) = 0.125$ ;  $P(X < 15.5) = 0.4375$   
 c.  $P(X > 14) = 0.75$
- 6.10  $f(x) = 0.11$ ;  $P(X > 10) = 0.67$
- 6.14 a.  $P(Z > 1.32) = 1 - P(Z \leq 1.32) = 1 - 0.9066 = 0.0934$   
 b.  $P(Z \leq -1.32) = 0.0934$   
 c.  $P(1.32 \leq Z \leq 2.37) = P(Z \leq 2.37) - P(Z < 1.32) = 0.9911 - 0.9066 = 0.0845$   
 d.  $P(-1.32 \leq Z \leq 2.37) = P(Z \leq 2.37) - P(Z < -1.32) = 0.9911 - 0.0934 = 0.8977$
- 6.16 a.  $P(-0.67 \leq Z \leq -0.23) = P(Z \leq -0.23) - P(Z < -0.67) = 0.4090 - 0.2514 = 0.1576$   
 b.  $P(0 \leq Z \leq 1.96) = P(Z \leq 1.96) - P(Z < 0) = 0.9750 - 0.5 = 0.4750$   
 c.  $P(-1.28 \leq Z \leq 0) = 0.5 - P(Z < -1.28) = 0.5 - 0.1003 = 0.3997$   
 d.  $P(Z > 4.2) = 1 - P(Z \leq 4.2) = 1 - 1 = 0$  (approximately)
- 6.18 a.  $z = -1.27$   
 b.  $P(z \leq Z \leq 0) = 0.1772$  implies that  $P(Z \leq z) = 0.5 - 0.1772 = 0.3228$ ;  $z = -0.46$   
 c.  $P(Z > z) = 0.9929$  implies that  $P(Z \leq z) = 1 - 0.9929 = 0.0071$ ;  $z = -2.45$   
 d.  $P(0.4 \leq Z \leq z) = 0.3368$  implies that  $P(Z \leq z) = 0.3368 + P(Z < 0.4) = 0.3368 + 0.6554 = 0.9922$ ;  $z = 2.42$
- 6.20 Let  $X$  represent the IQ score.  
 a. Since 84 and 116 represent plus or minus one standard deviation from the mean, about 68.26% of people scored between 84 and 116.  
 b. An IQ score of 68 is two standard deviations below the mean ( $68 = 100 - 32$ ). Since about 95.44% of the observations fall within two standard deviations of the mean, half of 4.56% ( $100 - 95.44$ ) = 2.28% of people scored less than 68.
- 6.22 Let  $X$  equal points scored in a game.  
 a. Since 60 and 100 represent plus or minus two standard deviations from the mean, about 95.44% of scores are between 60 and 100 points.  
 b. A score of 100 is two standard deviations above the mean ( $100 = 80 + 2 \times 10$ ). Since about 95.44% of the observations fall within two standard deviations of the mean, half of 4.56% ( $100 - 95.44$ ) = 2.28% of scores are more than 100 points. If there are 82 games in the regular season, we expect the team to score more than 100 points in approximately 2 games ( $82 \times 0.0228 = 1.87$ ).
- 6.24 a.  $P(X \leq 0) = P\left(Z \leq \frac{0-10}{4}\right) = P(Z \leq -2.5) = 0.0062$   
 b.  $P(X > 2) = P\left(Z > \frac{2-10}{4}\right) = P(Z > -2) = 1 - P(Z \leq -2) = 1 - 0.0228 = 0.9772$   
 c.  $P(4 \leq X \leq 10) = P\left(\frac{4-10}{4} \leq Z \leq \frac{10-10}{4}\right) = P(-1.5 \leq Z \leq 0) = P(Z \leq 0) - P(Z < -1.5) = 0.5 - 0.0668 = 0.4332$   
 d.  $P(6 \leq X \leq 14) = P\left(\frac{6-10}{4} \leq Z \leq \frac{14-10}{4}\right) = P(-1 \leq Z \leq 1) = P(Z \leq 1) - P(Z < -1) = 0.8413 - 0.1587 = 0.6826$
- 6.26 a.  $P(X > 7.6) = P\left(Z > \frac{7.6-2.5}{2}\right) = P(Z > 2.55) = 1 - P(Z \leq 2.55) = 1 - 0.9946 = 0.0054$   
 b.  $P(7.4 \leq X \leq 10.6) = P\left(\frac{7.4-2.5}{2} \leq Z \leq \frac{10.6-2.5}{2}\right) = P(2.45 \leq Z \leq 4.05) = P(Z \leq 4.05) - P(Z < 2.45) = 1 - 0.9929 = 0.0071$   
 c. Given  $P(Z > z) = 0.025$ , or  $P(Z \leq z) = 1 - 0.025 = 0.9750$ , we find  $z = 1.96$ . Therefore,  $x = 2.5 + 1.96(2) = 6.42$



- d.  $P(x \leq X \leq 2.5) = P(X \leq 2.5) - P(X < x) = 0.4943$ ;  
 $P(X < x) = P(X \leq 2.5) - 0.4943 = 0.50 - 0.4943 = 0.0057$ ;  
 using  $z = -2.53$ , we derive  $x = 2.5 - 2.53(2) = -2.56$
- 6.28 a.  $P(X \leq 150) = 1 - 0.10 = 0.90$ ;  $z = 1.28$ ;  $\mu + 1.28(15) = 150$ ;  
 $\mu = 130.8$   
 b.  $z = 1.28$ ;  $\mu + 1.28(25) = 150$ ;  $\mu = 118$   
 c.  $z = 1.28$ ;  $136 + 1.28 \sigma = 150$ ;  $\sigma = 10.9375$   
 d.  $z = 1.28$ ;  $128 + 1.28 \sigma = 150$ ;  $\sigma = 17.1875$
- 6.32 Let  $X$  represent sleep time on weekdays.
- a.  $P(X > 8) = P\left(Z > \frac{8-6.2}{1.2}\right) = P(Z > 1.5) = 1 - P(Z \leq 1.5) = 1 - 0.9332 = 0.0668$   
 b.  $P(X < 6) = P\left(Z < \frac{6-6.2}{1.2}\right) = P(Z < -0.17) = 0.4325$   
 c.  $P(6 \leq X \leq 8) = P\left(\frac{6-6.2}{1.2} \leq Z \leq \frac{8-6.2}{1.2}\right) = P(-0.17 \leq Z \leq 1.5) = P(Z \leq 1.5) - P(Z < -0.17) = 0.9332 - 0.4325 = 0.5007$
- 6.38 For both distributions, let  $X$  represent the number of weeks to find a job.
- a.  $P(X > 19) = P\left(Z > \frac{19-22}{2}\right) = P(Z > -1.5) = 1 - P(Z \leq -1.5) = 1 - 0.0668 = 0.9332$   
 b.  $P(X > 19) = P\left(Z > \frac{19-22}{2}\right) = P(Z > -1.5) = 1 - P(Z \leq -1.5) = 1 - 0.9332 = 0.0668$   
 c.  $P(23 \leq X \leq 25) = P\left(\frac{23-22}{2} \leq Z \leq \frac{25-22}{2}\right) = P(0.5 \leq Z \leq 1.5) = P(Z \leq 1.5) - P(Z < 0.5) = 0.9332 - 0.6915 = 0.2417$   
 d.  $P(23 \leq X \leq 25) = P\left(\frac{23-16}{2} \leq Z \leq \frac{25-16}{2}\right) = P(3.5 \leq Z \leq 4.5) = P(Z \leq 4.5) - P(Z < 3.5) = 1 - 0.9998 = 0.0002$
- 6.40 Let  $X$  represent the time required to assemble an electronic component.
- a.  $P(10 \leq X \leq 20) = P\left(\frac{10-16}{4} \leq Z \leq \frac{20-16}{4}\right) = P(-1.5 \leq Z \leq 1) = P(Z \leq 1) - P(Z < -1.5) = 0.8413 - 0.0668 = 0.7745$   
 b.  $P(X > 24) + P(X < 6) = P\left(Z > \frac{24-16}{4}\right) + P\left(Z < \frac{6-16}{4}\right) = P(Z > 2) + P(Z < -2.5) = 0.0228 + 0.0062 = 0.029$
- 6.42 Since  $\mu = 25 = (22 + 28)/2$ ,  $P(22 \leq X \leq 28) = 0.95$  implies  $P(X \leq 28) = 0.975$ . From the  $z$  table we infer that  $P(Z \leq z) = 0.975$  implies  $z = 1.96$ . We then use the inverse transformation  $x = \mu + z\sigma$  to solve for  $\sigma$  as  $\sigma = \frac{x - \mu}{z}$ . Therefore,  $\sigma = \frac{28 - 25}{1.96} = 1.53$ .
- 6.44 Let  $X$  represent the score on a marketing exam.
- a.  $P(50 \leq X \leq 80) = P\left(\frac{50-60}{20} \leq Z \leq \frac{80-60}{20}\right) = P(-0.5 \leq Z \leq 1) = P(Z \leq 1) - P(Z < -0.5) = 0.8413 - 0.3085 = 0.5328$   
 b.  $P(20 \leq X \leq 40) = P\left(\frac{20-60}{20} \leq Z \leq \frac{40-60}{20}\right) = P(-2 \leq Z \leq -1) = P(Z \leq -1) - P(Z < -2) = 0.1587 - 0.0228 = 0.1359$   
 c.  $P(X \geq x) = 0.15$  is equivalent to  $P(X < x) = 0.85$ . Since  $P(Z < 1.04) = 0.85$ , we find  $x = 60 + 1.04(20) = 80.80$   
 d.  $P(X < x) = 0.10$ ;  $z = -1.28$ ;  $x = 60 - 1.28(20) = 34.40$
- 6.46  $P(X > 0) = 0.90$  is equivalent to  $P(X \leq 0) = 0.10$ . We first use this cumulative probability to find the corresponding  $z = -1.28$ . We then use the inverse transformation  $x = \mu + z\sigma$  to solve for  $\sigma$  as  $\sigma = \frac{x - \mu}{z}$ . Therefore,  $\sigma = \frac{0 - 5.6}{-1.28} = 4.375$ .
- 6.48 Let  $X$  represent the return on a mutual fund.
- a. For the riskier fund:  $P(X < 0) = P\left(Z < \frac{0-8}{14}\right) = P(Z < -0.57) = 0.2843$   
 For the less risky fund:  $P(X < 0) = P\left(Z < \frac{0-4}{5}\right) = P(Z < -0.8) = 0.2119$ .

You should pick the less risky fund because it gives you a lower likelihood of earning a negative return ( $21.19\% < 28.43\%$ ).

b. For the riskier fund:  $P(X > 8) = P\left(Z > \frac{8-8}{14}\right) = P(Z > 0) = 0.5$

For the less risky fund:  $P(X > 8) = P\left(Z > \frac{8-4}{5}\right) = P(Z > 0.8) = 1 - P(Z \leq 0.8) = 1 - 0.7881 = 0.2119$

You should pick the riskier fund because it gives you a higher likelihood of earning a return above 8% ( $50\% > 21.19\%$ ).

- 6.50 Let  $X$  represent the life of the car battery (in months).
- a.  $P(X < 24) = P(Z < (24 - 40)/16) = P(Z < -1) = 0.1587$   
 b.  $P(X > 24) = 1 - P(X \leq 24) = 1 - 0.1587 = 0.8413$ ;  $E(X) = (-10)(0.1587) + (20)(0.8413) = 15.239$   
 c.  $E(500X) = 500E(X) = 500(15.239) = 7,619.5$
- 6.56 a.  $\lambda = \frac{1}{E(X)} = \frac{1}{25} = 0.04$ ;  $SD(X) = E(X) = 25$   
 b.  $P(20 \leq X \leq 30) = P(X \leq 30) - P(X < 20) = (1 - e^{-0.04(30)}) - (1 - e^{-0.04(20)}) = 0.6988 - 0.5507 = 0.1481$   
 c.  $P(15 \leq X \leq 35) = P(X \leq 35) - P(X < 15) = (1 - e^{-0.04(35)}) - (1 - e^{-0.04(15)}) = 0.7534 - 0.4512 = 0.3022$
- 6.58 a.  $P(X \leq 1) = 0.3935$   
 Excel command: '=EXPON.DIST(1,0.5,1)'  
 b.  $P(2 < X < 4) = P(X < 4) - P(X \leq 2) = 0.8647 - 0.6321 = 0.2326$  Excel command: '=EXPON.DIST(4,0.5,1) - EXPON.DIST(2,0.5,1)'  
 c.  $P(X > 10) = 1 - P(X \leq 10) = 1 - 0.9933 = 0.0067$  Excel command: '=1 - EXPON.DIST(10,0.5,1)'
- 6.60 a.  $\mu_Y = \exp\left(\frac{2(3)+2}{2}\right) = 54.60$ ;  $\sigma_Y^2 = (\exp(2) - 1)\exp(2(3) + 2) = 19,046$   
 b.  $\mu_Y = \exp\left(\frac{2(5)+2}{2}\right) = 403.43$ ;  $\sigma_Y^2 = (\exp(2) - 1)\exp(2(5) + 2) = 1,039,849$   
 c.  $\mu_Y = \exp\left(\frac{2(5)+3}{2}\right) = 665.14$ ;  $\sigma_Y^2 = (\exp(3) - 1)\exp(2(5) + 3) = 8,443,697$
- 6.62 a.  $P(Y \leq 7.5) = P(\ln(Y) \leq \ln(7.5)) = P(X \leq 2.01) = P\left(Z < \frac{2.01-1.8}{\sqrt{64}}\right) = P(Z \leq 0.26) = 0.6026$   
 b.  $P(8 < Y < 9) = P(\ln(8) < \ln(Y) < \ln(9)) = P(2.08 < X < 2.20) = P\left(\frac{2.08-1.8}{0.8} < Z < \frac{2.20-1.8}{0.8}\right) = P(0.35 < Z < 0.50) = P(Z < 0.50) - P(Z < 0.35) = 0.6915 - 0.6368 = 0.0547$   
 c.  $P(Y < y) = 0.90$  is equivalent to  $P(\ln(Y) < \ln(y)) = P(X < x) = 0.90$ ;  $z = 1.28$ ,  $x = 1.28(0.80) + 1.8 = 2.824$ ;  $y = \exp(2.824) = 16.84$
- 6.64 Let  $X$  represent the time between eating mosquitoes.
- a.  $\mu$  (Poisson) = 10;  $\mu$  (Exponential) =  $\frac{60}{10} = 6$   
 b.  $\lambda = \frac{1}{\mu} = \frac{1}{6} = 0.1667$ ;  $P(X > 15) = e^{-0.1667(15)} = 0.0821$   
 c.  $P(15 \leq X \leq 20) = P(X \leq 20) - P(X < 15) = e^{-0.1667(20)} - e^{-0.1667(15)} = 0.0821 - 0.0357 = 0.0464$
- 6.72 Let  $Y$  represent household income.
- a.  $\mu = \ln\left(\frac{\mu_Y^2}{\sqrt{\mu_Y^2 + \sigma_Y^2}}\right) = \ln\left(\frac{39,626^2}{\sqrt{39,626^2 + 10,000^2}}\right) = \ln(38,421.44) = 10.5564$   

$$\sigma = \sqrt{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)} = \sqrt{\ln\left(1 + \frac{10,000^2}{39,626^2}\right)} = 0.248$$
  
 b.  $P(Y > 39,626) = P(\ln(Y) > \ln(39,626)) = P\left(X > 10.5872\right) = P\left(Z > \frac{10.5872 - 10.5564}{0.2485}\right) = P(Z > 0.12) = 1 - P(Z \leq 0.12) = 1 - 0.5478 = 0.4522$



- c.  $P(Y < 20,000) = P(\ln(Y) < \ln(20,000)) = P(X < 9.9035) = P\left(Z < \frac{9.9035 - 10.5564}{0.2485}\right) = P(Z < -2.63) = 0.0043$
- 6.74 Let  $X$  represent the delivery time.
- a.  $\mu = \frac{1+5}{2} = 3$ ;  $\text{Var}(X) = \frac{(5-1)^2}{12} = \frac{16}{12} = 1.3333$
- b.  $f(x) = \frac{1}{5-1} = 0.25$ ;  $P(X > 4) = (5-4)(0.25) = 0.25$ . 25% of deliveries are made after 4:00 pm.
- c.  $P(X < 2.5) = (2.5-1)(0.25) = 0.375$ . 37.5% of deliveries are made prior to 2:30 pm.
- 6.76 Let  $X$  represent diastolic (a) and systolic readings (b).
- a.  $P(80 \leq X \leq 90) = P\left(\frac{80-79}{10} \leq Z \leq \frac{90-79}{10}\right) = P(0.1 \leq Z \leq 1.1) = P(Z \leq 1.1) - P(Z < 0.1) = 0.8643 - 0.5398 = 0.3245$
- b.  $P(120 \leq X \leq 139) = P\left(\frac{120-125}{17} \leq Z \leq \frac{139-125}{17}\right) = P(-0.29 \leq Z \leq 0.82) = P(Z \leq 0.82) - P(Z < -0.29) = 0.7939 - 0.3859 = 0.4080$
- 6.78 Let  $X$  represent the amount spent on St. Patrick's Day.
- a.  $P(X > 50) = P\left(Z > \frac{50-43.87}{3}\right) = P(Z > 2.04) = 1 - P(Z \leq 2.04) = 1 - 0.9793 = 0.0207$
- b.  $P(X > 50) = P\left(Z > \frac{50-29.54}{11}\right) = P(Z > 1.86) = 1 - P(Z \leq 1.86) = 1 - 0.9686 = 0.0314$
- c. Women are slightly more likely to spend over \$50, with a 3.14% likelihood as opposed to 2.07% likelihood for men.
- 6.94 Let  $X$  represent the relief time.
- a.  $P(X < 4) = P\left(Z < \frac{4-6}{2}\right) = P(Z < -1) = 0.1587$
- b.  $\mu = \ln\left(\frac{\mu_Y^2}{\sqrt{\mu_Y^2 + \sigma_Y^2}}\right) \ln\left(\frac{6^2}{\sqrt{6^2 + 2^2}}\right) = 1.7391$
- $$\sigma = \sqrt{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)} = \sqrt{\ln\left(1 + \left(\frac{2}{6}\right)^2\right)} = 0.3247$$
- $$P(Y < 4) = P(\ln(Y) < \ln(4)) = P(X < 1.3863) = P\left(Z < \frac{1.3863 - 1.7391}{0.3247}\right) = P(Z < -1.09) = 0.1379$$
- ## Chapter 7
- 7.2 Nonresponse bias if some people are less likely to stop at the booth. Selection bias since the booth is only open on the weekend.
- 7.4 a. Nonresponse bias if the people who respond are systematically different from those who do not respond.  
b. Selection bias since those who frequent the store in the morning are likely to prefer an earlier opening time.  
c. Selection bias since not everyone reads a newspaper. Nonresponse bias since people who respond may be systematically different.
- 7.6 a. Both sample means will be normally distributed since the population is normally distributed.  
b. Yes  
c.  $n = 20$ :  $P(\bar{X} < 12.5) = P(Z < 1.49) = 0.9319$   
 $n = 40$ :  $P(\bar{X} < 12.5) = P(Z < 2.11) = 0.9826$
- 7.8 a.  $E(\bar{X}) = 80$ ;  $SD(\bar{X}) = \frac{14}{\sqrt{100}} = 1.4$   
b.  $P(77 \leq \bar{X} \leq 85) = P(-2.14 \leq Z \leq 3.57) = 0.9998 - 0.0162 = 0.9836$   
c.  $P(\bar{X} > 84) = P(Z > 2.86) = 1 - 0.9979 = 0.0021$
- 7.10 a.  $P(\bar{X} > 105) = P(Z > 1.77) = 1 - 0.9616 = 0.0384$   
b.  $P(\bar{X} < 95) = P(Z < -1.77) = 0.0384$   
c.  $P(95 \leq \bar{X} \leq 105) = 0.9616 - 0.0384 = 0.9232$
- 7.12 a.  $P(\bar{X} \geq 18) = P(Z \geq 1.85) = 1 - 0.9678 = 0.0322$   
b.  $P(\bar{X} \geq 17.5) = P(Z \geq 2.03) = 1 - 0.9788 = 0.0212$   
c. Janice; her findings are more likely if a representative sample is used.
- 7.14 a. The sample mean has a normal distribution because the population is normally distributed.  
b.  $P(\bar{X} > 25) = P(Z > 2.4) = 1 - 0.9918 = 0.0082$   
c.  $P(18 \leq \bar{X} \leq 24) = P(-3.20 \leq Z \leq 1.60) = 0.9452 - 0.0007 = 0.9445$
- 7.16 a.  $P(\bar{X} > 25,000) = P\left(Z > \frac{25,000 - 27,200}{7000/\sqrt{4}}\right) = P(Z > -0.63) = 1 - 0.2643 = 0.7357$   
b.  $P(\bar{X} > 30,000) = P\left(Z > \frac{30,000 - 27,200}{7000/\sqrt{4}}\right) = P(Z > 0.80) = 1 - 0.7881 = 0.2119$
- 7.18 a.  $P(X > 1,000,000) = P(Z > (1,000,000 - 800,000)/250,000) = P(Z > 0.80) = 1 - 0.7881 = 0.2119$   
b.  $P\left(\sum_{i=1}^4 X_i > 4,000,000\right) = P(\bar{X} > 1,000,000) = P\left(Z > \frac{1,000,000 - 800,000}{250,000/\sqrt{4}}\right) = P(Z > 1.60) = 1 - 0.9452 = 0.0548$
- 7.24 a.  $\bar{p} = \frac{60}{200} = 0.30$ ,  $p = 0.26$  and  $n = 200$ ; Therefore  
 $P(\bar{P} < 0.30) = P\left(Z < \frac{0.30 - 0.26}{\sqrt{\frac{0.26(1-0.26)}{200}}}\right) = P(Z < 1.29) = 0.9015$   
b.  $p = 0.26$  is the proportion of French people who approve, therefore  $1 - 0.26 = 0.74$  is the population proportion of disapproval.  
 $\bar{p} = \frac{150}{200} = 0.75$ ;  $P(\bar{P} > 0.75) = P\left(Z > \frac{0.75 - 0.74}{\sqrt{\frac{0.74(1-0.74)}{200}}}\right) = P(Z > 0.32) = 1 - 0.6255 = 0.3745$
- 7.26 a. The sampling distribution of  $\bar{P}$  has  $E(\bar{P}) = p = 0.17$  and  $se(\bar{P}) = \sqrt{\frac{0.17(1-0.17)}{200}} = 0.0266$ ; The normal approximation criteria are met because  $np = 200(0.17) = 34 > 5$  and  $n(1-p) = 200(1-0.17) = 166 > 5$ . Therefore, it is appropriate to use the normal distribution approximation for the sample proportion.  
b.  $P(\bar{P} > 0.20) = P\left(Z > \frac{0.20 - 0.17}{\sqrt{\frac{0.17(1-0.17)}{200}}}\right) = P(Z > 1.13) = 1 - 0.8708 = 0.1292$
- 7.28 You would choose 50 balls because with larger sample sizes the standard deviation of  $\bar{P}$  is *reduced*. The proportion of green balls is  $60/100 = 0.6$ . Therefore your probability of getting 70% green balls is slightly higher with a smaller sample because of the increased standard deviation. If you are unsure about this you can calculate  $P(\bar{P} > 0.70)$  with  $n = 50$  and  $n = 100$  to confirm that the probability is higher for  $n = 50$ .  
For  $n = 50$ ,  $P(\bar{P} > 0.70) = P\left(Z > \frac{0.70 - 0.60}{\sqrt{0.60(1-0.60)/50}}\right) = P(Z > 1.44) = 1 - 0.9251 = 0.0749$ .  
For  $n = 100$ ,  $P(\bar{P} > 0.70) = P\left(Z > \frac{0.70 - 0.60}{\sqrt{0.60(1-0.60)/100}}\right) = P(Z > 2.04) = 1 - 0.9793 = 0.0207$
- 7.30 a.  $E(\bar{X}) = \mu = -45$ ;  $se(\bar{X}) = \sqrt{\frac{81}{100}} = 0.90$ . It is not necessary to apply the finite population correction because the sample constitutes less than 5 percent of the population:  $n = 100 < 125 = 2500(0.05)$ .  
b.  $P(-47 \leq \bar{X} \leq -43) = P\left(\frac{-47 - (-45)}{0.90} \leq Z \leq \frac{-43 - (-45)}{0.90}\right) = P(-2.22 \leq Z \leq 2.22) = 0.9868 - 0.0132 = 0.9736$   
c.  $P(\bar{X} > -44) = P\left(Z > \frac{-44 - (-45)}{0.90}\right) = P(Z > 1.11) = 1 - 0.8665 = 0.1335$

7.32 a.  $E(\bar{P}) = p = 0.34$ ;  $se(\bar{P}) = \sqrt{\frac{0.34(1-0.34)}{100}} = 0.047$ ; there is no need to apply the finite population correction because the sample size accounts for less than 5 percent of the population size:  $n = 100 < 0.05(3,000)$ .

b.  $P(\bar{P} > 0.37) = P\left(Z > \frac{0.37 - 0.34}{0.047}\right) = P(Z > 0.63) = 1 - 0.7357 = 0.2643$

7.34 First define:  $n = 120$ ,  $N = 1,000$ ,  $p = \frac{2}{3} = 0.67$ . Since the sample accounts for more than 5 percent of the population size ( $n = 120 > 0.05(1,000) = 50$ ), we need to apply the finite population correction; we then find  $E(\bar{P}) = 0.67$  and  $se(\bar{P}) =$

$\sqrt{\frac{0.67(1-0.67)}{120}} \sqrt{\frac{1,000-120}{1,000-1}} = 0.0403$ . Then,  $\frac{75}{120} = 0.625$ , so we compute  $P(\bar{P} > 0.625) = P\left(Z > \frac{0.625 - 0.67}{0.0403}\right) = P(Z > -1.12) = 1 - 0.1314 = 0.8686$

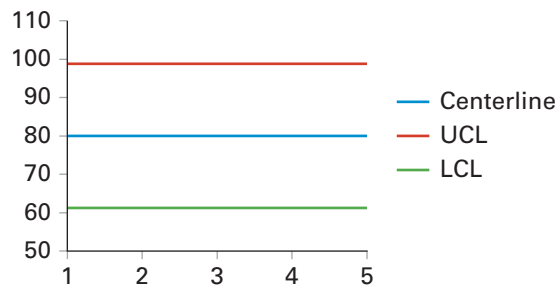
7.36 a. No, it is not necessary because the sample size accounts for less than 5 percent of the population size:  $n = 12 < 0.05(500) = 25$ .

b. We cannot assume the sampling distribution of the sample mean is approximately normally distributed because we do not know if the population has a normal distribution and the sample size is not sufficiently large to assume so ( $n < 30$ ).

c.  $E(\bar{X}) = 10.32$ ;  $se(\bar{X}) = \frac{9.78}{\sqrt{12}} = 2.8232$

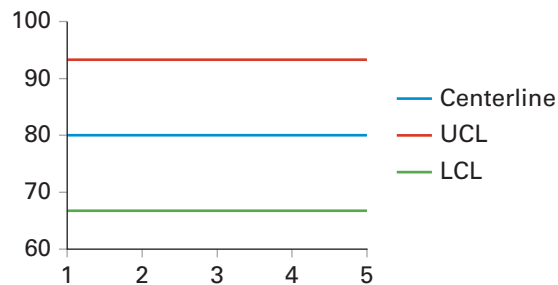
d. The normal approximation is not justified (see part b).

7.38 a.



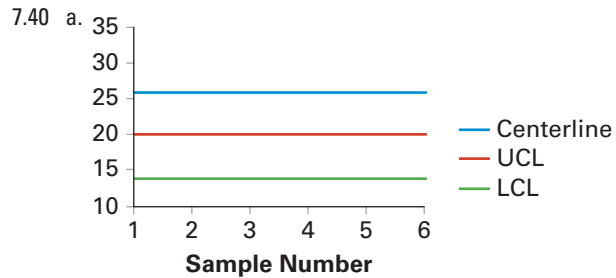
Centerline:  $\mu = 80$   
 $UCL = 80 + 3\left(\frac{14}{\sqrt{5}}\right) = 98.78$   
 $LCL = 80 - 3\left(\frac{14}{\sqrt{5}}\right) = 61.22$

b.



Centerline:  $\mu = 80$   
 $UCL = 80 + 3\left(\frac{14}{\sqrt{10}}\right) = 80 + 13.28 = 93.28$   
 $LCL = 80 - 3\left(\frac{14}{\sqrt{10}}\right) = 80 - 13.28 = 66.72$

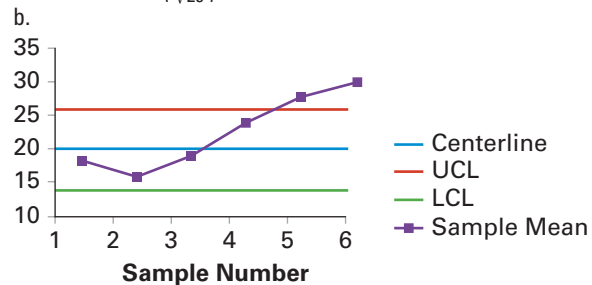
The larger sample size gives narrower control limits due to the smaller standard deviation.



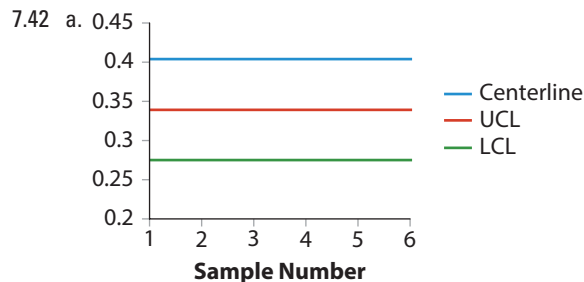
Centerline:  $\mu = 20$

$UCL = 20 + 3\left(\frac{10}{\sqrt{25}}\right) = 20 + 6 = 26$

$LCL = 20 - 3\left(\frac{10}{\sqrt{25}}\right) = 20 - 6 = 14$



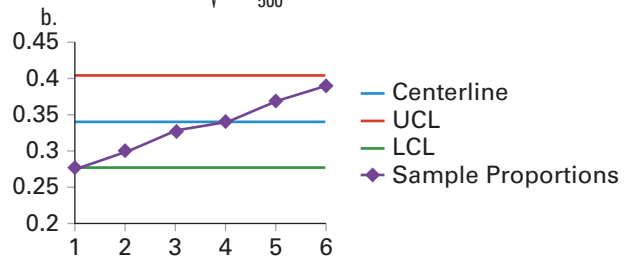
c. The last two points are outside the upper control limit. There is also an upward trend, suggesting the process is becoming increasingly out of control. The process should be adjusted.



Centerline:  $p = 0.34$

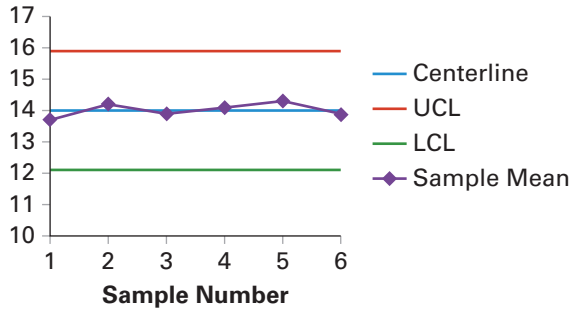
$UCL = 0.34 + 3\sqrt{\frac{0.34(1-0.34)}{500}} = 0.34 + 0.064 = 0.404$

$LCL = 0.34 - 3\sqrt{\frac{0.34(1-0.34)}{500}} = 0.34 - 0.064 = 0.276$



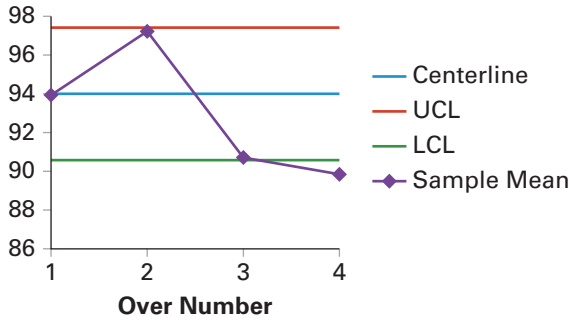
c. There are no points outside the control limits, so the process is under control. However, the positive trend suggests that the process may become out of control if the upward trend continues.

7.44 a.



All the sample means randomly lie within the control limits. Therefore, we can conclude that the production process is in control and operating properly.

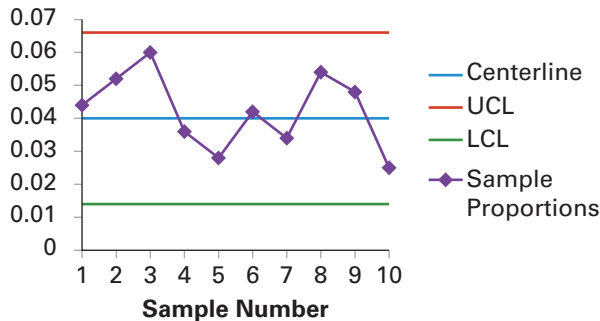
7.46 a.



To plot the average speed, take the average of each over:  
 Over 1:  $\bar{x} = \frac{96.8 + 99.5 + 88.8 + 81.9 + 100.1 + 96.8}{6} = 93.98$   
 Similarly,  
 Over 2:  $\bar{x} = 97.23$   
 Over 3:  $\bar{x} = 90.70$   
 Over 4:  $\bar{x} = 89.85$

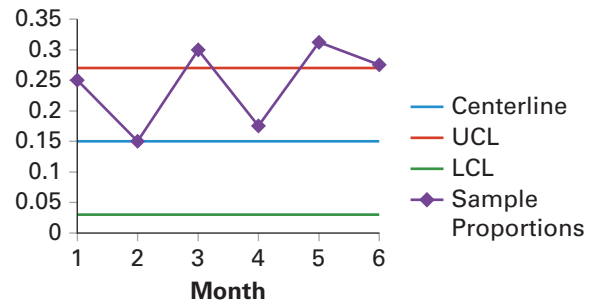
b. Kalwant's average speed is out of the control limits on 1 out of 4 of his overs, which rather justifies his coach's concern that he is not very consistent.

7.48 a.



b. All sample proportions are within the control limits and there is no apparent trend, which suggests that the machine is operating properly.

7.50 a.



b. Find the proportion of complaints each month:

| Month | Sample Proportion |
|-------|-------------------|
| 1     | 20/80 = 0.25      |
| 2     | 12/80 = 0.15      |
| 3     | 24/80 = 0.30      |
| 4     | 14/80 = 0.175     |
| 5     | 25/80 = 0.3125    |
| 6     | 22/80 = 0.275     |

We plot each sample proportion on the control chart (shown above) to see that 3 out of 6 months were out of the control limits, which is a good justification for why Dell chose to direct customers away from India call centers.

7.52

- As one example, use a random number table or a random number generator (in Excel, for instance) to randomly select individuals into the sample from the list of all residents of Miami. Then conduct the survey by contacting those selected.
- To get a stratified random sample, you could create strata based on ethnicity; for example under white, black, Hispanic, Asian, and then randomly select adults in each group and ask whether or not they walk regularly.
- To get a cluster sample, you could choose a number of representative neighborhoods in Miami and randomly select adults within these neighborhoods and ask whether or not they walk regularly.

$$7.54 \text{ a. } P(\bar{P} > 0.50) = P\left(Z > \frac{0.50 - 0.45}{\sqrt{\frac{0.45(1-0.45)}{50}}}\right) = P(Z > 0.71)$$

$$= 1 - 0.7611 = 0.2389$$

$$\text{b. } P(\bar{P} > 0.50) = P\left(Z > \frac{0.50 - 0.48}{\sqrt{\frac{0.48(1-0.48)}{100}}}\right) = P(Z > 0.28)$$

$$= 1 - 0.6103 = 0.3897$$

$$7.56 \text{ a. } P(X < 79) = P\left(Z < \frac{79 - 80}{2}\right) = P(Z < -0.5) = 0.3085$$

$$\text{b. } P(\bar{X} < 79) = P\left(Z < \frac{79 - 80}{2/\sqrt{10}}\right) = P(Z < -1.58) = 0.0571$$

$$\text{c. } P(\bar{X} < 79) = P\left(Z < \frac{79 - 80}{2/\sqrt{30}}\right) = P(Z < -2.74) = 0.0031$$

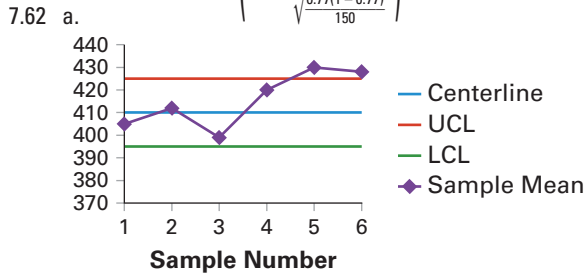
$$7.58 \text{ a. } P(X > 500) = P\left(Z > \frac{500 - 470.73}{50}\right) = P(Z > 0.59) = 1 - 0.7224 = 0.2776$$

$$\text{b. } P(\bar{X} > 500) = P\left(Z > \frac{500 - 470.73}{50/\sqrt{4}}\right) = P(Z > 1.17) = 1 - 0.8790 = 0.1210$$

- c. Since the amounts spent on the lottery are independent, the probability of all four randomly selected Georgia residents spending more than \$500 on lottery is:  
 $P(X > 500)^4 = (0.2776)^4 = 0.0059$ .

7.60 a.  $\frac{120}{150} = 80$ ;  $P(\bar{P} > 0.80) = P\left(Z > \frac{0.80 - 0.77}{\sqrt{\frac{0.77(1 - 0.77)}{150}}}\right) = P(Z > 0.87)$   
 $= 1 - 0.8078 = 0.1922$

b.  $P(\bar{P} < 0.70) = P\left(Z < \frac{0.70 - 0.77}{\sqrt{\frac{0.77(1 - 0.77)}{150}}}\right) = P(Z < -2.04) = 0.0207$



Centerline:  $\mu = 410$

$UCL = 410 + 3\left(\frac{25}{\sqrt{25}}\right) = 410 + 15 = 425$

$LCL = 410 - 3\left(\frac{25}{\sqrt{25}}\right) = 410 - 15 = 395$

- b. Two of the sample means are above the upper control limit, indicating that the advertised amount of sodium content is not accurate.

7.64 a.  $P(\bar{P} > 0.15) = P\left(Z > \frac{0.15 - 0.10}{\sqrt{\frac{0.10(1 - 0.10)}{50}}}\right) = P(Z > 1.18) = 1 - 0.8810 = 0.1190$

$P(\bar{P} > 0.15) = P\left(Z > \frac{0.15 - 0.10}{\sqrt{\frac{0.10(1 - 0.10)}{100}}}\right) = P(Z > 1.67) = 1 - 0.9525 = 0.0475$

## Chapter 8

- 8.2 a. For 89%,  $\alpha = 0.11$ ;  $z_{\alpha/2} = z_{0.055} = 1.598$   
 b. For 92%,  $\alpha = 0.08$ ;  $z_{\alpha/2} = z_{0.04} = 1.751$   
 c. For 96%,  $\alpha = 0.04$ ;  $z_{\alpha/2} = z_{0.02} = 2.054$

- 8.4 a.  $\bar{x}$  is approximately normally distributed because the sample size  $n$  is sufficiently large ( $n \geq 30$ ).  
 b. We use  $\alpha = 0.05$  to find  $z_{\alpha/2} = z_{0.025} = 1.96$ . The margin of error, with  $n = 64$ , for the 95% confidence interval is  $1.96 \frac{26.8}{\sqrt{64}} = 6.57$ .  
 c. The margin of error with  $n = 225$ , for the 95% confidence interval is  $1.96 \frac{26.8}{\sqrt{225}} = 3.50$ .  
 d. The smaller sample size of  $n = 64$  will lead to a higher margin of error, which will lead to a wider interval.

- 8.6 a. The sample mean,  $\bar{x} = 78.1$ , is the point estimate of the population mean.  
 b.  $z_{\alpha/2} = z_{0.05} = 1.645$ ; the margin of error is:  $1.645 \frac{4.5}{\sqrt{50}} = 1.05$ .  
 c. The 90% confidence interval is  $78.1 \pm 1.05$  or  $[77.05, 79.15]$

- 8.8 a. For the 95% confidence interval,  $z_{\alpha/2} = z_{0.025} = 1.96$ ; the confidence interval is  $6.4 \pm 1.96 \frac{1.8}{\sqrt{80}} = 6.4 \pm 0.39$  or  $[6.01, 6.79]$   
 b. Yes, we can conclude with 95% confidence that the mean sleep in this Midwestern town is not 7 hours because the value 7 does not fall within the confidence interval.

- 8.10 a. For the 99% confidence interval, the margin of error is  $2.576 \frac{500}{\sqrt{100}} = 128.80$

- b. The 99% confidence interval is  $7,790 \pm 128.80$  or  $[7,661.20, 7,918.80]$

8.20 a.  $\bar{x} = \frac{22 + 18 + 14 + 25 + 17 + 28 + 15 + 21}{8} = 20$   
 $s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{(22 - 20)^2 + (18 - 20)^2 + \dots + (21 - 20)^2}{7} = 24$   
 $s = \sqrt{s^2} = \sqrt{24} = 4.90$

- b. We use  $\alpha = 0.20$ ,  $df = 8 - 1 = 7$ , to find  $t_{0.10,7} = 1.415$ . The 80% confidence interval is  $20 \pm 1.415 \frac{4.90}{\sqrt{8}} = 20 \pm 2.45$  or  $[17.55, 22.45]$ .

- c. We use  $\alpha = 0.10$ ,  $df = 7$  to find  $t_{0.05,7} = 1.895$ . The 90% confidence interval is  $20 \pm 1.895 \frac{4.90}{\sqrt{8}} = 20 \pm 3.28$  or  $[16.72, 23.28]$ .

As the confidence level increases, the interval becomes wider.

- 8.22 a. Since weight loss is believed to be normally distributed, we can consider  $\bar{x}$  to be normally distributed. We use the  $t_{df}$  distribution. We use  $\alpha = 0.05$ ,  $df = 18 - 1 = 17$  to find  $t_{0.025,17} = 2.110$ . For the 95% confidence interval, the margin of error is  $2.11 \frac{9.2}{\sqrt{18}} = 4.58$ .

- b.  $12.5 \pm 4.58$  or  $[7.92, 17.08]$

- 8.24 a. Since the sample size is sufficiently large ( $n \geq 30$ ), we consider  $\bar{X}$  to be approximately normally distributed, and apply the  $t_{df}$  distribution. We use  $\alpha = 0.01$ ,  $df = 36 - 1 = 35$  to find  $t_{0.005,35} = 2.724$ . For the 99% confidence interval, the margin of error is  $2.724 \frac{10}{\sqrt{36}} = 4.54$ .

- b.  $100 \pm 4.54$  or  $[95.46, 104.54]$

8.26 a.  $\bar{x} = \frac{14 + 7 + 17 + 20 + 18 + 15 + 19 + 28}{8} = 17.25$   
 $s^2 = \frac{(14 - 17.25)^2 + (7 - 17.25)^2 + \dots + (28 - 17.25)^2}{8 - 1} = 35.36$   
 $s = \sqrt{s^2} = \sqrt{35.36} = 5.95$

We use  $\alpha = 0.01$ ,  $df = 8 - 1 = 7$  to find  $t_{0.005,7} = 3.499$ . The 99% confidence interval is  $17.25 \pm 3.499 \frac{5.95}{\sqrt{8}} = 17.25 \pm 7.36$  or  $[9.89, 24.61]$ .

- b. In order for the above confidence interval to be valid,  $\bar{X}$  must be normally distributed. Therefore, we must assume that the population has a normal distribution since  $n$  is smaller than 30.

8.30 a.  $\bar{x} = \frac{71 + 73 + 76 + 78 + 81 + 75}{6} = 75.67$   
 $s^2 = 12.67$   
 $s = \sqrt{s^2} = \sqrt{12.67} = 3.56$

- b. We use  $\alpha = 0.10$ ,  $df = 6 - 1 = 5$ , to find  $t_{0.05,5} = 2.015$ . The 90% confidence interval is:  $75.67 \pm 2.015 \frac{3.56}{\sqrt{6}} = 75.67 \pm 2.93$  or  $[72.74, 78.60]$ .

- c. The margin of error increases as the confidence level increases, and therefore the confidence interval becomes wider.

- 8.32 We use  $\alpha = 0.10$ ,  $df = n - 1 = 25 - 1 = 24$  to find  $t_{0.05,24} = 1.711$ .

The 90% confidence interval is  $\bar{x} \pm 1.711 s/\sqrt{25} = [1,690, 1,810]$ .

Therefore,  $\bar{x} = \frac{1,690 + 1,810}{2} = 1,750$ , and we can find  $s$  from  $1750 + 2.064 s/\sqrt{25} = 1810$ .

Thus,  $s = \frac{1810 - 1750}{1.711} \sqrt{25} = 175.34$ .

- 8.38 We use  $\alpha = 0.10$ ,  $df = 26 - 1 = 25$  to find  $t_{0.05,25} = 1.708$ .  
 Debt payments:  $\bar{x} = 983.46$ ; Excel command: '=AVERAGE(D2:D27)'

$s = 124.61$ ; Excel command: '=STDEV.S(D2:D27)'

The 90% confidence interval is:

$$983.46 \pm 1.708 \frac{124.61}{\sqrt{26}} = 983.46 \pm 41.74 \text{ or } [941.70, 1025.20].$$

Similarly, we use  $t_{\alpha/2, 25} = t_{0.025, 25} = 2.060$  to compute a 95% confidence interval as  $983.46 \pm 2.060 \frac{124.61}{\sqrt{26}} = 983.46 \pm 50.34$  or  $[933.12, 1033.80]$ . The 95% confidence interval is wider because its confidence level is higher.

- 8.40 a. We use  $\alpha = 0.05$  to find  $z_{\alpha/2} = z_{0.025} = 1.960$ . With  $n = 50$ , the 95% confidence interval for the population proportion is  $0.6 \pm 1.960 \sqrt{\frac{0.6(1-0.6)}{50}} = 0.6 \pm 0.136 = [0.464, 0.736]$ .
- b. With  $n = 200$ , the 95% confidence interval for the population proportion is  $0.6 \pm 1.960 \sqrt{\frac{0.6(1-0.6)}{200}} = 0.6 \pm 0.068$  or  $[0.532, 0.668]$ . Here, since  $n$  is larger, the interval is narrower and therefore more precise.
- 8.42 a.  $\bar{p} = \frac{40}{100} = 0.40$  is the point estimate of the population proportion.
- b. We use  $\alpha = 0.10$  to find  $z_{\alpha/2} = z_{0.05} = 1.645$ . Therefore the 90% confidence interval estimate is  $0.40 \pm 1.645 \sqrt{\frac{0.40(1-0.40)}{100}} = 0.40 \pm 0.081$  or  $[0.319, 0.481]$ .
- We use  $\alpha = 0.01$  to find  $z_{\alpha/2} = z_{0.005} = 2.576$ . Therefore the 99% confidence interval estimate is  $0.40 \pm 2.576 \sqrt{\frac{0.40(1-0.40)}{100}} = 0.40 \pm 0.126$  or  $[0.274, 0.526]$ .
- c. Yes, with 90% confidence, we can conclude that the population proportion differs from 0.5 because the value 0.5 does not fall within the interval.
- d. No, since the value 0.5 falls within the interval, we cannot conclude with 99% confidence that the population proportion differs from 0.5.
- 8.44 The population parameter of interest is the proportion of Americans who support Arizona's new immigration enforcement law. We assume  $n = 1,079$ ,  $\bar{p} = 0.51$ , and use  $\alpha = 0.05$  to find  $z_{\alpha/2} = z_{0.025} = 1.960$ . The 95% confidence interval is  $0.51 \pm 1.960 \sqrt{\frac{0.51(1-0.51)}{1,079}} = 0.51 \pm 0.030$  or  $[0.480, 0.540]$ .
- 8.46 a.  $\bar{p} = 0.37$ ,  $\alpha = 0.10$ , thus  $z_{\alpha/2} = z_{0.05} = 1.645$
- The 90% confidence interval is  $0.37 \pm 1.645 \sqrt{\frac{0.37(1-0.37)}{5,324}} = 0.37 \pm 0.011$  or  $[0.359, 0.381]$
- b.  $\bar{p} = 0.37$ ,  $\alpha = 0.01$ , thus  $z_{\alpha/2} = z_{0.005} = 2.576$
- The 99% confidence interval is  $0.37 \pm 2.576 \sqrt{\frac{0.37(1-0.37)}{5,324}} = 0.37 \pm 0.017$  or  $[0.353, 0.387]$
- c. The margin of error in part b is greater because it uses a higher confidence level.
- 8.48 a.  $\bar{p} = 0.44$ . We use  $\alpha = 0.10$  to find  $z_{\alpha/2} = z_{0.05} = 1.645$ . The 90% confidence interval for the population proportion is  $0.44 \pm 1.645 \sqrt{\frac{0.44(1-0.44)}{1000}} = 0.44 \pm 0.026$  or  $[0.414, 0.466]$ .
- b. The margin of error is 0.026.
- c. Using  $\alpha = 0.01$  and  $z_{\alpha/2} = z_{0.005} = 2.576$ , the margin of error increases to  $2.576 \sqrt{\frac{0.44(1-0.44)}{1000}} = 0.040$ .
- 8.52 a.  $\bar{p} = \frac{308}{1,026} = 0.300$ ,  $\alpha = 0.10$ , thus  $z_{\alpha/2} = z_{0.05} = 1.645$ . The 90% confidence interval for the population proportion is:  $0.300 \pm 1.645 \sqrt{\frac{0.30(1-0.30)}{1,026}} = 0.300 \pm 0.024$  or  $[0.276, 0.324]$ .

- b.  $\bar{p} = \frac{103}{1,026} = 0.100$ . The 90% confidence interval for the population proportion is:  $0.100 \pm 1.645 \sqrt{\frac{0.10(1-0.10)}{1,026}} = 0.100 \pm 0.015$  or  $[0.085, 0.115]$ .
- 8.54 a.  $\bar{p} = \frac{20}{80} = 0.25$ .
- b. We use  $\alpha = 0.05$  to find  $z_{\alpha/2} = z_{0.025} = 1.960$ . The 95% confidence interval for the population proportion is:  $0.25 \pm 1.960 \sqrt{\frac{0.25(1-0.25)}{80}} = 0.25 \pm 0.095$  or  $[0.155, 0.345]$ .
- c. No, the mayor's claim cannot be justified with 95% confidence since the national average value 0.20 ( $=1/5$ ) falls within the interval.
- 8.56 Given  $E = 10$ ,  $\sigma = 40$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$ ,
- $$n = \left( \frac{1.96 \times 40}{10} \right)^2 = 61.47$$
- , which is rounded up to 62.
- 8.58 For a 90% confidence interval,  $z_{\alpha/2} = z_{0.05} = 1.645$ . Given  $E = 1.2$ ,  $\hat{\sigma} = 3.5$ ,  $n = \left( \frac{z_{\alpha/2} \hat{\sigma}}{E} \right)^2 = \left( \frac{1.645 \times 3.5}{1.2} \right)^2 = 23.02$ , which is rounded up to 24. If the margin of error decreases to  $E = 0.7$ , then,
- $$n = \left( \frac{1.645 \times 3.5}{0.7} \right)^2 = 67.65$$
- , which is rounded up to 68.
- 8.62  $E = 20$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$ ,  $\hat{\sigma} = \frac{\text{range}}{4} = \frac{850 - 300}{4} = 137.50$ . Thus,
- $$n = \left( \frac{1.96 \times 137.50}{20} \right)^2 = 181.58$$
- , which is rounded up to 182.
- 8.64 a.  $E = 0.04$ ,  $\sigma_A = 0.206$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$ , thus
- $$n = \left( \frac{1.96 \times 0.206}{0.04} \right)^2 = 101.89$$
- , which is rounded up to 102.
- b.  $E = 0.04$ ,  $\sigma_B = 0.128$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$ , thus
- $$n = \left( \frac{1.96 \times 0.128}{0.04} \right)^2 = 39.34$$
- , which is rounded up to 40.
- c. Since the population standard deviation for Fund A is higher than for Fund B, it leads to a higher margin of error. Therefore, in order to achieve the same margin of error for both funds, Fund A requires a larger sample size.
- 8.70 Given  $E = 0.06$ ,  $\hat{p} = 2/5 = 0.40$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$ ,
- $$n = \left( \frac{1.96}{0.06} \right)^2 0.40(1 - 0.40) = 256.11$$
- , which is rounded up to 257.
- 8.72  $df = 9$ ,  $t_{\alpha/2, df} = t_{0.025, 9} = 2.262$ ; The 95% confidence interval is:
- $$10 \pm 2.262 \frac{15}{\sqrt{10}} = 10 \pm 10.73 \text{ or } [-0.73, 20.73]$$
- 8.74 a.  $df = 224$
- $$t_{\alpha/2, df} = t_{0.025, 224} = 1.971$$
- The 95% confidence interval is:  $16 \pm 1.971 \frac{12}{\sqrt{225}} = 16 \pm 1.58$  or  $[14.42, 17.58]$
- b. Yes, we can conclude with 95% confidence that the average worker does not take 14 days of vacation because the interval does not include the value 14.
- 8.78 With 90% confidence,  $z_{\alpha/2} = z_{0.05} = 1.645$ . Since  $\hat{\sigma}$  is not given, calculate it as  $\hat{\sigma} = \frac{\text{range}}{4} = \frac{800 - 200}{4} = 150$ .
- Given  $E = 15$ ,  $n = \left( \frac{1.645 \times 150}{15} \right)^2 = 270.60$ , which is rounded up to 271.
- 8.80 a.  $\bar{x} = \frac{13 + -2 + \dots + -14}{5} = 3.6$ ,
- $$s = \sqrt{\frac{(13 - 3.6)^2 + (-2 - 3.6)^2 + \dots + (-14 - 3.6)^2}{5 - 1}} = \sqrt{159.3} = 12.62$$
- ,
- $$df = 4$$
- .
- b.  $t_{\alpha/2, df} = t_{0.025, 4} = 2.776$ ; The 95% confidence interval is:  $3.6 \pm 2.776 \frac{12.62}{\sqrt{5}} = 3.6 \pm 15.67$  or  $[-12.07, 19.27]$ .
- c. You must assume that the annual returns at Vanguard Energy Fund follow a normal distribution.
- 8.86 a.  $z_{0.025} = 1.96$ ;  $1.96 \sqrt{\frac{0.121(1-0.121)}{1235}} = 0.018$
- b. The 95% confidence interval is:  $0.121 \pm 0.018$  or  $[0.103, 0.139]$ .



- 8.88 Given  $E = 0.05$ ,  $\hat{p} = 1/5 = 0.20$ ,  $z_{\alpha/2} = z_{0.05} = 1.645$ ,  
 $n = \left(\frac{1.645}{0.05}\right)^2 0.20(1 - 0.20) = 173.19$ , which is rounded up to 174.  
 This is assuming that  $\hat{p} = 0.20$ , based on prior studies, is a reasonable estimate of  $p$  in the planning stage.

## Chapter 9

- 9.2 a. Invalid. The test is about the population parameter  $\mu$ .  
 b. Valid  
 c. Valid  
 d. Invalid. The null hypothesis must include some form of the equality sign.
- 9.4 a. Incorrect. We never accept the null hypothesis.  
 b. Correct.  
 c. Incorrect. We cannot establish a claim because the null hypothesis is not rejected.  
 d. Correct.
- 9.6 a. Type I error is to incorrectly conclude the mean weight is different from 18 ounces. Type II error is to incorrectly conclude the mean weight does not differ from 18 ounces.  
 b. Type I error is to incorrectly conclude that the stock price increases on more than 60 percent of trading days. Type II error is to incorrectly conclude that the price does not increase on more than 60 percent of trading days.  
 c. Type I error is to incorrectly conclude that Americans sleep less than 7 hours a day. Type II error is to incorrectly conclude that Americans do not sleep less than 7 hours a day.
- 9.8 a. Type I error is to incorrectly conclude that the majority of voters support the candidate. Type II error is to incorrectly conclude that the majority of the voters do not support the candidate.  
 b. Type I error is to incorrectly conclude that the average pizza is less than 10 inches. Type II error is to incorrectly conclude that the average pizza is not less than 10 inches.  
 c. Type I error is to incorrectly conclude that the average tablet does not contain 250 mg. Type II error is to incorrectly conclude that an average tablet contains 250 mg.
- 9.10 a. 3%  
 b. 2%  
 c. Type I error is to incorrectly conclude that an individual has the disease. Type II error is to incorrectly conclude that an individual does not have the disease.  
 d. We do not prove that the individual is free of disease if we do not reject the null hypothesis.
- 9.14 With  $\alpha = 0.10$ , the critical value is  $z_{0.10} = 1.28$ .  
 With  $n = 25$ , the value of the test statistic,  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{13.4 - 12.6}{\frac{3.2}{\sqrt{25}}} = 1.25$ . We do not reject  $H_0$  because  $z = 1.25 < 1.28$ .  
 With  $n = 100$ , the  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{13.4 - 12.6}{\frac{3.2}{\sqrt{100}}} = 2.5$ . Here, we reject  $H_0$  because  $z = 2.5 > 1.28$ .
- 9.16 The value of the test statistic,  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{144 - 150}{28/\sqrt{80}} = -1.92$   
 and the  $p$ -value  $= P(Z \leq -1.92) = 0.0274$ . If  $\alpha = 0.01$ , we do not reject  $H_0$  because the  $p$ -value  $= 0.0274 > 0.01 = \alpha$ . If  $\alpha = 0.05$ , we reject  $H_0$  because the  $p$ -value  $= 0.0274 < 0.05 = \alpha$ .
- 9.22 a.  $H_0: \mu = 120$ ;  $H_A: \mu \neq 120$   
 b. The value of the test statistic,  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{114 - 120}{22/\sqrt{36}} = -1.64$ . The  $p$ -value  $= 2P(Z \leq -1.64) = 2(0.0505) = 0.101$ .  
 c. Since  $0.101 > 0.01 = \alpha$ , we do not reject  $H_0$ . The average braking distance is not significantly different from 120 miles.
- d. With  $\alpha = 0.01$ ,  $z_{0.005} = 2.576$ ; the critical values are  $-2.576$  and  $2.576$ . Since  $z = -1.64$  falls between  $-2.576$  and  $2.576$ , we do not reject  $H_0$ . We reach the same conclusion as in c.
- 9.24 a.  $H_0: \mu \leq 90$ ;  $H_A: \mu > 90$   
 b. The value of the test statistic,  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{95 - 90}{20/\sqrt{40}} = 1.58$ .  
 The  $p$ -value  $= P(Z \geq 1.58) = P(Z \leq -1.58) = 0.0571$ .  
 c. Since the  $p$ -value  $= 0.0571 > 0.01 = \alpha$ , we do not reject  $H_0$ . The manager's claim is not supported by the sample data.  
 d. With  $\alpha = 0.01$ , the critical value,  $z_{0.01} = 2.33$ . Since  $z = 1.58 < 2.33$ , we do not reject  $H_0$ . We reach the same conclusion as in c.
- 9.28 a.  $H_0: \mu = 30$ ;  $H_A: \mu \neq 30$   
 b. With  $\alpha = 0.05$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$ ; the critical values are  $-1.96$  and  $1.96$ .  
 Excel Command: `'=NORM.S.INV(0.975)'`  
 c. For the value of the test statistic, use Excel Command: `'=(average(B2:B27)-30)/(stdev.s(B2:B27)/sqrt(26))'`  $= 2.81$ .  
 d. Since  $2.81 > 1.96$  At the 5% significance level, the average weekly stock price of Home Depot is significantly different from \$30.
- 9.30 a.  $H_0: \mu \leq 27,200$ ;  $H_A: \mu > 27,200$   
 b. the critical value is 1.28.  
 c. For the value of the test statistic, use Excel Command: `'=(average(A2:A41)-27200)/(stdev.s(A2:A41)/sqrt(40))'`  $= 1.87$ .  
 d. Since  $-1.87 < 1.28$ , we do not reject  $H_0$ . At the 10% significance level, the average debt of recent undergraduates from Connecticut is not more than the national average.
- 9.32 a.  $0.05 < p$ -value  $< 0.10$ . At  $\alpha = 0.01$ , we do not reject  $H_0$  since the  $p$ -value  $> \alpha$ . At  $\alpha = 0.10$ , we reject  $H_0$  since the  $p$ -value  $< \alpha$ .  
 b.  $0.01 < p$ -value  $< 0.025$ . At  $\alpha = 0.01$ , we do not reject  $H_0$  since the  $p$ -value  $> \alpha$ . At  $\alpha = 0.10$ , we reject  $H_0$  since the  $p$ -value  $< \alpha$ .  
 c.  $0.01 < p$ -value  $< 0.025$ . At  $\alpha = 0.01$ , we do not reject  $H_0$  since the  $p$ -value  $> \alpha$ . At  $\alpha = 0.10$ , we reject  $H_0$  since the  $p$ -value  $< \alpha$ .  
 d.  $0.05 < p$ -value  $< 0.10$ . At  $\alpha = 0.01$ , we do not reject  $H_0$  since the  $p$ -value  $> \alpha$ . At  $\alpha = 0.10$ , we reject  $H_0$  since the  $p$ -value  $< \alpha$ .
- 9.34 a.  $0.05 < p$ -value  $< 0.10$ ; At  $\alpha = 0.01$ , we do not reject  $H_0$  since the  $p$ -value  $> \alpha$ . At  $\alpha = 0.10$ , we reject  $H_0$  since the  $p$ -value  $< \alpha$ .  
 b.  $0.05 < p$ -value  $< 0.10$ ; At  $\alpha = 0.01$ , we do not reject  $H_0$  since the  $p$ -value  $> \alpha$ . At  $\alpha = 0.10$ , we reject  $H_0$  since the  $p$ -value  $< \alpha$ .  
 c.  $0.02 < p$ -value  $< 0.05$ ; At  $\alpha = 0.01$ , we do not reject  $H_0$  since the  $p$ -value  $> \alpha$ . At  $\alpha = 0.10$ , we reject  $H_0$  since the  $p$ -value  $< \alpha$ .  
 d.  $0.01 < p$ -value  $< 0.02$ ; At  $\alpha = 0.01$ , we do not reject  $H_0$  since the  $p$ -value  $> \alpha$ . At  $\alpha = 0.10$ , we reject  $H_0$  since the  $p$ -value  $< \alpha$ .
- 9.36  $H_0: \mu = 16$   $H_A: \mu \neq 16$   
 a.  $df = 31$ ;  $t_{31} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{15.2 - 16}{0.6/\sqrt{32}} = -7.54$ ;  $p$ -value  $= 2P(T_{31} \leq -7.54) < 2(0.005) = 0.01$ . Since the  $p$ -value  $< \alpha = 0.01$ , we reject  $H_0$ . The sample data suggest that the population mean is different from 16.



- b. With  $\alpha = 0.01$ ,  $t_{\alpha/2, df} = t_{0.005, 31} = 2.744$ ; the critical values are  $-2.744$  and  $2.744$ . Since  $t_{31} = 7.54 > 2.744$ , we reject  $H_0$ ; same conclusion as in a.
- 9.40  $\bar{x} = \frac{554}{6} = 92.33$ ;  $s = \sqrt{\frac{311.33}{5}} = 7.89$ ;  $df = 5$ ;  $t_5 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{92.33 - 100}{7.89/\sqrt{6}} = -2.38$ .
- With  $\alpha = 0.01$ , the critical value,  $-t_{0.01, 5} = -3.365$ ; we reject  $H_0$  if  $t_5 < -3.365$ . Since  $-2.38 > -3.365$ , we do not reject  $H_0$ .
- 9.42 a.  $H_0: \mu \leq 5$ ;  $H_A: \mu > 5$
- b.  $df = 6$ ;  $t_6 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5.53 - 5}{2.18/\sqrt{7}} = 0.64$ .
- It is necessary to assume that the population is normally distributed given that the small sample size of 7 observations is less than 30.
- c. With  $\alpha = 0.10$ , the critical value,  $t_{\alpha, df} = t_{0.10, 6} = 1.440$ . Since  $t_6 = 0.64 < 1.440$ , we do not reject  $H_0$ . The average waiting time is not significantly more than 5 minutes at the 10% level. There is no need to hire an additional employee.
- d.  $p\text{-value} = P(T_6 \geq 0.64) > 0.20$ ; since the  $p\text{-value} > \alpha = 0.10$ , we do not reject  $H_0$ . Same conclusion as in c.
- 9.44 a.  $H_0: \mu = 12$ ;  $H_A: \mu \neq 12$
- b. It is not necessary to assume that the underlying population is normally distributed. The sample size  $n = 48$  is larger than 30, so the sample mean is normally distributed according to Central Limit Theorem.
- c. With  $\alpha = 0.05$ ,  $t_{\alpha/2, df} = t_{0.025, 47} = 2.012$ ; the critical values are 2.012 and -2.012. The decision rule is to reject  $H_0$  if  $t_{47} > 2.012$  or  $t_{47} < -2.012$ .
- d.  $t_{47} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{11.80 - 12}{0.8/\sqrt{48}} = -1.73$ . Since  $t_{47} = 1.73$  falls between  $-2.012$  and  $2.012$ , we do not reject  $H_0$ . The bottling process has not fallen out of adjustment.
- 9.46 We will conduct the test with the  $p\text{-value}$  approach.
- $H_0: \mu \geq 6$ ;  $H_A: \mu < 6$
- Based on the performance of 12 cars, we use  $\bar{x} = 5.92$ ,  $s = 0.09$ , and  $df = 11$  to compute  $t_{11} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5.92 - 6}{0.09/\sqrt{12}} = -3.08$ ; the  $p\text{-value} = P(t \leq -3.08)$ ;  $0.005 < p\text{-value} < 0.01$ .
- Since the  $p\text{-value} < \alpha = 0.05$ , we reject  $H_0$ . At the 5% level of significance, we can conclude that the average clock time of all cars is less than 6 seconds.
- 9.50 a.  $H_0: \mu = 95$ ;  $H_A: \mu \neq 95$
- b.  $df = 25 - 1 = 24$ ;  $t_{24} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{96.52 - 95}{10.70/\sqrt{25}} = 0.71$
- We use Excel to compute the  $p\text{-value}$  as  $'=T.DIST.2T(0.71, 24)' = 0.48$ .
- c. Since the  $p\text{-value} = 0.48 > 0.05 = \alpha$ , we do not reject  $H_0$ . The average MPG is not significantly different from 95.
- 9.54 a.  $\bar{p} = \frac{22}{74} = 0.30$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.30 - 0.38}{\sqrt{\frac{0.38(1-0.38)}{74}}} = -1.42$ ; the  $p\text{-value} = P(Z \leq -1.42) = 0.0778$ .
- b.  $\bar{p} = \frac{110}{300} = 0.37$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.37 - 0.38}{\sqrt{\frac{0.38(1-0.38)}{300}}} = -0.36$ ; the  $p\text{-value} = P(Z \leq -0.36) = 0.3594$ .
- c.  $\bar{p} = 0.34$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.34 - 0.38}{\sqrt{\frac{0.38(1-0.38)}{50}}} = -0.58$ ; the  $p\text{-value} = P(Z \leq -0.58) = 0.2810$ .
- d.  $\bar{p} = 0.34$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.34 - 0.38}{\sqrt{\frac{0.38(1-0.38)}{400}}} = -1.65$ ; the  $p\text{-value} = P(Z \leq -1.65) = 0.0495$ .

- 9.56 a.  $\bar{p} = \frac{20}{66} = 0.30$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.30 - 0.32}{\sqrt{\frac{0.32(1-0.32)}{66}}} = -0.35$ ; the  $p\text{-value} = 2P(Z \leq -0.35) = 2(0.3632) = 0.7264$ .
- b.  $\bar{p} = \frac{100}{264} = 0.38$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.38 - 0.32}{\sqrt{\frac{0.32(1-0.32)}{264}}} = 2.09$ ; the  $p\text{-value} = 2P(Z \geq 2.09) = 2(0.0183) = 0.0366$ .
- c.  $\bar{p} = 0.40$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.40 - 0.32}{\sqrt{\frac{0.32(1-0.32)}{40}}} = 1.08$ ; the  $p\text{-value} = 2P(Z \geq 1.08) = 2(0.1401) = 0.2802$ .
- d.  $\bar{p} = 0.38$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.38 - 0.32}{\sqrt{\frac{0.32(1-0.32)}{180}}} = 1.73$ ; the  $p\text{-value} = 2P(Z \geq 1.73) = 2(0.0418) = 0.0836$ .
- 9.58 a. With  $\alpha = 0.05$ , the critical value,  $z_\alpha = z_{0.05} = 1.645$ .
- b. With  $\alpha = 0.05$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$ . The critical values are  $-1.96$  and  $1.96$ .
- c. With  $\alpha = 0.05$ , the critical value is  $-z_\alpha = -z_{0.05} = -1.645$ .
- 9.60 a.  $\bar{p} = \frac{128}{320} = 0.40$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.40 - 0.45}{\sqrt{\frac{0.45(1-0.45)}{320}}} = -1.80$ ; With  $\alpha = 0.01$ ,  $-z_\alpha = -z_{0.01} \approx -2.33$ ; the decision rule is to reject  $H_0$  if  $z < -2.33$ . Since  $z = -1.80 > -2.33$ , we do not reject  $H_0$ .
- b.  $\bar{p} = \frac{128}{320} = 0.40$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.40 - 0.45}{\sqrt{\frac{0.45(1-0.45)}{320}}} = -1.80$ ; With  $\alpha = 0.01$ ,  $-z_{\alpha/2} = z_{0.05} = 2.576$ ; the critical values are  $-2.576$  and  $2.576$ . The decision rule is to reject  $H_0$  if  $z < -2.576$  or  $z > 2.576$ . Since  $z = 1.80$  falls between  $-2.576$  and  $2.576$ , we do not reject  $H_0$ .
- 9.62  $H_0: p \leq 0.50$ ;  $H_A: p > 0.50$
- $\bar{p} = \frac{13}{20} = 0.65$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.65 - 0.50}{\sqrt{\frac{0.50(1-0.50)}{20}}} = 1.34$ ;
- The  $p\text{-value} = P(Z \geq 1.34) = 1 - 0.9099 = 0.0901$ . Since  $0.0901 > 0.05 = \alpha$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that more than 50% of the observations in a population are below 10.
- 9.64 a.  $H_0: p \leq 0.20$ ;  $H_A: p > 0.20$
- b.  $\bar{p} = \frac{50}{190} = 0.263$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.263 - 0.20}{\sqrt{\frac{0.20(1-0.20)}{190}}} = 2.17$ ;
- The  $p\text{-value} = P(Z \geq 2.17) = P(Z \leq -2.17) = 0.015$ .
- c. Since the  $p\text{-value} = 0.015 < 0.05 = \alpha$ , we reject  $H_0$ . At the 5% significance level, the economist's concern is supported by the sample data.
- 9.66 a.  $H_0: p \leq 0.30$ ;  $H_A: p > 0.30$
- $\bar{p} = \frac{68}{200} = 0.34$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.34 - 0.30}{\sqrt{\frac{0.30(1-0.30)}{200}}} = 1.23$ .
- The  $p\text{-value} = P(Z \geq 1.23) = 1 - 0.8907 = 0.1093$ . Since the  $p\text{-value} = 0.1093 > 0.05 = \alpha$ , we do not reject  $H_0$ .
- b. Since the  $p\text{-value} = 0.1093 > 0.10 = \alpha$ , we do not reject  $H_0$ .
- c. Based on the sample data, we do not reject  $H_0$  at both the 5% and 10% significance levels. The production company's expectation of more than 30% of viewers returning to the theaters for the same movie is not supported by the sample data.
- 9.68  $H_0: p \leq 0.50$ ;  $H_A: p > 0.50$
- $\bar{p} = \frac{24}{40} = 0.60$ ;  $z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.60 - 0.50}{\sqrt{\frac{0.50(1-0.50)}{40}}} = 1.26$ .

The  $p$ -value =  $P(Z \geq 1.26) = 1 - 0.8962 = 0.1038$ . Since the  $p$ -value =  $0.1038 > 0.05 = \alpha$ , we do not reject  $H_0$ . At the 5% significance level, the politician's claims are not supported by the data.

9.70  $H_0: p \leq 0.60$ ;  $H_A: p > 0.60$

$$\bar{p} = \frac{90}{140} = 0.643; z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.64 - 0.60}{\sqrt{\frac{0.60(1-0.60)}{140}}} = 0.97.$$

The  $p$ -value =  $P(Z \geq 0.97) = P(Z \leq -0.97) = 0.1660$ . Since the  $p$ -value =  $0.1660 > 0.01 = \alpha$ ,

we do not reject  $H_0$ . The sample evidence does not support the claim that more than 60% of seniors have made serious adjustments to their lifestyle.

9.74 a.  $H_0: \mu \leq 10$ ;  $H_A: \mu > 10$

b. With  $\alpha = 0.05$ , the critical value,  $t_{\alpha, df} = t_{0.05, 17} = 1.740$ ; the decision rule is to reject  $H_0$  if  $t_{17} > 1.740$ .

$$c. df = 17; t_{17} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{10.8 - 10}{\frac{2.4}{\sqrt{18}}} = 1.41.$$

d. Since  $t_{17} = 1.41 < 1.740$ , we do not reject  $H_0$ . The claim by the weight loss clinic is not supported by the sample data.

9.78  $H_0: p \geq 0.35$ ;  $H_A: p < 0.35$

Case 1:  $n = 1000$

$$\bar{p} = 0.33; z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.33 - 0.35}{\sqrt{\frac{0.35(1-0.35)}{1000}}} = -1.33.$$

The  $p$ -value =  $P(Z \leq -1.33) = 0.0918$ .

Since the  $p$ -value =  $0.0918 > 0.05 = \alpha$ , we do not reject  $H_0$ . At the 5% significance level, the sample evidence suggests that the percentage of Americans who feel that the country is headed in the right direction is not below 35%.

Case 2:  $n = 2000$

$$\bar{p} = 0.33; z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.33 - 0.35}{\sqrt{\frac{0.35(1-0.35)}{2000}}} = -1.88.$$

The  $p$ -value =  $P(Z \leq -1.88) = 0.0301$ .

Since the  $p$ -value =  $0.0301 < 0.05 = \alpha$ , we reject  $H_0$ . At the 5% significance level, the sample evidence suggests that the percentage of Americans who feel that the country is headed in the right direction is below 35%.

9.82 a.  $H_0: p = 0.23$ ;  $H_A: p \neq 0.23$

$$b. \bar{p} = \frac{51}{200} = 0.26; z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.26 - 0.23}{\sqrt{\frac{0.23(1-0.23)}{200}}} = 1.01.$$

The  $p$ -value =  $2P(Z \geq 1.01) = 2(1 - 0.8438) = 2(0.1562) = 0.3124$ .

c. Since the  $p$ -value =  $0.3124 > 0.05 = \alpha$ , we do not reject  $H_0$ . At the 5% significance level, the sample data does support Pew Research 2010 findings. In other words, the percentage of Americans who only use cell phones does not differ from 23%.

## Chapter 10

10.2 a.  $t_{0.025, 33} = 2.035$ ;  $s_p^2 = 8.71$ ;  $6.3 \pm 2.05$ , or  $[4.25 \text{ to } 8.35]$

b.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_A: \mu_1 - \mu_2 \neq 0$

c. The interval does not contain 0; reject  $H_0$ .

10.4 a.  $t_{0.05, 20} = 1.725$ ;  $s_p^2 = 13.46$ ;

$$t_{20} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = 1.72; t_{20} = 1.72 < 1.725 = t_{0.05, 20};$$

do not reject  $H_0$ .

b.  $t_{20} = 1.72 > 1.325 = t_{0.10, 20}$ ; reject  $H_0$ .

$$10.6 \text{ a. } s_p^2 = 358.81; t_{38} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = 1.49$$

b.  $0.10 < p\text{-value} < 0.20$ ; do not reject  $H_0$ .

c.  $-2.024 < t_{38} < 2.024$ ; do not reject  $H_0$ .

10.8 a.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_A: \mu_1 - \mu_2 \neq 0$

$$b. t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -1.67$$

$$c. df = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left( \frac{s_1^2}{n_1} \right)^2 + \left( \frac{s_2^2}{n_2} \right)^2} = 8.74,$$

which is rounded down to 8;  $0.10 < p\text{-value} < 0.20$ .

d. Since  $p\text{-value} > 0.10 = \alpha$ , do not reject  $H_0$ . Cannot conclude that the population means differ.

10.10 a.  $H_0: \mu_1 - \mu_2 \geq 0$ ;  $H_A: \mu_1 - \mu_2 < 0$

$$b. z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = -5.81;$$

$p\text{-value} \approx 0$

c. Reject  $H_0$  since  $p\text{-value} < \alpha = 0.05$ ; yes

10.12 Sample 1 represents condominiums and Sample 2 represents apartment buildings.

a.  $H_0: \mu_1 - \mu_2 = 0$ ;  $H_A: \mu_1 - \mu_2 \neq 0$

$$b. z = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{244,200 - 235,800}{\sqrt{\frac{22,500^2}{30} + \frac{20,000^2}{30}}} = 1.53;$$

the  $p\text{-value} = 2 * P(Z > 1.53) = 2 * (1 - P(Z \leq 1.53)) = 2 * (1 - 0.9370) = 0.126$

c. Since the  $p$ -value is greater than 0.05 and 0.10, do not reject  $H_0$ . At either the 5% or 10% significance levels, we cannot conclude the mean profitability differs between condominiums and apartment buildings.

10.14 Sample 1 is the output rates of the new process and Sample 2 is the output rates of the old process.

a.  $H_0: \mu_1 - \mu_2 \leq 0$ ;  $H_A: \mu_1 - \mu_2 > 0$

$$b. s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(8 - 1)90.78^2 + (10 - 1)148.22^2}{8 + 10 - 2} =$$

15963.098, With  $df = 8 + 10 - 2 = 16$ ,  $t_{df} = t_{16} =$

$$\frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(2613.63 - 2485.10) - 0}{\sqrt{15963.098 \left( \frac{1}{8} + \frac{1}{10} \right)}} = 2.14.$$

c. With  $\alpha = 0.05$  and  $df = 16$ ,  $t_{\alpha, df} = t_{0.05, 16} = 1.746$ . Since  $2.14 > 1.746$ , reject  $H_0$ . We can conclude the mean output rate of the new process exceeds that of the old process.

d. With  $\alpha = 0.01$  and  $df = 16$ ,  $t_{\alpha, df} = t_{0.01, 16} = 2.583$ . Since  $2.14 < 2.583$ , do not reject  $H_0$ . We cannot conclude the mean output rate of the new process exceeds that of the old process.

10.16 Let Sample 1 be the sample of SUVs and Sample 2 be the sample of small cars.

a.  $H_0: \mu_1 - \mu_2 = 30$ ,  $H_A: \mu_1 - \mu_2 \neq 30$

$$b. s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(18 - 1)32.00^2 + (38 - 1)24.00^2}{18 + 38 - 2} = 717.04$$

With  $df = n_1 + n_2 - 2 = 18 + 38 - 2 = 54$ ,

$$t_{df} = t_{54} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(95.00 - 48.00) - 30}{\sqrt{717.04 \left( \frac{1}{18} + \frac{1}{38} \right)}} = 2.22.$$

- c. With  $\alpha = 0.10$  and  $df = 54$ ,  $t_{\alpha/2, df} = t_{0.05, 54} = 1.674$ . Since  $t_{54} = 2.22 > 1.674$ , we reject  $H_0$ .  
The sample data contradicts the claim that it takes 30 days longer to sell SUVs compared to smaller cars at the 10% significance level.
- 10.24 a. With  $\alpha = 1 - 0.90 = 0.10$  and  $df = n - 1 = 20 - 1 = 19$ ,  $t_{\alpha/2, df} = t_{0.05, 19} = 1.729$ .  $\bar{d} \pm t_{\alpha/2, df} \frac{s_D}{\sqrt{n}} = 1.3 \pm 1.729 \frac{1.61}{\sqrt{20}} = 1.3 \pm 0.62$ , or  $[0.68, 1.92]$ .
- b. Since the value zero is not included in the  $[0.68, 1.92]$  interval, we reject  $H_0$ . At the 10% significance level, we can conclude that the mean difference differs from zero.
- 10.26 a. With  $\alpha = 0.05$  and  $df = n - 1 = 12 - 1 = 11$ , the critical value is  $-t_{\alpha, df} = -t_{0.05, 11} = -1.796$ .
- b.  $t_{df} = \frac{\bar{d} - d_0}{s_D / \sqrt{n}} = \frac{-2.8 - 0}{5.7 / \sqrt{12}} = -1.702$ .
- c. Since  $t_{11} = -1.702 > -1.796 = -t_{0.05, 11}$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that the mean difference is less than 0.
- 10.28 a.  $H_0: \mu_D \leq 0$ ;  $H_A: \mu_D > 0$
- b. With  $df = n - 1 = 35 - 1 = 34$ ,  $t_{34} = \frac{\bar{d} - d_0}{s_D / \sqrt{n}} = \frac{1.2 - 0}{3.8 / \sqrt{35}} = 1.87$ .  
The  $p$ -value  $= P(T_{34} \geq 1.87)$ ;  $0.025 < p$ -value  $< 0.05$ .
- c. Since the  $p$ -value  $< \alpha = 0.05$ , we reject  $H_0$ . At the 5% significance level, we can conclude that the mean difference is greater than 0.
- d. With  $\alpha = 0.05$  and  $df = n - 1 = 35 - 1 = 34$ , the critical value is  $t_{\alpha, df} = t_{0.05, 34} = 1.691$ . Since  $t_{34} = 1.87 > 1.691 = t_{0.05, 34}$ , we reject  $H_0$ . We reach the same conclusion as in c.
- 10.30 a.  $H_0: \mu_D = 0$ ;  $H_A: \mu_D \neq 0$
- b.  $\bar{d} = \frac{\sum d_i}{n} = \frac{-13}{7} = -1.86$ ;  $s_D = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} = \sqrt{\frac{32.8571}{6}} = 2.34$   
With  $df = n - 1 = 7 - 1 = 6$ ,  $t_6 = \frac{\bar{d} - d_0}{s_D / \sqrt{n}} = \frac{-1.86 - 0}{2.34 / \sqrt{7}} = -2.10$
- c. With  $\alpha = 0.10$  and  $df = 6$ ,  $t_{\alpha/2, df} = t_{0.05, 6} = 1.943$ . The decision rule is to reject  $H_0$  if  $t_6 > 1.943$  or  $t_6 < -1.943$ .
- d. Since  $t_6 = -2.10 < -1.943$ , we reject  $H_0$ . The manager's assertion is supported by the data at the 5% significance level.
- 10.32 a.  $H_0: \mu_D = 0$ ;  $H_A: \mu_D \neq 0$
- b. With  $\alpha = 0.05$  and  $df = 5$ ,  $t_{\alpha/2, df} = t_{0.025, 5} = 2.571$ . The decision rule is to reject  $H_0$  if  $t_5 > 2.571$  or  $t_5 < -2.571$ .
- c.  $\bar{d} = \frac{\sum d_i}{n} = \frac{-13,000}{6} = -2,166.67$ ;  $s_D = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} = \sqrt{\frac{190,833,333.33}{5}} = 6177.92$   
With  $df = n - 1 = 6 - 1 = 5$ ,  $t_5 = \frac{\bar{d} - d_0}{s_D / \sqrt{n}} = \frac{-2,166.67 - 0}{6,177.92 / \sqrt{6}} = -0.86$ .
- d. Since  $t_5 = -0.86$  falls between  $-2.571$  and  $2.571$ , we do not reject  $H_0$ . We cannot conclude that the appraisers are inconsistent in their estimates since the test result shows that the mean difference is not different from 0 at the 5% significance level.
- 10.34 a. Is there difference in processing time required for new and old processors?  
 $H_0: \mu_D \geq 0$ ;  $H_A: \mu_D < 0$
- b. With  $\alpha = 0.05$  and  $df = 6$ ,  $t_{\alpha, df} = t_{0.05, 6} = 1.943$ . The decision rule is to reject  $H_0$  if  $t_6 < -1.943$ .

- c.  $\bar{d} = -0.2771$ ;  $s_D = 0.1991$ ,  $df = 6$

$$t_6 = \frac{\bar{d} - d_0}{s_D / \sqrt{n}} = \frac{-0.2771 - 0}{0.1991 / \sqrt{7}} = -3.68$$

Since  $t_{56} = -3.68 < -1.943$ , we reject  $H_0$ . We can conclude the mean difference between new processing time and the existing processor time is less than zero. Yes, there is evidence the new processor is faster than the old processor.

- 10.36 a. Are the differences in premiums?

$$H_0: \mu_D \leq 100; H_A: \mu_D > 100$$

- b. Excel Output:

| t-Test: Paired Two Sample for Means |                      |                     |
|-------------------------------------|----------------------|---------------------|
|                                     | Competitor's Premium | "Insure-Me" Premium |
| Mean                                | 985.56               | 788.5               |
| Variance                            | 84185.11             | 81995.97            |
| Observations                        | 50                   | 50                  |
| Pearson Correlation                 | -0.186               |                     |
| Hypothesized Mean Difference        | 100                  |                     |
| df                                  | 49                   |                     |
| t Stat                              | 1.545975             |                     |
| P(T ≤ t) one-tail                   | 0.064273             |                     |
| t Critical one-tail                 | 1.677                |                     |
| P(T ≤ t) two-tail                   | 0.128546             |                     |
| t Critical two-tail                 | 2.010                |                     |

Since the  $p$ -value is greater than 0.054, do not reject  $H_0$ . At the 5% significance level, we cannot conclude the mean difference between the competitor's premium and the "Insure Me" premium is more than \$100. At the 10% significance level, reject  $H_0$ . We can conclude the mean difference between the competitor's premium and the "Insure Me" premium is more than \$100.

- 10.42 a.  $\bar{p}_1 = \frac{x_1}{n_1} = \frac{50}{200} = 0.25$ ;  $\bar{p}_2 = \frac{x_2}{n_2} = \frac{70}{250} = 0.28$ ;  $\alpha = 0.05$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$

$$b. (\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}} = (0.25 - 0.28) \pm 1.96 \sqrt{\frac{0.25(1 - 0.25)}{200} + \frac{0.28(1 - 0.28)}{250}} = -0.03 \pm 0.0818 \text{ or } [-0.1118, 0.0518].$$

Since the 95% confidence interval contains the value 0, we cannot conclude that there is any difference between the population proportions at the 5% significance level.

- 10.44 a.  $\bar{p}_1 = \frac{x_1}{n_1} = \frac{100}{250} = 0.40$ ;  $\bar{p}_2 = \frac{x_2}{n_2} = \frac{172}{400} = 0.43$ ,  $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{272}{650} = 0.4185$ .

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1 - \bar{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0.40 - 0.43}{\sqrt{0.4185(1 - 0.4185) \left( \frac{1}{250} + \frac{1}{400} \right)}} = -0.75$$

- b. For  $z = -0.75$ , the  $p$ -value  $= 2P(Z \leq -0.75) = 2(0.2266) = 0.4532$

- c. Since the  $p$ -value  $= 0.4532 > 0.05 = \alpha$ , we do not reject  $H_0$ . We cannot conclude that the population proportions differ at the 5% significance level.

- d. With  $\alpha = 0.05$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$ . The decision rule is to reject  $H_0$  if  $z > 1.96$  or  $z < -1.96$ . Since  $z = -0.75$  falls between  $-1.96$  and  $1.96$ , we do not reject  $H_0$ ; we reach the same conclusion as in part c.

10.46 a.  $\bar{p}_1 = \frac{x_1}{n_1} = \frac{150}{250} = 0.60$ ;  $\bar{p}_2 = \frac{x_2}{n_2} = \frac{130}{400} = 0.325$ .

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - d_0}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}} = \frac{(0.600 - 0.325) - 0.20}{\sqrt{\frac{0.600(1-0.600)}{250} + \frac{0.325(1-0.325)}{400}}} = 1.93$$

b. The  $p$ -value =  $2P(Z \geq 1.93) = 2(0.0268) = 0.0536$ .

c. Since the  $p$ -value =  $0.0536 > 0.05 = \alpha$ , we do not reject  $H_0$ . The difference between population proportions does not differ from 0.20, at the 5% significance level.

d. With  $\alpha = 0.05$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$ . The decision rule is to reject  $H_0$  if  $z > 1.96$  or  $z < -1.96$ .

Since  $z = 1.93$  falls between  $-1.96$  and  $1.96$ , we do not reject  $H_0$ ; we reach the same conclusion as in part c.

10.50 Let  $p_1$  represent the population proportion of boys and  $p_2$  the population proportion of girls.

a.  $H_0: p_1 - p_2 \leq 0$ ;  $H_A: p_1 - p_2 > 0$ .

$$\bar{p}_1 = 0.27, n_1 = 500; \bar{p}_2 = 0.14, n_2 = 500 \quad \bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{500(0.27) + 500(0.14)}{1000} = 0.205$$

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.27 - 0.14}{\sqrt{0.205(1-0.205)\left(\frac{1}{500} + \frac{1}{500}\right)}} = -5.10$$

Since the  $p$ -value =  $P(Z \geq 5.10) \approx 0 < 0.05 = \alpha$ , we reject  $H_0$ . The proportion of boys growing out of asthma is more than the corresponding proportion of girls, at the 5% significance level.

c.  $H_0: p_1 - p_2 \leq 0.10$ ;  $H_A: p_1 - p_2 > 0.10$

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - d_0}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}} = \frac{(0.27 - 0.14) - 0.10}{\sqrt{\frac{0.27(1-0.27)}{500} + \frac{0.14(1-0.14)}{500}}} = 1.19.$$

With  $\alpha = 0.05$ , the critical value  $z_{0.05} = 1.645$ . Since  $z = 1.19 < 1.645$ , we do not reject  $H_0$ . We cannot conclude that the proportion of boys who grow out of asthma exceeds by more than 0.10 that of girls, at the 5% significance level.

10.54.a. Let  $p_1$  represent the population proportion of obese African-American men and  $p_2$  the population proportion of obese Caucasian men.

$H_0: p_1 - p_2 \geq 0$ ;  $H_A: p_1 - p_2 < 0$

$$\bar{p}_1 = \frac{x_1}{n_1} = \frac{36}{130} = 0.2769; \bar{p}_2 = \frac{x_2}{n_2} = \frac{62}{180} = 0.3444;$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{98}{310} = 0.3161$$

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.2769 - 0.3444}{\sqrt{0.3161(1-0.3161)\left(\frac{1}{130} + \frac{1}{180}\right)}} = -1.26.$$

Since the  $p$ -value =  $P(Z \leq -1.26) = 0.1038 > 0.05 = \alpha$ , we do not reject  $H_0$ . We cannot conclude that the proportion of obese African-American men is significantly less than the proportion of their Caucasian counterparts at the 5% significance level.

b. Let  $p_1$  represent the population proportion of obese African-American women and  $p_2$  the population proportion of obese Caucasian women.

$H_0: p_1 - p_2 \leq 0$ ;  $H_A: p_1 - p_2 > 0$

$$\bar{p}_1 = \frac{x_1}{n_1} = \frac{35}{90} = 0.3889; \bar{p}_2 = \frac{x_2}{n_2} = \frac{31}{120} = 0.2583; \bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$= \frac{66}{210} = 0.3143$$

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.3889 - 0.2583}{\sqrt{0.3143(1-0.3143)\left(\frac{1}{90} + \frac{1}{120}\right)}} = 2.02.$$

With  $\alpha = 0.05$ , the critical value  $z_{0.05} = 1.645$ . Since  $z = 2.02 > 1.645$ , we reject  $H_0$ . The proportion of obese African-

American women is greater than the proportion of their Caucasian counterparts at the 5% significance level.

c. Let  $p_1$  represent the population proportion of obese African-Americans and  $p_2$  the population proportion of obese Caucasians.

$H_0: p_1 - p_2 = 0$ ;  $H_A: p_1 - p_2 \neq 0$

$$\bar{p}_1 = \frac{x_1}{n_1} = \frac{71}{220} = 0.3227; \bar{p}_2 = \frac{x_2}{n_2} = \frac{93}{300} = 0.3100; \bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{164}{520} = 0.3154$$

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.3227 - 0.3100}{\sqrt{0.3154(1-0.3154)\left(\frac{1}{220} + \frac{1}{300}\right)}} = 0.31.$$

With  $\alpha = 0.05$ ,  $z_{\alpha/2} = z_{0.025} = 1.96$ . The decision rule is to reject  $H_0$  if  $z > 1.96$  or  $z < -1.96$ .

Since  $z = 0.31$  falls between  $-1.96$  and  $1.96$ , we do not reject  $H_0$ . The proportion of obese African-American adults is not significantly different from the proportion of their Caucasian counterparts at the 5% significance level.

10.60 Let men refer to population 1 and women refer to population 2.

a.  $H_0: \mu_1 - \mu_2 \leq 0$ ;  $H_A: \mu_1 - \mu_2 > 0$

$$b. z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{43.87 - 29.54}{\sqrt{\frac{32^2}{100} + \frac{25^2}{100}}} = 3.53$$

c. The  $p$ -value =  $P(Z \geq 3.53) = P(Z \leq -3.53) = 0.0002$ .

d. Since the  $p$ -value =  $0.0002 < 0.01 = \alpha$ , we reject  $H_0$ . We conclude that on average men spend more money than women on St. Patrick's Day, at the 1% significance level.

10.64 a.  $H_0: \mu_D = 0$ ;  $H_A: \mu_D \neq 0$

$$b. \bar{d} = \frac{\sum d_i}{n} = \frac{-4}{6} = -0.67; s_D = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{13.3334}{5}} = 1.63$$

$$\text{With } df = n - 1 = 6 - 1 = 5, t_5 = \frac{\bar{d} - d_0}{s_D/\sqrt{n}} = \frac{-0.67 - 0}{1.63/\sqrt{6}} = -1.01.$$

c. With  $\alpha = 0.05$  and  $df = 5$ ,  $t_{\alpha/2, df} = t_{0.025, 5} = 2.571$ .

d. Since  $t_5 = -1.01$  falls between  $-2.571$  and  $2.571$ , we do not reject  $H_0$ . At the 5% significance level, the crop yield with the new fertilizer is not significantly different from the crop yield with the old fertilizer. There is no need for the farmer to be concerned.

10.68 Let  $p_1$  represent the proportion on time flights at JFK and  $p_2$  the proportion of on time flights at O'Hare.

a.  $H_0: p_1 - p_2 \leq 0.05$ ;  $H_A: p_1 - p_2 > 0.05$

b.  $\bar{p}_1 = 0.70$ ,  $n_1 = 200$ ;  $\bar{p}_2 = 0.63$ ,  $n_2 = 200$

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - d_0}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}} = \frac{(0.70 - 0.63) - 0.05}{\sqrt{\frac{0.70(1-0.70)}{200} + \frac{0.63(1-0.63)}{200}}} = 0.42$$

The  $p$ -value =  $P(Z \geq 0.42) = 1 - 0.6628 = 0.3372$ .

c. Since the  $p$ -value =  $0.3372 > 0.05 = \alpha$ , we do not reject  $H_0$ . The proportion of on-time flights at JFK is not more than 5 percentage points higher than that of O'Hare, at the 5% significance level.

d. With  $\alpha = 0.05$ , the critical value  $z_{0.05} = 1.645$ . Since  $z = 0.42 < 1.645 = z_{0.05}$ , we do not reject  $H_0$ .

We reach the same conclusion as in c.

## Chapter 11

11.2 a. 39.997

b. 37.556



- c. 7.434  
d. 8.260
- 11.4 a.  $\left[ \frac{(25-1)14.44}{36.415}, \frac{(25-1)14.44}{13.848} \right] = [9.52, 25.03]$   
b.  $\left[ \frac{(25-1)14.44}{45.558}, \frac{(25-1)14.44}{9.886} \right] = [7.61, 35.06]$   
c. The width of the interval increases with the confidence level.
- 11.6 a.  $\chi^2_{20} = \frac{(21-1)75}{50} = 30 > 28.412 = \chi^2_{0.10,20}$ ; reject  $H_0$   
b.  $\chi^2_{0.05,20} = 31.410$ ,  $\chi^2_{0.95,20} = 10.851$ ;  $10.851 < \chi^2_{20} = 30 < 31.410$ ; do not reject  $H_0$
- 11.8 a.  $H_0: \sigma^2 \leq 2$ ;  $H_A: \sigma^2 > 2$ ;  $s^2 = 2.89$ ;  
 $\chi^2_9 = \frac{(10-1)2.89}{2} = 13$ ; using Excel  $p$ -value =  
 $0.16 > 0.10 = \alpha$ ; do not reject  $H_0$ . Cannot conclude that variance is greater than 2.  
b.  $\chi^2_9 = 13 < 14.684 = \chi^2_{0.10,9}$ ; do not reject  $H_0$
- 11.10 a.  $\left[ \frac{(20-1)0.03}{32.852}, \frac{(20-1)0.03}{8.907} \right] = [0.02, 0.06]$   
b. 0.05 is included in interval; cannot conclude that specification is being violated.
- 11.12 a.  $s^2 = 1295.48$ ,  $s = 35.99$   
b.  $\left[ \frac{(5-1)1295.48}{11.143}, \frac{(5-1)1295.48}{0.484} \right]$   
 $= [465.04, 10,706.45]$ ; the interval for the population standard deviation is [21.56, 103.47].
- 11.14 a.  $H_0: \sigma^2 \geq 90,000$ ;  $H_A: \sigma^2 < 90,000$   
b. The critical value with  $\alpha = 0.01$  is  $\chi^2_{0.99,6} = 0.872$ .  
c. Given  $s^2 = 30,696.81$  and  $df = n - 1 = 7 - 1 = 6$ , the value of the test statistic,  $\chi^2_6 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(7-1)30,696.81}{90,000} = 2.05$ .  
d. Since  $\chi^2_6 = 2.05 > 0.872 = \chi^2_{0.99,6}$ , we do not reject  $H_0$  at the 1% significance level. The sample data do not support the restaurant's owner hope that the standard deviation is less than 300 (or variance is less than 90,000) at the 1% significance level.  
e. The critical value with  $\alpha = 0.10$  is  $\chi^2_{0.90,6} = 2.204$ . Since  $\chi^2_6 = 2.05 < 2.204 = \chi^2_{0.90,6}$ , we reject  $H_0$  at the 10% significance level. The sample data do suggest that the standard deviation is less than 300 (or variance is less than 90,000) at the 10% significance level.
- 11.20 a.  $H_0: \sigma^2 \leq 10,000$  (thousands)<sup>2</sup>;  $H_A: \sigma^2 > 10,000$  (thousands)<sup>2</sup>  
b. Using Excel we find  $s^2 = 10,527.97$ . Given  $df = n - 1 = 36 - 1 = 35$ , the value of the test statistic,  
 $\chi^2_{35} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(36-1)10,527.97}{10,000} = 36.85$ .  
c. Using Excel's function **CHISQ.DIST.RT** (36.85,35), the  $p$ -value = 0.383.  
d. Since the  $p$ -value = 0.383  $>$  0.05 =  $\alpha$ , we do not reject  $H_0$ . Therefore, the realtor's claim is not supported by the data at the 5% significance level.
- 11.22 a. Using Excel we calculate the standard deviation of rentals as 176.11 for Ann-Arbor, Michigan and 297.64 for Davis, California.  
b. Given  $df = n - 1 = 10 - 1 = 9$ ,  $\alpha = 0.05$ ,  $\frac{\alpha}{2} = 0.025$ ,  $1 - \frac{\alpha}{2} = 0.975$ , we find  $\chi^2_{0.025,9} = 19.023$  and  $\chi^2_{0.975,9} = 2.700$ . For Ann-Arbor, Michigan, the confidence interval for the population variance is

$$\left[ \frac{(n-1)s^2}{\chi^2_{0.025,9}}, \frac{(n-1)s^2}{\chi^2_{0.975,9}} \right] = \left[ \frac{(10-1)176.11^2}{19.023}, \frac{(10-1)176.11^2}{2.700} \right] = [14,673, 103,382].$$

and for the population standard deviation is

$$\left[ \sqrt{\frac{(n-1)s^2}{\chi^2_{0.025,9}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{0.975,9}}} \right] = [121,322].$$

For Davis, California, the confidence interval for the population variance is

$$\left[ \frac{(n-1)s^2}{\chi^2_{0.025,9}}, \frac{(n-1)s^2}{\chi^2_{0.975,9}} \right] = \left[ \frac{(10-1)297.64^2}{19.023}, \frac{(10-1)297.64^2}{2.700} \right] =$$

[41,913, 295,299] and for the population standard deviation is

$$\left[ \sqrt{\frac{(n-1)s^2}{\chi^2_{0.025,9}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{0.975,9}}} \right] = [205,543].$$

- c. With  $\alpha = 0.05$ , we can infer that the standard deviation of rent differs from 200 only for Davis, California since only for Davis, the confidence interval does not include the value 200.
- 11.24 a.  $P(F_{(10,8)} \geq 3.35) = 0.05$ .  
b.  $P(F_{(10,8)} < 0.42) = P(F_{(8,10)} \geq \frac{1}{0.42}) = P(F_{(8,10)} \geq 2.38) = 0.10$ .  
c.  $P(F_{(10,8)} \geq 4.30) = 0.025$   
d.  $P(F_{(10,8)} < 0.26) = P(F_{(8,10)} \geq \frac{1}{0.26}) = P(F_{(8,10)} \geq 3.85) = 0.025$ .
- 11.26 a. We use  $\alpha = 0.05$ ,  $df_1 = n_1 - 1 = 20 - 1 = 19$ ,  $df_2 = n_2 - 1 = 15 - 1 = 14$  to derive the 95% confidence interval as  
 $\left[ \frac{s_1^2}{s_2^2} \frac{1}{F_{\alpha/2, df_1, df_2}}, \frac{s_1^2}{s_2^2} F_{\alpha/2, df_1, df_2} \right] = \left[ \left( \frac{220}{196} \right), \frac{1}{2.86} \left( \frac{220}{196} \right) 2.65 \right] =$   
[0.39, 2.97]  
b. The 95% confidence interval for the ratio of the population variances  $\sigma_1^2/\sigma_2^2$  ranges from 0.39 to 2.97, and contains the value 1. Thus, we cannot conclude that the population variances differ at the 5% significance level.
- 11.28 Given  $df_1 = n_1 - 1 = 14 - 1 = 13$ ,  $df_2 = n_2 - 1 = 11 - 1 = 10$ , the value of the test statistic,  $F_{(df_1, df_2)} = F_{(13,10)} = \frac{s_1^2}{s_2^2} = \frac{935}{812} = 1.15$ . The critical value with  $\alpha = 0.05$ ,  $F_{\alpha(df_1, df_2)} = F_{0.05, (13,10)}$ , lies between 2.85 and 2.98. Since  $F_{(13,10)} = 1.15 < 2.85$ , we do not reject  $H_0$  at the 5% significance level. We cannot conclude that the variance in population 1 is more than the variance in population 2. The above test is based on the assumption that the two samples are drawn independently from normally distributed populations.
- 11.30 a. We specify the hypotheses as ( $H_0: \sigma_2^2/\sigma_1^2 = 1$ ,  $H_A: \sigma_2^2/\sigma_1^2 \neq 1$  since  $s_2^2 > s_1^2$ ).  
b. Given  $df_2 = df_1 = n_2 - 1 = n_1 - 1 = 15 - 1 = 14$ , the value of the test statistic,  $F_{(df_2, df_1)} = F_{(14,14)} = \frac{s_2^2}{s_1^2} = \frac{0.48}{0.35} = 1.37$ .  
c. The critical value with  $\alpha = 0.05$ ,  $F_{\alpha/2, (df_2, df_1)} = F_{(0.025, (14,14))}$ , is approximately 2.86. The decision rule is to reject  $H_0$  if  $F_{(14,14)} > 2.86$ .  
d. Since  $F_{(14,14)} = 1.37 < 2.86$ , we do not reject  $H_0$  at the 5% significance level. We cannot conclude that the population variances differ at the 5% significance level. Therefore, the company can go ahead and adopt the new cost-cutting technology.

| Summary statistics |            |           |
|--------------------|------------|-----------|
|                    | Electronic | Utilities |
| $s^2$              | 428.5      | 42.2      |
| $n$                | 5          | 5         |

We specify the hypotheses as  $H_0: \sigma_1^2/\sigma_2^2 = 1$ ,  $H_A: \sigma_1^2/\sigma_2^2 \neq 1$ . Given  $df_1 = df_2 = n_1 - 1 = n_2 - 1 = 5 - 1 = 4$ , the

value of the test statistic,  $F_{(df_1, df_2)} = F_{(4,4)} = \frac{s_1^2}{s_2^2} = \frac{428.5}{42.2} =$

10.15. The critical value with  $\alpha = 0.05$  is  $F_{0.025, (4,4)} = 9.60$ .

Since  $F_{(4,4)} = 10.15 > F_{0.025, (4,4)} = 9.60$ , we reject  $H_0$ . The population variances do differ at the 5% significance level. The above test is based on the assumption that the two samples are drawn independently from normally distributed populations.

11.36 a.  $H_0: \sigma_1^2/\sigma_2^2 = 1$ ,  $H_A: \sigma_1^2/\sigma_2^2 \neq 1$

b. It is necessary to assume that the sample data on monthly closing stock prices are drawn independently from normally distributed populations.

c. Using Excel's function **FTEST**, the  $p$ -value = 0.1045.

d. Since the  $p$ -value = 0.1045  $>$  0.05 =  $\alpha$ , we do not reject  $H_0$ . The sample data do not suggest that the variances of price differ for the two firms at the 5% significance level.

11.38 We specify the hypotheses as  $H_0: \sigma_2^2/\sigma_1^2 \leq 1$ ,  $H_A: \sigma_2^2/\sigma_1^2 > 1$  since  $s_2^2 > s_1^2$ . ('Davis, California' is population 1 and 'Ann Arbor, Michigan' is population 2).

In order to implement the test, we assume that the sample variances are computed from independently drawn samples from the normally distributed populations.

We use Excel to calculate the standard deviations as 176.11 for Ann-Arbor, Michigan and 297.64 for Davis, California.

Given  $df_1 = df_2 = n_1 - 1 = n_2 - 1 = 10 - 1 = 9$ , the value of the test statistic,  $F_{(df_2, df_1)} = F_{(9,9)} = \frac{s_2^2}{s_1^2} = \frac{297.64^2}{176.11^2} = 2.86$ . We use Excel to calculate the  $p$ -value as 0.0667. At  $\alpha = 0.05$ , we do not reject  $H_0$  and thus we cannot conclude that the variability of rents in Davis, California is significantly higher than that of Ann Arbor, Michigan.

11.40 a.  $H_0: \sigma^2 \leq 5$ ;  $H_A: \sigma^2 > 5$ .

We use the sample data to compute  $s^2 = 11.57$ .

Given  $df = n - 1 = 7 - 1 = 6$ , the value of the test statistic,

$$\chi_6^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(7-1)11.57}{5} = 13.88. \text{ The } p\text{-value} = P(\chi_6^2 \geq$$

13.88) lies between 0.025 and 0.05. Since the  $p$ -value  $>$   $\alpha = 0.01$ , we do not reject  $H_0$ . The sample data do not suggest that the variance exceeds 5 at the 1% significance level. There is no cause for concern for the advocacy group.

b. The above analysis is valid only under the assumption that the generic drug prices are normally distributed.

11.44 a.  $H_0: \sigma^2 \leq 1,225 (\%)^2$ ,  $H_A: \sigma^2 > 1,225 (\%)^2$

b. For a valid statistical inference, it is necessary to assume that the population of Fidelity's Select Automotive Fund is normally distributed.

c. We calculate the sample variance as  $s^2 = 1057.67$ . Given

$df = n - 1 = 27 - 1 = 826$  the value of the test statistic,

$$\chi_{26}^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(27-1)1057.67}{1,225} = 22.45.$$

d. Using Excel's function **CHIDIST** (22.45,26) the  $p$ -value = 0.6638.

e. Since the  $p$ -value = 0.6638  $>$  0.05 =  $\alpha$ , we do not reject  $H_0$ . The sample data do not suggest that the variance is greater than  $1,225 (\%)^2$  (the standard deviation is greater than 35%) at the 5% significance level.

11.46 a.  $H_0: \sigma_1^2/\sigma_2^2 = 1$ ,  $H_A: \sigma_1^2/\sigma_2^2 \neq 1$

( $\sigma_1^2$  is variance for Hasbro,  $\sigma_2^2$  is the variance for Mattel)

b. It is necessary to assume that the growth rates are normally distributed.

c. Given  $df_1 = df_2 = n_1 - 1 = 5 - 1 = 4$ , and  $\alpha = 0.05$ , the critical values are computed as  $F_{(0.025, (4,4))} = 9.60$  and  $\frac{1}{F_{0.025, (4,4)}} = \frac{1}{9.60} = 0.10$ . The decision rule is to reject  $H_0$  if the value of the test statistic is greater than 9.60 or less than 0.10.

d. The value of the test statistic,  $F_{(df_2, df_1)} = F_{(4,4)} = \frac{s_2^2}{s_1^2} = \frac{74.16}{43.00} = 1.72$ . Since  $F_{(df_2, df_1)} = 1.72$  falls between 0.10 and 9.60, we do not reject  $H_0$ . The variance in the growth rates of the two firms is not different at the 5% significance level.

11.50 a.  $H_0: \sigma_1^2/\sigma_2^2 \leq 1$ ,  $H_A: \sigma_1^2/\sigma_2^2 > 1$

( $\sigma_1^2$  is variance for Asia,  $\sigma_2^2$  is the variance for Latin America)

b. Given  $df_1 = df_2 = n_1 - 1 = n_2 - 1 = 5 - 1 = 4$ , the value of the test statistic,  $F_{(df_1, df_2)} = F_{(4,4)} = \frac{s_1^2}{s_2^2} = \frac{219,354.8}{78,780.2} = 2.78$

c. Using Excel function **FTEST**( ), the  $p$ -value = 0.173.

d. Since the  $p$ -value = 0.173  $>$  0.05 =  $\alpha$ , we do not reject  $H_0$ . The sample data do not support the claim that the variance in revenues is higher in Asia than in Latin America, at the 5% significance level.

## Chapter 12

12.2 a.  $H_A$ : At least one of the  $p_i$  ( $i = 1, 2, 3$ ) differs from its hypothesized value.

b.  $\chi^2 = 1.043$ ;  $p$ -value  $>$  0.10.

c.  $p$ -value  $>$   $\alpha = 0.05$ ; do not reject  $H_0$ . Cannot conclude that some proportions differ from hypothesized values.

12.4  $\chi^2 = 2.271$ ;  $p$ -value  $>$   $\alpha = 0.01$ ; do not reject  $H_0$ .

12.6 a.  $H_0: p_1 = 0.37, p_2 = 0.17, p_3 = 0.28, p_4 = 0.18$

$H_A$ : At least one of the  $p_i$  ( $i = 1, 2, 3, 4$ ) differs from its hypothesized value.

b.  $\chi_{0.05,3}^2 = 7.815$

c.  $\chi_3^2 = 11.16$

d. Reject  $H_0$  since  $\chi_3^2 = 11.16 > 7.815$ . Can conclude that the proportions from the initial study have changed.

12.8 a.  $H_0: p_1 = p_2 = p_3 = 1/3$ ;  $H_A$ : At least one of the  $p_i$  ( $i = 1, 2, 3$ ) differs from  $1/3$ .

b.  $\chi^2 = 1.02 < 5.991 = \chi_{0.05,2}^2$ ; do not reject  $H_0$ . No.

c.  $p$ -value  $>$   $\alpha = 0.05$ ; do not reject  $H_0$ .

12.10  $\chi^2 = 2.18$ ;  $p$ -value  $>$   $\alpha = 0.05$ ; do not reject  $H_0$ . No.

12.14 a. Given = 0.025 and  $df = (r-1)(c-1) = (5-1)(2-1) = 4$ ,

$$\chi_{(\alpha, df)}^2 = \chi_{0.025,4}^2 = 11.143.$$

b. Given = 0.01 and  $df = (r-1)(c-1) = (3-1)(5-1) = 8$ ,

$$\chi_{(\alpha, df)}^2 = \chi_{0.01,8}^2 = 20.090.$$



12.16  $H_0$ : The two categories are independent

$H_A$ : The two categories are dependent

| $O_{ij}$ | $e_{ij}$ | $O_{ij} - e_{ij}$ | $(O_{ij} - e_{ij})^2$ | $(O_{ij} - e_{ij})^2/e_{ij}$ |
|----------|----------|-------------------|-----------------------|------------------------------|
| 120      | 115.65   | 4.35              | 18.933                | 0.164                        |
| 112      | 108.36   | 3.64              | 13.240                | 0.122                        |
| 100      | 104.56   | -4.56             | 20.786                | 0.199                        |
| 110      | 113.43   | -3.43             | 11.771                | 0.104                        |
| 127      | 127.16   | -0.16             | 0.026                 | 0.000                        |
| 115      | 119.15   | -4.15             | 17.209                | 0.144                        |
| 120      | 114.97   | 5.03              | 25.324                | 0.220                        |
| 124      | 124.72   | -0.72             | 0.522                 | 0.004                        |
| 118      | 122.19   | -4.19             | 17.556                | 0.144                        |
| 115      | 114.49   | 0.51              | 0.260                 | 0.002                        |
| 110      | 110.47   | -0.47             | 0.224                 | 0.002                        |
| 124      | 119.85   | 4.15              | 17.251                | 0.144                        |
| Total    |          |                   |                       | 1.25                         |

With  $df = (r - 1)(c - 1) = (3 - 1)(4 - 1) = 6$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_6 = 1.25$ .

a.  $p$ -value approach:

The  $p$ -value =  $P(\chi^2_6 > 1.25)$  is more than 0.95.

Since the  $p$ -value  $> \alpha = 0.01$ , we do not reject  $H_0$ . At the 1% significance level, we cannot conclude that Category 1 and Category 2 are dependent.

b. Critical value approach

The critical value with  $\alpha = 0.01$  and  $df = 6$  is  $\chi^2_{\alpha, df} = \chi^2_{0.01, 6} = 16.812$ . We reject  $H_0$  if  $\chi^2_6 > 16.812$ . Since  $\chi^2_6 = 1.25 < 16.812$ , we do not reject  $H_0$ , which is the same conclusion as in a.

12.18 a.  $H_0$ : Color preference is independent of gender

$H_A$ : Color preference is dependent on gender

b.

| $O_{ij}$ | $e_{ij}$ | $O_{ij} - e_{ij}$ | $(O_{ij} - e_{ij})^2$ | $(O_{ij} - e_{ij})^2/e_{ij}$ |
|----------|----------|-------------------|-----------------------|------------------------------|
| 470      | 465.84   | 4.16              | 17.318                | 0.037                        |
| 280      | 284.16   | -4.16             | 17.318                | 0.061                        |
| 535      | 509.32   | 25.68             | 659.628               | 1.295                        |
| 285      | 310.68   | -25.68            | 659.628               | 2.123                        |
| 495      | 524.84   | -29.84            | 890.707               | 1.697                        |
| 350      | 320.16   | 29.84             | 890.707               | 2.782                        |
| Total    |          |                   |                       | 8.00                         |

The critical value with  $\alpha = 0.01$  and  $df = 2$  is  $\chi^2_{\alpha, df} = \chi^2_{0.01, 2} = 9.210$ .

c. With  $df = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_2 = 8.00$ .

d. Since  $\chi^2_2 = 8.00 < 9.210$ , we do not reject  $H_0$ . At the 1% significance level, we cannot conclude that color preference is dependent on gender. Thus, there is no need for gender-targeted advertisement.

12.22 a.  $H_0$ : Optimism in China and age are independent

$H_A$ : Optimism in China and age are dependent

b.

| $O_{ij}$ | $e_{ij}$ | $O_{ij} - e_{ij}$ | $(O_{ij} - e_{ij})^2$ | $(O_{ij} - e_{ij})^2/e_{ij}$ |
|----------|----------|-------------------|-----------------------|------------------------------|
| 23       | 30.23    | -7.23             | 52.201                | 1.727                        |
| 50       | 43.23    | 6.78              | 45.901                | 1.062                        |
| 18       | 17.55    | 0.45              | 0.202                 | 0.012                        |
| 51       | 34.88    | 16.13             | 260.016               | 7.456                        |
| 38       | 49.88    | -11.88            | 141.016               | 2.827                        |
| 16       | 20.25    | -4.25             | 18.063                | 0.892                        |
| 19       | 27.90    | -8.90             | 79.210                | 2.839                        |
| 45       | 39.90    | 5.10              | 26.010                | 0.652                        |
| 20       | 16.20    | 3.80              | 14.440                | 0.891                        |
| Total    |          |                   |                       | 18.36                        |

With  $df = (r - 1)(c - 1) = (3 - 1)(3 - 1) = 4$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_4 = 18.36$ . Using Excel's function, **CHISQ.DIST.RT** (18.36, 4), the  $p$ -value = 0.001.

c. Since the  $p$ -value is less than  $\alpha$  ( $0.001 < 0.01$ ), we reject  $H_0$ .

d. At the 1% significance level, we can conclude that optimism among Chinese is dependent on age.

12.24 Critical value approach.

$H_0$ : Breakup reasons and gender are independent

$H_A$ : Breakup reasons and gender are dependent

| $O_{ij}$ | $e_{ij}$ | $O_{ij} - e_{ij}$ | $(O_{ij} - e_{ij})^2$ | $(O_{ij} - e_{ij})^2/e_{ij}$ |
|----------|----------|-------------------|-----------------------|------------------------------|
| 54       | 61.20    | -7.20             | 51.840                | 0.847                        |
| 48       | 40.80    | 7.20              | 51.840                | 1.271                        |
| 378      | 342.00   | 36.00             | 1296.000              | 3.789                        |
| 192      | 228.00   | -36.00            | 1296.000              | 5.684                        |
| 324      | 352.80   | -28.80            | 829.440               | 2.351                        |
| 264      | 235.20   | 28.80             | 829.440               | 3.527                        |
| 504      | 489.60   | 14.40             | 207.360               | 0.424                        |
| 312      | 326.40   | -14.40            | 207.360               | 0.635                        |
| 540      | 554.40   | -14.40            | 207.360               | 0.374                        |
| 384      | 369.60   | 14.40             | 207.360               | 0.561                        |
| Total    |          |                   |                       | 19.46                        |

With  $df = (r - 1)(c - 1) = (5 - 1)(2 - 1) = 4$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_4 = 19.46$ . The critical value with  $\alpha = 0.01$  and  $df = 4$  is  $\chi^2_{\alpha, df} = \chi^2_{0.01, 4} = 13.277$ . We reject  $H_0$  since  $\chi^2_4 = 19.46 > 13.277$ . At the 1% significance level, we conclude that breakup reason is dependent on gender.

12.26  $p$ -value approach.

$H_0$ : The data are normally distributed with a mean of -3.5 and a standard deviation of 9.7

$H_A$ : The data are not normally distributed with a mean of -3.5 and a standard deviation of 9.7

$P(X < -10) = P(Z < -0.67) = 0.2514$

$P(-10 \leq X < 0) = P(-0.67 \leq Z < 0.36) = 0.3892$

$P(0 \leq X < 10) = P(0.36 \leq Z < 1.39) = 0.2771$

$P(X \geq 10) = P(Z \geq 1.39) = 0.0823$

| Class         | Observed ( $O_i$ ) | $p_i$ if normal | Expected ( $e_i$ ) | $(O_i - e_i)^2/e_i$ |
|---------------|--------------------|-----------------|--------------------|---------------------|
| Less than -10 | 70                 | 0.2514          | 50.28              | 7.7343              |
| -10 up to 0   | 40                 | 0.3892          | 77.84              | 18.3950             |
| 0 up to 10    | 80                 | 0.2771          | 55.42              | 10.9018             |
| 10 or more    | 10                 | 0.0823          | 16.46              | 2.5353              |
| Total         | 200                |                 |                    | 39.57               |

With  $k = 4$ ,  $df = k - 3 = 1$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_1 = 39.57$ . The  $p$ -value  $= P(\chi^2_1 > 39.57)$  is less than 0.001 (almost zero). We reject  $H_0$  since the  $p$ -value  $< \alpha = 0.01$ . At the 1% significance level, we conclude that the data are not normally distributed with a mean of  $-3.5$  and a standard deviation  $9.7$ .

- 12.28 a.  $H_0$ : The final grades are normally distributed with a mean of 72 and a standard deviation of 10.  
 $H_A$ : The final grades are not normally distributed with a mean of 72 and a standard deviation of 10.
- b. The critical value with  $\alpha = 0.05$  and  $df = 2$  is  $\chi^2_{\alpha, df} = \chi^2_{0.05, 2} = 5.991$ .
- c.
- $P(X < 50) = P(Z < -2.20) = 0.0139$   
 $P(50 \leq X < 70) = P(-2.20 \leq Z < -0.20) = 0.4068$   
 $P(70 \leq X < 80) = P(-0.20 \leq Z < 0.80) = 0.3674$   
 $P(80 \leq X < 90) = P(0.80 \leq Z < 1.80) = 0.1760$   
 $P(X \geq 90) = P(Z \geq 1.80) = 0.0359$

| Class        | Observed<br>( $o_i$ ) | $p_i$ if<br>normal | Expected<br>( $e_i$ ) | $(o_i - e_i)^2/e_i$ |
|--------------|-----------------------|--------------------|-----------------------|---------------------|
| Less than 50 | 5                     | 0.0139             | 4.17                  | 0.1652              |
| 50 up to 70  | 135                   | 0.4068             | 122.04                | 1.3763              |
| 70 up to 80  | 105                   | 0.3674             | 110.22                | 0.2472              |
| 80 up to 90  | 45                    | 0.1760             | 52.80                 | 1.1523              |
| 90 or above  | 10                    | 0.0359             | 10.77                 | 0.0551              |
| Total        | $n = 300$             |                    |                       | 3.00                |

With  $k = 5$ ,  $df = k - 3 = 2$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_2 = 3.00$ .

- d. We do not reject  $H_0$  since  $\chi^2_2 = 3.00 < 5.991$ . At the 5% significance level, we cannot reject the professor's claim that the final grades are distributed using the normal distribution.

#### 12.32 $p$ -value approach.

- a. Goodness-of-fit test for normality:

$H_0$ : CEO compensation is normally distributed with a mean of \$19.03 million and a standard deviation of \$27.61 million.

$H_A$ : CEO compensation is not normally distributed with a mean of \$19.03 million and a standard deviation of \$27.61 million.

$$P(X < 5) = P(Z < -0.51) = 0.3050$$

$$P(5 \leq X < 10) = P(-0.51 \leq Z < -0.33) = 0.0657$$

$$P(10 \leq X < 15) = P(-0.33 \leq Z < -0.15) = 0.0697$$

$$P(15 \leq X < 20) = P(-0.15 \leq Z < 0.04) = 0.0756$$

$$P(X \geq 20) = P(Z \geq 0.04) = 0.4840$$

| Class       | Observed<br>( $o_i$ ) | $p_i$ if<br>normal | Expected<br>( $e_i$ ) | $(o_i - e_i)^2/e_i$ |
|-------------|-----------------------|--------------------|-----------------------|---------------------|
| Less than 5 | 43                    | 0.3050             | 72.59                 | 12.0618             |
| 5 up to 10  | 65                    | 0.0657             | 15.64                 | 155.8360            |
| 10 up to 15 | 32                    | 0.0697             | 16.59                 | 14.3177             |
| 15 up to 20 | 38                    | 0.0756             | 17.99                 | 22.2471             |
| 20 or more  | 60                    | 0.4840             | 115.19                | 26.4442             |
| Total       | 238                   |                    |                       | 230.91              |

With  $k = 5$ ,  $df = k - 3 = 2$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_2 = 230.91$ . The  $p$ -value  $= P(\chi^2_2 > 230.91)$  is less than 0.001 (almost zero). We reject  $H_0$  since  $p$ -value  $< \alpha = 0.01$ . At the 1% significance level, we conclude that CEO compensation is not normally distributed with a mean of \$19.03 million and a standard deviation of \$27.61 million.

- b. Jarque-Bera test for normality:

$$H_0: S = 0 \text{ and } K = 0$$

$$H_A: S \neq 0 \text{ or } K \neq 0$$

$$JB = \chi^2_2 = \left(\frac{n}{6}\right) \left[S^2 + \frac{K^2}{4}\right] = (238/6)[5.26^2 + 35.53^2/4] =$$

13616.09. We reject  $H_0$  since the  $p$ -value  $= P(\chi^2_2 > 13616.09)$  is almost zero. We reach the same conclusion as in a.

- c. Both tests show that at the 5% significance level, total compensation for CEOs is not normally distributed.

- 12.34 a. Jarque-Bera test for normality:

$$H_0: S = 0 \text{ and } K = 0$$

$$H_A: S \neq 0 \text{ or } K \neq 0$$

- b. Using Excel functions **SKEW** and **KURT**,  $S = -0.11$  and  $K = -0.95$ .

The value of the test statistic,  $JB = \chi^2_2 = \left(\frac{n}{6}\right) \left[S^2 + \frac{K^2}{4}\right] = (26/6)[(-0.11)^2 + (-0.95)^2/4] = 1.03$ . Using Excel the  $p$ -value  $= P(\chi^2_2 > 1.03) = 0.60$ .

- c. Since the  $p$ -value  $> \alpha$  ( $0.60 > 0.05$ ), we do not reject  $H_0$ . At the 5% significance level, we cannot reject the hypothesis that Home Depot's stock prices are normally distributed.

- 12.36 a.  $H_0: p_1 = 0.40, p_2 = 0.30, p_3 = 0.20, p_4 = 0.10$

$H_A$ : At least one of the  $p_i$  ( $i = 1, 2, 3, 4$ ) differs from its hypothesized value.

- b. The critical value with  $\alpha = 0.01$  is  $\chi^2_{\alpha, df} = \chi^2_{0.01, 3} = 11.345$ .
- c.

| Firm  | Frequency             |                       |             |                 |                     |
|-------|-----------------------|-----------------------|-------------|-----------------|---------------------|
|       | Observed<br>( $o_i$ ) | Expected<br>( $e_i$ ) | $o_i - e_i$ | $(o_i - e_i)^2$ | $(o_i - e_i)^2/e_i$ |
| 1     | 200                   | 220.00                | -20.00      | 400.00          | 1.818               |
| 2     | 180                   | 165.00                | 15.00       | 225.00          | 1.364               |
| 3     | 100                   | 110.00                | -10.00      | 100.00          | 0.909               |
| 4     | 70                    | 55.00                 | 15.00       | 225.00          | 4.091               |
| Total | 550                   | 550.00                |             |                 | 8.182               |

With  $k = 4$ ,  $df = k - 1 = 3$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_3 = 8.182$ .

- d. Since  $\chi^2_3 = 8.182 < 11.345$ , we do not reject  $H_0$ . At the 1% significance level we cannot conclude that at least one  $p_i$  is different from its hypothesized value. In other words, the market shares in 2011 have not changed from what they were in 2010.

- 12.38 a.  $H_0: p_A = 0.60, p_B = 0.30, p_C = 0.10$

$H_A$ : At least one of the  $p_i$  ( $i = A, B, C$ ) differs from its hypothesized value.

- b.

| Candidate | Frequency             |                       |             |                 |                     |
|-----------|-----------------------|-----------------------|-------------|-----------------|---------------------|
|           | Observed<br>( $o_i$ ) | Expected<br>( $e_i$ ) | $o_i - e_i$ | $(o_i - e_i)^2$ | $(o_i - e_i)^2/e_i$ |
| A         | 350                   | 300.00                | 50.00       | 2500.00         | 8.333               |
| B         | 125                   | 150.00                | -25.00      | 625.00          | 4.167               |
| C         | 25                    | 50.00                 | -25.00      | 625.00          | 12.500              |
| Total     | 500                   | 500.00                |             |                 | 25.000              |

The value of the test statistic,  $\chi^2_{df} = \chi^2_2 = 25.0$ . The  $p$ -value  $= P(\chi^2_2 > 25.0)$  is almost zero. Since the  $p$ -value  $< \alpha = 0.01$ , we reject  $H_0$ . At the 1% significance level we conclude that at least one  $p_i$  is different from its hypothesized value. This suggests that, contrary to the TV station claim, voter preference has changed.

- 12.40 a.  $H_0$ : Surviving for discharge is independent of the time of the cardiac arrest  
 $H_A$ : Surviving for discharge is dependent of the time of cardiac arrest  
b. The critical value with  $\alpha = 0.01$  is  $\chi^2_{\alpha, df} = \chi^2_{0.01, 1} = 6.635$ .  
c.

| $o_{ij}$ | $e_{ij}$ | $o_{ij} - e_{ij}$ | $(o_{ij} - e_{ij})^2$ | $(o_{ij} - e_{ij})^2/e_{ij}$ |
|----------|----------|-------------------|-----------------------|------------------------------|
| 11,604   | 10633.44 | 970.56            | 941987.925            | 88.587                       |
| 46,989   | 47959.56 | -970.56           | 941987.925            | 19.641                       |
| 4,139    | 5109.56  | -970.56           | 941987.925            | 184.358                      |
| 24,016   | 23045.44 | 970.56            | 941987.925            | 40.875                       |
| Total    |          |                   |                       | 333.46                       |

With  $df = (r-1)(c-1) = (2-1)(2-1) = 1$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_1 = 333.46$ .

- d. Since  $\chi^2_1 = 333.46 > 6.635$ , we reject  $H_0$ . At the 1% significance level, the sample data suggests that a patient surviving a cardiac arrest is dependent on the time that it happens. Given that patients do not have control over the timing of cardiac arrest, hospitals need to put in place resources that ensure that patients have equal chances of surviving a cardiac arrest, regardless of when it happens.
- 12.42  $H_0$ : Effect on ADHD is independent of the treatment  
 $H_A$ : Effect on ADHD is dependent on the treatment

| $o_{ij}$ | $e_{ij}$ | $o_{ij} - e_{ij}$ | $(o_{ij} - e_{ij})^2$ | $(o_{ij} - e_{ij})^2/e_{ij}$ |
|----------|----------|-------------------|-----------------------|------------------------------|
| 12       | 13       | -1                | 1                     | 0.077                        |
| 15       | 14       | 1                 | 1                     | 0.071                        |
| 14       | 13       | 1                 | 1                     | 0.077                        |
| 13       | 14       | -1                | 1                     | 0.071                        |
| Total    |          |                   |                       | 0.297                        |

Critical value approach.

With  $df = (r-1)(c-1) = (2-1)(2-1) = 1$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_1 = 0.297$ . The critical value with  $\alpha = 0.05$  and  $df = 1$  is  $\chi^2_{\alpha, df} = \chi^2_{0.05, 1} = 3.841$ . Since  $\chi^2_1 = 0.297 < 3.841$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that the effect of ADHD depends on the treatment by St. John's wort.

- 12.44  $H_0$ : A household's delinquency in payment is independent of the type of heating  
 $H_A$ : A household's delinquency in payment is dependent of the type of heating

| Delinquent in Payment? | Type of Heating (Observed $o_{ij}$ ) |             |             |         |       |
|------------------------|--------------------------------------|-------------|-------------|---------|-------|
|                        | Natural Gas                          | Electricity | Heating Oil | Propane | Total |
| Yes                    | 50                                   | 20          | 15          | 10      | 95    |
| No                     | 240                                  | 130         | 20          | 15      | 405   |
| Total                  | 290                                  | 150         | 35          | 25      | 500   |

| Delinquent in Payment? | Type of Heating (Expected $e_{ij}$ ) |             |             |         |       |
|------------------------|--------------------------------------|-------------|-------------|---------|-------|
|                        | Natural Gas                          | Electricity | Heating Oil | Propane | Total |
| Yes                    | 55.10                                | 28.50       | 6.65        | 4.75    | 95    |
| No                     | 234.90                               | 121.50      | 28.35       | 20.25   | 405   |
| Total                  | 290                                  | 150         | 35          | 25      | 500   |

| $o_{ij}$ | $e_{ij}$ | $o_{ij} - e_{ij}$ | $(o_{ij} - e_{ij})^2$ | $(o_{ij} - e_{ij})^2/e_{ij}$ |
|----------|----------|-------------------|-----------------------|------------------------------|
| 50       | 55.10    | -5.10             | 26.010                | 0.472                        |
| 20       | 28.50    | -8.50             | 72.250                | 2.535                        |
| 15       | 6.65     | 8.35              | 69.723                | 10.485                       |
| 10       | 4.75     | 5.25              | 27.563                | 5.803                        |
| 240      | 234.90   | 5.10              | 26.010                | 0.111                        |
| 130      | 121.50   | 8.50              | 72.250                | 0.595                        |
| 20       | 28.35    | -8.35             | 69.723                | 2.459                        |
| 15       | 20.25    | -5.25             | 27.563                | 1.361                        |
| Total    |          |                   |                       | 23.820                       |

Critical value approach.

With  $df = (r-1)(c-1) = (2-1)(4-1) = 3$ , the value of the test statistic,  $\chi^2_{df} = \chi^2_3 = 23.82$ . The critical value with  $\alpha = 0.05$  and  $df = 3$  is  $\chi^2_{\alpha, df} = \chi^2_{0.05, 3} = 7.815$ . Since  $\chi^2_3 = 23.82 > 7.815$ , we reject  $H_0$ . At the 5% significance level, the sample data suggest that the type of heating that a household uses is dependent on whether or not the household is delinquent on its bill payment.

- 12.46  $p$ -value approach.

$H_0$ :  $S = 0$  and  $K = 0$

$H_A$ :  $S \neq 0$  or  $K \neq 0$

- a. Jarque-Bera test for house prices:

The value of the test statistic,  $JB = \chi^2_2 = \left(\frac{n}{6}\right) \left[S^2 + \frac{K^2}{4}\right] = (36/6) [(0.68)^2 + (-0.06)^2/4] = 2.78$ . We use Excel to find the  $p$ -value =  $P(\chi^2_2 > 2.78) = 0.2491$ . Since the  $p$ -value = 0.2491  $> \alpha = 0.05$ , we do not reject  $H_0$ . At the 5% significance level, the sample data do not suggest that the house prices in Arlington, MA are not normally distributed.

- b. Jarque-Bera test for square footage:

The value of the test statistic,  $JB = \chi^2_2 = \left(\frac{n}{6}\right) \left[S^2 + \frac{K^2}{4}\right] = (36/6) [(0.45)^2 + (-0.56)^2/4] = 1.69$ . We use Excel to find the  $p$ -value =  $P(\chi^2_2 > 1.69) = 0.4305$ . We do not reject  $H_0$  since the  $p$ -value = 0.4305  $> \alpha = 0.05$ . At the 5% significance level, the sample data do not suggest that square footage in Arlington, MA is not normally distributed.

## Chapter 13

- 13.2 a.  $\bar{x} = \frac{-174}{15} = -11.6$

b.  $SSTR = \sum_{i=1}^c n_i(\bar{x}_i - \bar{x})^2 = 16.78$ ;  $MSTR = \frac{16.78}{4-1} = 5.59$

c.  $SSE = \sum_{i=1}^c (n_i - 1)s_i^2 = 60.56$ ;  $MSE = \frac{60.56}{15-4} = 5.51$

- d.  $H_0: \mu_A = \mu_B = \mu_C = \mu_D$ ;  $H_A$ : Not all population means are equal

e.  $F_{(3,11)} = \frac{MSTR}{MSE} = \frac{5.59}{5.51} = 1.01$

- f.  $p$ -value  $> 0.10$

- g.  $p$ -value  $> \alpha = 0.10$ ; do not reject  $H_0$ ; No significant differences in population means.

- 13.4

| Source of Variation | SS    | df | MS   | F    | F crit at 5% |
|---------------------|-------|----|------|------|--------------|
| Between Groups      | 11.34 | 3  | 3.78 | 3.58 | 2.77         |
| Within Groups       | 59.13 | 56 | 1.06 |      |              |
| Total               | 70.47 | 59 |      |      |              |

$F_{(3,56)} = 3.58 > 2.77 = F_{0.05,(3,56)}$ ; reject  $H_0$ . Some population means differ.

13.6 a.

| Source of Variation | SS       | df | MS     | F    | p-value | F crit at 10% |
|---------------------|----------|----|--------|------|---------|---------------|
| Between Groups      | 548.37   | 5  | 109.67 | 1.37 | 0.250   | 1.96          |
| Within Groups       | 4,321.11 | 54 | 80.02  |      |         |               |
| Total               | 4,869.48 | 59 |        |      |         |               |

- b.  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ ;  $H_A$ : Not all populations means are equal  
c.  $p\text{-value} = 0.250 > 0.10 = \alpha$ , do not reject  $H_0$ . No significant differences in population means.

13.8 a.

| Source of Variation | SS     | df | MS    | F    | p-value | F crit at 5% |
|---------------------|--------|----|-------|------|---------|--------------|
| Between Groups      | 174.72 | 2  | 87.36 | 2.81 | 0.083   | 3.47         |
| Within Groups       | 652.40 | 21 | 31.07 |      |         |              |
| Total               | 827.12 | 23 |       |      |         |              |

- b.  $H_0: \mu_1 = \mu_2 = \mu_3$ ;  $H_A$ : Not all populations means are equal  
c.  $p\text{-value} = 0.083 > 0.05 = \alpha$ ; do not reject  $H_0$ . No significant differences in the average whitening effectiveness of the detergents.

13.10 a.  $H_0: \mu_{\text{Snork}} = \mu_{\text{Sail}} = \mu_{\text{NBoard/Windsurf}} = \mu_{\text{Bowl}} = \mu_{\text{On-road tri}} = \mu_{\text{Off-road tri}}$   
 $H_A$ : Not all population mean incomes are equal.

b.

| Source of Variation | SS       | df | MS     | F     | p-value | F crit at 5% |
|---------------------|----------|----|--------|-------|---------|--------------|
| Between Groups      | 4,895.15 | 5  | 979.03 | 37.85 | 0.000   | 2.62         |
| Within Groups       | 620.80   | 24 | 25.87  |       |         |              |
| Total               | 5,515.95 | 29 |        |       |         |              |

- c.  $F_{0.05,(5,24)} = 2.62$   
d.  $F_{(5,24)} = 37.85 > 2.62 = F_{0.05,(5,24)}$ ; reject  $H_0$ . Some mean incomes differ.

13.12 a.

| Source of Variation | SS            | df | MS           | F     | P-value |
|---------------------|---------------|----|--------------|-------|---------|
| Between Groups      | 7,531,769.00  | 3  | 2,510,589.67 | 69.01 | 0.000   |
| Within Groups       | 3,492,385.00  | 96 | 36,379.01    |       |         |
| Total               | 11,024,154.00 | 99 |              |       |         |

Since the  $p\text{-value} \approx 0 < 0.01 = \alpha$ , we reject  $H_0$ . At the 1% significance level, we can conclude that the average annual energy bills vary by region.

13.14 a. The completed ANOVA table is below.

| Source of Variation | SS     | df  | MS      | F       | p-value   |
|---------------------|--------|-----|---------|---------|-----------|
| Between Groups      | 0.5293 | 2   | 0.2647  | 10.5912 | 0.0000592 |
| Within Groups       | 2.9238 | 117 | 0.02499 |         |           |
| Total               | 3.4531 | 119 |         |         |           |

Note:  $p\text{-value}$  is obtained from Excel using **FDIST.RT()**

- b.  $H_0: \mu_{\text{Low}} = \mu_{\text{Medium}} = \mu_{\text{High}}$   
 $H_A$ : Not all population means are equal.  
c. Since the  $p\text{-value}$  is less than 0.01 and 0.05, reject  $H_0$ . At the both the 1% and 5% significance levels, we can conclude that the mean fill volumes are not the same for the three pressures.

13.18 ANOVA: Single Factor

SUMMARY

| Groups  | Count | Sum  | Average  | Variance |
|---------|-------|------|----------|----------|
| Field 1 | 30    | 2438 | 81.26667 | 35.16782 |
| Field 2 | 30    | 2249 | 74.96667 | 36.44713 |
| Field 3 | 30    | 2400 | 80       | 15.44828 |

ANOVA

| Source of Variation | SS       | df | MS       | F       | P-value  | F crit @ 10% |
|---------------------|----------|----|----------|---------|----------|--------------|
| Between Groups      | 666.2889 | 2  | 333.1444 | 11.4794 | 3.76E-05 | 2.364616     |
| Within Groups       | 2524.833 | 87 | 29.02107 |         |          |              |
| Total               | 3191.122 | 89 |          |         |          |              |

$H_0: \mu_1 = \mu_2 = \mu_3$   
 $H_A$ : Not all population means are equal.  
At  $\alpha = 0.10$ ,  $F_{\alpha,(df_1, df_2)} = F_{0.10,(2,87)} = 2.365$ . Since  $F_{(2,87)} = 11.479 > 2.365$ , we reject  $H_0$ . At the 10% significance level, we can conclude that the average job satisfaction differs by field.

13.20 ANOVA: Single Factor

SUMMARY

| Groups      | Count | Sum   | Average | Variance |
|-------------|-------|-------|---------|----------|
| August 31   | 10    | 37576 | 3757.6  | 995333.6 |
| November 30 | 10    | 33669 | 3366.9  | 802284.1 |
| February 28 | 10    | 34301 | 3430.1  | 949813.9 |
| May 31      | 10    | 38624 | 3862.4  | 913229.6 |

ANOVA

| Source of Variation | SS       | df | MS        | F     | P-value | F crit @ 5% |
|---------------------|----------|----|-----------|-------|---------|-------------|
| Between Groups      | 1768209  | 3  | 589402.97 | 0.644 | 0.5918  | 2.866       |
| Within Groups       | 32945951 | 36 | 915165.29 |       |         |             |
| Total               | 34714160 | 39 |           |       |         |             |

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4;$$

$H_A$ : Not all population means are equal.

At  $\alpha = 0.05$ ,  $F_{\alpha(df_1, df_2)} = F_{0.05(3, 36)} = 2.866$ . Since  $F_{(3, 36)} = 0.644 < 2.866$ , we do reject  $H_0$ . At the 5% significance level, we cannot conclude that the average revenue for the four quarters are not equal.

#### 13.24 a. Fisher LSD approach

| Population Mean Differences | Confidence Interval: $(\bar{x}_i - \bar{x}_j) \pm t_{\frac{\alpha}{2}, n_T - c} \sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}$ |
|-----------------------------|--|
| $\mu_1 - \mu_2$             | $(56 - 66) \pm 2.020 \sqrt{(7.79)(\frac{1}{18} + \frac{1}{12})} = [-12.10, -7.90]^*$                                       |
| $\mu_1 - \mu_3$             | $(56 - 63) \pm 2.020 \sqrt{(7.79)(\frac{1}{18} + \frac{1}{14})} = [-9.01, -4.99]^*$  |
| $\mu_2 - \mu_3$             | $(66 - 63) \pm 2.020 \sqrt{(7.79)(\frac{1}{12} + \frac{1}{14})} = [0.78, 5.22]^*$  |

At  $\alpha = 0.05$ ,  $\alpha/2 = 0.025$ , and  $n_T - c = 44 - 3 = 41$ , so  $t_{\frac{\alpha}{2}, n_T - c} = t_{0.025, 41} = 2.020$

The asterisk \* shows that the confidence interval does not include the value zero, thus indicating the corresponding means are statistically different at the 5% significance level.

#### b. Tukey's HSD approach

| Population Mean Differences | Confidence Interval: $(\bar{x}_i - \bar{x}_j) \pm q_{\alpha(c, n_T - c)} \sqrt{MSE/2(\frac{1}{n_i} + \frac{1}{n_j})}$ |
|-----------------------------|---|
| $\mu_1 - \mu_2$             | $(56 - 66) \pm \frac{3.44}{\sqrt{2}} \sqrt{7.79(\frac{1}{18} + \frac{1}{12})} = [-12.53, -7.47]^*$                    |
| $\mu_1 - \mu_3$             | $(56 - 63) \pm \frac{3.44}{\sqrt{2}} \sqrt{7.79(\frac{1}{18} + \frac{1}{12})} = [-9.42, -4.58]^*$                     |
| $\mu_2 - \mu_3$             | $(66 - 63) \pm \frac{3.44}{\sqrt{2}} \sqrt{7.79(\frac{1}{18} + \frac{1}{12})} = [0.33, 5.67]^*$                       |

At  $\alpha = 0.05$ ,  $c = 3$ ,  $n_T - c = 44 - 3 = 41$ , so  $q_{\alpha(c, n_T - c)} = q_{0.05(3, 41)} = 3.44$

All the confidence intervals do not contain the value zero, thus indicating the corresponding means are different at the 5% significance level.

c. The Tukey's HSD approach is preferred over Fisher's because it reduces the probability of Type I error (the likelihood of incorrectly rejecting the null hypothesis).

d. The data appear to support that the 'lean' technic improve efficiency; however, more research should be conducted since significant differences exist even in stores where no change in technique occurred.

$$13.30 \text{ a. } SST = \sum_{i=1}^c \sum_{j=1}^r (x_{ij} - \bar{x})^2 = (5 - 1.33)^2 + (15 - 1.33)^2 + \dots + (-8 - 1.33)^2 = 894.67$$

$$SSA = r \sum_{i=1}^c (\bar{x}_i - \bar{x})^2 = 4[(1.00 - 1.33)^2 + (6 - 1.33)^2 + (0.50 - 2.50)^2] = 8.67$$

$$SSB = c \sum_{j=1}^r (\bar{x}_j - \bar{x})^2 = 3[(10.67 - 1.33)^2 + \dots + (-3.67 - 1.33)^2] = 702$$

$$SSE = SST - (SSA + SSB) = 894.67 - (8.67 + 107.00) = 184.0$$

b.

$$MSA = \frac{SSA}{c-1} = \frac{8.67}{3-1} = 4.33$$

$$MSB = \frac{SSB}{r-1} = \frac{702.0}{4-1} = 234.0$$

$$MSE = \frac{SSE}{n_T - c - r - 1} = \frac{184.0}{12 - 3 - 4 + 1} = 30.67$$

c.

| Source of Variation | SS     | df | MS     | F    | F crit @ 1% |
|---------------------|--------|----|--------|------|-------------|
| Rows                | 702.00 | 3  | 234.00 | 7.63 | 9.78        |
| Columns             | 8.67   | 2  | 4.33   | 0.14 | 10.92       |
| Error               | 184.00 | 6  | 30.67  |      |             |
| Total               | 894.67 | 11 |        |      |             |

d. At  $\alpha = 0.01$ ,  $F_{\alpha(df_1, df_2)} = F_{0.01(2, 6)} = 10.92$  and  $F_{(2, 6)} = 0.14 <$

10.92; thus, we do not reject  $H_0$ . At the 1% significance level, we cannot conclude that the levels of the column means significantly differ.

At  $\alpha = 0.01$ ,  $F_{\alpha(df_1, df_2)} = F_{0.01(3, 6)} = 9.78$  and  $F_{(3, 6)} = 7.63 < 9.78$ ; thus, we do not reject  $H_0$ . At the 1% significance level, we cannot conclude that the levels of Factor B significantly differ (we cannot conclude that blocking is necessary).

#### 13.32 a.

| Source of Variation | SS      | df | MS     | F    | F crit @ 1% |
|---------------------|---------|----|--------|------|-------------|
| Rows                | 532.3   | 2  | 266.15 | 4.26 | 10.92       |
| Columns             | 723.9   | 3  | 241.30 | 3.87 | 9.78        |
| Error               | 374.5   | 6  | 62.42  |      |             |
| Total               | 1,630.7 | 11 |        |      |             |

$$SSA = SST - (SSB + SSE) = 1,630.7 - (532.3 + 374.5) = 723.90$$

b. At  $\alpha = 0.01$ ,  $F_{\alpha(df_1, df_2)} = F_{0.01(3, 6)} = 9.78$  and  $F_{(3, 6)} = 3.87 < 9.78$ ; thus, we do not reject  $H_0$ . At the 1% significance level, we cannot conclude that the factor A (column) means significantly differ.

c. At  $\alpha = 0.01$ ,  $F_{\alpha(df_1, df_2)} = F_{0.01(2, 6)} = 10.92$  and  $F_{(2, 6)} = 4.26 < 10.92$ ; thus, we do not reject  $H_0$ . At the 1% significance level, we cannot conclude that the factor B (row) means significantly differ.

#### 13.34 a.

| Source of Variation | SS    | df | MS    | F    | P-value | F crit @ 5% |
|---------------------|-------|----|-------|------|---------|-------------|
| Rows                | 1057  | 5  | 211.4 | 6.41 | 0.0064  | 3.326       |
| Columns             | 7     | 2  | 3.5   | 0.11 | 0.9004  | 4.103       |
| Error               | 330   | 10 | 33    |      |         |             |
| Total               | 1,394 | 17 |       |      |         |             |

Note:  $p$ -value is obtained from Excel using **F.DIST.RT()**

b. Since the  $p$ -value = 0.9005  $> 0.05 = \alpha$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that the column means significantly differ.

c. Since the  $p$ -value = 0.0064  $< 0.05 = \alpha$ , we reject  $H_0$ . At the 5% significance level, we can conclude that the row means significantly differ.

#### 13.36 a.

| Source of Variation | SS     | df | MS   | F    | P-value | F crit @ 5% |
|---------------------|--------|----|------|------|---------|-------------|
| Rows                | 93.2   | 4  | 23.3 | 2.19 | 0.1316  | 3.26        |
| Columns             | 52.15  | 3  | 17.4 | 1.63 | 0.2335  | 3.49        |
| Error               | 127.6  | 12 | 10.6 |      |         |             |
| Total               | 272.95 | 19 |      |      |         |             |

Note:  $p$ -value is obtained from Excel using **F.DIST.RT()**



- b. Since the  $p$ -value = 0.2335 > 0.05 =  $\alpha$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that average scores differ from round to round.
- c. Since the  $p$ -value = 0.1316 > 0.05 =  $\alpha$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that the average scores by the five players differ.

13.42 a.

ANOVA: Two-Factor With Replication

| Source of Variation | SS    | df | MS   | F     | P-value  | F crit at 5% |
|---------------------|-------|----|------|-------|----------|--------------|
| Sample (Rows)       | 1,000 | 2  | 500  | 85.71 | 1.44E-16 | 3.19         |
| Columns             | 1,200 | 3  | 400  | 68.57 | 2.25E-17 | 2.80         |
| Interaction         | 20    | 6  | 3.33 | 0.57  | 0.7510   | 2.29         |
| Within (Error)      | 280   | 48 | 5.83 |       |          |              |
| Total               | 2,500 | 59 |      |       |          |              |

Note: The  $p$ -values are obtained from Excel using **F.DIST.RT()**; the critical values are obtained from Excel using **F.INV.RT()**.

- b. Since the  $p$ -value (interaction) = 0.751 > 0.05 =  $\alpha$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that there is interaction between factors A and B.
- c. Since the  $p$ -value (columns)  $\approx 0 < 0.05 = \alpha$ , we reject  $H_0$ . At the 5% significance level, average values differ for factor A.
- d. Since the  $p$ -value (rows)  $\approx 0 < 0.05 = \alpha$ , we reject  $H_0$ . At the 5% significance level, average values differ for factor B.
- 13.44 a. Since the  $p$ -value (interaction) = 0.4629 > 0.01 =  $\alpha$ , we do not reject  $H_0$ . At the 1% significance level, we cannot conclude that there is interaction between the two factors.
- b. Since the interaction between the two factors is not significant, we can conduct tests based on the main effects.
- Since the  $p$ -value (columns) =  $1.13 \times 10^{-6} \approx 0.00 < 0.01 = \alpha$ , we reject  $H_0$ . At the 1% significance level, the column means differ.
  - Since the  $p$ -value (rows) = 0.0031 < 0.01 =  $\alpha$ , we reject  $H_0$ . At the 1% significance level, the row means differ.

13.50 a. Relevant Excel Output

ANOVA

| Source of Variation | SS      | df | MS       | F      | P-value | F crit @ 5% |
|---------------------|---------|----|----------|--------|---------|-------------|
| Sample              | 383.63  | 2  | 191.815  | 1.844  | 0.187   | 3.555       |
| Columns             | 2591.19 | 2  | 1295.593 | 12.453 | 0.000   | 3.555       |
| Interaction         | 622.59  | 4  | 155.648  | 1.496  | 0.245   | 2.928       |
| Within              | 1872.67 | 18 | 104.037  |        |         |             |
| Total               | 5470.07 | 26 |          |        |         |             |

- b. Since the  $p$ -value (interaction) = 0.245 > 0.05 =  $\alpha$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that there is interaction between industry and work experience.
- c. Since the  $p$ -value (columns)  $\approx 0 < 0.05 = \alpha$ , we reject  $H_0$ . At the 5% significance level, we can conclude that job satisfaction differs by industry.
- d. Since the  $p$ -value (rows) = 0.187 > 0.05 =  $\alpha$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude job satisfaction differs depending on work experience.

13.58 a.

| Source of Variation | SS       | df | MS       | F    | P-value | F crit @ 5% |
|---------------------|----------|----|----------|------|---------|-------------|
| Between Groups      | 258.8192 | 2  | 129.4096 | 9.40 | 0.0035  | 3.89        |
| Within Groups       | 165.2234 | 12 | 13.76861 |      |         |             |
| Total               | 424.0426 | 14 |          |      |         |             |

b.  $H_0: \mu_A = \mu_B = \mu_C$

$H_A$ : Not all population means are equal.

Since the  $p$ -value = 0.0035 < 0.05 =  $\alpha$ , we reject  $H_0$ . At the 5% significance level, we can conclude that average P/E ratios of these three industries differ.

c. Tukey's HSD approach

| Population Mean Difference | Confidence Interval: $(\bar{x}_i - \bar{x}_j) \pm q_{\alpha, (c, n_{T-c})} \sqrt{\frac{MSE}{n}}$ |
|----------------------------|--|
| $\mu_A - \mu_B$            | $(10.48 - 13.15) \pm 3.77 \sqrt{\frac{13.77}{5}} = [-8.93, 3.59]$                                |
| $\mu_A - \mu_C$            | $(10.48 - 20.32) \pm 3.77 \sqrt{\frac{13.77}{5}} = [-16.10, -3.58] *$                            |
| $\mu_B - \mu_C$            | $(13.15 - 20.32) \pm 3.77 \sqrt{\frac{13.77}{5}} = [-13.43, -0.91] *$                            |

At  $\alpha = 0.05$ ,  $c = 3$ ,  $n_{T-c} = 15 - 3 = 12$ , so  $q_{\alpha, (c, n_{T-c})} = q_{0.05, (3, 12)} = 3.77$ . Intervals not containing the value zero are indicated by the asterisk\*. At the 5% significance level, average P/E ratios for A and C, and B and C are differ.

13.68 a. Relevant Excel Output

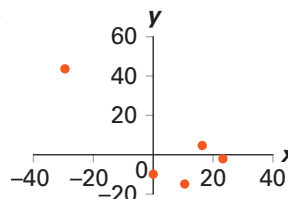
ANOVA

| Source of Variation | SS      | df | MS      | F     | P-value | F crit @ 5% |
|---------------------|---------|----|---------|-------|---------|-------------|
| Sample              | 98.296  | 2  | 49.148  | 2.035 | 0.160   | 3.555       |
| Columns             | 289.852 | 2  | 144.926 | 6.002 | 0.010   | 3.555       |
| Interaction         | 153.037 | 4  | 38.259  | 1.584 | 0.221   | 2.928       |
| Within              | 434.667 | 18 | 24.148  |       |         |             |
| Total               | 975.852 | 26 |         |       |         |             |

- b. Since the  $p$ -value (interaction) = 0.05 =  $\alpha$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that there is interaction between fuel type and hybrid type.
- c. Since the  $p$ -value (columns) = 0.010 < 0.05 =  $\alpha$ , we reject  $H_0$ . At the 5% significance level, we can conclude that average fuel consumption differs by fuel type.
- d. Since the  $p$ -value (rows) = 0.160 > 0.05 =  $\alpha$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude average fuel consumption differs by type of hybrid.

## Chapter 14

14.2 a.



x and y appear to have a negative relationship.



b.  $s_{xy} = \frac{-1519.6}{5-1} = -379.9$ ;  
negative linear relationship

c.  $r_{xy} = \frac{-379.9}{(20.69)(23.42)} = -0.78$ ;  
strong negative linear relationship

14.4 a.  $t_{28} = \frac{-0.60\sqrt{30-2}}{\sqrt{1-(-0.60)^2}} = -3.97$

b.  $p\text{-value} < 0.005$

c.  $p\text{-value} < \alpha = 0.05$ ; reject  $H_0$ ; variables significantly, negatively correlated

14.6 a.  $r_{xy} = \frac{-1.75}{(2)(5)} = -0.175$ ; weak negative linear relationship

b.  $H_0: \rho_{xy} = 0$ ;  $H_A: \rho_{xy} \neq 0$

c.  $t_{23} = \frac{-0.175\sqrt{25-2}}{\sqrt{1-(-0.175)^2}} = -0.85$ ;  $-2.069 < t_{23} < 2.069$ ; do not reject  $H_0$ ; not significantly correlated.

14.8 a.  $r_{mc} = 0.57$ ;  $r_{mb} = 0.63$ ;  $r_{mg} = 0.82$ ;  $r_{cb} = 0.04$ ;  $r_{cg} = 0.63$ ;  $r_{bg} = 0.46$

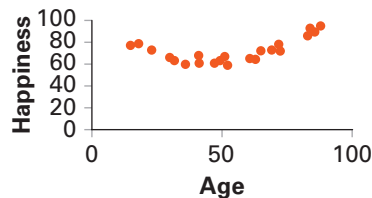
b. Coca-Cola

c. Coca-Cola and Bank of America ( $r_{cb} = 0.04$ )

14.10 a.  $r_{xy} = \frac{146.30}{(22.68)(11.29)} = 0.57$ ; positive relationship

b.  $H_0: \rho_{xy} = 0$ ;  $H_A: \rho_{xy} \neq 0$ ;  $t_{22} = \frac{0.57\sqrt{24-2}}{\sqrt{1-(0.57)^2}} = 3.25$ ;  
 $p\text{-value} < 0.01 = \alpha$ , reject  $H_0$ ; variables significantly correlated.

c. Relationship is nonlinear.



14.12 a.  $b_1 = \frac{1250}{925} = 1.35$

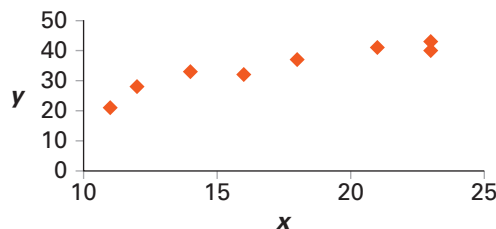
b.  $b_0 = 44 - (1.35)(34) = -1.90$

c.  $\hat{y} = -1.90 + 1.35x$ ;  
 $\hat{y} = -1.90 + (1.35)(40) = 52.10$

14.14 a.  $\hat{y} = 15 + (2.5)(10) = 40$

b. If  $x$  doubles,  $x = 2(10) = 20$ , and  $\hat{y} = 15 + 2.5(20) = 65$ . Thus, when  $x$  doubles from 10 to 20,  $\hat{y}$  increases by 25 ( $65 - 40$ ).

14.16 a.



There appears to be a linear relationship between  $x$  and  $y$ ; a linear regression model is appropriate.

b.

| $(x - \bar{x})$                   | $(y - \bar{y})$                    | $(x - \bar{x})(y - \bar{y})$                   | $(x - \bar{x})^2$                  |
|-----------------------------------|------------------------------------|--|------------------------------------|
| 12 - 17.25<br>= -5.25             | 28 - 34.375<br>= -6.375            | (-5.25)(-6.375)<br>= 33.47                     | -5.25 <sup>2</sup><br>= 27.56      |
| ⋮                                 | ⋮                                  | ⋮  | ⋮                                  |
| 16 - 17.25<br>= -1.25             | 32 - 34.375<br>= -2.375            | (-1.25)(-2.375)<br>= 2.97                      | -1.25 <sup>2</sup><br>= 1.56       |
| $\bar{x} = 17.25$<br>$s_x = 4.77$ | $\bar{y} = 34.375$<br>$s_y = 7.41$ | $\Sigma(x - \bar{x})(y - \bar{y}) =$<br>233.25 | $\Sigma(x - \bar{x})^2 =$<br>159.5 |

$b_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{233.25}{159.5} = 1.46$

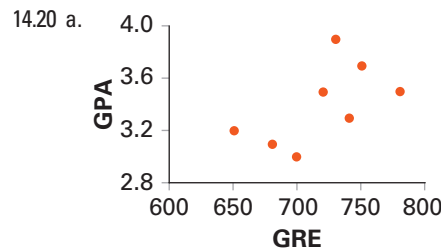
$b_0 = \bar{y} - b_1 \bar{x} = 34.375 - 1.46(17.25) = 9.19$

The sample regression equation is  $\hat{y} = 9.19 + 1.46x$

c. If  $x = 10$ ,  $\hat{y} = 9.19 + 1.46(10) = 23.79$

If  $x = 15$ ,  $\hat{y} = 9.19 + 1.46(15) = 31.09$

If  $x = 20$ ,  $\hat{y} = 9.19 + 1.46(20) = 38.39$



b.  $b_1 = \frac{52}{11887.5} = 0.0044$ ;

$b_0 = 3.4 - (0.0044)(718.75) = 0.24$ ;

$\widehat{\text{GPA}} = 0.24 + 0.0044\text{GRE}$

c.  $\widehat{\text{GPA}} = 3.36$

14.24 a. Choose Data>Data Analysis>Regression from the menu. Select the Consumption data as your Input  $y$  range, and Disposable Income as your Input  $x$  range. Click OK.

| Regression Statistics |         |
|-----------------------|---------|
| Multiple R            | 0.9949  |
| R Square              | 0.9899  |
| Adj. R Square         | 0.9894  |
| Std. Error            | 742.859 |
| Observations          | 22      |

| ANOVA      |    |            |            |          |                |
|------------|----|------------|------------|----------|----------------|
|            | df | SS         | MS         | F        | Significance F |
| Regression | 1  | 1.08E + 09 | 1.08E + 09 | 1965.385 | 1.9E - 21      |
| Residual   | 20 | 11036796   | 551839.8   |          |                |
| Total      | 21 | 1.1E + 09  |            |          |                |

|                   | Coefficients | Standard Error | t Stat  | p-value  | Lower 95% | Upper 95% |
|-------------------|--------------|----------------|---------|----------|-----------|-----------|
| Intercept         | 8550.675     | 593.92         | 14.3969 | 5.12E-12 | 7311.77   | 9789.58   |
| Disposable Income | 0.6860       | 0.0155         | 44.3327 | 1.9E-21  | 0.6538    | 0.7183    |

- b.  $\widehat{\text{Consumption}} = 8550.675 + 0.686 (\text{Disposable Income})$   
 c. As disposable income increases by \$1, consumption is predicted to increase by \$0.686. In other words, the marginal propensity to consume is \$0.686 per \$1 of disposable income.  
 d.  $\widehat{\text{Consumption}} = 8550.675 + 0.686(57,000) = 47,652.675$ .

14.26 a. Excel Output:

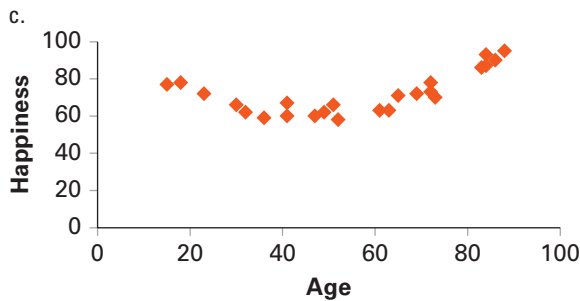
| Regression Statistics |        |
|-----------------------|--------|
| Multiple R            | 0.5716 |
| R Square              | 0.3267 |
| Adj. R Square         | 0.2961 |
| Std. Error            | 9.4696 |
| Observations          | 24     |

| ANOVA      |    |         |        |       |                |
|------------|----|---------|--------|-------|----------------|
|            | df | SS      | MS     | F     | Significance F |
| Regression | 1  | 957.19  | 957.19 | 10.67 | 0.0035         |
| Residual   | 22 | 1972.81 | 89.67  |       |                |
| Total      | 23 | 2930    |        |       |                |

|           | Coefficients | Standard Error | t Stat  | P-value  | Lower 95% | Upper 95% |
|-----------|--------------|----------------|---------|----------|-----------|-----------|
| Intercept | 56.1772      | 5.2145         | 10.7732 | 3.06E-10 | 45.3630   | 66.9914   |
| Age       | 0.2844       | 0.0871         | 3.2671  | 0.0035   | 0.1039    | 0.4650    |

The estimated model is  $\widehat{\text{Happiness}} = 56.18 + 0.28 \text{ Age}$

- b. When Age = 25,  $\widehat{\text{Happiness}} = 56.18 + 0.28(25) = 63.18$   
 When Age = 50,  $\widehat{\text{Happiness}} = 56.18 + 0.28(50) = 70.18$   
 When Age = 75,  $\widehat{\text{Happiness}} = 56.18 + 0.28(75) = 77.18$



- d. According to the scatterplot, happiness seems to start higher, be lowest between ages 40–50, and then increase. In other words, the relationship between Age and Happiness is not linear. However, our linear model predicts a person aged 50 to be happier than a person aged 25. Therefore, our predictions based on our linear model are not accurate.

- 14.32 a. If  $x_1 = 40$ , and  $x_2 = -10$ ,  $\hat{y} = -8 + 2.6(40) - 47.2(-10) = 568$   
 b. As  $x_2$  increases by one unit,  $y$  is predicted to decrease by 47.2 units, holding  $x_1$  constant.  
 14.34 a. The point estimate of  $\beta_1$  is  $-2.53$ . As  $x_1$  increases by 1 unit,  $y$  is predicted to decrease by 2.53 units, holding  $x_2$  as constant.  
 b. The sample regression equation is,  $\hat{y} = 13.83 - 2.53x_1 + 0.29x_2$   
 c. If  $x_1 = -9$  and  $x_2 = 25$ ,  $\hat{y} = 13.83 - 2.53x_1 + 0.29x_2 = 13.83 - 2.53(-9) + 0.29(25) = 43.85$   
 14.40 a. The slope coefficient for Income is not as expected, since according to the sociologist's hypothesis we would expect it to have a negative sign. The positive sign for the Poverty coefficient is as expected.

- b. The slope coefficient of 53.16 suggests that as Poverty increases by one percent, the crime rate is predicted to rise by 53.16 crimes per 100,000 residents, holding Income constant.  
 c. If Poverty = 20 and Income = 50, then  $\widehat{\text{Crime}} = -301.62 + 53.16(20) + 4.95(50) = 1009.08$  (crimes per 100,000 residents).

- 14.50 a.  $s_e = \sqrt{s_e^2} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - k - 1}} = \sqrt{\frac{35}{50 - 2 - 1}} = \sqrt{0.74} = 0.86$ .  
 b. Given  $SSE = 35$  and  $SST = 90$ , we compute  $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{35}{90} = 0.6111$ .  
 14.54 a.  $s_e = \sqrt{s_e^2} = \sqrt{MSE} = \sqrt{6,969.03} = 83.48$ .  
 b.  $R^2 = \frac{SSR}{SST} = \frac{161478.4}{349642.2} = 0.4618$ . Therefore, 46.18% of the sample variation in  $y$  is explained by the estimated regression model.  
 c. From the ANOVA table we get  $k = 2$  and  $n = 29 + 1 = 30$ . Thus, adjusted  $R^2 = 1 - (1 - R^2) \left( \frac{n-1}{n-k-1} \right) = 1 - (1 - 0.4618) \left( \frac{30-1}{30-2-1} \right) = 0.4219$   
 14.58 a. Model 1:  $s_e = \sqrt{MSE} = \sqrt{0.1253} = 0.3540$   
 Model 2:  $s_e = \sqrt{MSE} = \sqrt{0.1718} = 0.4145$   
 b. Model 1:  $R^2 = \frac{SSR}{SST} = \frac{1.4415}{2.4440} = 0.5898$ .  
 Model 2:  $R^2 = \frac{SSR}{SST} = \frac{1.0699}{2.4440} = 0.4378$ .  
 c. Model 1 provides a better fit. It has a lower standard error of the estimate and a higher coefficient of determination.  
 14.60 a.  $s_e = \sqrt{MSE} = \sqrt{163.066} = 12.77$ .  
 b.  $R^2 = \frac{SSR}{SST} = \frac{918.746}{5,321.532} = 0.1726$ .  
 c. From the ANOVA table we get  $k = 2$  and  $n = 29 + 1 = 30$ . Thus, adjusted  $R^2 = 1 - (1 - R^2) \left( \frac{n-1}{n-k-1} \right) = 1 - (1 - 0.1726) \left( \frac{30-1}{30-2-1} \right) = 0.1113$

14.64 a.

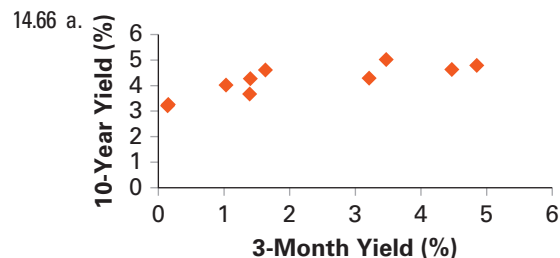
| Explanatory Variables | Adj ROA (Model 1) | Adj Return (Model 2) | Total Assets (Model 3) |
|-----------------------|-------------------|----------------------|------------------------|
| $s_e$                 | 9.79              | 9.78                 | 9.00                   |
| $R^2$                 | 0.0004            | 0.0024               | 0.1560                 |

Since Model 3 has the lowest  $s_e$  and the highest  $R^2$ , Model 3 is the best model.

b.

| Explanatory Variables | Adj. ROA & Adj Return | Adj. ROA & Total Assets | Adj. Return & Total Assets | All 3 Variables |
|-----------------------|-----------------------|-------------------------|----------------------------|-----------------|
| $s_e$                 | 9.79                  | 8.99                    | 8.99                       | 8.98            |
| Adjusted $R^2$        | -0.0018               | 0.1567                  | 0.1567                     | 0.1585          |

Based on the highest Adjusted  $R^2$ , and lowest  $s_e$ , the model using all 3 explanatory variables is the best.



The scatterplot suggests a positive correlation between the two yields.

b.

| $(x - \bar{x})$                  | $(y - \bar{y})$                  | $(x - \bar{x})(y - \bar{y})$                 | $(x - \bar{x})^2$                  |
|----------------------------------|----------------------------------|--|------------------------------------|
| 3.47 - 2.17<br>= 1.3             | 5.02 - 4.18<br>= 0.84            | (1.3)(0.84) =<br>1.09                        | 1.3 <sup>2</sup> = 1.68            |
| ⋮                                | ⋮                                | ⋮  | ⋮                                  |
| 0.14 - 2.17<br>= -2.03           | 3.21 - 4.18<br>= -0.97           | (-2.03)(-0.97)<br>= 1.97                     | -2.03 <sup>2</sup> = 4.14          |
| $\bar{x} = 2.17$<br>$s_x = 1.71$ | $\bar{y} = 4.18$<br>$s_y = 0.63$ | $\Sigma(x - \bar{x})(y - \bar{y})$<br>= 7.98 | $\Sigma(x - \bar{x})^2$<br>= 26.24 |

$$s_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{7.98}{10 - 1} = 0.89$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{0.89}{(1.71)(0.63)} = 0.81 \text{ (slightly different from Excel's}$$

estimate of 0.83, due to rounding)

To test if the population correlation is statistically significant, set up the hypotheses:

$$H_0: \rho_{xy} = 0$$

$$H_A: \rho_{xy} \neq 0$$

$$df = n - 2 = 8$$

$$\text{The value of the test statistic } t_8 = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0.81 \sqrt{10-2}}{\sqrt{1-(0.81)^2}} = 3.91$$

With  $\alpha = 0.05$ ,  $t_{\alpha/2, df} = t_{0.025, 8} = 2.306$ . Thus, we reject  $H_0$  if  $t_8 > 2.306$  or  $t_8 < -2.306$ . Since  $3.91 > 2.306$ , we reject  $H_0$ . At the 5% significance level, we can conclude that the correlation is significantly different from zero.

$$c. b_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{7.98}{26.24} = 0.30$$

$$b_0 = \bar{y} - b_1 \bar{x} = 4.18 - (0.30)(2.17) = 3.53$$

The sample regression equation is  $\hat{y} = 3.53 + 0.3x$

14.68 a. Excel Output:

| Regression Statistics |         |
|-----------------------|---------|
| Multiple R            | 0.6346  |
| R Square              | 0.4028  |
| Adj. R Square         | 0.3585  |
| Standard Error        | 13.6377 |
| Observations          | 30      |

| ANOVA      |    |         |         |      |                |
|------------|----|---------|---------|------|----------------|
|            | Df | SS      | MS      | F    | Significance F |
| Regression | 2  | 3386.46 | 1693.23 | 9.10 | 0.00           |
| Residual   | 27 | 5021.63 | 185.99  |      |                |
| Total      | 29 | 8408.09 |         |      |                |

|           | Coefficients | Standard Error | t Stat  | P-value | Lower 95% | Upper 95% |
|-----------|--------------|----------------|---------|---------|-----------|-----------|
| Intercept | -33.3966     | 12.7798        | -2.6132 | 0.0145  | -59.6185  | -7.1747   |
| P/E       | 3.9674       | 0.9587         | 4.1381  | 0.0003  | 2.0002    | 5.9346    |
| P/S       | -3.3681      | 2.6294         | -1.2809 | 0.2111  | -8.7632   | 2.0271    |

The signs are as expected. As P/E increases, the predicted returns increase, and as P/S decreases, the predicted returns increase.

b. As the P/S ratio increases by 1 unit, the predicted return of the firm decreases by 3.37%, holding P/E constant.

$$c. \widehat{\text{Return}} = -33.40 + 3.97(10) - 3.37(2) = -0.44\%$$

d.  $s_e = 13.64$  (from above)

$\frac{s_e}{\bar{y}} = \frac{13.64}{9.46} = 1.44$ . Since the ratio is greater than 0.20, this model is not very promising.

$R^2 = 0.4028$ . This means that 40.28% of the sample variation in  $y$  is explained by the sample regression equation.

14.70 a. Excel Output:

| Regression Statistics |        |
|-----------------------|--------|
| Multiple R            | 0.8466 |
| R Square              | 0.7168 |
| Adj. R Square         | 0.7106 |
| Standard Error        | 3.1373 |
| Observations          | 143    |

| ANOVA      |     |         |         |        |                |
|------------|-----|---------|---------|--------|----------------|
|            | df  | SS      | MS      | F      | Significance F |
| Regression | 3   | 3462.07 | 1154.02 | 117.25 | 0.00           |
| Residual   | 139 | 1368.14 | 9.84    |        |                |
| Total      | 142 | 4830.21 |         |        |                |

|           | Coefficients | Standard Error | t Stat  | P-value | Lower 95% | Upper 95% |
|-----------|--------------|----------------|---------|---------|-----------|-----------|
| Intercept | 0.4190       | 0.5124         | 0.8178  | 0.4149  | -0.5940   | 1.4321    |
| Research  | 0.0087       | 0.0012         | 7.1332  | 0.0000  | 0.0063    | 0.0111    |
| Patents   | 0.0517       | 0.0199         | 2.5967  | 0.0104  | 0.0123    | 0.0910    |
| Duration  | -0.0194      | 0.0236         | -0.8238 | 0.4115  | -0.0661   | 0.0272    |

Using the Excel output, the estimated model is

$$\widehat{\text{Startups}} = 0.4190 + 0.0087 \text{ Research} + 0.0517 \text{ Patents} - 0.0194 \text{ Duration}$$

b.  $\widehat{\text{Startups}} = 0.4190 + 0.0087(120) + 0.0517(8) - 0.0194(20) = 1.49$  startups

c. A \$1 million increase in research expenditure results in a predicted increase in the number of startups by 0.0087, holding everything else constant. Thus, approximately \$114.94 million ( $\frac{1}{0.0087} = \$114.94$ ) in additional research

expenditures would be needed to have 1 additional predicted startup, everything else being the same. Note that  $\$114.94 \times 0.0087$  equals (approximately) 1.

## Chapter 15

15.2 a.  $t_{0.05, 23} = 1.714$

$$b. t_{23} = \frac{0.5 - 0}{0.3} = 1.67$$

c.  $t_{23} = 1.67 < 1.714$ , we do not reject  $H_0$ ;  $\beta_1$  not positive

15.4 a.  $H_0: \beta_0 = 0$ ;  $H_A: \beta_0 \neq 0$ ;  $p\text{-value} \approx 0 < 0.05 = \alpha$ ; reject  $H_0$ ; the intercept differs from zero

b.  $0.1223 \pm (2.101)(0.1794)$ ;  $[-0.2546, 0.4992]$ ; not significant since the interval contains 0

15.6 a.  $H_0: \beta_1 = 0$ ;  $H_A: \beta_1 \neq 0$ ;  $p\text{-value} \approx 0 < 0.05 = \alpha$ ; reject  $H_0$ ;  $x_1$  and  $y$  are related

b. reported as  $[-1.67, 7.14]$ ; no, since the interval contains 0

c.  $H_0: \beta_1 \geq 20$ ;  $H_A: \beta_1 < 20$ ;  $t_{27} = \frac{12.91 - 20}{2.68} = -2.65 < -1.703 = -t_{0.05, 27}$ ; reject  $H_0$ ; yes, the slope is less than 20

15.10 a.  $\widehat{\text{Time}} = 13.353 - 0.0477 \text{ Height}$

b.  $H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

c.  $df = n - k - 1 = 7 - 1 - 1 = 5;$

$$t_5 = \frac{b_1 - \beta_{10}}{se(b_1)} = \frac{-0.0477 - 0}{0.0163} = -2.93$$

d. The reported  $p$ -value is 0.0332, which is less than  $\alpha = 0.05$ . Therefore, we reject  $H_0$ . At the 5% significance level, Height is significant in explaining Time.

15.14 a.  $\widehat{\text{Return}} = -12.0243 + 0.1459\text{P/E} + 5.4417\text{P/S}$

b.  $H_0: \beta_1 = \beta_2 = 0$

$H_A: \text{At least one } \beta_j \neq 0$

$df_1 = k = 2; df_2 = n - k - 1 = 27$

$$F_{(2,27)} = \frac{MSR}{MSE} = \frac{459.3728}{163.0661} = 2.8171$$

Since the reported  $p$ -value of 0.0774 is less than  $\alpha = 0.10$ , we reject  $H_0$ . At the 10% significance level, the two explanatory variables are jointly significant.

c. For the first explanatory variable P/E, the hypotheses are

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

The test statistic value and its  $p$ -value are reported in the regression results, and they are 0.1459 and 0.7383, respectively. Thus, since the  $p$ -value is greater than  $\alpha = 0.10$ , we do not reject  $H_0$ . At the 10% significance level, we cannot conclude that P/E is significant in explaining Return.

For the second explanatory variable, P/S, we state the hypotheses as

$H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

The reported  $p$ -value associated with P/S is equal to 0.0250. Thus, since the  $p$ -value is less than  $\alpha = 0.10$ , we reject  $H_0$ . At the 10% significance level, we can conclude that P/S is significant in explaining Return.

15.20 a.  $\widehat{\text{Cost}} = 14039.187 + 92.783 \text{ Temp} + 446.141 \text{ Work} - 27.003 \text{ Tons}$

b.  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

$H_A$ : At least one  $\beta_j \neq 0$

The reported  $p$ -value is 0.0262. Since the  $p$ -value is less than 0.10, reject  $H_0$ . At the 10% significance level, we can conclude the explanatory variables are jointly significant in explaining the electricity costs.

- c. For the average temperature, the hypotheses are

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

The reported  $p$ -value is 0.0114. Since the  $p$ -value is less than 0.10, reject  $H_0$ . At the 10% significance level, we can conclude the average temperature is significant in explaining the electricity costs.

For the number of work days, the hypotheses are

$H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

The reported  $p$ -value is 0.5494. Since the  $p$ -value is greater than 0.10, do not reject  $H_0$ . At the 10% significance level, we cannot conclude the number of work days is significant in explaining the electricity costs.

For the tons produced, the hypotheses are

$H_0: \beta_3 = 0$

$H_A: \beta_3 \neq 0$

The reported  $p$ -value is 0.8103. Since the  $p$ -value is greater than 0.10, do not reject  $H_0$ . At the 10% significance level, we cannot conclude the tons produced is significant in explaining the electricity costs.

- 15.26 Restricted Model:  $y = \beta_0 + \beta_1 x_1 + (1 - \beta_1)x_2 + \varepsilon = \beta_0 + \beta_1 x_1 + x_2 - \beta_1 x_2 + \varepsilon = \beta_0 + \beta_1 (x_1 - x_2) + x_2 + \varepsilon$

Note that we need to estimate only one regression coefficient in the restricted model, which we estimate by using  $y - x_2$  as the response variable and  $(x_1 - x_2)$  as the explanatory variable.

Unrestricted Model:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

- 15.28 a.  $H_0: \beta_2 = \beta_3 = 0$

$H_A$ : At least one of the coefficients is nonzero.

- b.  $df_1 = 2$  (because there are 2 restrictions,  $\beta_2 = \beta_3 = 0$ )

$$df_2 = n - k - 1 = 50 - 3 - 1 = 46$$

$$F_{(2,46)} = \frac{(SSE_R - SSE_U)/df_1}{SSE_U/df_2} = \frac{(56,944 - 48,074)}{2/48,074/46} = 4.24$$

- c. With  $\alpha = 0.05$ ,  $df_1 = 2$ ,  $df_2 = 46$ , the critical value is  $F_{0.05,(2,46)} = 3.20$ . The decision rule is to reject  $H_0$  if  $F_{(2,46)} > 3.20$ .

Since  $F_{(2,46)} = 4.24 > 3.20$ , we reject  $H_0$ . At the 5% significance level, we can conclude that Beds and Baths are jointly significant in explaining Price.

- 15.30 The hypotheses to test the joint significance of Patents and Duration are:

$H_0: \beta_2 = \beta_3 = 0$

$H_A$ : At least one of the coefficients is nonzero.

$df_1 = 2$  (because there are 2 restrictions,  $\beta_2 = \beta_3 = 0$ )

$$df_2 = n - k - 1 = 143 - 3 - 1 = 139$$

$$F_{(2,139)} = \frac{(SSE_R - SSE_U)/df_1}{SSE_U/df_2} = \frac{(1434.78 - 1368.14)}{2/1368.14/139} = 3.39$$

With  $\alpha = 0.05$ , the critical value is  $F_{\alpha,(df_1,df_2)} = F_{0.05,(2,139)} = 3.06$ . Thus, we reject  $H_0$  if  $F_{(2,139)} > 3.06$ . Since  $F_{(2,139)} = 3.39 > 3.06$ , we reject  $H_0$ . At the 5% level, Patents and Duration are jointly significant and Lisa should add both variables in the model predicting Startups.

- 15.32 a.  $\widehat{Sale} = 23.30 + 0.5184 \text{ MaleHours} + 0.6779 \text{ FemaleHours}$

- b.  $H_0: \beta_1 = \beta_2 = 0$

$H_A$ : At least one  $\beta_j \neq 0$

The reported  $p$ -value is about 0. Since the  $p$ -value is less than 0.05, reject  $H_0$ . At the 5% significance level, we can

conclude the explanatory variables are jointly significant in explaining the sales of mobile phone contracts.

For the number of hours clocked by male employees, test

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

The reported  $p$ -value is 3.096e-11. Since the  $p$ -value is less than 0.05, reject  $H_0$ . At the 5% significance level, we can conclude the number of hours clocked by male employees is significant in explaining the sales of model phone contracts.

For the number of hours clocked by female employees, the hypotheses are

$H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

The reported  $p$ -value is 7.156e-11. Since the  $p$ -value is less than 0.05, reject  $H_0$ . At the 5% significance level, we can conclude the number of hours clocked by female employees is significant in explaining the sales of model phone contracts.

- c. The completing hypotheses are

$H_0: \beta_1 = \beta_2$

$H_A: \beta_1 \neq \beta_2$

$df_1 = 1$  and  $df_2 = 49$ , the test statistic is given by

$$F_{(1,49)} = \frac{(SSE_R - SSE_U)/df_1}{(SSE_U/df_2)} = \frac{(70 - 66.914)/1}{(66.914/49)} = 2.2598.$$

The  $p$ -value is given by  $P(F_{(1,49)} > 2.2598) = 0.1392$ . Since the  $p$ -value is greater than 0.05, do not reject  $H_0$ . At the 5% significance level, we cannot conclude that there is a difference in productivity of male and female employees.

- 15.34 a.  $\widehat{Pay} = 10.307 + 0.2856 \text{ Job} - 0.0808 \text{ Company} + 0.0322 \text{ Union}$

- b.  $H_0: \beta_1 = \beta_2 = 3 = 0$

$H_A$ : At least one  $\beta_j \neq 0$

The reported  $p$ -value is about 8.15e-11. Since the  $p$ -value is less than 0.05, reject  $H_0$ . At the 5% significance level, we can conclude the explanatory variables are jointly significant in explaining the hourly pay.

For the job class, test

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

The reported  $p$ -value is 1.495e-11. Since the  $p$ -value is less than 0.05, reject  $H_0$ . At the 5% significance level, we can conclude the job class is significant in explaining the hourly pay.

For the number of years with the company, the hypotheses are

$H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

The reported  $p$ -value is 0.6402. Since the  $p$ -value is greater than 0.05, do not reject  $H_0$ . At the 5% significance level, we cannot conclude the number of years with the company significant in explaining the hourly pay.

For the number of years with the union, the hypotheses are

$H_0: \beta_3 = 0$

$H_A: \beta_3 \neq 0$

The reported  $p$ -value is 0.8513. Since the  $p$ -value is greater than 0.05, do not reject  $H_0$ . At the 5% significance level, we cannot conclude the number of years with the union significant in explaining the hourly pay.

- c.  $\widehat{Pay} = 10.307 + 0.2856 * 48 - 0.0808 * 18 + 0.0322 * 14 = \$23.01$

d. The completing hypotheses are

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_A: \text{At least one } \beta_j \neq 0$$

$df_1 = 2$  and  $df_2 = 46$ , the test statistic is given by

$$F_{(2,46)} = \frac{(SSE_R - SSE_Y)/df_1}{(SSE_Y/df_2)} = \frac{(296.193 - 289.973)/2}{(289.973/46)} = 0.4934.$$

The  $p$ -value is given by  $P(F_{(2,46)} > 0.4934) = 0.6137$ . Since the  $p$ -value is greater than 0.05, do not reject  $H_0$ . At the 5% significance level, we cannot conclude that the number of years at the company or the number of years in the union is jointly significant in explaining the hourly pay.

15.36 a. With  $df = n - k - 1 = 40 - 2 - 1 = 37$  and  $\alpha = 0.05$ , we find  $t_{\alpha/2, df} = t_{0.025, 37} = 2.026$ .

$\hat{y}^0 = 12.8 + (2.6)(15) - (1.2)(6) = 44.6$ ; thus, the 95% confidence interval is

$$\hat{y}^0 \pm t_{\alpha/2, df} se(\hat{y}^0) = 44.6 \pm (2.026)(2.20) = 44.6 \pm 4.4572.$$

Or, with 95% confidence,  $40.14 \leq E(y^0) \leq 49.06$ .

b.  $\hat{y}^0 \pm t_{\alpha/2, df} \sqrt{(se(\hat{y}^0))^2 + s_e^2} = 44.6 \pm 2.026 \sqrt{2.20^2 + 5.84^2} = 44.6 \pm 12.6435$ .

Or, with 95% confidence,  $31.96 \leq y^0 \leq 57.24$ .

c. The confidence interval is narrower because it assumes that the expected value of the error term is zero, whereas the prediction interval incorporates the non-zero error term.

15.38 a. Using Excel, the estimated regression line is  $\hat{y} = 22.81 + 0.85x_1 - 0.71x_2$ .

b. First, we need to derive  $\hat{y}^0$  and  $se(\hat{y}^0)$  by constructing the modified explanatory variables as  $x_1^* = x_1 - 50$  and  $x_2^* = x_2 - 20$ :

| $x_1$    | $x_2$    | $x_1^* = x_1 - 50$ | $x_2^* = x_2 - 20$ |
|----------|----------|--------------------|--------------------|
| 40       | 13       | -10                | -7                 |
| 48       | 28       | -2                 | 8                  |
| $\vdots$ | $\vdots$ | $\vdots$           | $\vdots$           |
| 29       | 14       | -21                | -6                 |

The relevant portion of the regression output with  $y$  as the response variable and  $x_1^*$  and  $x_2^*$  as the explanatory variables is:

| Regression Statistics |        |
|-----------------------|--------|
| Multiple R            | 0.9291 |
| R Square              | 0.8632 |
| Adj. R Square         | 0.8085 |
| Standard Error        | 4.6868 |
| Observations          | 8      |

|           | Coefficients | Standard Error | t Stat  | P-value | Lower 95% | Upper 95% |
|-----------|--------------|----------------|---------|---------|-----------|-----------|
| Intercept | 51.0041      | 2.1821         | 23.3742 | 0.0000  | 45.3949   | 56.6133   |
| $x_1^*$   | 0.8460       | 0.1523         | 5.5546  | 0.0026  | 0.4545    | 1.2376    |
| $x_2^*$   | -0.7053      | 0.2451         | -2.8773 | 0.0347  | -1.3353   | -0.0752   |

Therefore,  $\hat{y}^0 = 51.0041$  and  $se(\hat{y}^0) = 2.1821$ .

With  $df = n - k - 1 = 8 - 2 - 1 = 5$  and  $\alpha = 0.05$ , we find  $t_{\alpha/2, df} = t_{0.025, 5} = 2.571$ . The 95% confidence interval is

$$\hat{y}^0 \pm t_{\alpha/2, df} se(\hat{y}^0) = 51.0041 \pm (2.571)(2.1821) = 51.0041 \pm 5.6102.$$

Or, with 95% confidence,  $45.39 \leq E(y^0) \leq 56.61$ .

c. From the regression results above,  $\hat{y}^0 = 51.0041$ ,  $se(\hat{y}^0) = 2.1821$ , and  $s_e = 4.6868$ .

Thus, the 95% prediction interval is  $\hat{y}^0 \pm t_{\alpha/2, df}$

$$\sqrt{(se(\hat{y}^0))^2 + s_e^2} = 51.0041 \pm 2.571 \sqrt{2.1821^2 + 4.6868^2} = 51.0041 \pm 13.2918.$$

Or, with 95% confidence,  $37.71 \leq y^0 \leq 64.30$ .

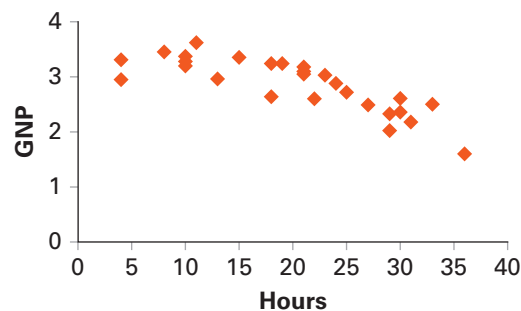
15.40 a. The sample regression equation is  $\widehat{Height} = 18.639 + 5.208 \text{ Fertilizer}$ .

The predicted height is  $\widehat{Height} = 18.639 + 5.208 * 3 = 34.263$  inches

b. With  $df = n - k - 1 = 30 - 1 - 1 = 28$  and  $\alpha = 0.10$ , we find  $t_{\alpha/2, df} = t_{0.05, 28} = 1.701$ . Also,  $\hat{y}^0 = 34.263$ . Thus, the 90% confidence interval is  $\hat{y}^0 \pm t_{\alpha/2, df} se(\hat{y}^0) = 34.263 \pm 1.701(1.566155) = 34.263 \pm 2.66403 = [31.599, 36.927]$ . Using this 90% confidence interval, we can state that the mean height of plants with 3 ounces of fertilizer is between 31.599 and 36.927 inches.

c. The prediction interval is given by  $\hat{y}^0 \pm t_{\alpha/2, df} \sqrt{(se(\hat{y}^0))^2 + s_e^2} = 34.263 \pm 1.701 \sqrt{(1.5662^2 + 8.5416^2)} = 34.263 \pm 14.77149 = [19.49, 49.04]$ . Using this 90% prediction interval, we can state that the height of a plant with 3 ounces of fertilizer is between 19.49 and 49.04 inches.

15.48 A scatterplot of GNP against Hours suggests that a simple linear regression model may not be the most appropriate model because of nonlinearities; that is, GPA seems to be positively related to Hours at lower levels of Hours yet negatively related to Hours at higher levels of Hours.



15.60 The relevant portion of the regression outcome is given below.

|           | Coefficients | Standard Error | t Stat  | P-value |
|-----------|--------------|----------------|---------|---------|
| Intercept | 78.9791      | 5.4855         | 14.3977 | 0.0000  |
| Income    | -0.0002      | 0.0001         | -1.7973 | 0.0785  |

a. The estimated equation is  $\widehat{Ownership} = 78.98 - 0.0001 \text{ Income}$

b. The hypotheses are:

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

Given  $n = n - k - 1 = 51 - 1 - 1 = 49$ ,  $t_{\alpha/2, df} = t_{0.025, 49} = 2.010$ . Thus, the decision rule is to reject  $H_0$  if  $t_{49} > 2.010$  or  $t_{49} < -2.010$ . Since  $-2.010 < t_{49} = -1.7973 < 2.010$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that the Ownership is linearly related to Income.

c. First, we need to derive  $\hat{y}^0$  and  $se(\hat{y}^0)$  by constructing the modified explanatory variable as  $x^* = x - 50000$ :



| $x$      | $x^* = x - 50000$ |
|----------|-------------------|
| 39,980   | - 10020           |
| 61,604   | 11604             |
| $\vdots$ | $\vdots$          |
| 52,470   | 2470              |

The relevant regression output with  $y$  as the response variable and  $x^*$  as the explanatory variable is:

| Regression Statistics |        |
|-----------------------|--------|
| Multiple R            | 0.2487 |
| R Square              | 0.0618 |
| Adj. R Square         | 0.0427 |
| Standard Error        | 5.7685 |
| Observations          | 51     |

|           | Coefficients | Standard Error | t Stat  | P-value |
|-----------|--------------|----------------|---------|---------|
| Intercept | 69.1995      | 0.8079         | 85.6533 | 0.0000  |
| $x^*$     | -0.0002      | 0.0001         | -1.7973 | 0.0785  |

Therefore,  $\hat{y}^0 = 69.1995$  and  $se(\hat{y}^0) = 0.8079$ .

With  $df = n - k - 1 = 51 - 1 - 1 = 49$  and  $\alpha = 0.05$ , we find  $t_{\alpha/2, df} = t_{0.025, 49} = 2.010$ . The 95% confidence interval is  $\hat{y}^0 \pm t_{\alpha/2, df} se(\hat{y}^0) = 69.1995 \pm (2.010)(0.8079) = 69.1995 \pm 1.6239$ .

Or, with 95% confidence,  $67.58 \leq E(y^0) \leq 70.82$ .

- d. From the regression results above,  $\hat{y}^0 = 69.1995$  and  $se(\hat{y}^0) = 0.8079$ , and  $s_e = 5.7685$ . Thus, the 95% prediction interval is

$$\hat{y}^0 \pm t_{\alpha/2, df} \sqrt{(se(\hat{y}^0))^2 + s_e^2} = 69.1995 \pm 2.010$$

$$\sqrt{0.8079^2 + 5.7685^2} = 69.1995 \pm 11.7078$$

Or, with 95% confidence,  $57.49 \leq y^0 \leq 80.91$ . The prediction interval is wider than the confidence interval since it accounts for the non-zero random error term.

- 15.62 Using  $y$  for Return and  $x_1$  and  $x_2$  for Turnover and Expense, respectively, the estimated model is  $\hat{y} = 2.5364 + 0.1047x_1 - 3.6056x_2$

Regression Output:

| Variable       | Model            |
|----------------|------------------|
| Intercept      | 2.5364 (0.1495)  |
| Turnover       | 0.1047 (0.0717)  |
| Expense        | -3.6056 (0.1189) |
| $s_e$          |                  |
| 1.5146         |                  |
| R <sup>2</sup> |                  |
| 0.5191         |                  |
| F (p-value)    | 2.6990 (0.1603)  |

Notes: Parameter estimates are in the top half of the table with the  $p$ -values in parentheses; \* represents significance at 5% level. The lower part of the table contains goodness-of-fit measures.

- a. For testing if Turnover is significant:

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

Since the associated  $p$ -value of 0.0717 is greater than 0.05, we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that Turnover is significant in explaining Return.

For testing if Expense is significant:

$$H_0: \beta_2 = 0$$

$$H_A: \beta_2 \neq 0$$

Since the associated  $p$ -value of 0.1189 is greater than 0.05, we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that Expense is significant in explaining Return.

For testing the variables' joint significance:

$$H_0: \beta_1 = \beta_2 = 0$$

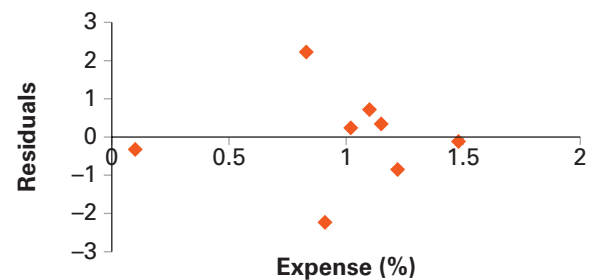
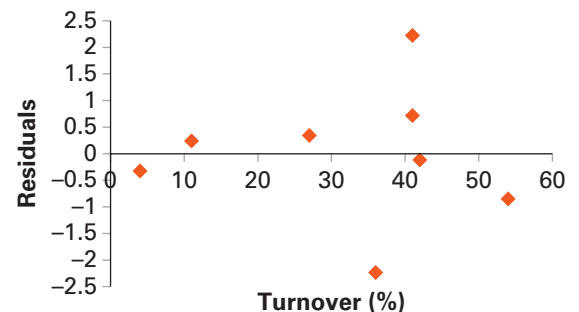
$$H_A: \text{At least one } \beta_i \neq 0$$

Since the associated  $p$ -value of 0.1603 is greater than 0.05, we do not reject  $H_0$ . Thus, at the 5% significance level, we cannot conclude that the explanatory variables are jointly significant.

George's theory does not appear to be valid.

- b. The sample correlation coefficient between Turnover and Expense is 0.6782. Since this is not above 0.80, there is not likely to be a severe problem of multicollinearity. The residual plots are shown below. Looking at the residual plots against

Turnover, it seems that the residuals vary more with larger turnover percentages than with lower. This suggests a possible problem heteroskedasticity.



- 15.64 a. The estimated model is:  $\widehat{\text{Return}} = -33.40 + 3.97 \text{ P/E} - 3.37 \text{ P/S}$

| Variable       | Model              |
|----------------|--------------------|
| Intercept      | -33.3966* (0.0145) |
| P/E            | 3.9674* (0.0003)   |
| P/S            | -3.3681 (0.2111)   |
| $s_e$          | 13.6377            |
| R <sup>2</sup> | 0.4028             |
| F (p-value)    | 9.1041 (0.0010)    |

Notes: Parameter estimates are in the top half of the table with the  $p$ -values in parentheses; \* represents significance at 5% level. The lower part of the table contains goodness-of-fit measures.

b.  $H_0: \beta_1 = \beta_2 = 0$

$H_A$ : At least one  $\beta_j \neq 0$

Since the reported  $p$ -value = 0.0010 < 0.05 =  $\alpha$ , we reject  $H_0$ . Thus, at the 5% significance level, we can conclude that the explanatory variables are jointly significant.

c. Using the  $p$ -values from the regression output, P/E has a  $p$ -value of 0.0003, which is less than  $\alpha = 0.05$ . Thus, at the 5% significance level we can conclude that P/E is significant in explaining Return. However, P/S has a  $p$ -value of 0.2111, which is greater than  $\alpha = 0.05$ . Thus, at the 5% significance level, we cannot conclude that P/S is individually significant.

d. For the confidence interval, we derive  $\hat{y}^0$  and  $se(\hat{y}^0)$  by constructing the modified explanatory variables as  $x_1^* = x_1 - 10$  and  $x_2^* = x_2 - 2$ :

| $x_1$    | $x_2$    | $x_1^* = x_1 - 10$ | $x_2^* = x_2 - 2$ |
|----------|----------|--------------------|-------------------|
| 14.37    | 2.41     | 4.37               | 0.41              |
| 11.01    | 0.78     | 1.01               | -1.22             |
| $\vdots$ | $\vdots$ | $\vdots$           | $\vdots$          |
| 13.94    | 1.94     | 3.94               | -0.06             |

The relevant regression output with  $y$  as the response variable and  $x_1^*$  and  $x_2^*$  as the explanatory variables is:

|           | Coefficients | Standard Error | t Stat  | P-value |
|-----------|--------------|----------------|---------|---------|
| Intercept | -0.4591      | 3.4099         | -0.1346 | 0.8939  |
| $x_1^*$   | 3.9674       | 0.9587         | 4.1381  | 0.0003  |
| $x_2^*$   | -3.3681      | 2.6294         | -1.2809 | 0.2111  |

Therefore,  $\hat{y}^0 = -0.4591$  and  $se(\hat{y}^0) = 3.4099$ .

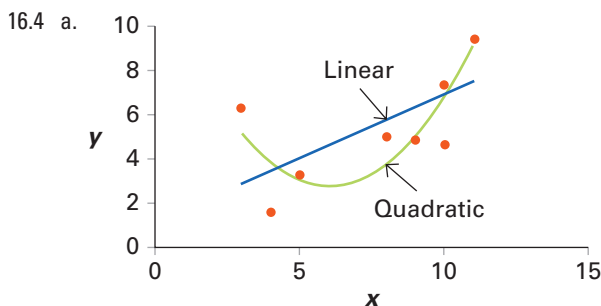
With  $df = n - k - 1 = 30 - 2 - 1 = 27$  and  $t_{\alpha/2, df} = t_{0.025, 27} = 2.052$ , the 95% confidence interval is

$\hat{y}^0 \pm t_{\alpha/2, df} se(\hat{y}^0) = -0.4591 \pm (2.052)(3.4099) = -0.4591 \pm 6.9971$ .

Or, with 95% confidence,  $-7.46 \leq E(y^0) \leq 6.54$ .

## Chapter 16

16.2  $x = 10, 15$ :  $\hat{y} = 92, 98$  (linear);  $\hat{y} = 161, 96.5$  (quadratic);  $\hat{y} = 120, 115$  (cubic)



$\hat{y} = 1.1006 + 0.5828x$ ;  $\hat{y} = 12.1338 - 3.1034x + 0.2565x^2$

b. The quadratic regression is better because of the higher adjusted  $R^2$  (0.78 > 0.42)

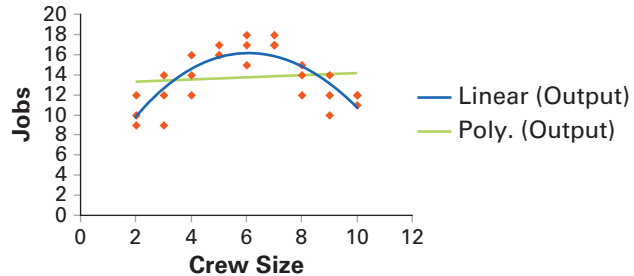
c.  $x = 4$ ,  $\hat{y} = 3.82$ ;  $x = 6$ ,  $\hat{y} = 2.75$ ;  $x = 12$ ,  $\hat{y} = 11.83$

d. minimum at  $x = \frac{-(-3.1034)}{2(0.2565)} = 6.05$

16.6 a.  $x = 2, 3$ :  $\hat{y} = 22.5, 23.85$  (linear);  $\hat{y} = 21.84, 21.79$  (quadratic);  $\hat{y} = 21.91, 21.57$  (cubic)

b. The quadratic model is the best (highest adjusted  $R^2$ )

16.8 a. The scatter plot is shown below. In general, crew sizes of 6 or 7 seem optimal.



b. Linear model:  $Jobs = 13.0741 + 0.1111 Crew Size$

Quadratic model:  $Jobs = 2.1111 + 4.5960 Crew Size - 0.3737 Crew Size^2$

Regression Summary Output:

| Variable                 | Linear Model       | Quadratic Model   |
|--------------------------|--------------------|-------------------|
| Intercept                | 13.0741*<br>(0.00) | 2.1111<br>(0.26)  |
| Crew Size                | 0.1111<br>(0.60)   | 4.5960*<br>(0.00) |
| (Crew Size) <sup>2</sup> | NA                 | -0.3737*          |
| Adjusted $R^2$           | -0.0284            | 0.6307            |

Notes: Parameter estimates are in the top part of the table with the  $p$ -values in parentheses; \* represents significance at the 5% level. The last row presents adjusted  $R^2$  for model comparison

The quadratic model provides a much better fit than the linear model. The linear model has a negative adjusted  $R^2$  and a statistically insignificant variable Crew Size. Conversely, the quadratic model has an adjusted  $R^2$  of 0.6307 and both the first order and second order Crew Size variables are significant (although the constant term is not significant). Moreover, the negative coefficient (-0.3737) of the second-order term in the quadratic model is indicative of the inverted U-shaped relation between jobs completed and crew size.

An intuitive justification for the quadratic model is as follows. Too small of a crew size can slow down jobs due to an insufficient number of available workers. Conversely, too large of a crew size can also slow down jobs due to excessive workers contributing to worker congestion and/or idle, unproductive workers.

c.  $\widehat{Jobs} = 2.1111 + 4.5960(5) - 0.3737(5)^2 = 15.75$  jobs/week for a crew size of 5 workers.

d. The cubic (third-order) model is as follows:

$\widehat{Jobs} = 0.6852 + 5.5407 Crew Size - 0.5505 Crew Size^2 + 0.0098 Crew Size^3$

However, this third-order (cubic) model provides a worse fit than the second-order (quadratic) model. Specifically,

adjusted  $R^2$  decreased from 0.6307 in the quadratic model to 0.6170 in the cubic model. Also, the second-order and third-order terms in the cubic model are insignificant based on  $p$ -values of 0.26 and 0.71, respectively.

16.12 Model 1:  $\hat{y} = 500 - 4.2x$

Model 2:  $\hat{y} = 1370 - 280 \ln(x)$

Model 3:  $\ln(\hat{y}) = 8.4 - 0.04x$

Model 4:  $\ln(\hat{y}) = 8 - 0.8 \ln(x)$

a. The slope coefficients are interpreted as:

Model 1: As  $x$  increases by one unit,  $\hat{y}$  decreases by 4.2 units.

Model 2: As  $x$  increases by one percent,  $\hat{y}$  decreases by about 2.8 units ( $= \frac{280}{100}$ ).

Model 3: As  $x$  increases by one unit,  $\hat{y}$  decreases by about 4% ( $= 0.04 \times 100$ ).

Model 4: As  $x$  increases by one percent,  $\hat{y}$  decreases by about 0.8%.

b. Model 1: When  $x = 100$ ,  $\hat{y} = 500 - 4.2 \times 100 = 80$ ; When  $x = 101$ ,  $\hat{y} = 75.80$ . Therefore, as  $x$  increases by 1,  $\hat{y}$  decreases by 4.2 ( $75.8 - 80 = -4.2$ ).

Model 2: When  $x = 100$ ,  $\hat{y} = 1370 - 280 \ln(100) = 80.55$ ; When  $x = 101$ ,  $\hat{y} = 77.77$ . Therefore, as  $x$  increases by 1%,  $\hat{y}$  decreases by 2.78 ( $77.77 - 80.55 = -2.78$ ).

Model 3: When  $x = 100$ ,  $\hat{y} = \exp(8.4 - 0.04(100) + \frac{0.13^2}{2}) = 82.14$ ; When  $x = 101$ ,  $\hat{y} = \exp(8.4 - 0.04(101) + \frac{0.13^2}{2}) = 78.92$ . Therefore, as  $x$  increases by 1,  $\hat{y}$  decreases by 3.22 ( $78.92 - 82.14$ ), or by 3.92% ( $\frac{78.92 - 82.14}{82.14} \times 100 = -3.92$ ).

Model 4: When  $x = 100$ ,  $\hat{y} = \exp(8 - 0.8 \ln(100) + \frac{0.11^2}{2}) = 75.33$ ; When  $x = 101$ ,  $\hat{y} = \exp(8 - 0.8 \ln(101) + \frac{0.11^2}{2}) = 74.74$ . Therefore, as  $x$  increases by 1%,  $\hat{y}$  decreases by 0.59 ( $74.74 - 75.33$ ), or by 0.78% ( $\frac{74.74 - 75.33}{75.33} \times 100 = -0.78$ ).

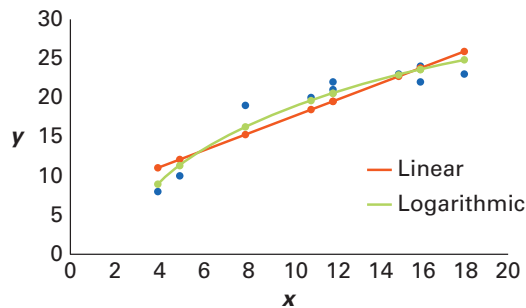
16.14 Model 1:  $\hat{y} = 240.42 + 4.68(100) = 708.42$

Model 2:  $\hat{y} = -69.75 + 162.51 \ln(100) = 678.64$

Model 3:  $\hat{y} = \exp(1.58 + 0.05(100) + \frac{0.12^2}{2}) = 725.74$

Model 4:  $\hat{y} = \exp(0.77 + 1.25 \ln(100) + \frac{0.09^2}{2}) = 685.75$

16.16 a.



It is difficult to tell by the graph alone whether the line or the logarithmic curve fit the data the best, although the curve shows a slightly better fit. Therefore the logarithmic model should be evaluated with the linear model.

b. The 2 models to evaluate are:

Model 1:  $y = \beta_0 + \beta_1 x + \varepsilon$

Model 2:  $y = \beta_0 + \beta_1 \ln(x) + \varepsilon$

In order to estimate the 2<sup>nd</sup> model, we must log-transform

$x$ . The model estimates are:

| Variable  | Model 1           | Model 2            |
|-----------|-------------------|--------------------|
| Intercept | 6.7904*<br>(0.01) | -5.6712*<br>(0.04) |
| $x$       | 1.0607*<br>(0.00) | NA                 |
| $\ln(x)$  | NA                | 10.5447*<br>(0.00) |
| $R^2$     | 0.8233            | 0.9341             |

Notes: Parameter estimates are in the top part of the table with the  $p$ -values in parentheses; \* represents significance at the 5% level. The last row presents the computer generated  $R^2$ .

Since Models 1 and 2 are both specified in terms of  $y$ , we can simply use the computer generated  $R^2$  to compare them. Model 2 is preferred since it has a higher  $R^2$  ( $0.9341 > 0.8233$ ).

c. Using Model 2, when  $x = 10$ ,  $\hat{y} = -5.6712 + 10.5447 \ln(10) = 18.61$

16.18 a. As the price increases by one unit, the consumption decreases by 1.25%.

b. The hypotheses are

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

The  $p$ -value is 0.0045. Since the  $p$ -value is less than 0.05, reject  $H_0$ . At the 5% significance level, we can conclude that price elasticity of demand for cigarettes is statistically significant.

c. As the annual income increases by one unit, the consumption increases by 0.18%.

d. The hypotheses are

$H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

The  $p$ -value is 0.3996. Since the  $p$ -value is greater than 0.05, do not reject  $H_0$ . At the 5% significance level, we cannot conclude that income elasticity of demand for cigarettes is statistically significant.

16.20 a. The sample regression equation is  $\widehat{\text{Watches}} = 35.909 - 0.0260 \text{ Time}$ . As the time increases by 1 second, the number of watches per shift is predicted to decrease by 0.0260. The predicted height is  $\widehat{\text{Watches}} = 35.909 - 0.0260 * 550 = 21.609$  watches per shift.

b. The sample regression equation is  $\widehat{\text{Watches}} = 123.11 - 16.21 \ln(\text{Time})$ . As the time increases by 1%, the number of watches per shift is predicted to decrease by 0.1621%. The predicted height is  $\widehat{\text{Watches}} = 123.11 - 16.21 \ln(550) = 20.8262$  watches per shift.

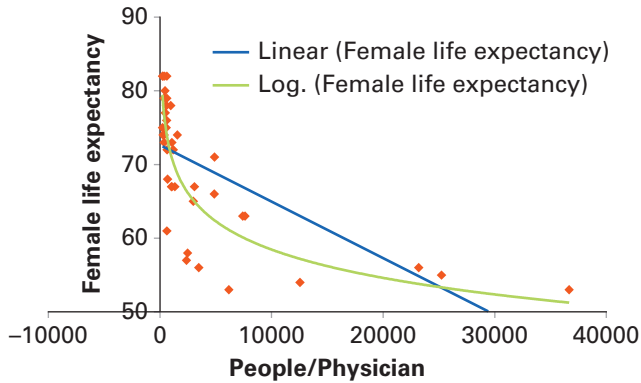
The  $R^2$  is slightly higher for the logarithmic model.

16.22 a.  $\widehat{\text{Cost}} = 14039.187 + 92.783 \text{ Temp} + 446.141 \text{ Work} - 27.003 \text{ Tons}$ ;  $\widehat{\text{Cost}} = 14039.187 + 92.783(65) + 446.141(23) - 27.003(76) = \$28,279.4$

b.  $\ln(\widehat{\text{Cost}}) = 9.70 + 0.0034 \text{ Temp} + 0.0181 \text{ Work} - 0.0012 \text{ Tons}$ ;  $\ln(\widehat{\text{Cost}}) = 9.70 + 0.0034(65) + 0.0181(23) - 0.0012(76) = 10.2461$ ; so the predicted cost is  $e^{10.2461} = \$28,172.45$

c. The  $R^2$  is slightly higher for the exponential model.

16.24 a.



The logarithmic model appears to have a better fit than the linear model.

- b. The regression results for both the linear and logarithmic models are below:

| Response Variable: Female Life Expectancy |                      |                       |
|---|----------------------|-----------------------|
| Variable                                  | Model 1              | Model 2               |
| Intercept                                 | 72.6441*<br>(0.0000) | 109.4952*<br>(0.0000) |
| People/Physician                          | -0.0008*<br>(0.0000) | NA                    |
| ln(People/Physician)                      | NA                   | -5.5414*<br>(0.0000)  |
| $R^2$                                     | 0.4126               | 0.6967                |

Notes: Parameter estimates are in the top part of the table with the  $p$ -values in parentheses; \* represents significance at the 5percent level. The last row presents the computer generated  $R^2$ .

When the people-to-physician ratio is 1000, female life expectancy is

$$\hat{y} = 72.6441 - 0.0008(1000) = 71.84.$$

When the people-to-physician ratio is 500,

$$\hat{y} = 72.6441 - 0.0008(500) = 72.24.$$

Therefore, as  $x$  decreases by 500,  $\hat{y}$  increases by  $72.24 - 71.84 = 0.40$  year.

- c. When the people-to-physician ratio is 1000, female life expectancy is

$$\hat{y} = 109.4952 - 5.5414 \ln(1000) = 71.22.$$

When the people-to-physician ratio is 500,

$$\hat{y} = 109.4952 - 5.5414 \ln(500) = 75.06.$$

Therefore, as  $x$  decreases by 500,  $\hat{y}$  increases by  $75.06 - 71.22 = 3.84$  years.

- d. Since both models have the same response variable, we can compare the computer generated  $R^2$  values directly. Since  $0.6967 > 0.4126$ , the logarithmic model provides a better fit.

## Chapter 17

17.2 a.  $\hat{y} = 160 + 15(1) + 32(1) = 207$

b.  $\hat{y} = 160 + 15(0) + 32(0) = 160$

- 17.4 a. The reference group for Model 1 is the employees who are female.

- b. The reference group for Model 2 is the employees who are female and without an MBA.

It would not matter if we set  $d_1 = 1$  for female employees. The inferences would not change.

- 17.10 a. Make two columns, one for players with nicknames and one for players without; a portion of the data used is shown below.

| Years with Nickname | Years with No Nickname |
|---------------------|------------------------|
| 74                  | 62                     |
| 62                  | 56                     |
| :                   | :                      |
| 68                  | 64                     |

The average lifespan for players with nicknames is 68.05, or approximately 68 years. Without nicknames, the average lifespan is 64.08, or approximately 64 years. The difference in lifespan for players with and without nicknames is  $68.05 - 64.08 = 3.97$ .

- b. If Nickname = 1 for players with a nickname, and Nickname = 0 for players without a nickname, the relevant regression results for  $Years = \beta_0 + \beta_1 \text{Nickname} + \epsilon$  are:

| Standard  |              |       |        |         |
|-----------|--------------|-------|--------|---------|
|           | Coefficients | Error | t Stat | P-value |
| Intercept | 64.08        | 1.80  | 35.55  | 0.0000  |
| Nickname  | 3.97         | 2.33  | 1.71   | 0.0989  |

For the players with a nickname,  $\widehat{Years} = 64.08 + 3.97(1) = 68.05$ .

For the players without a nickname,  $\widehat{Years} = 64.08 + 3.97(0) = 64.08$ .

Thus, the difference  $68.05 - 64.08 = 3.97$  is the same as in a.

- c. The hypotheses are:

$$H_0: \beta_1 \leq 0$$

$$H_A: \beta_1 > 0$$

For this one-tailed test, we must divide the reported  $p$ -value in half, so the  $p$ -value =  $0.0989/2 = 0.0494$ . Since  $0.0494 < 0.05 = \alpha$ , we reject  $H_0$ . At the 5% significance level, we conclude that players with a nickname do live longer than players without a nickname.

- 17.12 Relevant regression results for  $Math = \beta_0 + \beta_1 \text{GPA} + \beta_2 \text{Female} + \epsilon$ :

|           | Coefficients | Standard Error | t Stat | P-value |
|-----------|--------------|----------------|--------|---------|
| Intercept | 274.12       | 37.85          | 7.24   | 0.0000  |
| GPA       | 98.71        | 12.54          | 7.87   | 0.0000  |
| Female    | -21.10       | 10.55          | -2.00  | 0.0618  |

- a. For a male student with a GPA of 3.5,  $Math = 274.12 + 98.71(3.5) - 21.10(0) = 274.12 + 98.71(3.5) = 619.60$ .

$$\text{For a female student with a GPA of 3.5, } Math = 274.12 + 98.71(3.5) - 21.10(1) = 274.12 + 98.71(3.5) - 21.10 = 598.50.$$

- b. For the hypotheses  $H_0: \beta_2 = 0$ ,  $H_A: \beta_2 \neq 0$ , the  $p$ -value =  $0.0618 > 0.05 = \alpha$ , so we do not reject  $H_0$ . Therefore, we cannot conclude that there is a statistically significant gender difference in math scores at the 5% level.

- 17.14 a. The fourth dummy variable  $d_4$  would equal 1 if quarter 4 and 0 otherwise. However, the inclusion of  $d_4$  would cause perfect multicollinearity because  $d_1 + d_2 + d_3 + d_4 = 1$ .

Thus the model with the four dummy variables could not be estimated. We must use one less dummy variable than the number of categories, and quarter 4 was assumed to be the reference category.

- b. Since the  $p$ -values for all three variables are greater than 0.05, we cannot conclude that any of the dummy variables is individually significant at the 5% level. Therefore, the analyst cannot conclude that in any of the first three quarters the return differs from the return in quarter 4.
- c. To determine if the quarterly return is higher in quarter 2 than in quarter 3, one can consider one of the two equivalent models:  $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_4 d_4 + \varepsilon$  or  $y = \beta_0 + \beta_1 d_1 + \beta_3 d_3 + \beta_4 d_4 + \varepsilon$ .
- 17.18 a.  $\hat{y} = 5.2 + 0.9(10) + 1.4(1) + 0.2(1)(10) = 17.6$
- b.  $\hat{y} = 5.2 + 0.9(10) + 1.4(0) + 0.2(10)(0) = 5.2 + 0.9(10) = 14.2$

- 17.22 Model 1: Consumption =  $\beta_0 + \beta_1 x + \varepsilon$   
 Model 2: Consumption =  $\beta_0 + \beta_1 x + \beta_2 d + \varepsilon$   
 Model 3: Consumption =  $\beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd + \varepsilon$   
 Relevant regression results:

|                       | Model 1             | Model 2               | Model 3               |
|-----------------------|---------------------|-----------------------|-----------------------|
| Intercept             | 8160.34<br>(0.1551) | 13007.26*<br>(0.0385) | -1676.58<br>(0.8482)  |
| Income ( $x$ )        | 0.55*<br>(0.0000)   | 0.44*<br>(0.0000)     | 0.66*<br>(0.0000)     |
| Urban<br>(dummy $d$ ) | NA                  | 6544.43<br>(0.0717)   | 36361.71*<br>(0.0010) |
| $xd$                  | NA                  | NA                    | -0.38*<br>(0.0274)    |
| Adjusted $R^2$        | 0.5806              | 0.6006                | 0.6332                |

Notes: The top portion of the table contains parameter estimates with  $p$ -values in parentheses; \* represents significance at the 5 percent level; Adjusted  $R^2$ , reported in the last row, is used for model selection.

- a. The linear model, Model 1, is estimated as  $\hat{y} = 8160.34 + 0.55x$ .  
 For a family with income of \$75,000, the predicted consumption is  $\hat{y} = 8160.34 + 0.55(75000) = \$49,410.34$ .
- b. With a dummy, Model 2 is estimated as  $\hat{y} = 13007.26 + 0.44x + 6544.43d$ .  
 For a family with income of \$75,000 in an urban community, let  $d = 1$ . The predicted consumption is  $\hat{y} = (13007.26 + 6544.43) + 0.44(75000) = \$52,551.69$ .  
 For a comparable family in a rural community, let  $d = 0$ . The predicted consumption is  $\hat{y} = 13007.26 + 0.44(75000) = \$46,007.26$ .
- c. With a dummy and an interaction variable, Model 3 is estimated as  $\hat{y} = -1676.58 + 0.66x + 36361.71d - 0.38xd$ . For a family with income of \$75,000 in an urban community, let  $d = 1$ . The predicted consumption is  $\hat{y} = (-1676.58 + 36361.71) + (0.66 - 0.38)(75000) = \$55,685.13$ . For a comparable family in a rural community, let  $d = 0$ . The predicted consumption is  $\hat{y} = -1676.58 + 0.66(75000) = \$47,823.42$ .
- d. Since each model has a different number of explanatory variables, we use Adjusted  $R^2$  to compare the models. Model 3 has the highest value of 0.6332 and is therefore the most suitable model.

- 17.28 a. With  $x = 25$ ,  $\hat{y} = 0.92 - 0.02 \times 25 = 0.42$ , or 42%. With  $x = 40$ ,  $\hat{y} = 0.92 - 0.02 \times 40 = 0.12$ , or 12%.
- b. The estimated probability is negative if  $\hat{y} = 0.92 - 0.02x < 0$ , which is equivalent to  $x > 46$ . Since the explanatory variable  $x$  cannot exceed 50, the final answer is  $46 < x \leq 50$ .
- 17.30 a. The LPM model is estimated as  $\hat{y} = -0.40 + 0.32x$ .  
 When  $x = 1$ ,  $\hat{y} = -0.40 + 0.32 \times 1 = -0.08$ .  
 The logit model is estimated as  $\hat{y} = \frac{\exp(-4.5 + 1.54x)}{1 + \exp(-4.5 + 1.54x)}$ .  
 When  $x = 1$ ,  $\hat{y} = \frac{\exp(-4.5 + 1.54 \times 1)}{1 + \exp(-4.5 + 1.54 \times 1)} = 0.05$ .  
 The predictions for  $x = 2, 3, 4$  and 5 are made similarly and the results are shown below for both models:

| $x$ | LPM   | Logit |
|-----|-------|-------|
| 1   | -0.08 | 0.05  |
| 2   | 0.24  | 0.19  |
| 3   | 0.56  | 0.53  |
| 4   | 0.88  | 0.84  |
| 5   | 1.2   | 0.96  |

- b. The LPM model is not always appropriate as it can give negative probabilities and probabilities over 1 for small values and large values of  $x$ . The logit model always gives predicted probabilities between 0 and 1.
- 17.32 a. The LPM is estimated as  $\hat{y} = 1.31 - 0.04x$ . When  $x = 20$ ,  $\hat{y} = 1.31 - 0.04 \times 20 = 0.51$ .
- b. The competing hypotheses are specified as  $H_0: \beta_1 = 0$ ,  $H_A: \beta_1 \neq 0$ . Since the  $p$ -value = 0.0125 is less than  $\alpha = 0.05$ , we reject  $H_0$  and conclude that  $x$  is significant at the 5% level.
- 17.34 a. The logit model is estimated as  
 $\hat{y} = \frac{\exp(1.609 - 0.194x_1 + 0.202x_2 + 0.223x_3)}{1 + \exp(1.609 - 0.194x_1 + 0.202x_2 + 0.223x_3)}$ .  
 When  $x_1 = 15$ ,  $x_2 = 10$ , and  $x_3 = -2$ ,  
 $\hat{y} = \frac{\exp(1.609 - 0.194 \times 15 + 0.202 \times 10 + 0.223 \times (-2))}{1 + \exp(1.609 - 0.194 \times 15 + 0.202 \times 10 + 0.223 \times (-2))} = 0.57$   
 With a  $p$ -value of 0.01,  $x_3$  is the only variable that is significant at the 5% level. All of the other variables have  $p$ -values greater than 0.05 and are therefore not significant at the 5% level.
- 17.36 Relevant regression results from Minitab:

| Predictor | Coef    | SE     | Z     | P      |
|-----------|---------|--------|-------|--------|
| Constant  | 2.3069  | 1.6962 | 1.36  | 0.1740 |
| Age       | -0.1215 | 0.0609 | -1.99 | 0.0460 |

The model is estimated as  $\hat{y} = \frac{\exp(2.3069 - 0.1215\text{Age})}{1 + \exp(2.3069 - 0.1215\text{Age})}$ .

- a. For a 20-year-old customer,  
 $\hat{y} = \frac{\exp(2.3069 - 0.1215 \times 20)}{1 + \exp(2.3069 - 0.1215 \times 20)} = 0.47$   
 For a 30-year-old customer,  
 $\hat{y} = \frac{\exp(2.3069 - 0.1215 \times 30)}{1 + \exp(2.3069 - 0.1215 \times 30)} = 0.21$
- b. The competing hypotheses are specified as  $H_0: \beta_1 \geq 0$ ,  $H_A: \beta_1 < 0$ . Since the  $p$ -value =  $0.0460/2 = 0.0230$  is less than  $\alpha = 0.05$ , we reject  $H_0$  and conclude that  $\beta_1$  is less than 0 at the 5% level. Therefore, Annabel's belief that Under Armour attracts a younger clientele is again supported by the data.



17.38 Relevant regression results from Minitab:

| Predictor | Coef    | SE     | Z     | P     |
|-----------|---------|--------|-------|-------|
| Constant  | -9.6504 | 4.1710 | -2.31 | 0.021 |
| Premium%  | 0.0654  | 0.0287 | 2.28  | 0.023 |
| Income    | 0.1291  | 0.0553 | 2.34  | 0.020 |

The model is estimated as

$$\hat{y} = \frac{\exp(-9.6504 + 0.0654x_1 + 0.1291x_2)}{1 + \exp(-9.6504 + 0.0654x_1 + 0.1291x_2)}$$

For an individual with an income of \$60,000 and an employee contribution of 50 percent of the premium, let  $x_1 = 50$  and  $x_2 = 60$ . The predicted probability of having insurance coverage is  $\hat{P} = \frac{\exp(-9.6504 + 0.0654 \times 50 + 0.1291 \times 60)}{(1 + \exp(-9.6504 + 0.0654 \times 50 + 0.1291 \times 60))} = 0.80$ .

If the employer were to contribute 75 percent of the premium, let  $x_1 = 75$ . The estimated probability of coverage would be

$$\hat{y} = \frac{\exp(-9.6504 + 0.0654 \times 75 + 0.1291 \times 60)}{1 + \exp(-9.6504 + 0.0654 \times 75 + 0.1291 \times 60)} = 0.95.$$

17.40 Relevant regression results from Minitab:

| Predictor | Coef    | SE     | Z     | P     |
|-----------|---------|--------|-------|-------|
| Constant  | -0.8103 | 1.1594 | -0.70 | 0.485 |
| Age       | 2.5615  | 1.1783 | 2.17  | 0.030 |
| Income    | 0.0094  | 0.0184 | 0.51  | 0.608 |
| Children  | -1.2436 | 0.5964 | -2.09 | 0.037 |

The model is estimated as

$$\hat{y} = \frac{\exp(-0.8103 + 2.5615x_1 + 0.0094x_2 - 1.2436x_3)}{1 + \exp(-0.8103 + 2.5615x_1 + 0.0094x_2 - 1.2436x_3)}$$

a. The competing hypotheses are specified as  $H_0: \beta_1 \leq 0$ ,  $H_A: \beta_1 > 0$ . For this upper-tailed test, the  $p$ -value equals  $0.030/2 = 0.015$ , and is therefore individually significant at the 5% level. We conclude that the divorce rate is higher for this age group.

b. For an individual who is 27 years old, has \$60,000 in income and has one child, let  $x_1 = 1$ ,  $x_2 = 60$ , and  $x_3 = 1$ . The probability of divorce for an individual with these characteristics is

$$\hat{y} = \frac{\exp(-0.8103 + 2.5615 \times 1 + 0.0094 \times 60 - 1.2436 \times 1)}{1 + \exp(-0.8103 + 2.5615 \times 1 + 0.0094 \times 60 - 1.2436 \times 1)} = 0.74.$$

The corresponding probability with three children is

$$\hat{y} = \frac{\exp(-0.8103 + 2.5615 \times 1 + 0.0094 \times 60 - 1.2436 \times 3)}{1 + \exp(-0.8103 + 2.5615 \times 1 + 0.0094 \times 60 - 1.2436 \times 3)} = 0.20.$$

17.42 a. Relevant regression results for  $\text{Callback} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Caucasian} + \epsilon$ :

| Standard  |              |       |        |         |
|-----------|--------------|-------|--------|---------|
|           | Coefficients | Error | t Stat | P-value |
| Intercept | 25.1466      | 1.84  | 13.70  | 0.0000  |
| Age       | -0.3196      | 0.04  | -7.31  | 0.0000  |
| Caucasian | 9.4504       | 1.01  | 9.36   | 0.0000  |

b. For a 30-year-old applicant with a Caucasian name, set  $x = 30, d = 1$ . The predicted call-back rate is  $\hat{y} = (25.1466 + 9.4504) - 0.3196(30) = 25\%$ .

The corresponding call-back rate for a non-Caucasian name ( $d = 0$ ) is  $\hat{y} = 25.1466 - 0.3196(30) = 15.55\%$ .

c. To test for race discrimination, the hypotheses would be  $H_0: \beta_2 = 0$ ,  $H_A: \beta_2 \neq 0$ . With a  $p$ -value of approximately zero, we reject  $H_0$  and conclude that the *Caucasian*

dummy variable is significant at the 5% level. Therefore, the data suggest that there is race discrimination.

17.44 a. Relevant regression results for  $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon$ :

|                       | Coefficients | Standard Error | t Stat | P-value |
|-----------------------|--------------|----------------|--------|---------|
| Intercept             | 28.26        | 0.29           | 98.26  | 0.0000  |
| Female                | -3.45        | 0.61           | -5.67  | 0.0000  |
| Black                 | -1.31        | 0.43           | -3.03  | 0.0030  |
| Female $\times$ Black | 6.66         | 0.91           | 7.30   | 0.0000  |

The predicted BMI for white males, set  $d_1 = 0, d_2 = 0$ , and  $d_3 = 0$ .

Calculate  $\widehat{\text{BMI}} = 28.26$

The predicted BMI for white females, set  $d_1 = 1, d_2 = 0$ , and  $d_3 = 0$ .

Calculate  $\widehat{\text{BMI}} = 28.26 - 3.45 = 24.81$

The predicted  $\widehat{\text{BMI}}$  for black males, set  $d_1 = 0, d_2 = 1$ , and  $d_3 = 0$ .

Calculate  $\widehat{\text{BMI}} = 28.26 - 1.31 = 26.95$

The predicted  $\widehat{\text{BMI}}$  for black females, set  $d_1 = 1, d_2 = 1$ , and  $d_3 = 1$ .

Calculate  $\widehat{\text{BMI}} = 28.26 - 3.45 - 1.31 + 6.66 = 30.16$

b. To test for a difference between white females and white males, we look at  $d_1$ . With a  $p$ -value of approximately zero, we can conclude that the female dummy is significant and therefore there is a difference between white females and white males at the 5% level, holding all other variables constant.

c. To test for a difference between white males and black males, we look at  $d_2$ . With a  $p$ -value of 0.003, we can conclude that the black dummy is significant and therefore there is a difference between white males and black males at the 5% level, holding all other variables constant.

17.50 a. Relevant regression results:

|           | Coefficients | Standard Error | t Stat  | p-value |
|-----------|--------------|----------------|---------|---------|
| Intercept | -2.2461      | 0.6456         | -3.4792 | 0.0017  |
| SAT       | 0.0010       | 0.0002         | 4.3546  | 0.0002  |
| GPA       | 0.3780       | 0.1285         | 2.9428  | 0.0066  |

The estimated model is  $\hat{y} = -2.2461 + 0.0010 \text{ SAT} + 0.3780 \text{ GPA}$ . Since both *SAT* and *GPA* have  $p$ -values less than 0.05, we can conclude that the variables are individually significant at the 5% level.

b. The predicted probability of admission for an individual with a *GPA* of 3.5 and an *SAT* score of 1700,  $\hat{y} = -2.2461 + 0.0010(1700) + 0.3780(3.5) = 0.78$ .

c. With an *SAT* score of 1800, the predicted probability is  $\hat{y} = -2.2461 + 0.0010(1800) + 0.3780(3.5) = 0.88$ .

17.52 a. Relevant regression results:

|           | Coefficients | Standard Error | t Stat  | P-value |
|-----------|--------------|----------------|---------|---------|
| Intercept | 1.3863       | 0.3691         | 3.7557  | 0.0016  |
| Age       | -0.0360      | 0.0099         | -3.6363 | 0.0020  |
| Gender    | 0.2511       | 0.1734         | 1.4479  | 0.1658  |

b. *Age* has a  $p$ -value of 0.0020, and is therefore significant in explaining the probability of returning to crime. However, *Gender* has a  $p$ -value of 0.1658 and is therefore not



significant at even the 10% level. Therefore, the claim that women are less likely to re-offend is not supported.

- c. For a 25 year-old male, let  $x_1 = 25$  and  $x_2 = 1$ . The predicted probability of re-offending  $\hat{y} = 1.3863 - 0.0360(25) + 0.2511(1) = 0.74$ . The corresponding prediction for a female is  $\hat{y} = 1.3863 - 0.0360(25) + 0.2511(0) = 0.49$ .

## Chapter 18

- 18.2 a. The graph displays a jagged appearance, indicative of a significant random component in the series. The 5-period MA is much smoother.

b. 5-period MA

| $t$      | $y_t$    | $\bar{y}$ | $\hat{y}_t$ | $e_t = y_t - \hat{y}_t$ |
|----------|----------|-----------|-------------|-------------------------|
| 1        | 27       | —         | —           | —                       |
| 2        | 35       | —         | —           | —                       |
| 3        | 38       | 33.4      | —           | —                       |
| $\vdots$ | $\vdots$ | $\vdots$  | $\vdots$    | $\vdots$                |
| 6        | 39       | 36.8      | 33.4        | 5.6                     |
| $\vdots$ | $\vdots$ | $\vdots$  | $\vdots$    | $\vdots$                |
| 20       | 45       | —         | 36.00       | 9.0                     |
| 21       |          |           | <b>37</b>   |                         |

$$MSE = \frac{443.04}{15} = 29.54; MAD = \frac{68.80}{15} = 4.59$$

c.  $\hat{y}_{21} = 37$

- 18.4 a. 3-period MA

| $t$      | $y_t$    | $\bar{y}$ | $\hat{y}_t$ | $e_t = y_t - \hat{y}_t$ |
|----------|----------|-----------|-------------|-------------------------|
| 1        | 14       | —         | —           | —                       |
| 2        | 17       | 14.33     | —           | —                       |
| 3        | 12       | 15        | —           | —                       |
| 4        | 16       | 15.33     | 14.33       | 1.67                    |
| $\vdots$ | $\vdots$ | $\vdots$  | $\vdots$    | $\vdots$                |
| 20       | 20       | —         | 24          | -4.00                   |
| 21       |          |           | <b>23</b>   |                         |

$$MSE = \frac{155.33}{17} = 9.14; MAD = \frac{44.67}{17} = 2.63$$

b. Exponential smoothing with  $\alpha = 0.4$

| $t$      | $y_t$    | $A_t$    | $\hat{y}_t$  | $e_t = y_t - \hat{y}_t$ |
|----------|----------|----------|--------------|-------------------------|
| 1        | 14       | 14.00    | —            | —                       |
| 2        | 17       | 15.20    | 14.00        | 3.00                    |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$     | $\vdots$                |
| 20       | 20       | 22.13    | 23.55        | -3.55                   |
| 21       |          |          | <b>22.13</b> |                         |

$$MSE = \frac{156.4}{19} = 8.23; MAD = \frac{45.73}{19} = 2.41$$

c. The exponential smoothing method has lower  $MSE$  and a lower  $MAD$ ;  $\hat{y}_{21} = 22.13$ .

- 18.6 a. Exponential smoothing with  $\alpha = 0.4$

| $t$      | $y_t$    | $A_t$    | $\hat{y}_t$  | $e_t = y_t - \hat{y}_t$ |
|----------|----------|----------|--------------|-------------------------|
| 1991     | 34.8     | 34.80    | —            | —                       |
| 1992     | 31.6     | 33.52    | 34.80        | -3.20                   |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$     | $\vdots$                |
| 2008     | 31.8     | 31.32    | 31.00        | 0.80                    |
| 2009     |          |          | <b>31.32</b> |                         |

$$MSE = \frac{219.16}{17} = 12.89; \hat{y}_{2009} = 31.32$$

b. Exponential smoothing with  $\alpha = 0.6$

| $t$      | $y_t$    | $A_t$    | $\hat{y}_t$  | $e_t = y_t - \hat{y}_t$ |
|----------|----------|----------|--------------|-------------------------|
| 1991     | 34.8     | 34.80    | —            | —                       |
| 1992     | 31.6     | 32.88    | 34.80        | -3.20                   |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$     | $\vdots$                |
| 2008     | 31.8     | 31.95    | 32.18        | -0.38                   |
| 2009     |          |          | <b>31.95</b> |                         |

$$MSE = \frac{178.79}{17} = 10.52; \hat{y}_{2009} = 31.95$$

c. Exponential smoothing with  $\alpha = 0.6$  since its  $MSE$  is smaller.

- 18.8 a. 3-period MA

| $t$      | $y_t$    | $\bar{y}$ | $\hat{y}_t$    | $e_t = y_t - \hat{y}_t$ |
|----------|----------|-----------|----------------|-------------------------|
| 1-Nov    | 1184.38  | —         | —              | —                       |
| 2-Nov    | 1193.57  | 1191.97   | —              | —                       |
| 3-Nov    | 1197.96  | 1204.2    | —              | —                       |
| 4-Nov    | 1221.06  | 1214.96   | 1191.97        | 29.09                   |
| $\vdots$ | $\vdots$ | $\vdots$  | $\vdots$       | $\vdots$                |
| 30-Nov   | 1180.55  | —         | 1191.84        | -11.29                  |
| 1-Dec    |          |           | <b>1185.90</b> |                         |

$$MSE = \frac{3616.579}{18} = 200.92; \hat{y}_{\text{December 1}} = 1185.90$$

b. Exponential smoothing with  $\alpha = 0.4$

| $t$      | $y_t$    | $A_t$    | $\hat{y}_t$    | $e_t = y_t - \hat{y}_t$ |
|----------|----------|----------|----------------|-------------------------|
| 1-Nov    | 1184.38  | 1184.38  |                |                         |
| 2-Nov    | 1193.57  | 1188.06  | 1184.38        | 9.19                    |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$       | $\vdots$                |
| 30-Nov   | 1180.55  | 1186.30  | 1190.14        | -9.59                   |
| 1-Dec    |          |          | <b>1186.30</b> |                         |

$$MSE = \frac{3562.81}{20} = 178.14; \hat{y}_{\text{December 1}} = 1186.30$$

c. Exponential smoothing method

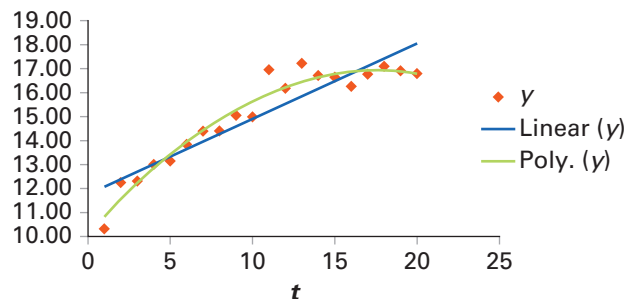
d. Yes, the December 1 absolute errors induced by the methods are 20.17 and 19.77, respectively.

- 18.10 a.  $\hat{y}_{21} = 13.54 + 1.08(21) = 36.22$

$$b. \hat{y}_{21} = 18.26 + 0.92(21) - 0.01(21)^2 = 33.19$$

$$c. \hat{y}_t = \exp\left(1.8 + 0.09(21) + \frac{0.01^2}{2}\right) = 40.05$$

- 18.12 a. Series  $y_t$  with linear and quadratic trends:



Compared to the linear trend, the quadratic trend appears to fit the data better.

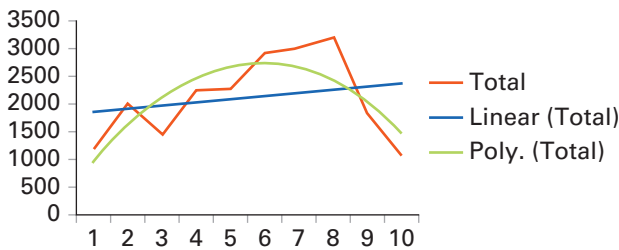
b. Summary Regression Output:

| Variable       | Linear Trend       | Quadratic Trend    |
|----------------|--------------------|--------------------|
| Intercept      | 11.7606*<br>(0.00) | 10.0651*<br>(0.00) |
| $t$            | 0.3144*<br>(0.00)  | 0.7768*<br>(0.00)  |
| $t^2$          | NA                 | -0.0220*<br>(0.00) |
| Adjusted $R^2$ | 0.8356             | 0.9481             |

Notes: Parameter estimates are in the top part of the table with the  $p$ -values in parentheses; NA denotes not applicable; \* represents significance at the 5% level. The last row presents adjusted  $R^2$  for model comparison.

The adjusted  $R^2$  is higher for the quadratic trend model, 94.81% versus 83.56% for the linear trend model. Therefore, the quadratic trend model describes the data better, which confirms our initial guess.

18.14 a. Series  $y_t$  with superimposed linear and quadratic trends:



The data roughly have an inverted U-shaped trend. A quadratic trend seems more suitable.

b. Summary Regression Output:

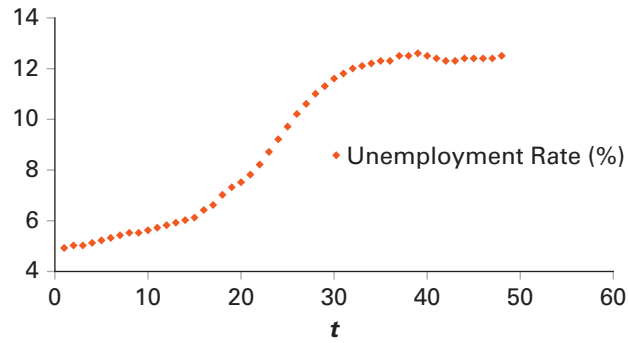
| Variable       | Linear Trend         | Quadratic Trend     |
|----------------|----------------------|---------------------|
| Intercept      | 1806.5333*<br>(0.01) | 153.7000<br>(0.81)  |
| $t$            | 57.7939<br>(0.52)    | 884.2106*<br>(0.01) |
| $t^2$          | NA                   | -75.1288*<br>(0.01) |
| Adjusted $R^2$ | -0.0643              | 0.5342              |

Notes: Parameter estimates are in the top part of the table with the  $p$ -values in parentheses; NA denotes not applicable; \* represents significance at 5% level. The last row presents adjusted  $R^2$  for model comparison.

As suggested by the above graph, the quadratic model is more suitable as it has an overwhelmingly greater adjusted  $R^2$  of 0.5342 versus for the linear model.

- c. For Week 27, or  $t = 11$ , and with the quadratic trend,  
 $\hat{y} = 153.7000 + 884.2106(11) - 75.1288(11)^2 = 789.43$

18.16 a. Scatterplot of monthly unemployment rate in California, Jan 2007–Dec. 2010



The cubic trend model seems to be the most appropriate.

| Summary Regression Output for Linear, Quadratic, and Cubic trends |                   |                    |                    |
|---|-------------------|--------------------|--------------------|
| Variable  | Linear            | Quadratic          | Cubic              |
| Intercept   | 3.9919*<br>(0.00) | 3.3808*<br>(0.00)  | 5.5027*<br>(0.00)  |
| $t$   | 0.2087*<br>(0.00) | 0.2820*<br>(0.00)  | -0.2123*<br>(0.00) |
| $t^2$   | NA                | -0.0015*<br>(0.03) | 0.0235*<br>(0.00)  |
| $t^3$   | NA                | NA                 | -0.0003*<br>(0.00) |
| Adjusted $R^2$  | 0.9242            | 0.9302             | 0.9879             |

Notes: Parameter estimates are in the top part of the table with the  $p$ -values in parentheses; NA denotes not applicable; \* represents significance at 5% level. The last row presents adjusted  $R^2$  for model comparison.

Compared to the linear and quadratic trends, the cubic model is most suitable as it has highest adjusted  $R^2$  of 98.79%.

- c. The cubic trend is used to forecast unemployment rate in California for January 2011, or  $t = 49$  as

$$\hat{y} = 5.5027 - 0.2123(49) + 0.0235(49)^2 - 0.0003(49)^3 = 11.48\%$$

18.18 a. Third and fourth quarter series are 14% below and 12% above the average quarterly level, respectively.

- b. Q1:  $\hat{y}_{33} = \exp(2.80 + 0.03(33) + 0.08^2/2) \times 0.94 = 41.73$   
 Q2:  $\hat{y}_{34} = \exp(2.80 + 0.03(34) + 0.08^2/2) \times 1.08 = 49.41$   
 Q3:  $\hat{y}_{35} = \exp(2.80 + 0.03(35) + 0.08^2/2) \times 0.86 = 40.54$   
 Q4:  $\hat{y}_{36} = \exp(2.80 + 0.03(36) + 0.08^2/2) \times 1.12 = 54.41$

18.20 a–b. Calculating the ratio-to-moving averages:

| Year | Quarter | $t$ | $y$   | $\bar{y}$ | $y/\bar{y}$ |
|------|---------|-----|-------|-----------|-------------|
| 1    | 1       | 1   | 8.37  | —         | —           |
| 1    | 2       | 2   | 12.78 | —         | —           |
| 1    | 3       | 3   | 8.84  | 11.69     | 0.7564      |
| ⋮    | ⋮       | ⋮   | ⋮     | ⋮         | ⋮           |
| 5    | 2       | 18  | 10.58 | 11.59     | 0.9130      |
| 5    | 3       | 19  | 13.35 | —         | —           |
| 5    | 4       | 20  | 19.77 | —         | —           |

c. Seasonal indices:

| Quarter    | Quarter 1 | Quarter 2 | Quarter 3 | Quarter 4 |
|------------|-----------|-----------|-----------|-----------|
| Unadjusted | 0.7412    | 0.8995    | 0.8145    | 1.4629    |
| Adjusted   | 0.7567    | 0.9183    | 0.8315    | 1.4935    |

First quarter series is 24.33% below its average quarterly level; fourth quarter series is 49.35% above its average quarterly level.

18.22 b. Calculating the ratio-to-moving averages:

| Year   | Quarter   | t  | y     | $\bar{y}$ | $y/\bar{y}$ |
|--------|-----------|----|-------|-----------|-------------|
| Year 1 | Quarter 1 | 1  | 6.49  | —         | —           |
| Year 1 | Quarter 2 | 2  | 7.34  | —         | —           |
| Year 1 | Quarter 3 | 3  | 7.11  | 8.01      | 0.8878      |
| ⋮      | ⋮         | ⋮  | ⋮     | ⋮         | ⋮           |
| Year 5 | Quarter 2 | 18 | 10.08 | 10.78     | 0.9353      |
| Year 5 | Quarter 3 | 19 | 9.78  | —         | —           |
| Year 5 | Quarter 4 | 20 | 14.88 | —         | —           |

Seasonal indices:

| Quarter    | Quarter 1 | Quarter 2 | Quarter 3 | Quarter 4 |
|------------|-----------|-----------|-----------|-----------|
| Unadjusted | 0.8484    | 0.9345    | 0.8880    | 1.3278    |
| Adjusted   | 0.8487    | 0.9348    | 0.8883    | 1.3282    |

Estimated exponential trend model with seasonal indices:

$$\hat{y}_t = \exp(2.017531 + 0.019950t + s_e^2/2) \times s_t, \text{ where}$$

$$s_e = 0.0018867; MSE = \frac{0.0041}{20} = 0.0002$$

c. Estimated exponential trend model with seasonal dummies:

$$\hat{y}_t = \exp(2.301381 - 0.449995d_1 - 0.350265d_2 - 0.400063d_3 + 0.019921t + s_e^2/2), \text{ where } s_e = 0.000971;$$

$$MSE = \frac{0.00093}{20} = 0.00005$$

d. The estimated exponential trend model with dummies has a lower MSE. Forecasts:

| Year | Quarter | $d_1$ | $d_2$ | $d_3$ | t  | $\hat{y}_t$ |
|------|---------|-------|-------|-------|----|-------------|
| 6    | 1       | 1     | 0     | 0     | 21 | 9.68        |
| 6    | 2       | 0     | 1     | 0     | 22 | 10.91       |
| 6    | 3       | 0     | 0     | 1     | 23 | 10.59       |
| 6    | 4       | 0     | 0     | 0     | 24 | 16.11       |

18.24 a. Estimated linear trend model with seasonal dummies:

$$\hat{y}_t = 390.6250 - 1683.9410d_1 - 1967.4736d_2 - 1743.4063d_3 - 1590.5389d_4 - 1400.8715d_5 - 1616.2042d_6 - 1480.5368d_7 - 1071.6694d_8 - 1506.0021d_9 - 995.1347d_{10} - 505.2674d_{11} + 118.9326t;$$

$$MSE = \frac{59373548.39}{60} = 989559.14;$$

$$MAD = \frac{46905.90}{60} = 781.77$$

b. Estimated exponential trend model with seasonal dummies:

$$\hat{y}_t = \exp(6.3047 - 0.4901d_1 - 0.6134d_2 - 0.4474d_3 - 0.3839d_4 - 0.3035d_5 - 0.3803d_6 - 0.3195d_7 - 0.2112d_8 - 0.3568d_9 - 0.2304d_{10} - 0.1188d_{11} + 0.0502t + s_e^2/2), \text{ where } s_e = 0.042759;$$

$$MSE = \frac{159484.55}{60} = 26580.74;$$

$$MAD = \frac{5573.33}{60} = 92.89$$

The exponential model has lower MSE. Forecasts:

| Year | Month | $d_1$ | $d_2$ | t  | $\hat{y}_t$ |
|------|-------|-------|-------|----|-------------|
| 6    | Jan   | 1     | 0     | 61 | 7186.15     |
| 6    | Feb   | 0     | 1     | 62 | 6679.55     |

18.26 b. Calculating the ratio-to-moving averages:

| Year | Month | t  | y    | $\bar{y}$ | $y/\bar{y}$ |
|------|-------|----|------|-----------|-------------|
| 2006 | Jan   | 1  | 4.65 | —         | —           |
| ⋮    | ⋮     | ⋮  | ⋮    | —         | —           |
| 2006 | Jul   | 7  | 5.25 | 5.01      | 1.0490      |
| ⋮    | ⋮     | ⋮  | ⋮    | —         | —           |
| 2010 | Jun   | 54 | 3.95 | 4.04      | 0.9779      |
| ⋮    | ⋮     | ⋮  | ⋮    | —         | —           |
| 2010 | Dec   | 60 | 4.16 | —         | —           |

Seasonal indices:

| Month      | Jan    | Feb    | Mar    | Apr    | May    | Jun    |
|------------|--------|--------|--------|--------|--------|--------|
| Unadjusted | 0.9582 | 0.9935 | 0.9852 | 1.0100 | 1.0243 | 1.0595 |
| Adjusted   | 0.9550 | 0.9902 | 0.9819 | 1.0066 | 1.0209 | 1.0559 |

| Month      | Jul    | Aug    | Sep    | Oct    | Nov    | Dec    |
|------------|--------|--------|--------|--------|--------|--------|
| Unadjusted | 1.0588 | 1.0332 | 0.9955 | 1.0071 | 0.9831 | 0.9322 |
| Adjusted   | 1.0552 | 1.0297 | 0.9921 | 1.0037 | 0.9798 | 0.9290 |

$$c. \hat{T}_t = 5.160547 - 0.02228t$$

| Year | Month | t  | $\hat{S}_t$ | $\hat{T}_t$ | $\hat{y}_t = \hat{T}_t \times \hat{S}_t$ |
|------|-------|----|-------------|-------------|--|
| 2011 | Jan   | 61 | 0.9550      | 3.8017      | 3.6306                                   |
| 2011 | Feb   | 62 | 0.9902      | 3.7794      | 3.7422                                   |
| 2011 | Mar   | 63 | 0.9819      | 3.7572      | 3.6892                                   |

18.28 a. Estimated linear trend model with seasonal dummies:

$$\hat{y}_t = 1947.5250 + 196.9188d_1 + 324.6125d_2 + 55.7063d_3 + 64.9063t;$$

$$MSE = \frac{115438.38}{20} = 5771.92;$$

$$MAD = \frac{1164.85}{20} = 58.24$$

b. Estimated exponential trend model with seasonal dummies:

$$\hat{y}_t = \exp(7.615911 + 0.069376d_1 + 0.112936d_2 + 0.020109d_3 + 0.023939t + s_e^2/2), \text{ where } s_e = 0.037138;$$

$$MSE = \frac{147176.96}{20} = 7358.85;$$

$$MAD = \frac{1372.20}{20} = 68.61$$

c. The linear trend model outperforms the exponential model.

18.30 a. Estimated linear trend model with seasonal dummies:

$$\hat{y}_t = 1.4702 + 0.0174d_1 - 0.0729d_2 - 0.0630d_3 - 0.0137t; \text{ adjusted } R^2 = 0.8560;$$

- b. Estimated quadratic trend model with seasonal dummies:

$$\hat{y}_t = 1.4155 + 0.0174d_1 - 0.0734d_2 - 0.0636d_3 - 0.0040t - 0.0003t^2; \text{ adjusted } R^2 = 0.8800$$

- c. The quadratic model is used for 2009 forecasts as it has greater adjusted  $R^2$ .

| Year | Quarter | $\hat{y}_t$ |
|------|---------|-------------|
| 2009 | 1       | 0.9802      |
| 2009 | 2       | 0.8656      |
| 2009 | 3       | 0.8511      |
| 2009 | 4       | 0.8898      |

18.32 a.  $\hat{y}_t = 86.071 + 1.1413t - 0.0630t^2 + 0.0006t^3 - 3.3372d_1 - 0.7017d_2 - 1.0219d_3 - 2.2938d_4 - 3.1546d_5 - 0.9083d_6 - 1.3086d_7 - 3.6762d_8 - 2.7149d_9 - 4.8952d_{10} - 4.0952d_{11}$ .

| Year | Month | $\hat{y}_t$ |
|------|-------|-------------|
| 2010 | Nov   | 77.055      |
| 2010 | Dec   | 83.178      |

18.34 a.  $\hat{y}_t = 11.4715 + 0.7569y_{t-1}$ ; adjusted  $R^2 = 0.8506$ ;  $\hat{y}_{25} = 48.24$

b.  $\hat{y}_t = 18.1852 + 0.3321y_{t-1} + 0.2819y_{t-2}$ ; adjusted  $R^2 = 0.7958$ ;  $\hat{y}_{25} = 47.88$

- c. Model estimated in a. has greater adjusted  $R^2$

18.36 a.  $\hat{y}_t = 113.6048 - 0.0064x_{t-1}$ ; adjusted  $R^2 = -0.0371$

b.  $\hat{y}_t = 3.9326 + 0.9201y_{t-1}$ ; adjusted  $R^2 = 0.7930$

c.  $\hat{y}_t = 56.8010 - 0.0057x_{t-1} + 0.9124y_{t-1}$ ; adjusted  $R^2 = 0.8324$

- d. Model in c. yields  $\hat{y}_{13} = 52.18$

18.38 a. Let  $y$  = Inflation:  $\hat{y}_t = 9.5846 - 0.2671y_{t-1}$ ; adjusted  $R^2 = -0.0351$

$\hat{y}_t = 9.6169 - 0.0381y_{t-1} - 0.0379y_{t-2}$ ; adjusted  $R^2 = -0.1098$

Given the negative adjusted  $R^2$  in both models, neither should be considered.

b. Let  $y$  = Unemployment:  $\hat{y}_t = 2.7465 + 0.7193y_{t-1}$ ; adjusted  $R^2 = 0.8902$

$\hat{y}_t = 3.0236 + 0.7224y_{t-1} - 0.0320y_{t-2}$ ; adjusted  $R^2 = 0.7987$

The estimated model with greater  $R^2$  yields  $\hat{y}_{24} = 9.80$ .

18.40 a.  $\hat{y}_{22} = 2.827$ ;  $MSE = \frac{0.141}{18} = 0.008$ ;  $MAD = \frac{1.241}{18} = 0.069$

b.  $\hat{y}_{22} = 2.819$ ;  $MSE = \frac{0.126}{20} = 0.006$ ;  $MAD = \frac{1.196}{20} = 0.060$

- c. The exponential smoothing method has slightly lower  $MSE$  and  $MAD$ .

- d. For  $t = 22$ ,  $|y_t - \hat{y}_t|$  are 0.133 and 0.141, respectively. Thus, rather surprisingly, the error induced by the 3-period MA method is smaller.

- 18.42 a. Seasonal indices found by the ratio-to-moving average method:

| Quarter    | Q1       | Q2       | Q3      | Q4       |
|------------|----------|----------|---------|----------|
| Unadjusted | 0.997561 | 1.000594 | 1.00194 | 0.999119 |
| Adjusted   | 0.9978   | 1.0008   | 1.0021  | 0.9993   |

b.  $\hat{T}_t = 636544.16 + 16324.26t - 1052.97t^2 + 18.60t^3$

- c.

| Year | Quarter | $t$ | $\hat{T}_t$ | $\hat{S}_t$ | $\hat{y}_t = \hat{T}_t \times \hat{S}_t$ |
|------|---------|-----|-------------|-------------|--|
| 2010 | 4       | 28  | 676354.7    | 0.9993      | 675892.8                                 |
| 2011 | 1       | 29  | 677982.8    | 0.9978      | 676463.1                                 |
| 2011 | 2       | 30  | 680740.9    | 1.0008      | 681279.6                                 |
| 2011 | 3       | 31  | 684740.7    | 1.0021      | 686201.3                                 |

- 18.44 a. The graph shows that loans were rising exponentially.

b.  $\hat{y}_t = -3.1320 + 74.5887t$ ;  $MSE = \frac{4080839.12}{36} = 113,356.64$

$\hat{y}_t = \exp(6.0545 + 0.0543t + s_e^2/2)$ , where  $s_e = 0.1082$ ;

$MSE = \frac{\sum e_t^2}{n} = \frac{1275107.37}{36} = 35,419.65$ . The estimated exponential trend model has lower  $MSE$ , and it yields the forecast  $\hat{y}_{37} = 3,189.63$  billion.

- c.  $\hat{y}_t = -37.2161 + 1.0961y_{t-1}$ ;  $MSE = 4104.42$ . The estimated AR(1) model has the lowest  $MSE$ , and it yields the forecast  $\hat{y}_{37} = 3,913.21$  billion.

- 18.46 a. The graph shows a relatively flat trend with a succession of almost identical spikes in July of each year, indicative of the presence of a strong seasonality in the series.

- b. Seasonal indices found by the ratio-to-moving average method:

| Month      | Jan    | Feb    | Mar    | Apr    | May    | Jun    |
|------------|--------|--------|--------|--------|--------|--------|
| Unadjusted | 0.8924 | 0.8471 | 1.0514 | 1.0065 | 1.0368 | 1.0998 |
| Adjusted   | 0.8933 | 0.8480 | 1.0525 | 1.0075 | 1.0379 | 1.1009 |

| Month      | Jul    | Aug    | Sep    | Oct    | Nov    | Dec    |
|------------|--------|--------|--------|--------|--------|--------|
| Unadjusted | 1.1528 | 1.1138 | 0.9049 | 0.9858 | 0.9303 | 0.9657 |
| Adjusted   | 1.1540 | 1.1150 | 0.9058 | 0.9869 | 0.9313 | 0.9667 |

- c. The cubic trend model is chosen to fit the seasonally adjusted series:  $\hat{T}_t = 45.8490 + 0.5940t - 0.0271t^2 + 0.0003t^3$ ; adjusted  $R^2 = 0.6924$

- d.

| Year | Month | $t$ | $\hat{T}_t$ | $\hat{S}_t$ | $\hat{y}_t = \hat{T}_t \times \hat{S}_t$ |
|------|-------|-----|-------------|-------------|--|
| 2010 | Oct   | 58  | 47.4182     | 0.9869      | 46.7951                                  |
| 2010 | Nov   | 59  | 47.9089     | 0.9313      | 44.6154                                  |
| 2010 | Dec   | 60  | 48.4509     | 0.9667      | 46.8383                                  |

18.48 a.  $\hat{y}_t = 7909.4250 + 1478.7938d_1 + 3159.8625d_2 + 982.9313d_3 + 176.3313t$

$MSE = \frac{9,634,206.98}{20} = 481,710.35$ ;  $MAD = \frac{11,258.05}{20} = 562.90$

b.  $\hat{y}_t = \exp(9.0170 + 0.1340d_1 + 0.2739d_2 + 0.0924d_3 + 0.0161t + s_e^2/2)$ , where  $s_e = 0.07293$ ;

$MSE = \frac{9099121.4346}{20} = 454,956.07$ ;

$MAD = \frac{10,953.9391}{20} = 547.70$

- c. Exponential model; smaller  $MSE$  and  $MAD$

| Year  | Quarter | Lowe's Net Sales |
|-------|---------|------------------|
| 2009  | 1       | 13237.42         |
| 2009  | 2       | 15471.72         |
| 2009  | 3       | 13112.69         |
| 2009  | 4       | 12148.44         |
| Total |         | 53970.26         |

- 18.50 a. The cubic trend model chosen to fit the series:  
 $\hat{y}_t = 8773.5228 + 287.3408t - 23.0026t^2 + 0.6474t^3$ ; adjusted  $R^2 = 0.9236$ ;  $\hat{y}_{21} = 10,659.50$   
 b.  $\hat{y}_t = 3564.9344 + 0.5974x_{t-1}$ ; adjusted  $R^2 = 0.8718$ ;  $\hat{y}_{21} = 10,444.34$   
 c. The cubic model has greater adjusted  $R^2$ .

## Chapter 19

19.2 a.  $\frac{\$60}{\$980} = 0.0612$ , or 6.12%

b.  $\frac{\$990 - \$980}{\$980} = 0.0102$ , or 1.02%

c.  $R_t = 7.14\%$

19.4 a. Year 2:  $\frac{\$24.80 - \$23.50 + \$0.18}{\$23.50} = 0.0630$ ;

Year 3:  $\frac{\$22.90 - \$24.80 + \$0.12}{\$24.80} = -0.0718$

b. Year 2:  $r = \frac{1 + 0.0630}{1 + 0.028} - 1 = 0.034$ ;

Year 3:  $r = (1 - 0.0718)/(1 + 0.016) - 1 = -0.0864$

19.6 a.  $R_t = \frac{\$30,480 - \$17,100 + \$520}{\$17,100} = 0.8129$ , or 81.29%

b. \$13,901

19.8

| Date   | Adjusted Close Price | Monthly Return |
|--------|----------------------|----------------|
| Dec-10 | \$21.48              | 0.0238         |
| Nov-10 | \$20.98              | 0.0634         |
| Oct-10 | \$19.73              | —              |

19.10 a.

| Year | Price | Price Index |
|------|-------|-------------|
| 1994 | 62    | 100.00      |
| 1995 | 60    | 96.77       |
| 1996 | 64    | 103.23      |
| 1997 | 67    | 108.06      |
| 1998 | 66    | 106.45      |
| 1999 | 70    | 112.90      |
| 2000 | 74    | 119.35      |
| 2001 | 72    | 116.13      |
| 2002 | 70    | 112.90      |

b. Increase of 6.45 percent.

19.12 a.

| Year | Simple Price Index |           |           |
|------|--------------------|-----------|-----------|
|      | Product 1          | Product 2 | Product 3 |
| 2008 | 100.00             | 100.00    | 100.00    |
| 2009 | 104.20             | 98.56     | 102.21    |
| 2010 | 108.39             | 99.28     | 98.90     |

b. Product 1's price increased each year, while product 2's price decreased each year. Product 3's price increased in 2009, but then decreased in 2010.

19.14 a.

| Month              | Jan    | Feb   | Mar    | Apr    | May    | Jun    |
|--------------------|--------|-------|--------|--------|--------|--------|
| Price              | 3.25   | 3.18  | 3.56   | 3.82   | 3.97   | 4.48   |
| Simple Price Index | 100.00 | 97.85 | 109.54 | 117.54 | 122.15 | 137.85 |

| Month              | Jul    | Aug    | Sep    | Oct    | Nov   | Dec   |
|--------------------|--------|--------|--------|--------|-------|-------|
| Price              | 4.46   | 4.16   | 3.79   | 3.39   | 2.46  | 1.82  |
| Simple Price Index | 137.23 | 128.00 | 116.62 | 104.31 | 75.69 | 56.00 |

b. The price rose by 37.85 percent.

19.16 a.

| Year | Tuition  | Simple Price Index (Base = 2004) |
|------|----------|----------------------------------|
| 2004 | \$36,850 | 100.00                           |
| 2005 | \$39,844 | 108.12                           |
| 2006 | \$42,634 | 115.70                           |
| 2007 | \$44,556 | 120.91                           |
| 2008 | \$46,784 | 126.96                           |
| 2009 | \$48,650 | 132.02                           |

b.

| Year | Simple Price index (Base = 2004) | Updated Price Index (Base = 2007) |
|------|----------------------------------|-----------------------------------|
| 2004 | 100.00                           | 82.70                             |
| 2005 | 108.12                           | 89.42                             |
| 2006 | 115.70                           | 95.69                             |
| 2007 | 120.91                           | 100.00                            |
| 2008 | 126.96                           | 105.00                            |
| 2009 | 132.02                           | 109.19                            |

c. Tuition increased by 20.91 percent from 2004 through 2007, but only 9.19 percent from 2007 through 2009.

19.18 a. Relative to 2007, the 2009 prices of omelet, pancake, and cereal increased by 5.26%, 28.57%, and 21.43%, respectively.

| Year | Simple Price Index |          |        |
|------|--------------------|----------|--------|
|      | Omelet             | Pancakes | Cereal |
| 2007 | 100.00             | 100.00   | 100.00 |
| 2008 | 110.53             | 121.43   | 114.29 |
| 2009 | 105.26             | 128.57   | 121.43 |

b. Relative to 2007, the prices of the three breakfast items increased by 14.89% and 17.02% in 2008 and 2009, respectively.

| Year | $\sum p_{it}$ | Unweighted Aggregate Price Index |
|------|---------------|----------------------------------|
| 2007 | 11.75         | 100.00                           |
| 2008 | 13.50         | 114.89                           |
| 2009 | 13.75         | 117.02                           |

19.20 a.

| Region    | Simple Price Index |       |       |
|-----------|--------------------|-------|-------|
|           | 2007               | 2008  | 2009  |
| Northeast | 100.00             | 94.24 | 83.55 |
| Midwest   | 100.00             | 93.25 | 88.29 |
| South     | 100.00             | 94.74 | 86.47 |
| West      | 100.00             | 80.61 | 65.46 |

- b. Home prices dropped significantly; more drastically in the West.

19.22

| Nominal Value | Price Index | Real Value |
|---------------|-------------|------------|
| 32            | 100         | 32.00      |
| 37            | 102         | 36.27      |
| 39            | 103         | 37.86      |
| 42            | 108         | 38.89      |

19.24 Real revenue increases to  $\frac{\$110,000}{104} \times 100 = \$105,769$ ,

resulting in an increase of  $\frac{105,769 - 100,000}{100,000} \times 100 = 5.77\%$ .

19.26 a-b.

| Year | Nominal Value | Price Index | Real Value |
|------|---------------|-------------|------------|
| 2009 | 38            | 100         | 38.00      |
| 2010 | 40            | 103         | 38.83      |
| 2011 | 42            | 112         | 37.50      |

2009–2010: Nominal values increase by 5.26 percent. Real values increase by 2.18 percent.

2010–2011: Nominal values increase by 5 percent. Real values decrease by 3.43 percent.

c. 2010:  $\frac{103 - 100}{100} \times 100 = 3.00\%$ ; 2011:  $\frac{112 - 103}{103} \times 100 = 8.74\%$

19.28

| Year | CPI   | Inflation Rate (%) |
|------|-------|--------------------|
| 2001 | 120.1 | —                  |
| 2002 | 119.0 | −0.92              |
| 2003 | 118.7 | −0.25              |
| 2004 | 118.7 | 0.00               |
| 2005 | 118.3 | −0.34              |
| 2006 | 118.7 | 0.34               |
| 2007 | 118.7 | 0.00               |
| 2008 | 120.3 | 1.35               |
| 2009 | 118.7 | −1.33              |

Inflation in Japan has been negative four times, which supports the deflation claim.

19.30 a.

| Year | Revenue      | PPI (1982 = 100) | Real Revenue    |
|------|--------------|------------------|-----------------|
| 2007 | \$35,510,000 | 172.7            | \$20,561,667.63 |
| 2008 | \$37,843,000 | 189.6            | \$19,959,388.19 |
| 2009 | \$36,149,000 | 172.9            | \$20,907,460.96 |

b. Nominal revenue increases in 2008, but decreases in 2009. Real revenue decreases in 2008, but increases in 2009.

19.32 The inflation rate in 2009 is  $\left(\frac{214.54 - 215.3}{215.3}\right) \times 100 = -0.35\%$ .  
The 2009 starting salary must be reduced to \$89,156(0.9965) = \$88,844.

19.34 a.

| Date   | Adjusted Close Price | Simple Price Index (Base = Oct 09) |
|--------|----------------------|------------------------------------|
| Oct-09 | 78.89                | 100.00                             |
| Nov-09 | 78.54                | 99.56                              |
| Dec-09 | 84.16                | 106.68                             |
| Jan-10 | 77.00                | 97.60                              |
| Feb-10 | 74.83                | 94.85                              |
| Mar-10 | 79.56                | 100.85                             |

b.

| Date   | Adjusted Close Price | Updated Index (Base = Jan 2010) |
|--------|----------------------|---------------------------------|
| Oct-09 | 78.89                | 102.46                          |
| Nov-09 | 78.54                | 102.00                          |
| Dec-09 | 84.16                | 109.30                          |
| Jan-10 | 77.00                | 100.00                          |
| Feb-10 | 74.83                | 97.19                           |
| Mar-10 | 79.56                | 103.33                          |

c. Increase by 6.68 percent.

d. Increase by 3.33 percent.

19.36 a.

| Year | Simple Price Index (Base = 2009) |           |           |
|------|----------------------------------|-----------|-----------|
|      | Product 1                        | Product 2 | Product 3 |
| 2009 | 100.00                           | 100.00    | 100.00    |
| 2010 | 105.26                           | 97.87     | 106.67    |
| 2011 | 110.53                           | 104.26    | 124.44    |

b.

| Year | $\sum p_{it}$ | Unweighted Aggregate Price Index |
|------|---------------|----------------------------------|
| 2009 | 177           | 100.00                           |
| 2010 | 180           | 101.69                           |
| 2011 | 196           | 110.73                           |

19.38 a.

| Year | $\sum p_{it}$ | Unweighted Aggregate Price Index |
|------|---------------|----------------------------------|
| 2005 | 233.09        | 100.00                           |
| 2006 | 475.34        | 203.93                           |
| 2007 | 553.66        | 237.53                           |

b.

| Year | $\sum p_{it} q_{i0}$ | Laspeyres Price Index |
|------|----------------------|-----------------------|
| 2005 | 33475                | 100.00                |
| 2006 | 59378                | 177.38                |
| 2007 | 69064                | 206.32                |



- c. The unweighted index shows higher increases due primarily to Google's rise, but since Lindsay did not buy as much Google stock as the others, the weighted index shows smaller increases.

19.40 a.

| Year | Net Revenue | PPI (1982 = 100) | Real Net Revenue |
|------|-------------|------------------|------------------|
| 2006 | 146.6       | 164.8            | 88.96            |
| 2007 | 159.2       | 172.7            | 92.18            |
| 2008 | 105.8       | 189.6            | 55.80            |
| 2009 | 111.0       | 172.9            | 64.20            |

b.

| Year | Net Income | CPI (1982–84 = 100) | Real Net Income |
|------|------------|---------------------|-----------------|
| 2006 | 21.2       | 201.59              | 10.52           |
| 2007 | 3.6        | 207.34              | 1.74            |
| 2008 | -27.7      | 215.30              | -12.87          |
| 2009 | -1.6       | 214.54              | -0.75           |

19.42 a.

| Date           | Adjusted Price | CPI (Base 1982–1984) | Real Adjusted Price | Real Return |
|----------------|----------------|----------------------|---------------------|-------------|
| January, 2008  | 8.94           | 212.23               | 4.21                | —           |
| February, 2008 | 8.29           | 212.70               | 3.90                | -0.0748     |
| March, 2008    | 6.01           | 213.54               | 2.81                | -0.2779     |

b.

| Date           | Adjusted Price | Nominal Return | CPI (Base 1982–1984) | Inflation | Real Interest with Fisher (%) |
|----------------|----------------|----------------|----------------------|-----------|-------------------------------|
| January, 2008  | 8.94           | —              | 212.23               | —         | —                             |
| February, 2008 | 8.29           | -0.0727        | 212.70               | 0.0023    | -0.0748                       |
| March, 2008    | 6.01           | -0.2750        | 213.54               | 0.0039    | -0.2779                       |

## Chapter 20

20.2 a. Reject  $H_0$  if  $T \leq T_L = 3$

b.  $T = T^+ = 3 < 3 = T_L$ ; reject  $H_0$ . Population median is less than 150.

20.4 a.  $H_0: m \geq 10$ ;  $H_A: m < 10$

b.  $T = T^+ = 1$

c. Reject  $H_0$  if  $T \leq T_L = 2$

d.  $T = 1 < 2 = T_L$ ; reject  $H_0$ . Population median is less than 10.

20.8 a.  $H_0: m \leq 25$ ;  $H_A: m > 25$

| $x$ | $d = x - 25$ | $ d $ | Rank | Ranks of Negative Differences | Ranks of Positive Differences |
|-----|--------------|-------|------|-------------------------------|-------------------------------|
| 45  | 20           | 20    | 10   |                               | 10                            |
| 42  | 17           | 17    | 9    |                               | 9                             |
| 30  | 5            | 5     | 6.5  |                               | 6.5                           |
| 30  | 5            | 5     | 6.5  |                               | 6.5                           |
| 27  | 2            | 2     | 3    |                               | 3                             |
| 24  | -1           | 1     | 1.5  | 1.5                           |                               |
| 24  | -1           | 1     | 1.5  | 1.5                           |                               |
| 21  | -4           | 4     | 4    | 4                             |                               |
| 20  | -5           | 5     | 6.5  | 6.5                           |                               |
| 20  | -5           | 5     | 6.5  | 6.5                           |                               |
|     |              |       |      | $T^- = 20$                    | $T^+ = 35$                    |

b. The value of the test statistic  $T = T^+ = 35$ . In addition,  $T$  is assumed normally distributed with  $\mu_T = \frac{n(n+1)}{4} = \frac{10(10+1)}{4} =$

$$27.5 \text{ and } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{10 \times (10+1)(2 \times 10+1)}{24}} = 9.81. \text{ Therefore, } z = \frac{T - \mu_T}{\sigma_T} = \frac{35 - 27.5}{9.81} = 0.76$$

c. The corresponding  $p$ -value for a right-tailed test is  $P(Z \geq 0.76) = 1 - 0.7764 = 0.2236$ .

d. Since the  $p$ -value  $= 0.2236 > \alpha = 0.01$ , we do not reject  $H_0$ . Thus, at the 1% significance level we cannot conclude that the median rent exceeds \$25 per square foot.

20.12 a. Given a left-tailed test,  $\alpha = 0.05$  and  $n = 9$ ,  $T_L = 8$ . So, we will reject  $H_0$  if  $T \leq T_L = 8$ .

b. Since  $T = T^+ = 5 < 8$ , we reject  $H_0$ . At the 5% significance level, the population median difference is less than 0.

20.14 a.  $H_0: m_D \geq 0$ ;  $H_A: m_D < 0$

b. With  $\alpha = 0.05$  and  $n = 8$ ,  $T_L = 5$ . So, we will reject  $H_0$  if  $T \leq T_L = 5$ .

c.

| Number | Sample 1 ( $x$ ) | Sample 2 ( $y$ ) | $d = x - y$ | $ d $ | Rank | Ranks of Negative Differences | Ranks of Positive Differences |
|--------|------------------|------------------|-------------|-------|------|-------------------------------|-------------------------------|
| 1      | 18               | 21               | -3          | 3     | 4.5  | 4.5                           |                               |
| 2      | 12               | 11               | 1           | 1     | 1    |                               | 1                             |
| 3      | 21               | 23               | -2          | 2     | 2.5  | 2.5                           |                               |
| 4      | 22               | 20               | 2           | 2     | 2.5  |                               | 2.5                           |
| 5      | 16               | 20               | -4          | 4     | 6.5  | 6.5                           |                               |
| 6      | 14               | 17               | -3          | 3     | 4.5  | 4.5                           |                               |
| 7      | 17               | 17               | 0           | 0     | -    |                               |                               |
| 8      | 18               | 22               | -4          | 4     | 6.5  | 6.5                           |                               |
|        |                  |                  |             |       |      | $T^- = 24.5$                  | $T^+ = 3.5$                   |

The value of the test statistic is  $T = T^+ = 3.5$

d. Since  $T = 3.5 < 5$ , we reject  $H_0$ . At the 5% significance level, we conclude that the median difference between Population 1 and Population 2 is less than zero.

- 20.16 a. With  $\alpha = 0.05$ ,  $n_1 = 7$ , and  $n_2 = 8$ , the lower and upper critical values are  $W_L = 39$  and  $W_U = 73$ . Thus, the decision rule is to reject  $H_0$  if  $W < 39$  or  $W > 73$ .
- b. Given that  $n_1 = 7 < n_2 = 8$ , the value of the test statistic is  $W = W_1 = 80$ .
- c. Since  $W = 80 > 73$ , we reject  $H_0$ . At the 5% significance level, the median of Population 1 differs from the median of Population 2.
- 20.18 a.  $H_0: m_1 - m_2 \geq 0$ ;  $H_A: m_1 - m_2 < 0$
- b. With  $\alpha = 0.05$ ,  $n_1 = 5$ , and  $n_2 = 6$ , the lower-tailed critical values is  $W_L = 20$ . Thus, the decision rule is to reject the  $H_0$  if  $W \leq 20$ .
- c.

| Pooled Sample | Sample of original | Rank | Sample 1 Ranks | Sample 2 Ranks |
|---------------|--------------------|------|----------------|----------------|
| 15            | Sample 1           | 1    | 1              |                |
| 19            | Sample 1           | 2    | 2              |                |
| 23            | Sample 1           | 3    | 3              |                |
| 25            | Sample 2           | 4    |                | 4              |
| 28            | Sample 2           | 5    |                | 5              |
| 30            | Sample 1           | 6    | 6              |                |
| 34            | Sample 1           | 7.5  | 7.5            |                |
| 34            | Sample 2           | 7.5  |                | 7.5            |
| 35            | Sample 2           | 9    |                | 9              |
| 37            | Sample 2           | 10   |                | 10             |
| 40            | Sample 2           | 11   |                | 11             |
|               |                    |      | $W_1 = 19.5$   | $W_2 = 46.5$   |

Since  $n_1 = 5 < n_2 = 6$ , the value of the test statistic is  $W = W_1 = 19.5$

- d. Since  $W = 19.5 < 20$ , we reject  $H_0$ . Therefore, at the 5% significance level, we conclude that the median of Population 1 is less than the median of Population 2.
- 20.20 a. Since  $n_1 = 25 > n_2 = 20$ ,  $W = W_2 = 700$ .  $W$  is approximately normally distributed with the mean  $\mu_W = \frac{n_2(n_2 + n_1 + 1)}{2} = \frac{20(20 + 25 + 1)}{2} = 460$  and  $\sigma_W = \sqrt{\frac{n_2 n_1 (n_2 + n_1 + 1)}{12}} = \sqrt{\frac{20 \times 25 \times (20 + 25 + 1)}{12}} = 43.78$ .
- b.  $H_0: m_1 - m_2 \leq 0$ ;  $H_A: m_1 - m_2 > 0$
- c. With  $\alpha = 0.05$ ,  $z_\alpha = z_{0.05} = 1.645$ . Thus, the decision rule is to reject the  $H_0$  if  $z > 1.645$ .
- d. The value of the test statistic is  $z = \frac{W - \mu_W}{\sigma_W} = \frac{700 - 460}{43.78} = 5.48$ .
- e. Since  $z = 5.48 > 1.645$ , we reject  $H_0$ . Therefore, at the 5% significance level, the median of Population 1 is greater than the median of Population 2.
- 20.22 a. The differences are the real score minus the mock score  
 $H_0: m_D \leq 0$   
 $H_A: m_D > 0$
- b. With  $n = 8$  and  $\alpha = 0.05$ , Reject  $H_0$  if  $T^P > 31$ .
- c.  $T^P = 29$   
 Since  $29 < 31$ , do not reject  $H_0$ . At the 5% significance level, we cannot conclude the median score on the real SAT is greater than the median score on the mock SAT.
- 20.24 a.  $H_0: m_D = 0$ ;  $H_A: m_D \neq 0$

b.

| Number | Value from Appraiser 1 (x) | Value from Appraiser 2 (y) | $d = x - y$ | $ d $ | Rank | Ranks of Negative Differences | Ranks of Positive Differences |
|--------|----------------------------|----------------------------|-------------|-------|------|-------------------------------|-------------------------------|
| 1      | 235000                     | 239000                     | -4000       | 4000  | 5    | 5                             |                               |
| 2      | 195000                     | 190000                     | 5000        | 5000  | 6.5  |                               | 6.5                           |
| 3      | 264000                     | 271000                     | -7000       | 7000  | 8    | 8                             |                               |
| 4      | 315000                     | 310000                     | 5000        | 5000  | 6.5  |                               | 6.5                           |
| 5      | 435000                     | 437000                     | -2000       | 2000  | 2.5  | 2.5                           |                               |
| 6      | 515000                     | 525000                     | -10000      | 10000 | 10   | 10                            |                               |
| 7      | 350000                     | 352000                     | -2000       | 2000  | 2.5  | 2.5                           |                               |
| 8      | 225000                     | 224000                     | 1000        | 1000  | 1    |                               | 1                             |
| 9      | 437000                     | 440000                     | -3000       | 3000  | 4    | 4                             |                               |
| 10     | 575000                     | 583000                     | -8000       | 8000  | 9    | 9                             |                               |
|        |                            |                            |             |       |      | $T^- = 41$                    | $T^+ = 14$                    |

The value of the test statistic is  $T = T^+ = 14$ . Assuming that  $T$  is normally distributed with  $\mu_T = \frac{n(n+1)}{4} = \frac{10 \times (10+1)}{4} = 27.5$

$$\text{and } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{10(10+1)(2 \times 10 + 1)}{24}} = 9.81;$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{14 - 27.5}{9.81} = -1.38.$$

- c. The corresponding  $p$ -value  $= 2 \times P(Z \leq -1.38) = 2 \times 0.0838 = 0.1676$ .
- d. Since the  $p$ -value  $= 0.1676 > \alpha = 0.05$ , we do not reject  $H_0$ . Thus, at the 5% significance level, we cannot conclude that the median of the difference between the two appraisal values differs from zero. Therefore, the two appraisers are consistent.
- 20.26 a.  $H_0: m_1 - m_2 \leq 0$ ;  $H_A: m_1 - m_2 > 0$
- b. With  $\alpha = 0.05$ ,  $n_1 = 6$ , and  $n_2 = 7$ , the critical value is  $W_U = 54$ . Thus, the decision rule is to reject  $H_0$  if  $W \geq 54$ .

c.

| Pooled Sample | Sample of original | Rank | Sample 1 Ranks | Sample 2 Ranks |
|---------------|--------------------|------|----------------|----------------|
| 62            | Unmarried          | 1    |                | 1              |
| 63            | Unmarried          | 2    |                | 2              |
| 64            | Unmarried          | 3    |                | 3              |
| 66            | Unmarried          | 4    |                | 4              |
| 67            | Married            | 5.5  | 5.5            |                |
| 67            | Unmarried          | 5.5  |                | 5.5            |
| 68            | Unmarried          | 7    |                | 7              |
| 70            | Married            | 8    | 8              |                |
| 71            | Unmarried          | 9    |                | 9              |
| 75            | Married            | 10   | 10             |                |
| 76            | Married            | 11   | 11             |                |
| 83            | Married            | 12   | 12             |                |
| 84            | Married            | 13   | 13             |                |
|               |                    |      | $W_1 = 59.5$   | $W_2 = 31.5$   |

Since  $n_1 = 6 < n_2 = 7$ ,  $W = W_1 = 59.5$

- d. Since  $W = 59.5 > 54$ , we reject  $H_0$ . Therefore, at the 5% significance level, the median income of married men is higher than that of unmarried men. The claim is therefore supported by the sample data.
- 20.30 a.  $H_0: m_1 = m_2 = m_3 = m_4 = m_5$ ;  $H_A$ : Not all population medians are equal.
- b. With  $df = k - 1 = 5 - 1 = 4$ , the  $p$ -value  $= P(H > 12.4) = P(\chi^2_4 > 12.4)$  lies between 0.01 and 0.025.
- c. Since  $p$ -value  $< 0.05 = \alpha$ , we reject  $H_0$ . Therefore, we can conclude that the population medians are not all equal at the 5% significance level.

20.32 a.  $H_0: m_1 = m_2 = m_3 = m_4$ ;  $H_A$ : Not all population medians are equal.

b.

| Pooled Sample | Origin Sample | Ranks            | Sample 1 Ranks | Sample 2 Ranks | Sample 3 Ranks | Sample 4 Ranks |
|---------------|---------------|------------------|----------------|----------------|----------------|----------------|
| -15           | Sample 1      | 1                | 1              |                |                |                |
| -10           | Sample 1      | 2                | 2              |                |                |                |
| -6            | Sample 2      | 3                |                | 3              |                |                |
| -5            | Sample 2      | 4                |                | 4              |                |                |
| -4            | Sample 2      | 5                |                | 5              |                |                |
| -3            | Sample 2      | 6                |                | 6              |                |                |
| -2            | Sample 2      | 7                |                | 7              |                |                |
| 0             | Sample 1      | 8                | 8              |                |                |                |
| 2             | Sample 3      | 9                |                |                | 9              |                |
| 4             | Sample 3      | 10               |                |                | 10             |                |
| 5             | Sample 1      | 11.5             | 11.5           |                |                |                |
| 5             | Sample 4      | 11.5             |                |                |                | 11.5           |
| 6             | Sample 3      | 13               |                |                | 13             |                |
| 7             | Sample 4      | 14               |                |                |                | 14             |
| 8             | Sample 3      | 15               |                |                | 15             |                |
| 9             | Sample 4      | 16               |                |                |                | 16             |
| 10            | Sample 1      | 17.5             | 17.5           |                |                |                |
| 10            | Sample 3      | 17.5             |                |                | 17.5           |                |
| 11            | Sample 4      | 19               |                |                |                | 19             |
| 13            | Sample 4      | 20               |                |                |                | 20             |
|               |               | $R_i$            | 40             | 25             | 64.5           | 80.5           |
|               |               | $R_i^2$          | 1600           | 625            | 4160.25        | 6480.25        |
|               |               | $R_i^2/n_i$      | 320            | 125            | 832.05         | 1296.05        |
|               |               | $\sum R_i^2/n_i$ | 2573.10        |                |                |                |

The value of the test statistic is

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1) = \left( \frac{12}{20 \times (20+1)} \times 2573.10 \right) - 3 \times (20+1) = 10.52.$$

- c. With  $df = k - 1 = 4 - 1 = 3$ , the  $p$ -value  $= P(H \geq 10.52) = P(\chi_3^2 \geq 10.52)$  lies strictly between 0.01 and 0.025.
- d. Since the  $p$ -value  $> 0.01 = \alpha$ , we do not reject  $H_0$ . Therefore, at the 1% significance level, we cannot conclude that the medians differ.

20.34 a.  $H_0: m_1 = m_2 = m_3$ ;  $H_A$ : Not all population medians are equal.

b.

| Pooled Sample | Origin Sample | Ranks            | Brand 1 Ranks | Brand 2 Ranks | Brand 3 Ranks |
|---------------|---------------|------------------|---------------|---------------|---------------|
| 280           | Brand 2       | 1                |               | 1             |               |
| 290           | Brand 2       | 2                |               | 2             |               |
| 300           | Brand 2       | 3                |               | 3             |               |
| 325           | Brand 2       | 4                |               | 4             |               |
| 350           | Brand 2       | 5.5              |               | 5.5           |               |
| 350           | Brand 3       | 5.5              |               |               | 5.5           |
| 375           | Brand 1       | 7                | 7             |               |               |
| 380           | Brand 3       | 8                |               |               | 8             |
| 400           | Brand 1       | 9                | 9             |               |               |
| 405           | Brand 3       | 10               |               |               | 10            |
| 410           | Brand 1       | 11               | 11            |               |               |
| 415           | Brand 3       | 12               |               |               | 12            |
| 420           | Brand 1       | 13               | 13            |               |               |
| 425           | Brand 1       | 14.5             | 14.5          |               |               |
| 425           | Brand 3       | 14.5             |               |               | 14.5          |
|               |               | $R_i$            | 54.5          | 15.5          | 50            |
|               |               | $R_i^2$          | 2970.25       | 240.25        | 2500          |
|               |               | $R_i^2/n_i$      | 594.05        | 48.05         | 500           |
|               |               | $\sum R_i^2/n_i$ | 1142.10       |               |               |

The value of the test statistic is

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1) = \left( \frac{12}{15 \times (15+1)} \times 1142.10 \right) - 3 \times (15+1) = 9.11.$$

- c. With  $df = k - 1 = 3 - 1 = 2$ , the  $p$ -value  $= P(H \geq 9.11) = P(\chi_2^2 \geq 9.11)$  lies between 0.01 and 0.025.
- d. Since the  $p$ -value  $< 0.10 = \alpha$ , we reject  $H_0$ . Therefore, at the 10% significance level we conclude that some differences exist in the median length of life of the three brands.

20.36 a.  $H_0: m_1 = m_2 = m_3$ ;  $H_A$ : Not all population medians are equal.

b.

| Pooled Sample | Origin Sample | Ranks            | Detergent 1 Ranks | Detergent 2 Ranks | Detergent 3 Ranks |
|---------------|---------------|------------------|-------------------|-------------------|-------------------|
| 74            | Detergent 2   | 1                |                   | 1                 |                   |
| 77            | Detergent 3   | 2                |                   |                   | 2                 |
| 78            | Detergent 2   | 3.5              |                   | 3.5               |                   |
| 78            | Detergent 3   | 3.5              |                   |                   | 3.5               |
| 79            | Detergent 1   | 5                | 5                 |                   |                   |
| 80            | Detergent 3   | 6                |                   |                   | 6                 |
| 81            | Detergent 2   | 7                |                   | 7                 |                   |
| 84            | Detergent 1   | 8                | 8                 |                   |                   |
| 85            | Detergent 1   | 9                | 9                 |                   |                   |
| 86            | Detergent 2   | 10.5             |                   | 10.5              |                   |
| 86            | Detergent 2   | 10.5             |                   | 10.5              |                   |
| 87            | Detergent 1   | 12.5             | 12.5              |                   |                   |
| 87            | Detergent 3   | 12.5             |                   |                   | 12.5              |
| 91            | Detergent 3   | 14               |                   |                   | 14                |
| 94            | Detergent 1   | 15               | 15                |                   |                   |
|               |               | $R_i$            | 49.5              | 32.5              | 38                |
|               |               | $R_i^2$          | 2450.25           | 1056.25           | 1444              |
|               |               | $R_i^2/n_i$      | 490.05            | 211.25            | 288.8             |
|               |               | $\sum R_i^2/n_i$ | 990.10            |                   |                   |

The value of the test statistic is

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1) = \left( \frac{12}{15 \times (15+1)} \times 990.10 \right) - 3 \times (15+1) = 1.51.$$

- c. With  $df = k - 1 = 3 - 1 = 2$ , the  $p$ -value  $= P(H \geq 1.51) = P(\chi_2^2 \geq 1.51)$  is greater than 0.10.
- d. Since the  $p$ -value  $> 0.01 = \alpha$ , we do not reject  $H_0$ . Therefore, at the 1% significance level, we cannot conclude that some medians differ. This suggests that the cleansing action does not differ by the type of detergent.

20.40 a. With  $\alpha = 0.05$  and  $n = 9$ , the critical value is 0.600. The decision rule is to reject  $H_0$  if  $|r_s| > 0.600$ .

- b. The value of the test statistic is  $r_s = -0.64$ , and points to a moderate negative relationship between the two variables.
- c. Since  $|r_s| = 0.64 > 0.600$ , we reject  $H_0$ . Therefore, at the 5% significance level, the Spearman's rank correlation coefficient is less than 0.

20.42 a.

| x  | y   | Rank for x | Rank for y | $d_i = x - y$  | $d_i^2$           |
|----|-----|------------|------------|----------------|-------------------|
| -2 | -4  | 1          | 5          | -4             | 16                |
| 0  | -3  | 3          | 6          | -3             | 9                 |
| 3  | -8  | 4          | 3          | 1              | 1                 |
| -1 | -5  | 2          | 4          | -2             | 4                 |
| 4  | -9  | 5          | 2          | 3              | 9                 |
| 7  | -10 | 6          | 1          | 5              | 25                |
|    |     |            |            | $\sum d_i = 0$ | $\sum d_i^2 = 64$ |

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 64}{6 \times (6^2 - 1)} = -0.83$$

The Spearman's rank correlation coefficient of  $-0.83$  indicates a strong, negative relationship between  $x$  and  $y$ .

- b.  $H_0: \rho_S \geq 0$ ;  $H_A: \rho_S < 0$ ;  
 c. With  $\alpha = 0.01$  and  $n = 6$ , the critical value is  $0.943$ . The decision rule is to reject  $H_0$  if  $|r_S| > 0.943$ .  
 d. Since  $|r_S| = 0.83 < 0.943$ , we do not reject  $H_0$ . Therefore, at the 1% significance level, we cannot conclude that the correlation between the two variables,  $x$  and  $y$ , is less than 0.

20.44 a. Assuming the normal distribution,  $z = r_S \sqrt{n-1} = 0.64 \times \sqrt{50-1} = 4.48$ . The corresponding  $p$ -value  $= P(Z \geq 4.48) = 0.00$ .

- b. Since the  $p$ -value  $= 0.00 < 0.01 = \alpha$ , we reject  $H_0$ . Therefore, at the 1% significance level, the Spearman's rank correlation coefficient is greater than zero.

20.46 a.

| Country       | Per Capita GNP Rank (x) | Infant Mortality Rank (y) | $d_i = x - y$ | $d_i^2$ |
|---------------|-------------------------|---------------------------|---------------|---------|
| Luxembourg    | 1                       | 7                         | -6            | 36      |
| Norway        | 2                       | 4                         | -2            | 4       |
| Singapore     | 3                       | 1                         | 2             | 4       |
| United States | 4                       | 10                        | -6            | 36      |
| Ireland       | 5                       | 9                         | -4            | 16      |
| Switzerland   | 6                       | 5                         | 1             | 1       |
| Netherlands   | 7                       | 8                         | -1            | 1       |
| Austria       | 8                       | 6                         | 2             | 4       |
| Sweden        | 9                       | 2                         | 7             | 49      |
| Iceland       | 10                      | 3                         | 7             | 49      |
|               |                         |                           | 0             | 200     |

$$r_S = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 200}{10 \times (10^2 - 1)} = -0.21$$

The Spearman's rank correlation coefficient of  $-0.21$  indicates a weak negative relationship between the rankings of Per Capita GNP and the Infant Mortality.

- b.  $H_0: \rho_S = 0$ ;  $H_A: \rho_S \neq 0$ ;  
 c. With  $\alpha = 0.05$  and  $n = 10$ , the critical value is  $0.648$ . The decision rule is to reject  $H_0$  if  $|r_S| > 0.648$ .  
 d. Since  $|r_S| = 0.21 < 0.648$ , we do not reject  $H_0$ . Therefore, at the 5% significance level, we cannot conclude that the Per Capita GNP and Infant Mortality rankings are correlated.

20.48 a.

| Price (x) | Days to sell Home | Rank for x | Rank for y | $d_i = x - y$  | $d_i^2$             |
|-----------|-------------------|------------|------------|----------------|---------------------|
| 265       | 136               | 4          | 5          | -1             | 1                   |
| 225       | 125               | 3          | 4          | -1             | 1                   |
| 160       | 120               | 1          | 1          | 0              | 0                   |
| 325       | 140               | 5          | 6          | -1             | 1                   |
| 430       | 145               | 7          | 7.5        | -0.5           | 0.25                |
| 515       | 121               | 8          | 2          | 6              | 36                  |
| 180       | 122               | 2          | 3          | -1             | 1                   |
| 423       | 145               | 6          | 7.5        | -1.5           | 2.25                |
|           |                   |            |            | $\sum d_i = 0$ | $\sum d_i^2 = 42.5$ |

$$r_S = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 42.5}{8 \times (8^2 - 1)} = 0.49$$

The Spearman's rank correlation coefficient of  $0.49$  indicates a moderate positive relationship between the price of the home and the number of days it takes to sell it.

- b.  $H_0: \rho_S = 0$ ;  $H_A: \rho_S \neq 0$ ;  
 c. With  $\alpha = 0.05$  and  $n = 8$ , the critical value is  $0.738$ . The decision rule is to reject  $H_0$  if  $|r_S| > 0.738$ .  
 d. Since  $|r_S| = 0.49 < 0.738$ , we do not reject  $H_0$ . Therefore, at the 5% significance level, we cannot conclude that the price of the home and the number of days it takes to sell are correlated.

20.50 a.  $H_0: \rho_S = 0$ ;  $H_A: \rho_S \neq 0$ ;

- b. Assuming the normal distribution,  $z = r_S \sqrt{n-1} = 0.85 \times \sqrt{65-1} = 6.8$ . The corresponding  $p$ -value  $= 2 \times P(Z \geq 6.8) = 2 \times (1 - 1.0000) = 0.00$

- c. Since the  $p$ -value  $= 0.00 < \alpha = 0.05$ , we reject  $H_0$ . Therefore, at the 5% significance level educational attainment and salary are correlated.

20.54 a.  $z = \frac{\bar{p} - 0.50}{0.5/\sqrt{n}} = \frac{0.30 - 0.50}{0.5/\sqrt{40}} = -2.53$

- b. The corresponding  $p$ -value  $= 2 \times P(Z \leq -2.28) = 2 \times 0.0057 = 0.0114$ .

- c. Since the  $p$ -value  $= 0.0114 < 0.05 = \alpha$ , we reject  $H_0$ . Therefore, the population proportion of ordinal differences are significant, at the 5% significance level.

20.56 Let  $p$  represent the population proportion of plus signs.

- a.  $H_0: p = 0.50$ ;  $H_A: p \neq 0.50$

b.  $\bar{p} = \frac{15}{20} = 0.75$ ;  $z = \frac{\bar{p} - 0.50}{0.5/\sqrt{n}} = \frac{0.75 - 0.50}{0.5/\sqrt{20}} = 2.24$ .

- c. The corresponding  $p$ -value  $= 2 \times P(Z \geq 2.24) = 2 \times (1 - 0.9875) = 0.0250$ .

- d. Since the  $p$ -value  $= 0.0250 < 0.05 = \alpha$ , we reject  $H_0$ . Therefore, at the 5% significance level, the proportion of positive signs is different from that of negative signs.

20.58 Let  $p$  represents the population proportion of plus signs.

- a.  $H_0: p = 0.50$ ;  $H_A: p \neq 0.50$ ;

b.

| Consumer | Tap Water | Bottled Water | Difference | Sign |
|----------|-----------|---------------|------------|------|
| 1        | 4         | 5             | -1         | -    |
| 2        | 3         | 2             | 1          | +    |
| 3        | 5         | 4             | 1          | +    |
| 4        | 4         | 3             | 1          | +    |
| 5        | 3         | 5             | -2         | -    |
| 6        | 5         | 3             | 2          | +    |
| 7        | 2         | 1             | 1          | +    |
| 8        | 5         | 2             | 3          | +    |
| 9        | 3         | 4             | -1         | -    |
| 10       | 2         | 4             | -2         | -    |
| 11       | 5         | 4             | 1          | +    |
| 12       | 4         | 3             | 1          | +    |
| 13       | 5         | 2             | 3          | +    |
| 14       | 3         | 4             | -1         | -    |

$$\bar{p} = \frac{9}{14} = 0.64$$
;  $z = \frac{\bar{p} - 0.50}{0.5/\sqrt{n}} = \frac{0.64 - 0.50}{0.5/\sqrt{14}} = 1.05$

- c. The corresponding  $p$ -value  $= 2 \times P(Z \geq 1.05) = 2 \times (1 - 0.8531) = 0.2938$ .

- d. Since the  $p$ -value  $= 0.2938 > 0.05 = \alpha$ , we do not reject  $H_0$ . Therefore, at the 5% significance level, we cannot conclude that the proportions differ. There is no

significant difference in the preference between tap water and bottle water at the 5% significance level.

20.60 Let  $p$  represents the population proportion of plus signs.

- a.  $H_0: p = 0.50$ ;  $H_A: p \neq 0.50$ ;  
b. At  $\alpha = 0.10$ ,  $z_{\alpha/2} = z_{0.05} = 1.645$ . The decision rule is to reject  $H_0$  if  $z < -1.645$  or  $z > 1.645$ .

| Candidate | Faculty A's Rating | Faculty B's Rating | Difference | Sign |
|-----------|--------------------|--------------------|------------|------|
| 1         | 5                  | 6                  | -1         | -    |
| 2         | 7                  | 8                  | -1         | -    |
| 3         | 8                  | 5                  | 3          | +    |
| 4         | 7                  | 7                  | 0          |      |
| 5         | 9                  | 10                 | -1         | -    |
| 6         | 4                  | 3                  | 1          | +    |
| 7         | 2                  | 2                  | 0          |      |
| 8         | 8                  | 9                  | -1         | -    |
| 9         | 9                  | 10                 | -1         | -    |
| 10        | 6                  | 4                  | 2          | +    |
| 11        | 8                  | 9                  | -1         | -    |
| 12        | 6                  | 8                  | -2         | -    |

For positive signs,  $\bar{p} = \frac{3}{10} = 0.30$ ;  $z = \frac{\bar{p} - 0.50}{0.5/\sqrt{n}} = \frac{0.30 - 0.50}{0.5/\sqrt{10}} = -1.26$

- d. Since  $-1.645 < z = 1.26 < 1.645$ , we do not reject  $H_0$ . Therefore, at the 10% significance level, we cannot conclude that differences exist between the two faculty ratings of the PhD applicants.

20.62 a.  $H_0$ : The elements occur randomly;

$H_A$ : The elements do not occur randomly

$$b. \mu_R = \frac{2n_1 n_2}{n} + 1 = \frac{2 \times 24 \times 28}{52} + 1 = 26.85; \sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}} = \sqrt{\frac{2 \times 24 \times 28 \times (2 \times 24 \times 28 - 52)}{(52)^2 \times (52-1)}} = 3.55, \text{ and } z = \frac{R - \mu_R}{\sigma_R} = \frac{18 - 26.85}{3.55} = -2.49$$

- c. The  $p$ -value  $= 2 \times P(Z \leq -2.49) = 2 \times 0.0064 = 0.0128$   
d. Since the  $p$ -value  $= 0.0128 < 0.05 = \alpha$ , we reject  $H_0$ . Therefore, at the 5% significance level, the observations do not occur randomly.

20.64  $H_0$ : The outcomes occur randomly;

$H_A$ : The outcomes do not occur randomly

$R = 10 + 10 = 20$ ,  $n_1 = n_B = 17$ ,  $n_2 = n_A = 14$ ,  $n = n_1 + n_2 = 17 + 14 = 31$

$$\mu_R = \frac{2n_1 n_2}{n} + 1 = \frac{2 \times 17 \times 14}{31} + 1 = 16.35; \sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}} = \sqrt{\frac{2 \times 17 \times 14 \times (2 \times 17 \times 14 - 31)}{31^2 \times (31-1)}} = 2.71, \text{ and } z = \frac{R - \mu_R}{\sigma_R} = \frac{20 - 16.35}{2.71} = 1.35.$$

The corresponding  $p$ -value  $= 2 \times P(Z \geq 1.34) = 2 \times 0.0885 = 0.1770$ . Since the  $p$ -value  $= 0.1802 > 0.05 = \alpha$ , we do not reject  $H_0$ . Therefore, at the 5% significance level, we cannot conclude that the outcomes occur non-randomly.

20.66 a.  $H_0$ : Even and odd numbers occur randomly

$H_A$ : Even and odd numbers do not occur randomly

- b. At  $\alpha = 0.01$ ,  $z_{\alpha/2} = z_{0.005} = 2.58$ . The decision rule is to reject  $H_0$  if  $z < -2.58$  or  $z > 2.58$ .

c.  $R = 4 + 4 = 8$ ,  $n_1$  (odd)  $= 12$ ,  $n_2$  (even)  $= 10$ ,  $n = n_1 + n_2 = 12 + 10 = 22$

$$\mu_R = \frac{2n_1 n_2}{n} + 1 = \frac{2 \times 12 \times 10}{22} + 1 = 11.91; \sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}} = \sqrt{\frac{2 \times 12 \times 10 \times (2 \times 12 \times 10 - 22)}{(22)^2 \times (22-1)}} = 2.27, \text{ and } z = \frac{R - \mu_R}{\sigma_R} = \frac{8 - 11.91}{2.27} = -1.72.$$

- d. Since  $-2.58 < z = -1.72 < 2.58$ , we do not reject  $H_0$ . Therefore, we cannot reject the null hypothesis of randomness at the 1% significance level. Consequently, the computer program is operating properly.

20.68  $H_0$ : GDP growth rate in India is random;

$H_A$ : GDP growth rate in India is not random

The median GDP growth rate in India is 6%.

| Year | GDP  | GDP - 6 | Runs |
|------|------|---------|------|
| 1980 | 6.74 | 0.74    | A    |
| 1981 | 6    | 0       |      |
| 1982 | 3.47 | -2.53   | B    |
| 1983 | 7.3  | 1.3     | A    |
| 1984 | 3.82 | -2.18   | B    |
| 1985 | 5.23 | -0.77   | B    |
| 1986 | 4.77 | -1.23   | B    |
| 1987 | 3.96 | -2.04   | B    |
| 1988 | 9.64 | 3.64    | A    |
| 1989 | 5.95 | -0.05   | B    |
| 1990 | 5.53 | -0.47   | B    |
| 1991 | 1.06 | -4.94   | B    |
| 1992 | 5.48 | -0.52   | B    |
| 1993 | 4.77 | -1.23   | B    |
| 1994 | 6.65 | 0.65    | A    |
| 1995 | 7.57 | 1.57    | A    |
| 1996 | 7.56 | 1.56    | A    |
| 1997 | 4.05 | -1.95   | B    |
| 1998 | 6.19 | 0.19    | A    |
| 1999 | 7.39 | 1.39    | A    |
| 2000 | 4.03 | -1.97   | B    |
| 2001 | 5.22 | -0.78   | B    |
| 2002 | 3.77 | -2.23   | B    |
| 2003 | 8.37 | 2.37    | A    |
| 2004 | 8.28 | 2.28    | A    |
| 2005 | 9.35 | 3.35    | A    |
| 2006 | 9.67 | 3.67    | A    |
| 2007 | 9.06 | 3.06    | A    |
| 2008 | 6.07 | 0.07    | A    |

$R = 6 + 5 = 11$ ,  $n_1 = n_A = 14$ ,  $n_2 = n_B = 14$ ,  $n = n_1 + n_2 = 14 + 14 = 28$

$$\mu_R = \frac{2n_1 n_2}{n} + 1 = \frac{2 \times 14 \times 14}{28} + 1 = 15.0; \sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}} = \sqrt{\frac{2 \times 14 \times 14 \times (2 \times 14 \times 14 - 28)}{(28)^2 \times (28-1)}} = 2.60, \text{ and } z = \frac{R - \mu_R}{\sigma_R} = \frac{11 - 15}{2.60} = -1.54.$$

The corresponding  $p$ -value  $= 2 \times P(Z < -1.54) = 2 \times 0.0618 = 0.1236$ .

Since the  $p$ -value  $= 0.1236 > 0.05 = \alpha$ , we do not reject  $H_0$ . Therefore, we cannot conclude that India GDP growth rate is non-random at the 5% significance level.

- 20.70 a.  $H_0$ : Amgen stock price follows a random walk;  
 $H_A$ : Amgen stock price does not follow a random walk.  
 b. Given the large sample size ( $n = 252$ ), we use Minitab for this exercise.

Minitab Results ( $K$  denotes the median)

**Runs Test: Adj Close**

Runs test for Adj Close

Runs above and below  $K = 55.9$

The observed number of runs = 30

The expected number of runs = 127

126 observations above  $K$ , 126 below

$P$ -value = 0.000

Since the  $p$ -value =  $0.00 < 0.05 = \alpha$ , we reject  $H_0$ . We can therefore conclude that Amgen stock price does not follow a random walk at the 5% significance level.

- 20.72 a.  $H_0: m_D = 0$ ;  $H_A: m_D \neq 0$   
 b. With  $\alpha = 0.05$  and  $n = 6$ ,  $T_L = 0$  and  $T_U = 21$ . The decision rule is to reject  $H_0$  if  $T \leq T_L = 0$  or  $T \geq T_U = 21$ .

c.

| Lot | Crop yield using old fertilizer(x) | Crop yield using new fertilizer(y) | $d = x - y$ | $ d $ | Rank | Ranks of Negative Differences | Ranks of Positive Differences |
|-----|------------------------------------|------------------------------------|-------------|-------|------|-------------------------------|-------------------------------|
| 1   | 10                                 | 12                                 | -2          | 2     | 4    | 4                             |                               |
| 2   | 11                                 | 10                                 | 1           | 1     | 2    |                               | 2                             |
| 3   | 10                                 | 13                                 | -3          | 3     | 5.5  | 5.5                           |                               |
| 4   | 12                                 | 9                                  | 3           | 3     | 5.5  |                               | 5.5                           |
| 5   | 12                                 | 11                                 | 1           | 1     | 2    |                               | 2                             |
| 6   | 11                                 | 12                                 | -1          | 1     | 2    | 2                             |                               |
|     |                                    |                                    |             |       |      | $T^- = 11$                    | $T^+ = 9.5$                   |

The value of the test statistic is  $T = T^+ = 9.5$ .

- d. Since  $0 < T = 9.5 < 21$ , we do not reject  $H_0$ . At the 5% significance level, we cannot conclude that the median difference between the crop yields differs from zero. As a result, the farmer should not be concerned.
- 20.74 a.  $H_0: m_A - m_B = 0$ ;  $H_A: m_A - m_B \neq 0$   
 b. The value of the test statistic was found using Minitab

**Mann-Whitney Test and CI: Brand A, Brand B**

|         | N  | Median |
|---------|----|--------|
| Brand A | 40 | 16.000 |
| Brand B | 40 | 17.000 |

Point estimate for ETA1 - ETA2 is -1.000

95.1 Percent CI for ETA1 - ETA2 is (-2.000, 0.000)

**W = 1492.0**

Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.2199

The test is significant at 0.2148 (adjusted for ties)

Since  $n_1 = n_2 = 40$ ,  $W = W_A = 1492$

- c. With  $n_1 = n_2 = 40$ , the mean and the standard deviation for  $W$  are:  $\mu_W = \frac{n_2(n_2 + n_1 + 1)}{2} = \frac{40 \times (40 + 40 + 1)}{2} = 1620.00$  and  $\sigma_W = \sqrt{\frac{n_2 n_1 (n_2 + n_1 + 1)}{12}} = \sqrt{\frac{40 \times 40 \times (40 + 40 + 1)}{12}} = 103.92$ .  
 Under normality, the value of the test statistic  $z = \frac{W - \mu_W}{\sigma_W} = \frac{1492 - 1620}{103.92} = -1.23$ . The corresponding  $p$ -value =  $2 \times P(Z \leq -1.23) = 0.2186$  (these results differ slightly from Minitab's results due to rounding).

- d. Since  $p$ -value =  $0.2186 > \alpha = 0.05$ , we do not reject  $H_0$ . We cannot conclude that the median longevity between the two brands differs at the 5% significance level.

- 20.76 a.  $H_0: m_1 = m_2 = m_3$ ;  
 $H_A$ : Not all population medians are equal.  
 b. With  $\alpha = 0.05$  and with  $df = k - 1 = 3 - 1 = 2$ ,  $\chi^2_{\alpha, df} = \chi^2_{0.05, 2} = 5.991$ .  
 c.

| Pooled Sample | Origin Sample | Ranks            | Industry A Ranks | Industry B Ranks | Industry C Ranks |
|---------------|---------------|------------------|------------------|------------------|------------------|
| 7.28          | Industry A    | 1                | 1                |                  |                  |
| 9.32          | Industry B    | 2                |                  | 2                |                  |
| 9.41          | Industry B    | 3                |                  | 3                |                  |
| 9.96          | Industry A    | 4                | 4                |                  |                  |
| 10.51         | Industry A    | 5                | 5                |                  |                  |
| 12.19         | Industry A    | 6                | 6                |                  |                  |
| 12.44         | Industry A    | 7                | 7                |                  |                  |
| 14.34         | Industry B    | 8                |                  | 8                |                  |
| 14.9          | Industry B    | 9                |                  | 9                |                  |
| 16.7          | Industry C    | 10               |                  |                  | 10               |
| 16.87         | Industry C    | 11               |                  |                  | 11               |
| 16.88         | Industry C    | 12               |                  |                  | 12               |
| 17.8          | Industry B    | 13               |                  | 13               |                  |
| 24.75         | Industry C    | 14               |                  |                  | 14               |
| 26.38         | Industry C    | 15               |                  |                  | 15               |
|               |               | $R_i$            | 23               | 35               | 62               |
|               |               | $R_i^2$          | 529              | 1225             | 3844             |
|               |               | $R_i^2/n_i$      | 105.80           | 245              | 768.8            |
|               |               | $\sum R_i^2/n_i$ | 1119.60          |                  |                  |

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1) = \left( \frac{12}{15 \times (15+1)} \times 1119.60 \right) - 3 \times (15+1) = 7.98$$

- d. Since  $H = 7.98 > 5.991$ , we reject  $H_0$ . Therefore, at the 5% significance level, some median P/E ratios differ by industry.
- 20.78 a.  $H_0: \rho_S = 0$ ;  $H_A: \rho_S \neq 0$ ;  
 b. Assuming the normal distribution,  $z = r_S \sqrt{n-1} = 0.45 \times \sqrt{65-1} = 3.60$ . The corresponding  $p$ -value =  $2 \times P(Z \geq 3.60) = 2 \times (1 - 0.9998) = 0.0004$ .  
 c. Since the  $p$ -value =  $0.0004 < \alpha = 0.05$ , we reject  $H_0$ . Therefore, advertising and sales are related at the 5% significance level.
- 20.80  $H_0$ : GDP growth rate in China is random;  
 $H_A$ : GDP growth rate in China is not random  
 The median GDP growth rate in China is 10%.

| Year | GDP | GDP - 10 | Runs |
|------|-----|----------|------|
| 1980 | 7.8 | -2.2     | B    |
| 1981 | 5.2 | -4.8     | B    |
| 1982 | 9.1 | -0.9     | B    |
| :    | :   | :        | :    |
| 2008 | 9   | -1       | B    |

$$R = 5 + 4 = 9, n_1 = n_A = 14, n_2 = n_B = 14, n = n_1 + n_2 = 14 + 14 = 28$$

$$\mu_R = \frac{2n_1 n_2}{n} + 1 = \frac{2 \times 14 \times 14}{28} + 1 = 15.0; \sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}} = \sqrt{\frac{2 \times 14 \times 14 \times (2 \times 14 \times 14 - 28)}{(28^2 \times (28-1))}} = 2.6; z = \frac{R - \mu_R}{\sigma_R} = \frac{9 - 15}{2.60} = -2.31$$

The corresponding  $p$ -value =  $2 \times P(Z \leq -2.31) = 2 \times 0.0104 = 0.0208$



Since the  $p$ -value = 0.0208 < 0.05 =  $\alpha$ , we reject  $H_0$ .  
Therefore, China GDP growth rate is non-random at the 5% significance level.

20.82  $H_0$ : US CPI is random;

$H_A$ : US CPI is not random

The median US CPI is 3.3.

| Year | CPI  | CPI – Median | Runs |
|------|------|--------------|------|
| 1980 | 12.5 | 9.2          | A    |
| 1981 | 8.9  | 5.6          | A    |
| 1982 | 3.8  | 0.5          | A    |
| 1983 | 3.8  | 0.5          | A    |
| 1984 | 3.9  | 0.6          | A    |
| 1985 | 3.8  | 0.5          | A    |
| 1986 | 1.1  | -2.2         | B    |
| 1987 | 4.4  | 1.1          | A    |
| 1988 | 4.4  | 1.1          | A    |
| 1989 | 4.6  | 1.3          | A    |
| 1990 | 6.1  | 2.8          | A    |
| 1991 | 3.1  | -0.2         | B    |
| 1992 | 2.9  | -0.4         | B    |
| 1993 | 2.7  | -0.6         | B    |
| 1994 | 2.7  | -0.6         | B    |
| 1995 | 2.5  | -0.8         | B    |
| 1996 | 3.3  | 0            |      |
| 1997 | 1.7  | -1.6         | B    |
| 1998 | 1.6  | -1.7         | B    |
| 1999 | 2.7  | -0.6         | B    |
| 2000 | 3.4  | 0.1          | A    |
| 2001 | 1.6  | -1.7         | B    |
| 2002 | 2.4  | -0.9         | B    |
| 2003 | 1.9  | -1.4         | B    |
| 2004 | 3.3  | 0            |      |
| 2005 | 3.4  | 0.1          | A    |
| 2006 | 2.5  | -0.8         | B    |
| 2007 | 4.1  | 0.8          | A    |
| 2008 | 0.1  | -3.2         | B    |

$R = 5 + 5 = 10$ ,  $n_1 = n_A = 13$ ,  $n_2 = n_B = 14$ ,  $n = n_1 + n_2 = 13 + 14 = 27$ .

$$\mu_R = \frac{2n_1 n_2}{n} + 1 = \frac{2 \times 13 \times 14}{27} + 1 = 14.48; \sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}}$$

$$= \sqrt{\frac{2 \times 13 \times 14 \times (2 \times 13 \times 14 - 27)}{27^2 \times (27-1)}} = 2.54, \text{ and } z =$$

$$\frac{R - \mu_R}{\sigma_R} = \frac{10 - 14.48}{2.54} = -1.76$$

The corresponding  $p$ -value =  $2 \times P(Z \leq -1.76) = 2 \times 0.0392 = 0.0784$

Since the  $p$ -value = 0.0784 > 0.05 =  $\alpha$ , we do not reject  $H_0$ .

Therefore, we cannot conclude that the US CPI is non-random at the 5% significance level.



## A

**Acceptance sampling** A statistical quality control technique in which a portion of the completed products is inspected.

**Addition rule** The probability that  $A$  or  $B$  occurs, or that at least one of these events occurs, is equal to the probability that  $A$  occurs, plus the probability that  $B$  occurs, minus the probability that both  $A$  and  $B$  occur, that is,  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .

**Adjusted close price** Stock price data adjusted using appropriate dividend and split multipliers.

**Adjusted  $R^2$**  A modification of the coefficient of determination that imposes a penalty for using additional explanatory variables in the linear regression model. It is used to compare models with different numbers of explanatory variables; the higher the adjusted  $R^2$ , the better the model.

**Aggregate price index** A representation of relative price movements for a group of items.

**Alpha** The probability of Type I error in hypothesis testing.

**Alternative hypothesis ( $H_A$ )** In a hypothesis test, the alternative hypothesis contradicts the default state or status quo specified in the null hypothesis. Generally, whatever we wish to establish is placed in the alternative hypothesis.

**Analysis of variance (ANOVA)** A statistical technique used to determine if differences exist between three or more population means.

**Annualized return** A measure equivalent to the geometric mean return.

**Arithmetic mean** The primary measure of central location, also referred to as the mean or the average.

**Assignable variation** In a production process, the variation that is caused by specific events or factors that can usually be identified and eliminated.

**Autoregressive model** A regression model where lagged values of the response variable are used as explanatory variables.

**Average** See *Arithmetic mean*.

**Average growth rate** For growth rates  $g_1, g_2, \dots, g_n$ , the average growth rate  $G_g$  is computed as

$G_g = \sqrt[n]{(1 + g_1)(1 + g_2) \cdots (1 + g_n)} - 1$ , where  $n$  is the number of multi-period growth rates.

## B

**Balanced data** A completely randomized ANOVA design with an equal number of observations in each sample.

**Bar chart** A graph that depicts the frequency or relative frequency of each category of qualitative data as a series of horizontal or vertical bars, the lengths of which are proportional to the values that are to be depicted.

**Bayes' theorem** The rule for updating probabilities is

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}, \text{ where } P(B) \text{ is the}$$

prior probability and  $P(B|A)$  is the posterior probability.

**Bell-shaped distribution** See *normal distribution*.

**Bernoulli process** A series of  $n$  independent and identical trials of an experiment such that each trial has only two possible outcomes, conventionally labeled success and failure, and each

time the trial is repeated, the probabilities of success and failure remain the same.

**Between-treatments variance** In ANOVA, a measure of the variability between sample means.

**Bias** The tendency of a sample statistic to systematically overestimate or underestimate a population parameter.

**Binary choice models** Regression models that use a dummy (binary) variable as the response variable. Also called discrete choice or qualitative response models.

**Binomial distribution** A description of the probabilities associated with the possible values of a binomial random variable.

**Binomial random variable** The number of successes achieved in the  $n$  trials of a Bernoulli process.

**Box plot** A graphical display of the minimum value, quartiles, and the maximum value of a data set.

## C

**c chart** A control chart that monitors the count of defects per item in statistical quality control.

**Capital gains yield** The gain or loss resulting from the increase or decrease in the value of an asset.

**Causal forecasting models** Quantitative forecasts based on a regression framework, where the variable of interest is related to a single or multiple explanatory variables.

**Centered Moving Averages CMA** In time series analysis, a smoothing technique based on computing the average from a fixed number  $m$  of the most recent observations, where  $m$  is even.

**Centerline** In a control chart, the centerline represents a variable's expected value when the production process is in control.

**Central limit theorem (CLT)** The CLT states that the sum or mean of a large number of independent observations from the same underlying distribution has an approximate normal distribution. The approximation steadily improves as the number of observations increases.

**Chance variation** In a production process, the variation that is caused by a number of randomly occurring events that are part of the production process.

**Changing variability** In regression analysis, a violation of the assumption that the variance of the error term is the same for all observations. It is also referred to as heteroskedasticity.

**Chebyshev's theorem** For any data set, the proportion of observations that lie within  $k$  standard deviations from the mean will be at least  $1 - 1/k^2$ , where  $k$  is any number greater than 1.

**Chi-square test of a contingency table** See *test for independence*.

**Chi-square ( $X^2$ ) distribution** A family of distributions where each distribution depends on its particular degrees of freedom  $df$ . It is positively skewed, with values ranging from zero to infinity, but becomes increasingly symmetric as  $df$  increase.

**Classes** Intervals for a frequency distribution of quantitative data.

**Classical probability** A probability often used in games of chance. It is based on the assumption that all outcomes are equally likely.

**Cluster sampling** A population is first divided up into mutually exclusive and collectively exhaustive groups of observations, called clusters. A cluster sample includes observations from randomly selected clusters.

**Coefficient of determination ( $R^2$ )** The proportion of the sample variation in the response variable that is explained by the sample regression equation; used as a goodness-of-fit measure in regression analysis.

**Coefficient of variation (CV)** The ratio of the standard deviation of a data set to its mean; a relative measure of dispersion.

**Complement** The complement of event  $A$ , denoted  $A^c$ , is the event consisting of all outcomes in the sample space that are not in  $A$ .

**Complement rule** The probability of the complement of an event is one minus the probability of the event, that is,  $P(A^c) = 1 - P(A)$ .

**Conditional probability** The probability of an event given that another event has already occurred.

**Confidence coefficient** The probability that a given confidence interval will contain the population parameter of interest.

**Confidence interval** A range of values that, with a certain level of confidence, contains the population parameter of interest.

**Combination formula** The number of ways to choose  $x$  objects from a total of  $n$  objects, where the order in which the  $x$  objects is listed *does not matter*, is  ${}_nC_x = \binom{n}{x} = \frac{n!}{(n-x)!x!}$ .

**Consistency** An estimator is consistent if it approaches the unknown population parameter being estimated as the sample size grows larger.

**Consumer price index (CPI)** A monthly weighted aggregate price index, computed by the U.S. Bureau of Labor Statistics, based on the prices paid by urban consumers for a representative basket of goods and services.

**Contingency table** A table that shows frequencies for two qualitative (categorical) variables,  $x$  and  $y$ , where each cell represents a mutually exclusive combination of the pair of  $x$  and  $y$  values.

**Continuous (random) variable** A variable that assumes uncountable values in an interval.

**Continuous uniform distribution** A distribution describing a continuous random variable that has an equally likely chance of assuming a value within a specified range.

**Control chart** A plot of statistics of a production process over time. If the statistics randomly fall in an expected range, then the production process is in control. If the statistics reveal an undesirable trend, then adjustment of the production process is likely necessary.

**Correlated observations** In regression analysis, a violation of the assumption that the observations are uncorrelated. It is also referred to as serial correlation.

**Correlation coefficient** A measure that describes the direction and strength of the linear relationship between two variables.

**Covariance** A measure that reveals the direction of the linear relationship between two variables.

**Critical value** In a hypothesis test, the critical value is a point that separates the rejection region from the nonrejection region.

**Cross-sectional data** Values of a characteristic of many subjects at the same point in time or approximately the same point in time.

**Cubic regression model** Allows two sign changes in the influence of an explanatory variable on the response variable.

**Cumulative distribution function** A probability that the value of a random variable  $X$  is less than or equal to a particular value  $x$ ,  $P(X \leq x)$ .

**Cumulative frequency distribution** A distribution of quantitative data recording the number of observations that falls below the upper limit of each class.

**Cumulative relative frequency distribution** A distribution of quantitative data recording the fraction (proportion) of observations that falls below the upper limit of each class.

**Cyclical component** Wave-like fluctuations or business cycles of a time series, often caused by expansion and contraction of the economy.

## D

**Decision rule** In hypothesis testing, the decision rule specifies when to reject and when not to reject the null hypothesis.

**Decomposition analysis** A method of estimating trend and seasonal components from a time series and then recomposing them to make forecasts.

**Deflated time series** A series obtained by adjusting economic time series for changes in prices, or inflation. The two most commonly used price indices for deflating economic time series are the Consumer Price Index, CPI, and the Producer Price Index, PPI.

**Degrees of freedom** The number of independent pieces of information that go into the calculation of a given statistic. Many probability distributions are identified by the degrees of freedom.

**Dependent events** The occurrence of one event is related to the probability of the occurrence of the other event.

**Descriptive statistics** The summary of a data set in the form of tables, graphs, or numerical measures.

**Detection approach** In statistical quality control, the detection approach determines at which point the production process does not conform to specifications.

**Deterministic relationship** A relationship in which the value of the response variable is uniquely determined by the values of the explanatory variables.

**Discrete choice models** See *binary choice models*.

**Discrete uniform distribution** A symmetric distribution where the random variable assumes a finite number of values and each value is equally likely.

**Discrete (random) variable** A variable that assumes a countable number of values.

**Dummy variable** A variable that takes on values of 0 or 1. It is commonly used to describe a qualitative variable with two categories.

**Dummy variable trap** A linear regression model where the number of dummy variables equals the number of categories of a qualitative variable. In a correctly specified model, the number of dummy variables is one less than the number of categories.

## E

**Efficiency** An unbiased estimator is efficient if its standard error is lower than that of other unbiased estimators.

**Empirical probability** A probability value based on observing the relative frequency with which an event occurs.

**Empirical rule** Given a sample mean  $\bar{x}$ , a sample standard deviation  $s$ , and a relatively symmetric and bell-shaped distribution, approximately 68% of all observations fall in the interval  $\bar{x} \pm s$ ; approximately 95% of all observations fall in the interval  $\bar{x} \pm 2s$ ; and almost all observations fall in the interval  $\bar{x} \pm 3s$ .

**Endogeneity** See *Excluded variables*.

**Estimate** A particular value of an estimator.

**Estimator** A statistic used to estimate a population parameter.

**Event** A subset of the sample space.

**Excluded variables** In regression analysis, a situation where important explanatory variables are excluded from the regression. It often leads to the violation of the assumption that the error is uncorrelated with the (included) explanatory variables.

**Exhaustive events** When all possible outcomes of an experiment are included in the events.

**Expected return of a portfolio** A weighted average of the expected returns of the assets comprising the portfolio.

**Expected value** A weighted average of all possible values of a random variable.

**Experiment** A process that leads to one of several possible outcomes.

**Explanatory variables** In regression analysis, the variables that we assume influence the response variable. They are also called the independent variables, predictor variables, control variables, or regressors.

**Exponential distribution** A continuous, nonsymmetric probability distribution used to describe the time that has elapsed *between* occurrences of an event.

**Exponential model** A regression model in which only the response variable is transformed into natural logs.

**Exponential smoothing** In time series analysis, a smoothing technique based on a weighted average where the weights decline exponentially as they become more distant.

**Exponential regression model** A regression model in which only the response variable is transformed into natural logs. In forecasting, it is used for a time series that grows by an increasing amount each time period.

**Exponential trend model** A regression model used for a time series that grows (declines) by an increasing (decreasing) amount each period.

## F

**F distribution** A family of distributions where each distribution depends on two degrees of freedom: the numerator degrees of freedom  $df_1$  and the denominator degrees of freedom  $df_2$ . It is

positively skewed, with values ranging from zero to infinity, but becomes increasingly symmetric as  $df_1$  and  $df_2$  increase.

**Factorial formula** The number of ways to assign every member of a group of size  $n$  to  $n$  slots is  $n! = n \times (n - 1) \times (n - 2) \times (n - 3) \times \cdots \times 1$ .

**Finite population correction factor** A correction factor that accounts for the added precision gained by sampling a larger percentage of the population. It is recommended when the sample constitutes at least 5% of the population.

**Fisher equation** A theoretical relationship between nominal returns, real returns, and the expected inflation rate.

**Fisher's least difference (LSD) method** In ANOVA, a test that determines which means significantly differ by computing all pairwise differences of the means.

**Frequency distribution** A table that groups qualitative data into categories, or quantitative data into intervals called classes, where the number of observations that fall into each category or class is recorded.

## G

**Geometric mean return** For multiperiod returns  $R_1, R_2, \dots, R_n$ , the geometric mean return  $G_R$  is computed as  $G_R = \sqrt[n]{(1 + R_1)(1 + R_2) \cdots (1 + R_n)} - 1$ , where  $n$  is the number of multiperiod returns.

**Goodness-of-fit test** A test, using the chi-square statistic, to determine if the sample proportions resulting from a multinomial experiment differ from the hypothesized population proportions specified in the null hypothesis.

**Goodness-of-fit test for normality** A chi-square test used to determine if sample data are drawn from the normally distributed population.

**Grand mean** In ANOVA, the sum of all observations in a data set divided by the total number of observations.

## H

**Heteroskedasticity** See *Changing variability*.

**Histogram** A graphical depiction of a frequency or relative frequency distribution; it is a series of rectangles where the width and height of each rectangle represent the class width and frequency (or relative frequency) of the respective class.

**Hypergeometric distribution** A description of the probabilities associated with the possible values of a hypergeometric random variable.

**Hypergeometric random variable** The number of successes achieved in the  $n$  trials, each with two outcomes, of an experiment where the trials cannot be assumed to be independent.

**Hypothesis test** A statistical procedure to resolve conflicts between two competing claims (hypotheses) on a particular population parameter of interest.

## I

**Income yield** The direct cash payments from an underlying asset, such as dividends, interest, or rental income.

**Independent events** The occurrence of one event does not affect the probability of the occurrence of the other event.

**Independent random samples** Two (or more) random samples are considered independent if the process that generates

one sample is completely separate from the process that generates the other sample.

**Index number** An easy-to-interpret numerical value that reflects a percentage change in price or quantity from a base value. The most common example is a price index.

**Indicator variable** See *dummy variable*.

**Inexact relationship** In regression analysis, a relationship in which the explanatory variables do not exactly predict the response variable.

**Inferential statistics** The practice of extracting useful information from a sample to draw conclusions about a population.

**Inflation rate** The percentage rate of change of a price index over time.

**In-sample criteria** Model selection criteria showing how well a forecasting model predicts values within the given sample.

**Interaction variable** In a regression model, an explanatory variable that takes account of the joint variation between two variables. It is constructed as the product of two different explanatory variables.

**Intersection** The intersection of two events  $A$  and  $B$ , denoted  $A \cap B$ , is the event consisting of all outcomes in  $A$  and  $B$ .

**Interval data** Values of a quantitative variable that can be categorized and ranked, and in which differences between values are meaningful.

**Interval estimate** See *Confidence interval*.

**Interval scale** Data that are measured on an interval scale can be categorized and ranked and the differences between scale values are meaningful.

**Interquartile range (IQR)** The difference between the third and first quartiles.

**Inverse transformation** A standard normal variable  $Z$  can be transformed to the normally distributed random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$  as  $X = \mu + Z\sigma$ .

**Investment return** The net gain or loss in value of an investment over a time period.

## J

**Jarque-Bera test** Uses the skewness and kurtosis coefficients to determine if sample data are drawn from the normally distributed population.

**Joint probabilities** The values in the interior of the joint probability table, representing the probabilities of the intersection of two events.

**Joint probability table** A contingency table whose frequencies have been converted to relative frequencies.

## K

**Kruskal-Wallis test** A nonparametric test to determine whether differences exist between several population medians.

**Kurtosis coefficient** A measure of whether data is more or less peaked than a normal distribution.

## L

**Laspeyres price index** A weighted aggregate price index based on quantities evaluated in the base period.

**Law of large numbers** In probability theory, if an experiment is repeated a large number of times, its empirical probability approaches its classical probability.

**Left-tailed test** In hypothesis testing, when the null hypothesis is rejected if the value of the test statistic falls in the left tail of the distribution.

**Linear probability model (LPM)** A linear regression model applied to a binary response variable.

**Linear trend model** A regression model used for a time series that grows by a fixed amount each time period.

**Logarithmic regression model** A regression model in which only the explanatory variable is transformed into natural logs.

**Logit model** A nonlinear regression model that ensures that the predicted probability of the binary response variable falls between zero and one.

**Log-log regression model** A regression model in which both the response variable and the explanatory variable are transformed into natural logs.

**Lognormal distribution** A continuous nonsymmetric probability distribution used to describe random variables that are known to be positively skewed.

**Lower control limit** In a control chart, the lower control limit indicates excessive deviation below the expected value of the variable of interest.

## M

**Mann-Whitney test** See *Wilcoxon rank-sum test*.

**Margin of error** A value that accounts for the standard error of the estimator and the desired confidence level of the interval.

**Marginal probabilities** The values in the margins of a joint probability table that represent unconditional probabilities.

**Matched-pairs sample** When samples are matched or paired in some way. They are commonly employed in “before” and “after” studies.

**Maximum likelihood estimation (MLE)** An estimation technique used to estimate models such as the logit models.

**Mean** The average value of a data set; the most commonly used measure of central location.

**Mean absolute deviation (MAD)** The average of the absolute differences between the observations and the mean.

**Mean square error (MSE)** The average of the sum of squares due to error (residual), where residual is the difference between the observed and the predicted value of a variable.

**Mean square regression** The average of the sum of squares due to regression.

**Mean-variance analysis** The idea that the performance of an asset is measured by its rate of return, and this rate of return is evaluated in terms of its reward (mean) and risk (variance).

**Median** The middle value of a data set.

**Method of least squares** See *Ordinary Least squares (OLS)*.

**Method of runs above and below the median** A nonparametric test to determine randomness with quantitative data.

**Mode** The most frequently occurring value in a data set.

**Moving average (MA)** In time series analysis, a smoothing technique based on computing the average from a fixed number  $m$  of the most recent observations.



**Moving average method** In time series analysis, a smoothing technique based on computing the average from a fixed number of the most recent observations.

**Multicollinearity** In regression analysis, a situation where two or more explanatory variables are linearly related.

**Multinomial experiment** A series of  $n$  independent and identical trials, such that on each trial: there are  $k$  possible outcomes, called categories; the probability  $p_i$  associated with the  $i$ th category remains the same; and, the sum of the probabilities is one.

**Multiple linear regression model** In regression analysis, more than one explanatory variable is used to explain the variability in the response variable.

**Multiplication rule** The probability that  $A$  and  $B$  both occur is equal to the probability that  $A$  occurs given that  $B$  has occurred times the probability that  $B$  occurs, that is,  $P(A \cap B) = P(A|B)P(B)$ .

**Mutually exclusive events** Events that do not share any common outcome of an experiment.

## N

**Negatively skewed (left-skewed) distribution** A distribution in which extreme values are concentrated in the left tail of the distribution.

**Nominal data** Values of a qualitative variable that differ merely by name or label.

**Nominal returns** Returns that have not been adjusted for a change in purchasing power due to inflation.

**Noncausal forecasting models** Quantitative forecasts that do not present any explanation of the mechanism generating the variable of interest and simply provide a method for projecting historical data. Also called purely time-series models.

**Nonparametric tests** Statistical tests that rely on fewer assumptions concerning the distribution of the underlying population. These tests are often used when the underlying distribution is not normal and the sample size is small.

**Nonresponse bias** A systematic difference in preferences between respondents and nonrespondents of a survey or a poll.

**Normal curve** A graph depicting the normal probability density function; also referred to as the bell curve.

**Normal distribution** The most extensively used probability distribution in statistical work and the cornerstone of statistical inference. It is symmetric and bell-shaped and is completely described by the mean and the variance.

**Null hypothesis ( $H_0$ )** In a hypothesis test, the null hypothesis corresponds to a presumed default state of nature or status quo.

## O

**Ogive** A graph of the cumulative frequency or cumulative relative frequency distribution in which lines connect a series of neighboring points, where each point represents the upper limit of each class and its corresponding cumulative frequency or cumulative relative frequency.

**One-tailed hypothesis test** A test in which the null hypothesis is rejected only on one side of the hypothesized value of the population parameter.

**One-way ANOVA** A statistical technique that analyzes the effect of one categorical variable (factor) on the mean.

**Ordinal data** Values of a qualitative variable that can be categorized and ranked.

**Ordinary least squares (OLS)** A regression technique for fitting a straight line that is “closest” to the data. Also known as the method of least squares. It chooses the line whereby the sum of squares due to error (residual) is minimized.

**Outliers** Extreme small or large data values.

## P

**$\bar{p}$  chart** A control chart that monitors the proportion of defectives (or some other characteristic) of a production process.

**$p$ -value** In a hypothesis test, the likelihood of observing a sample mean that is at least as extreme as the one derived from the given sample, under the assumption that the null hypothesis is true.

**Paasche price index** A weighted aggregate price index based on quantities evaluated in the current period.

**Parameter** See *Population parameter*.

**Partial  $F$  test** See *test of linear restrictions*.

**Percentile** The  $p$ th percentile divides a data set into two parts: approximately  $p$  percent of the observations have values less than the  $p$ th percentile and approximately  $(100 - p)$  percent of the observations have values greater than the  $p$ th percentile.

**Percent frequency** The percent (%) of observations in a category; it equals the relative frequency of the category multiplied by 100.

**Permutation formula** The number of ways to choose  $x$  objects from a total of  $n$  objects, where the order in which the  $x$  objects is listed *does matter*, is  ${}_nP_x = \frac{n!}{(n-x)!}$ .

**Pie chart** A segmented circle portraying the categories and relative sizes of some qualitative variable.

**Point estimate** The value of the point estimator derived from a given sample.

**Point estimator** A function of the random sample used to make inferences about the value of an unknown population parameter.

**Poisson distribution** A description of the probabilities associated with the possible values of a Poisson random variable.

**Poisson process** An experiment in which the number of successes within a specified time or space interval equals any integer between zero and infinity; the numbers of successes counted in nonoverlapping intervals are independent from one another; and the probability that success occurs in any interval is the same for all intervals of equal size and is proportional to the size of the interval.

**Poisson random variable** The number of successes over a given interval of time or space in a Poisson process.

**Polygon** A graph of a frequency or relative frequency distribution in which lines connect a series of neighboring points, where each point represents the midpoint of a particular class and its associated frequency or relative frequency.

**Polynomial regression models** Regression models that allow sign changes of the slope capturing the influence of an explanatory variable on the response variable. A linear model is used for no sign change, a quadratic model for one sign change, and a cubic model for two sign changes of the slope.

**Population** The complete collection of items of interest in a statistical problem.

**Population parameter** A characteristic of a population.

**Portfolio** A collection of assets. The expected return of a portfolio depends on the portfolio weights and the expected return of the individual assets. The variance of the portfolio depends on the portfolio weights, the variance of the individual assets, and the covariance between the assets.

**Portfolio weights** The relative share of assets comprising the portfolio.

**Positively skewed (right-skewed) distribution** A distribution in which extreme values are concentrated in the right tail of the distribution.

**Posterior probability** The updated probability, conditional on the arrival of new information.

**Prediction interval** In regression analysis, an interval that pertains to the individual value of the response variable defined for specific explanatory variables.

**Prior probability** The unconditional probability before the arrival of new information.

**Probability** A numerical value between 0 and 1 that measures the likelihood that an event occurs.

**Probability density function** The probability density function provides the probability that a continuous random variable falls within a particular range of values.

**Probability distribution** Every random variable is associated with a probability distribution that describes the variable completely. It is used to compute probabilities associated with a random variable.

**Probability mass function** The probability mass function provides the probability that a discrete random variable takes on a particular value.

**Probability tree** A graphical representation of the various possible sequences of an experiment.

**Producer price index (PPI)** A monthly weighted aggregate price index, computed by the U.S. Bureau of Labor Statistics, based on prices measured at the wholesale or producer level.

## Q

**Quadratic regression model** Allows U-shaped or inverted U-shaped curves, capturing the influence of an explanatory variable on the response variable.

**Quadratic trend model** In time series analysis, a model that captures either a U-shaped trend or an inverted U-shaped trend.

**Qualitative forecasting** Forecasts based on the judgment of the forecaster using prior experience and expertise.

**Qualitative variable** A variable that uses labels or names to identify the distinguishing characteristics of observations.

**Qualitative response models** See *binary choice models*.

**Quantitative forecasts** Forecasts based on a formal model using historical data for the variable of interest.

**Quantitative variable** A variable that assumes meaningful numerical values for observations.

**Quartiles** Any of the three values that divide the ordered data into four equal parts, where the first, second, and third quartiles refer to the 25th, 50th, and 75th percentiles, respectively.

## R

**R chart** A control chart that monitors the variability of a production process.

**Random variable** A function that assigns numerical values to the outcomes of an experiment.

**Random error** In regression analysis, random error is due to the omission of relevant factors that influence the response variable.

**Randomized block design** In ANOVA, allowing the variation in the means to be explained by two factors.

**Range** The difference between the maximum and the minimum values in a data set.

**Ratio data** Values of a quantitative variable that can be categorized and ranked, and in which differences between values are meaningful; in addition, a true zero point (origin) exists.

**Ratio-to-moving average** In time series analysis, a method used to isolate seasonal variations of a time series.

**Real return** Investment return that is adjusted for the change in purchasing power due to inflation.

**Regression analysis** A statistical method for analyzing the relationship between variables. The method assumes that one variable, called the response variable, is influenced by other variables, called the explanatory variables.

**Rejection region** In a hypothesis test, a range of values such that if the value of the test statistic falls into this range, then the decision is to reject the null hypothesis.

**Relative frequency distribution** A frequency distribution that shows the fraction (proportion) of observations in each category of qualitative data or class of quantitative data.

**Residual ( $e$ )** In regression analysis, the difference between the observed and the predicted of the response variable value, that is,  $e = y - \hat{y}$ .

**Residual plots** In regression analysis, the residuals are plotted sequentially or against an explanatory variable to identify model inadequacies. The model is adequate if the residuals are randomly dispersed around the zero value.

**Response variable** In regression analysis, the variable that we assume is influenced by the explanatory variable(s). It is also called the dependent variable, the explained variable, the predicted variable, or the regressand.

**Restricted model** A regression model that imposes restrictions on the coefficients.

**Right-tailed test** In hypothesis testing, when the null hypothesis is rejected if the value of the test statistic falls in the right tail of the distribution.

**Risk-averse consumer** Someone who takes risk only if it entails a suitable compensation and may decline a risky prospect even if it offers a positive expected gain.

**Risk-loving consumer** Someone who may accept a risky prospect even if the expected gain is negative.

**Risk-neutral consumer** Someone who is indifferent to risk and makes his/her decisions solely on the basis of the expected gain.

**Runs test** A nonparametric test to examine whether the elements in a sequence appear in random order.

## S

**s chart** A control chart that monitors the variability of a production process.

**Sample** A subset of a population of interest.

**Sample correlation coefficient** A sample measure that describes both the direction and strength of the linear relationship between two variables.

**Sample space** A record of all possible outcomes of an experiment.

**Sample statistic** A random variable used to estimate the unknown population parameter of interest.

**Sampling distribution** The probability distribution of an estimator.

**Scatterplot** A graphical tool that helps in determining whether or not two variables are related in some systematic way. Each point in the diagram represents a pair of known or observed values of the two variables.

**Seasonal component** Repetitions of a time series over a one-year period.

**Seasonal dummy variables** Dummy variables used to capture the seasonal component from a time series.

**Seasonal index** A measure of the seasonal variation within a time series used to deseasonalize time series data.

**Seasonally adjusted series** A time series that is free of seasonal variations.

**Selection bias** A systematic underrepresentation of certain groups from consideration for a sample.

**Serial correlation** See *Correlated observations*.

**Sharpe ratio** A ratio calculated by dividing the difference of the mean return from the risk-free rate by the asset's standard deviation.

**Sign test** A nonparametric test to determine whether significant differences exist between two populations using matched-pairs sampling with ordinal data.

**Simple linear regression model** In regression analysis, one explanatory variable is used to explain the variability in the response variable.

**Simple price index** For any item, the ratio of the price in a given time period to the price in the base period, expressed as a percentage.

**Simple random sample** A sample of  $n$  observations that has the same probability of being selected from the population as any other sample of  $n$  observations. Most statistical methods presume simple random samples.

**Skewness coefficient** A measure that determines if the data are symmetric about the mean. Symmetric data have a skewness coefficient of zero.

**Smoothing techniques** In time series analysis, methods to provide forecasts if short-term fluctuations represent random departures from the structure with no discernible systematic patterns.

**Spearman rank correlation test** A nonparametric test to determine whether two variables are correlated.

**Standard deviation** The positive square root of the variance; a common measure of dispersion.

**Standard error** The standard deviation of an estimator.

**Standard error of the estimate** The standard deviation of the residual; used as a goodness-of-fit measure for regression analysis.

**Standard normal distribution** A special case of the normal distribution with a mean equal to zero and a standard deviation (or variance) equal to one.

**Standard normal table** See  $z$  table.

**Standard transformation** A normally distributed random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$  can be transformed into the standard normal random variable  $Z$  as  $Z = (X - \mu)/\sigma$ .

**Standardize** A technique used to convert a value into its corresponding  $z$ -score.

**Statistic** See *Sample statistic*.

**Statistical quality control** Statistical techniques used to develop and maintain a firm's ability to produce high-quality goods and services.

**Stem-and-leaf diagram** A visual method of displaying quantitative data where each value of a data set is separated into two parts: a stem, which consists of the leftmost digits, and a leaf, which consists of the last digit.

**Stratified random sampling** A population is first divided up into mutually exclusive and collectively exhaustive groups, called strata. A stratified sample includes randomly selected observations from each stratum. The number of observations per stratum is proportional to the stratum's size in the population. The data for each stratum are eventually pooled.

**Sum of squares due to treatments, SST** In ANOVA, a weighted sum of squared differences between the sample means and the overall mean of the data.

**Subjective probability** A probability value based on personal and subjective judgment.

**Sum of squares due to regression (SSR)** In regression analysis, it measures the explained variation in the response variable.

**Sum of squares due to error (SSE)** In ANOVA, a measure of the degree of variability that exists even if all population means are the same. Also known as within-sample variance. In regression analysis, it measures the unexplained variation in the response variable.

**Symmetric distribution** A distribution where one side of the mean is just the mirror image of the other side.

**Systematic patterns** In time series, patterns caused by a set of identifiable components: the trend, seasonal, and the cyclical components.

**Symmetry** When one side of a distribution is a mirror image of the other side.

## T

**$t$  distribution** A family of distributions that are similar to the  $z$  distribution except that they have broader tails. They are identified by their degrees of freedom  $df$ ; as  $df$  increase, the  $t$  distribution resembles the  $z$  distribution.

**Test of independence** A goodness-of-fit test analyzing the relationship between two qualitative variables. Also called a chi-square test of a contingency table.

**Test of individual significance** In regression analysis, a test that determines whether an explanatory variable has an individual statistical influence on the response variable.

**Test of joint significance** In regression analysis, a test to determine whether the explanatory variables have a joint statistical influence on the response variable; it is often regarded as a test of the overall usefulness of a regression model.

**Test of linear restrictions** In regression analysis, a test to determine if the restrictions specified in the null hypothesis are invalid.

**Test Statistic** A sample-based measure used in hypothesis testing.

**Time series** A set of sequential observations of a variable over time.

**Time series data** Values of a characteristic of a subject over time.

**Total probability rule** A rule that expresses the unconditional probability of an event,  $P(A)$ , in terms of probabilities conditional on various mutually exclusive and exhaustive events. The total probability rule conditional on two events  $B$  and  $B^c$  is  $P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c)$ .

**Total sum of squares (SST)** In regression analysis, it measures the total variation in the response variable. It can be decomposed into explained and unexplained variations.

**Trend** The trend refers to a long-term upward or downward movement of a time series.

**Tukey's honestly significant differences (HSD) method** In ANOVA, a test that determines which means significantly differ by comparing all pairwise differences of the means.

**Two-tailed hypothesis test** A test in which the null hypothesis can be rejected on either side of the hypothesized value of the population parameter.

**Two-tailed test** In hypothesis testing, when the null hypothesis is rejected if the value of the test statistic falls in either the left tail or the right tail of the distribution.

**Two-way ANOVA test** A test that simultaneously examines the effect of two factors on the mean.

**Two-way ANOVA test with interaction** A two-way ANOVA test that captures the possible relationship between the two factors.

**Two-way ANOVA tests without interaction** A two-way ANOVA test that does not capture the possible relationship between the two factors.

**Type I error** In a hypothesis test, this error occurs when the decision is to reject the null hypothesis when the null hypothesis is actually true.

**Type II error** In a hypothesis test, this error occurs when the decision is to not reject the null hypothesis when the null hypothesis is actually false.

## U

**Unbalanced data** A completely randomized ANOVA design where the number of observations are not the same for each sample.

**Unbiased** An estimator is unbiased if its expected value equals the unknown population parameter being estimated.

**Unconditional probability** The probability of an event without any restriction.

**Union** The union of two events  $A$  and  $B$ , denoted  $A \cup B$ , is the event consisting of all outcomes in  $A$  or  $B$ .

**Unrestricted model** A regression model that imposes no restrictions on the coefficients.

**Unsystematic patterns** In time series, patterns caused by the presence of a random error term.

**Unweighted aggregate price index** An aggregate price index based entirely on aggregate prices with no emphasis placed on quantity.

**Upper control limit** In a control chart, the upper control limit indicates excessive deviation above the expected value of the variable of interest.

## V

**Variable** A general characteristic being observed on a set of people, objects, or events, where each observation varies in kind or degree.

**Variance** The average of the squared differences from the mean; a common measure of dispersion.

## W

**Wald-Wolfowitz runs test** A nonparametric test to determine whether the elements in a sequence appear in a random order.

**Weighted aggregate price index** An aggregate price index that gives higher weight to the items sold in higher quantities.

**Weighted mean** When some observations contribute more than others in the calculation of an average.

**Wilcoxon rank-sum test** A nonparametric test to determine whether two population medians differ under independent sampling. Also known as the Mann-Whitney test.

**Wilcoxon signed-rank test** A nonparametric test to determine whether a sample could have been drawn from a population having a hypothesized value as its median; this test can also be used to determine whether the median difference differs from zero under matched-pairs sampling.

**Within-treatments variance** In ANOVA, a measure of the variability within each sample.

## X

**$\bar{x}$  chart** A control chart that monitors the central tendency of a production process.

## Z

**z-score** The relative position of a value within a data set; it is also used to detect outliers.

**z table** A table providing cumulative probabilities for positive or negative values of the standard normal random variable  $Z$ .



## PHOTO CREDITS

### Chapter 1

**Opener:** © Randy Lincks/All Canada Photos/Getty Images; p. 7: © Comstock/Stockbyte/Getty Images RF; p. 9: © Agencja Fotograficzna Caro/Alamy Stock Photo; p. 13: © Dennis Welsh/UpperCut Images/Getty Images RF.

### Chapter 2

**Opener:** © Mitch Diamond/Photodisc/Getty Images RF; p. 18: © sbk\_20d pictures/Moment/Getty Images RF; p. 37: © Brand X Pictures/Stockbyte/Getty Images Plus/Getty Images RF; p. 47: © rubberball/Getty Images RF.

### Chapter 3

**Opener:** © Mark Bowden/iStock/Getty Images Plus/Getty Images RF; p. 64: © Sebastian Pfeutze/Taxi/Getty Images; p. 81: © Ingram Publishing RF; p. 96: © Mike Watson Images/moodboard/Getty Images Plus/Getty Images RF.

### Chapter 4

**Opener:** © Fab Fernandez/Image Source/Getty Images RF; p. 114: © Gene J. Puskar/AP Images; p. 129: © Digital Vision/Photodisc/Getty Images RF; p. 141: © Rolf Bruderer/Blend Images/Getty Images RF.

### Chapter 5

**Opener:** © Jewel Samad/AFP/Getty Images; p. 177: © Bubbles Photolibrary/Alamy; p. 182: © Image Source/Getty Images RF.

### Chapter 6

**Opener:** © Vision SRL/Photodisc/Getty Images RF; p. 210: © Natalia Lisovskaya/Shutterstock; p. 220: © Image Source, all rights reserved. RF.

### Chapter 7

**Opener:** © Chris Hondros/Getty Images News/Getty Images; p. 247: © Joe Raedle/Getty Images News/Getty Images; p. 257: © Ryan McVay/Photodisc/Getty Images RF.

### Chapter 8

**Opener:** © Uli Deck/picture-alliance/dpa/AP Images; p. 289: © McGraw-Hill Companies, Inc. Mark Dierker, photographer RF; p. 291(left): © The McGraw-Hill Companies, Inc./Andrew Resek, photographer RF; p. 291(right): © Paul Sakuma/AP Images.

### Chapter 9

**Opener:** © Ken Seet/Corbis Images/SuperStock RF; p. 322: © Asia Images Group/Getty Images RF; p. 330: © Ariel Skelley/Blend Images LLC RF.

### Chapter 10

**Opener:** © John Smock/SIPA/Newscom; p. 356: © Chris Hondros/Getty Images News/Getty Images; p. 366(top): © JGI/Blend Images LLC RF; p. 366(bottom): © STOCK4B-RF/Getty Images RF.

### Chapter 11

**Opener:** © Hero Images/Getty Images RF; p. 391: © Spencer Grant/PhotoEdit—All rights reserved.; p. 394: © Gerry Broome/AP Images.

### Chapter 12

**Opener:** © Hongqi Zhang/iStock/Getty Images Plus/Getty Images RF; p. 414: © Jean Baptiste Lacroix/WireImage/Getty Images; p. 422: © JGI/Blend Images/Getty Images RF.

### Chapter 13

**Opener:** © Jorge Garrido/Alamy; p. 447: © PBNJ Productions/Getty Images; p. 464: © Mitchell Funk/Photographer's Choice/Getty Images.

### Chapter 14

**Opener:** © Blend Images-JGI/Jamie Grill/Band X Pictures/Getty Images RF; p. 503: © David J. Phillip/AP Images; p. 506: © Kick Images/Photodisc/Getty Images RF.

### Chapter 15

**Opener:** © Rob Tringali/MLB Photos via Getty Images; p. 523: © Mark Cunningham/MLB Photos via Getty Images; p. 546: © Elsa/Getty Images Sport/Getty Images.

### Chapter 16

**Opener:** © Yellow Dog Productions/Digital Vision/Getty Images RF; p. 575: © Patrick Cooper/Shutterstock; p. 578: © Photodisc/Getty Images RF.

### Chapter 17

**Opener:** © Asia Images Group/Getty Images RF; p. 603: © Tom Stewart/Corbis/Getty Images; p. 613: © Mark J. Terrill, Pool/AP Images.

### Chapter 18

**Opener:** © Pietro Scozzari/age fotostock; p. 647: © David Paul Morris/Bloomberg via Getty Images; p. 653: © Evan Vucci/AP Images.

### Chapter 19

**Opener:** © imageBROKER/Alamy RF; p. 674: © Peter Titmuss/Alamy; p. 681: © Bettmann/Getty Images.

### Chapter 20

**Opener:** © monkeybusinessimages/iStock/Getty Images Plus/Getty Images RF; p. 709: © Ken Reid/Photographer's Choice/Getty Images; p. 719: © gkrphoto/Getty Images.





## A

Acceptance sampling, 251  
 Addition rule, 117–118  
 Adidas, 107, 126, 127, 403, 411, 414  
 Adjusted closing price, 665–666  
 Adjusted coefficient of determination, 502–503, 638  
 Adjusted seasonal index, 642–643  
 Aggregate price indices, 670  
     unweighted, 670–671  
     weighted, 671–673  
 Akiko Hamaguchi, 191, 210  
 Alpha, 519  
 Alstead, Troy, 4  
 Alternative hypotheses, 302, 303–305  
 American Public Transportation Association, 447  
 Analysis of variance (ANOVA); *see also* One-way ANOVA; Two-way ANOVA  
     ANOVA table, 437  
     defined, 434  
     uses of, 432, 434  
 Annualized return, 73–74  
 ANOVA; *see* Analysis of variance; One-way ANOVA; Two-way ANOVA  
 ANOVA table, 437  
 Anti-log function, 567  
 Arithmetic mean, 60–61, 74  
 Arizona Cardinals, 114  
 Asset returns; *see* Returns  
 Assets; *see* Portfolios  
 Assignable variation, 252  
 Associated Press, 330  
 Asymptotic distributions, 196  
 Autoregressive models, 650  
 Average growth rates, 74–75  
 Average study time, hypothesis tests, 301, 322  
 Averages; *see* Mean

## B

Balanced data, 444  
 Bar charts, 19, 21–22  
 Barnes, Valerie, 681  
 Bayes, Thomas, 134  
 Bayes' theorem, 131, 134–136  
 BEA; *see* Bureau of Economic Analysis  
 Beane, Jared, 269, 270, 289  
 Bell curve, 196; *see also* Symmetric distributions  
 Bell Telephone Laboratories, 252  
 Bell-shaped distribution, 86  
 Bernoulli, James, 166  
 Bernoulli process, 166, 404  
 Beta, 519  
 Between-treatments estimate of population variance, 435–436  
 Between-treatments variance, 435  
 Bias, 232  
 Binary choice models  
     defined, 606  
     linear probability model, 606–607, 608  
     logit model, 607–609

Binomial experiments, 166–169, 404  
 Binomial probability distributions  
     constructing, 166–168  
     defined, 166  
     with Excel, 171  
     formula, 168–169  
     normal approximation, 209  
     sampling distribution of sample proportion, 244  
 Binomial random variable  
     defined, 166, 168  
     expected values, 169  
     standard deviation, 169  
     variance, 169  
 Blocked outcomes, 451  
*Bloomberg Businessweek*, 247  
 BLS; *see* Bureau of Labor Statistics  
 Bonds; *see* Returns  
 Boston Celtics, 613  
*Boston Globe* poll, 4  
 Box plots, 70–71  
 Box-and-whisker plot, 70–71  
 Brewery, Guinness, 277  
 Brown, Scott, 4, 235–236  
 Bryant, Kobe, 613  
 Bureau of Economic Analysis (BEA), 7  
 Bureau of Labor Statistics (BLS), 6, 7, 677  
 Business cycles, 640

## C

*c* chart, 252  
 Capital asset pricing model (CAPM), 519–529  
 Capital gains yields, 664  
 CAPM; *see* Capital asset pricing model  
 Causal forecasting models, 624, 650–651  
 Causation, correlation and, 5, 481  
 Census Bureau, U.S., 7, 61  
 Centered moving average (CMA), 641–642  
 Centerline, 252  
 Central limit theorem (CLT)  
     sampling distribution of sample mean, 240–241  
     sampling distribution of sample proportion, 245  
 Central location measures  
     arithmetic mean, 60–61, 74  
     defined, 60  
     Excel calculations, 64–67  
     geometric mean, 73–75  
     median, 61–63  
     mode, 63–64  
     weighted mean, 66, 90  
 Chance variation, 251–252  
 Charts; *see* Control charts; Graphical displays  
 Chebyshev, Pavroty, 85  
 Chebyshev's theorem, 85–86, 87  
 Chi-square distribution  
     characteristics, 376–377  
     defined, 376  
     degrees of freedom, 376–377  
     locating values and probabilities, 377–379  
 Chi-square table, 378

- Chi-square tests
    - goodness-of-fit
      - for independence, 410–414
      - for multinomial experiments, 404–408
      - for normality, 416–418
    - Jarque-Bera, 419
  - Classes, 25–27, 28
  - Classical probability, 113
  - Clinton, Hillary, 234
  - CLT; *see* Central limit theorem
  - Cluster sampling, 235
  - Clusters, 235
  - CMA; *see* Centered moving average
  - Coakley, Martha, 4, 235–236
  - Coefficient of determination, 500–502, 574
  - Coefficient of variation (CV), 80
  - Combination formula, 139
  - Complement, of event, 110, 117
  - Complement rule, 117
  - Conditional probability, 119–120, 134
  - Confidence coefficient, 272
  - Confidence intervals
    - defined, 270
    - for differences in means, 340–342
    - individual significance tests, 516–519
    - interpreting, 272
    - margin of error, 270–271, 287, 288
    - for mean difference, 351–352
    - misinterpretation, 272
    - of population mean
      - with known standard deviation, 271–275
      - with unknown standard deviation, 277–281
    - of population proportion, 284–285, 288–289
    - of population variance, 379–380
    - for population variance ratio, 388
    - predicted values, 532–533
    - for proportion differences, 360–361
    - sample sizes and, 287–289
    - two-tailed hypothesis tests, 315–316
    - using Excel, 275, 281
    - width and precision, 273–274
  - Constants, 237
  - Consumer Price Index (CPI), 670, 677–678, 681; *see also* Inflation rates; Price indices
  - Consumers, risk and, 159–160
  - Contingency tables
    - chi-square tests, 410–411
    - defined, 126, 410
    - using, 126–129
  - Continuous probability distributions; *see also* Normal distribution
    - exponential, 213–215
    - lognormal, 216–218
    - uniform, 193
  - Continuous random variables, 9, 152, 192
  - Continuous uniform distribution, 193
  - Continuous variables, 9
  - Control charts
    - defined, 252
    - $\bar{p}$  charts, 252, 254–255
    - for qualitative data, 252, 254–255
    - for quantitative data, 252, 253–254
    - $\bar{x}$  charts, 252
  - Control variables; *see* Explanatory variables
  - Correlation
    - causation and, 5, 481
    - spurious, 5, 481
  - Correlation analysis
    - hypothesis tests, 480
    - limitations of, 481
  - Correlation coefficient
    - calculating, 94
    - defined, 93
    - hypothesis tests, 480
    - population, 93, 480
    - of portfolio returns, 164
    - sample, 93, 479
    - Spearman rank, 707
  - Correlation-to-causation fallacy, 5
  - Counting rules
    - combination formula, 139
    - factorial formula, 138
    - permutation formula, 139
  - Covariance
    - calculating, 94
    - defined, 92
    - population, 92
    - of portfolio returns, 164
    - sample, 92, 478
  - Cox, Sean, 433, 447
  - CPI; *see* Consumer Price Index
  - Critical values
    - defined, 312
    - four-step procedure and, 314
    - hypothesis testing, 312–315
  - Cross-sectional data, 6, 7
  - Cubic regression model, 563–564
    - predictions, 563
  - Cubic trend models, 638
  - Cumulative distribution function, 153, 154, 192
  - Cumulative frequency distribution
    - defined, 28
    - ogives, 35
  - Cumulative relative frequency distributions, 29
  - Curvilinear relationships, 44; *see also* Polynomial regression models
  - CV; *see* Coefficient of variation
  - Cyclical patterns, 626, 640
- ## D
- Data; *see also* Time series data; Variables
    - cross-sectional, 6, 7
    - measurement scales, 9–12
    - sources, 7
    - standardizing, 87
    - types, 6–7
    - websites, 7
  - Deciles

- Decision rules
  - critical value approach, 312, 313
  - $p$ -value approach, 309, 310
- Decomposition analysis, 640–641
  - extracting seasonality, 641–643
  - extracting trend, 643–644
  - forecasting, 644–645
- Deflated time series, 676–678
- Degrees of freedom ( $df$ ), 278
- Dependent events, 121
- Dependent variable; *see* Response variable
- Descriptive statistics, 5
- Detection approach, 251
- Deterministic relationships, 484
- $df$ ; *see* Degrees of freedom
- Discrete choice models; *see* Binary choice models
- Discrete probability distributions, 153–156
  - binomial
    - constructing, 166–168
    - defined, 166
    - with Excel, 171
    - formula, 168–169
    - sampling distribution of sample proportion, 244
  - discrete uniform distribution, 155
  - graphical displays, 154
  - hypergeometric, 178–180
  - Poisson, 173–176
  - properties of, 154
- Discrete random variables, 8–9, 152, 153
  - expected value, 158
  - standard deviation, 158
  - variance, 158
- Discrete uniform distribution, 155
- Discrete variables, 8–9
- Dispersion measures, 77
  - coefficient of variation, 80
  - Excel calculations, 80–81
  - mean absolute deviation, 77–78
  - range, 77
  - standard deviation, 78–79, 85–88
  - variance, 78–79
- Distribution-free tests; *see* Nonparametric tests
- DJIA; *see* Dow Jones Industrial Average
- Dow Jones Industrial Average (DJIA), 9
- Dummy variable trap, 594
- Dummy variables
  - defined, 590
  - interaction variables and, 599–603
  - for multiple categories, 593–596
  - seasonal, 640, 645–647
  - significance tests, 592, 600
  - use of, 590
- Duracel, 6

## E

- The Economist*, 7
- Edwards, Matthew, 17
- Empirical probability, 112, 113
- Empirical rule, 86–87, 202–204

- Endogeneity, 537
- Error sum of squares ( $SSE$ )
  - ordinary least squares, 486
  - in two-way ANOVA, 452–453, 460
  - within-treatments estimate of population variance, 436
- Errors; *see also* Margin of error; Standard error
  - random, 626
  - Type I errors, 305–306
  - Type II errors, 305–306
- espn.com, 7
- Estimates, 237
- Estimation; *see* Confidence intervals; Maximum likelihood estimation (MLE)
- Estimators, 237
- Events
  - complement of, 110, 117
  - defined, 108, 109
  - dependent, 121
  - exhaustive, 109
  - independent, 121, 123
  - intersection of, 109–110
  - mutually exclusive, 109, 119
  - probabilities, 108–110
  - union of, 109, 110, 117
- Excel
  - ANOVA problems, 437–439
  - bar charts, 21–22
  - central location measures, 64–67
  - chi-square tests, 406–408
  - confidence intervals, 275, 281
  - control charts, 255
  - correlation coefficient, 94, 480
  - covariance, 94, 480
  - dispersion measures, 80–81
  - exponential distribution, 215
  - exponential smoothing, 631
  - histogram, 31–33
  - hypergeometric probabilities, 180
  - hypothesis testing, 316–317
  - hypothesis tests for differences in means, 344–346
  - hypothesis tests for mean difference, 354–355
  - lognormal distribution, 218
  - moving averages, 631
  - multiple regression, 493–494
  - normal distribution, 209
  - ogives, 36
  - pie charts, 21
  - Poisson probabilities, 176
  - polygon, 34–35
  - $p$ -values, 321–322, 382, 390–391
  - regression statistics, 499–500
  - residual plots, 538
  - sample regression equation, 488–489
  - scatterplots, 45, 486–488
  - smoothing techniques, 631
  - standard error of the estimate, 499–500
  - trendlines, 486–488
  - two-way ANOVA
    - with interaction, 460–461
    - no interaction, 453–455

- Exhaustive events, 109
- Expected frequencies, 411–413
- Expected portfolio returns, 163–164
- Expected return of the portfolio, 163
- Expected value, 158
- Experiments; *see also* Events
  - binomial, 166–169, 404
  - defined, 108
  - multinomial, 404–408
  - outcomes, 108
- Explained variation, 500
- Explanatory variables; *see also* Dummy variables
  - assumption and, 544
  - defined, 483
  - multicollinearity, 540–541
  - in multiple regression, 492
  - qualitative, 590
  - quantitative, 590
  - in simple linear regression, 483
- Exponential function, 567
- Exponential model, 569, 570–574, 571
- Exponential probability distribution, 213–215
- Exponential smoothing, 628–631
- Exponential trend models, 634–637
  - estimating, 634–635
  - formula, 635
  - with seasonal dummy variables, 646

## F

- F* distribution
  - characteristics, 385–386
  - locating values and probabilities, 386–387
- F* table, 386–387
- F* test
  - one-tailed, 521
  - partial, 527–530
- Factorial formula, 138
- Fahrenheit scale, 12
- Farnham, Matthew, 506
- Federal Reserve, 106
- Federal Reserve Economic Data (FRED), 7
- Fidelity Investments, 59, 81, 375, 391, 687
- Finite population correction factor, 248–250
- Fisher, Irving, 666
- Fisher equation, 666
- Fisher's least significant difference (LSD) method, 442–444, 445, 455
- Fitted value, 625
- Forecasting; *see also* Smoothing techniques
  - causal models, 624, 650–651
  - decomposition analysis, 644–645
  - model selection, 625
  - noncausal models, 624
  - qualitative, 624
  - quantitative, 624
  - trend models
    - exponential, 634–637
    - linear, 633–634, 645
    - polynomial, 637–638
    - quadratic, 637–638

- Fortune*, 7
- Fraction defective charts, 254
- FRED; *see* Federal Reserve Economic Data
- Frequencies, expected, 411–413
- Frequency distributions
  - classes, 25–27, 28
  - constructing, 26–27
  - cumulative, 28
  - defined, 28
  - graphical displays
    - charts, 19–23
    - histograms, 30–33
    - ogives, 35–36
    - polygons, 33–35
  - qualitative data, 18–23
  - quantitative data, 25–36
  - relative, 19, 29

## G

- Gallup Poll, 5
- Geometric mean, 73–75
- Geometric mean return, 73–74
- Goodness-of-fit tests
  - adjusted coefficient of determination, 502–503
  - coefficient of determination, 500–502
  - for independence, 410–414
  - for multinomial experiments, 404–408
  - for normality, 416–418
  - for regression analysis, 497
  - standard error of the estimate, 497–500
- Google, 7
- Gossett, William S., 277
- Grand mean, 435
- Graphical displays
  - bar charts, 19, 21–22
  - box plot, 70–71
  - guidelines, 22–23
  - histograms, 30–33
  - ogives, 35–36
  - pie charts, 19–21
  - polygons, 33–35
  - residual plots, 537
  - scatterplots, 43–46, 478, 486–488
  - stem-and-leaf diagrams, 41–43
- Great Depression, 232
- Grouped data, summarizing, 89–91
- Growth rates, average, 74–75

## H

- Heteroskedasticity, 537
- Histograms
  - constructing, 31–33
  - defined, 30
  - relative frequency, 30
  - shapes, 31
- Hoover, Herbert, 232
- HSD; *see* Tukey's honestly significant differences method
- Hypergeometric probability distributions, 178–180

Hypergeometric random variables, 179

Hypotheses

alternative, 302, 303–305

null, 302, 303–305

Hypothesis tests; *see also* Analysis of variance;

Nonparametric tests

for correlation coefficient, 480

for differences in means, 342–344

of individual significance, 516–519

interpreting results, 317

of joint significance, 521–522

left-tailed, 303, 314

for mean difference, 352–354

one-tailed, 303–305

of population mean, with known standard deviation,  
307–317

critical value approach, 312–315

*p*-value approach, 308–312

of population mean, with unknown standard deviation,  
319–322

of population median, 690

of population proportion, 325–328

for population variance, 380–381

for population variance ratio, 388–390

for proportion differences, 361–363

rejecting null hypothesis, 302–303

right-tailed, 303, 314

significance level, 309

two-tailed, 303–305, 314, 315–316

Type I and Type II errors, 305–306

using Excel, 316–317

## I

IBM, 6

Iced coffee, 247

Income yields, 664

Independence, chi-square test for, 410–414

Independent events, 121, 123

Independent random samples

defined, 340

Wilcoxon rank-sum test, 693

Index numbers; *see also* Indices

base periods, 668

defined, 668

Indicator variables; *see* Dummy variables

Indices; *see also* Price indices

base periods, 668

seasonal

adjusted, 642–643

unadjusted, 642

Individual significance, tests of, 516–519

Inferential statistics, 5

Inflation; *see also* Price indices

effects, 676, 677

returns and, 666

Inflation rates

calculating, 678–679

defined, 678

expected, 666

In-sample criteria, 625

Instrumental variables, 544

Interaction variables, 599–603

Interquartile range (IQR), 71, 77

Intersection, of events, 109–110

Interval, 270

Interval estimate, 270; *see also* Confidence intervals

Interval estimates, 532–535

Interval scale, 12

Inverse transformation, 207–209

Investment returns; *see also* Returns

calculating, 664

income component, 664

price change component, 664

IQR; *see* Interquartile range

## J

Jarque-Bera test, 419

Johnson, Rebecca, 687

Johnson & Johnson, 520

Joint probability, 128

Joint probability table, 128

Joint significance, tests of, 521–522

Jones, Anne, 151

## K

Kennedy, Ted, 4, 235

Knight, Susan, 301

Kruskal-Wallis test, 701–703

test statistics, 702

Kurtosis, 66

Kurtosis coefficients, 419

## L

Lagged regression models, 650–651

Landon, Alf, 232

Laspeyres price index, 671–672, 673

Law of large numbers, 113

LCL; *see* Lower control limit

Leach, Ben, 546

Left-tailed hypothesis, 303, 314

Linear probability model (LPM), 606–607, 608

Linear relationships; *see also* Multiple regression model;

Simple linear regression model

correlation coefficient, 93, 479, 481

covariance, 92, 478

multicollinearity, 540–541

negative, 92

positive, 92

sample regression equation, 485–486

scatterplots, 44, 478, 486–488

slopes, 484

Linear restrictions, general test of, 527–530

Linear trend models

forecasts, 633–634

formula, 633

with seasonal dummy variables, 645

*Literary Digest* polls, 232–233

Little Ginza, 191, 210

Logarithmic model, 569–570, 571

- Logarithms
  - exponential model, 569, 570–574, 571
  - logarithmic model, 569–570, 571
  - log-log model, 567–569, 571
  - natural, 567
  - semi-log model, 569, 571
- Logit model, 607–609
- Log-log model, 567–569, 571
- Lognormal distribution, 216–218
- Los Angeles Lakers, 613
- Lower control limit (LCL), 252–253
- LPM; *see* Linear probability model
- LSD; *see* Fisher's least significant difference method

## M

- MAD; *see* Mean absolute deviation
- Main effects, 461
- Major League Baseball (MLB), 515, 523, 546
- Mann-Whitney test; *see* Wilcoxon rank-sum test
- Margin of error, confidence interval, 270–271, 287, 288
- Marginal probability, 128
- Matched outcomes, 451
- Matched-pairs sampling
  - mean difference, 351–355
  - recognizing, 351
  - Wilcoxon signed-rank test, 355, 694
- Matthews, Robert, 481
- Maximum likelihood estimation (MLE), 608
- McCaffrey, Luke, 3
- McDonald's, 213
- Mean; *see also* Population mean; Sample mean
  - arithmetic, 60–61, 74
  - for frequency distribution, 89–91
  - geometric, 73–75
  - moving average, 626–628
  - weighted, 66, 90
- Mean absolute deviation (MAD), 77–78
  - forecasting models and, 625
- Mean difference
  - confidence intervals, 351–352
  - hypothesis test, 352–354
  - matched-pairs experiments, 351–355
- Mean square error (MSE), 436, 460, 521
  - forecasting models, 625
- Mean square for factor *A* (MSA), 452, 459
- Mean square for factor *B* (MSB), 452, 459
- Mean square for interaction (MSAB), 460
- Mean square for treatments (MSTR), 435
- Mean square regression (MSR), 521
- Mean-variance analysis, 83–84
- Measurement scales, 9
  - interval, 12
  - nominal, 9–10, 18
  - ordinal, 10–12, 18
  - ratio, 12
- Median, 61–63; *see also* Population median
  - calculating, 62–63
  - defined, 61
- Method of least squares; *see* Ordinary least squares (OLS)
- Method of runs above and below median, 716

- Michigan State University, 681
- Microsoft Corporation, 664
- MLB; *see* Major League Baseball
- MLE; *see* Maximum likelihood estimation
- Mode, 63–64
- Model selection, for forecasting, 625
- Moving average methods, 626–628, 631
- m*-period moving average, 626–628
- MSA; *see* Mean square for factor *A*
- MSAB; *see* Mean square for interaction
- MSB; *see* Mean square for factor *B*
- MSE; *see* Mean square error
- MSR; *see* Mean square regression
- MSTR; *see* Mean square for treatments
- Multicollinearity, 537, 540–541
- Multinomial experiments, 404
- Multiple *R*, 502
- Multiple regression model; *see also* Regression analysis
  - defined, 492
  - explanatory variables, 492
  - interval estimates for predictions, 532–535
  - response variable, 492
  - sample regression equation, 492–493
  - test of linear restrictions, 527–530
  - tests of individual significance, 516–519
  - tests of joint significance, 521–522
- Multiplication rule, 122–123
- Mutually exclusive events, 109, 119

## N

- n* factorial, 138
- Nasdaq; *see* National Association of Securities Dealers
  - Automated Quotations
- National Association of Securities Dealers Automated
  - Quotations (Nasdaq), 9
- National Basketball Association (NBA), 613
- Natural logarithms, 567
- NBA; *see* National Basketball Association
- Negatively skewed distributions, 31, 66
- New York Stock Exchange (NYSE), 9–10, 624
- The New York Times*, 7
- Nike, Inc., 107, 126, 127, 128, 403, 411, 414, 623, 641, 644–645, 647
- Nominal returns, 666; *see also* Returns
- Nominal scale, 9–10, 18
- Nominal terms, 676
- Noncausal models, 624
- Nonlinear models; *see* Polynomial regression models
- Nonparametric tests
  - compared to parametric tests, 688, 708
  - disadvantages, 688
  - Kruskal-Wallis test, 701–703
  - sign test, 711–713
  - Spearman rank correlation test, 705–708
  - Wilcoxon rank-sum test, 346, 355, 694–698
  - Wilcoxon signed-rank test
    - for matched-pairs sample, 694–695
    - for population median, 688–692
    - software, 698
- Nonresponse bias, 233



Normal curve, 197  
 Normal distribution  
   characteristics, 196–197  
   defined, 193  
   empirical rule, 86–87, 202–204  
   of error term, 537  
   probability density function, 196–197  
   standard, 198  
   using Excel, 209  
 Normal transformation, 205–207  
 Normality  
   goodness-of-fit test, 416–418  
   Jarque-Bera test, 419  
 Null hypotheses, 302, 303–305; *see also* Hypothesis tests  
   defined, 302  
   formulating, 303  
   rejecting, 302–303  
   Type I and Type II errors, 305–306  
 NYSE; *see* New York Stock Exchange

**O**

Obama, Barack, 5, 234  
 Objective probabilities, 113  
 Odds ratios, 114  
 Ogives, 35–36  
 Ohio State University, 5  
 OLS; *see* Ordinary least squares  
 One-tailed hypothesis, 303–305  
 One-way ANOVA  
   between-treatments estimate, 435–436  
   defined, 434  
   samples, 434  
   test statistic, 436  
   within-treatments estimate, 435, 436–437  
 Ordinal scale, 10–12, 18  
 Ordinary least squares (OLS)  
   lines fit by, 486  
   normal distribution, 516  
   properties, 537  
 Outliers, 61  
 Out-of-sample criteria, 625

**P**

Paasche index, 672, 673  
 Parameters; *see* Population parameters  
 Parametric tests; *see also* *F* test; *t* test  
   assumptions, 688  
   compared to nonparametric tests, 688, 708  
 Partial *F* test, 527–530  
 $\bar{p}$  charts, 252, 254–255  
 Pearson correlation coefficient; *see* Correlation coefficient  
 Percent defective chart, 254  
 Percent frequencies, 19  
 Percentiles  
   box plot, 70–71  
   calculating, 69–70  
   defined, 69  
 Permutation formula, 139  
 Pie charts, 19–21  
 Pinnacle Research, 269

Pittsburgh Steelers, 114  
 Poisson, Simeon, 173  
 Poisson probability distribution, 173–176  
 Poisson process, 174  
 Poisson random variables, 173, 174–176  
 Polls, 5, 232–233, 234–235  
 Polygons, 33–35  
   defined, 33  
 Polynomial regression models  
   cubic, 563–564  
   defined, 563  
   quadratic  
   compared to linear model, 558–563  
   formula, 558  
   predictions, 558  
 Polynomial trend models, 637–638  
 Pooled estimates, 341  
 Population  
   defined, 5, 232  
   samples and, 6, 232  
 Population coefficient of variation, 80  
 Population correlation coefficient, 93, 480  
 Population covariance, 92  
 Population mean, 158  
   confidence intervals  
   of differences, 340–342  
   with known standard deviation, 271–275  
   with unknown standard deviation, 277–281  
   differences between  
   confidence intervals, 340–342  
   Fisher's LSD method, 442–444, 445, 455  
   hypothesis test, 342–344  
   test statistic, 343  
   Tukey's HSD method, 442, 444–446, 455  
   expected value, 158  
   formula, 60  
   matched-pairs experiments, 351  
 Population mean absolute deviation, 78  
 Population median  
   Kruskal-Wallis test, 701–703  
   matched-pairs experiments, 694–695  
   Wilcoxon rank-sum test, 346, 355, 695–698  
   Wilcoxon signed-rank test  
   for matched-pairs sample, 694–695  
   for population median, 694–695  
   software, 698  
 Population parameters  
   constants, 237  
   defined, 5  
   inferences about, 6, 232, 237  
   mean as, 61  
 Population proportion  
   confidence intervals, 284–285, 288–289  
   differences between  
   applications, 359–360  
   confidence intervals, 360–361  
   hypothesis test, 361–363  
   test statistic, 362  
   hypothesis tests, 325–328

- Population standard deviation, 79
  - Population variance
    - between-treatments estimate, 435–436
    - confidence intervals, 379–380
    - formula, 79
    - hypothesis test, 380–381
    - quality control studies, 376
    - ratio between
      - confidence intervals, 388
      - estimating, 385
      - hypothesis test, 388–390
      - test statistic, 389
    - test statistic, 380
    - within-treatments estimate, 435, 436–437
  - Portfolio returns; *see also* Returns
    - correlation coefficient, 164
    - covariance, 164
    - expected, 163–164
    - standard deviation, 164
    - variance, 163–164
  - Portfolio risk, 164
  - Portfolio weights, 163
  - Portfolios, defined, 162
  - Positively skewed distributions, 31, 66
  - Posterior probability, 134
  - PPI; *see* Producer price index
  - Precision, confidence intervals, 274
  - Predicted value, 625
  - Prediction intervals, 532, 535
  - Predictions
    - cubic regression model, 563
    - quadratic regression model, 558
  - Predictor variables; *see* Explanatory variables
  - Price indices; *see also* Inflation rates
    - aggregate, 670
      - unweighted, 670–671
      - weighted, 671–673
    - base periods, 668
    - base values, 668
    - CPI, 670, 677–678, 681
    - deflated time series, 676–678
    - Laspeyres, 671–672, 673
    - Paasche, 672, 673
    - PPI, 670, 677–678
    - simple, 668
  - Prices
    - adjusted closing price, 665–666
    - capital gains yields, 664
  - Prior probability, 134
  - Probabilities
    - assigning, 111–113
    - classical, 113
    - concepts, 108–114
    - conditional, 119–120, 134
    - defined, 108
    - empirical, 112, 113
    - joint, 128
    - marginal, 128
    - objective, 113
    - odds ratios, 114
    - posterior, 134
    - prior, 134
    - properties, 111
    - rules (*see* Probability rules)
    - subjective, 111, 113
    - unconditional, 120, 134
  - Probability density function, 153, 192
  - Probability distributions, 153
    - continuous (*see* Continuous probability distributions)
    - discrete (*see* Discrete probability distributions)
  - Probability mass function, 153
  - Probability rules, 117–123
    - addition rule, 117–118
    - complement rule, 117
    - multiplication rule, 122–123
    - total, 131–134, 136
  - Probability tree, 132, 166, 168
  - Producer price index (PPI), 670, 677–678; *see also* Price indices
  - Proportion; *see* Population proportion; Sampling distribution of sample proportion
  - p*-values
    - defined, 308
    - four-step procedure and, 311
    - hypothesis testing, 308–312, 382
- ## Q
- Quadratic regression model
    - compared to linear model, 558–563
    - formula, 558
    - predictions, 558
  - Quadratic trend models, 637–638
  - Qualitative forecasting, 624; *see also* Forecasting
  - Qualitative response models; *see* Binary choice models
  - Qualitative variables; *see also* Binary choice models; Dummy variables; Population proportion
    - contingency tables, 126–129
    - control charts, 254–255
    - defined, 8
    - frequency distributions, 18–23
    - graphical displays, 19–23
    - nominal, 9–10, 18
    - ordinal, 10–12, 18
    - in regression, 590
  - Quality control; *see* Statistical quality control
  - Quantitative forecasting, 624; *see also* Forecasting
  - Quantitative variables
    - control charts, 252, 253–254
    - defined, 8
    - frequency distributions, 25–36
    - interval, 12
    - ratio, 12
    - in regression, 590
    - summarizing, 25–26
- ## R
- R* chart, 252
  - Random (irregular) error, 626

- Random samples
    - for ANOVA, 434
    - independent
      - defined, 340
    - simple, 233–234
    - stratified, 234–235
  - Random variables
    - binomial, 166, 168
    - continuous, 9, 152, 192
    - defined, 152
    - discrete, 8–9, 152
    - exponential, 213–219
    - hypergeometric, 179
    - lognormal, 216–217
    - normal transformation, 205–207
    - numerical outcomes and, 153
    - Poisson, 173, 174–176
    - properties of, 162
  - Randomized block designs, 451
  - Ranges
    - defined, 77
    - interquartile, 71, 77
  - Ratio scale, 12
  - Ratio-to-moving average, 642
  - Reagan, Ronald Wilson, 681–682
  - Real returns, 666
  - Real terms, 676
  - Regression analysis, 483; *see also* Multiple regression model; Simple linear regression model
    - assumptions, 537–538
    - comparing models, 574–575
    - goodness-of-fit tests, 497
    - qualitative variables, 590
    - quantitative variables, 590
    - reporting results, 522–523
    - violations of assumptions, 540–544
  - Rejection region, 312
  - Relative frequency distribution
    - cumulative, 29
    - qualitative data, 19
    - quantitative data, 29
    - summarizing, 91
  - Residual plots, 537
  - Residuals
    - model selection criteria, 625
    - in simple linear regression, 485
  - Response variable
    - binary choice models, 605–606
    - defined, 483
    - expected values, 484
    - in multiple regression, 492
    - in simple linear regression, 484
  - Restricted models, 527
  - Returns
    - adjusted closing price, 665–666
    - annualized, 73–74
    - calculating, 664
    - excess, 84
    - geometric mean, 73–74
    - historical, 665
    - mean-variance analysis, 83–84
    - nominal, 666
    - portfolio, 162–164
    - real, 666
    - risk and, 83–84, 162
    - risk-adjusted, 519
    - Sharpe ratio, 83–84
  - Right-tailed hypothesis, 303, 314
  - Right-tailed test, 412
  - Risk
    - measures, 376
    - portfolio, 164
    - returns and, 83–84, 162
  - Risk aversion, 159–160
  - Risk loving consumers, 160
  - Risk neutral consumers, 160
  - Risk-adjusted returns, 519
  - Roche, Jehanne-Marie, 664, 674
  - Roosevelt, Franklin D., 232, 233
  - Rules of probability; *see* Probability rules
  - Runs
    - above and below median, 716–718
    - defined, 715
    - Wald-Wolfowitz runs test, 715–716
- ## S
- s* chart, 252
  - Sample coefficient of variation, 80
  - Sample correlation coefficient, 93, 479
  - Sample covariance, 92, 478
  - Sample mean; *see also* Sampling distribution of sample mean
    - differences between, 340
    - formula, 60
    - weighted, 90
  - Sample mean absolute deviation, 78
  - Sample proportion, 360; *see also* Sampling distribution of sample proportion
  - Sample regression equation
    - multiple regression, 492–493
    - simple linear regression, 485–486
  - Sample spaces, 108; *see also* Events
  - Sample standard deviation, 79
  - Sample statistics
    - defined, 5, 232
    - estimators, 237
    - mean as, 61
    - random, 237
    - use of, 6, 232
  - Sample variance
    - formula, 79
    - properties, 376
    - sampling distribution of, 376–377
    - sampling distribution of ratio of, 385
  - Samples
    - defined, 5, 232
    - independent random, 340
    - matched-pairs
      - mean difference, 351–355

- Samples (*continued*)
  - recognizing, 351
  - population and, 6, 232
  - representative of population, 232
  - sizes, 234, 287–289
- Sampling
  - acceptance, 251
  - bias in, 6, 232–233
  - cluster, 235
  - need for, 6, 232
  - simple random samples, 233–234
  - stratified, 234–235
  - with and without replacement, 179
- Sampling distribution of sample mean
  - central limit theorem, 240–241
  - defined, 237
  - expected values, 238
  - hypothesis testing, 307
  - normal distribution, 239–240
  - standard deviation, 238
  - variance, 238
- Sampling distribution of sample proportion
  - central limit theorem, 245
  - defined, 244
  - expected values, 244
  - finite population correction factor, 249
  - standard deviation, 244–246
  - variance, 244
- Sampling distribution of sample variance, 376–377
- Scatterplots, 43–46, 478, 486–488
- Seasonal dummy variables, 640, 645–647
- Seasonal indices
  - adjusted, 642–643
  - unadjusted, 642
- Seasonal patterns, 626
- Seasonality, extracting, 641–643
- Seasonally adjusted series, 644
- Selection bias, 233
- Semi-log model, 569, 571
- Serial correlation, 537
- Seton Hall University, 589, 603
- Sharpe, William, 83
- Sharpe ratio, 83–84
- Shewhart, Walter A., 252
- Sign test, 711–713
  - test statistic, 712
- Significance levels, 309
- Significance tests
  - of dummies, 592, 600
  - of individual significance, 516–519
  - of interaction variables, 600
  - of joint significance, 521–522
- Simple linear regression model; *see also* Regression analysis
  - assumptions, 484, 537–538
  - defined, 485
  - explanatory variables, 483
  - goodness-of-fit tests, 497–500
  - response variable, 483
  - sample regression equation, 485–486
  - scatterplots, 486–488
- Simple price indices, 668
- Simple random samples, 233–234
- Skewed distributions, 31
- Skewness coefficients, 66, 419
- Smoothing techniques
  - exponential, 628–631
  - moving average methods, 626–628, 631
  - uses of, 626
- Spearman rank correlation test, 705–708
- Spurious correlation, 5, 481
- SSA; *see* Sum of squares for factor A
- SSAB; *see* Sum of squares for interaction of factor A and factor B
- SSB; *see* Sum of squares for factor B
- SSE; *see* Error sum of squares
- SSR; *see* Sum of squares due to regression
- SST; *see* Total sum of squares
- SSTR; *see* Sum of squares due to treatments
- Standard & Poor's 500 Index, 506
- Standard deviation
  - Chebyshev's theorem, 85–86, 87
  - defined, 78
  - of discrete random variable, 158
  - empirical rule, 86–87
  - interpreting, 85
  - population, 79
  - of portfolio returns, 164
  - sample, 79
  - z-score, 87–88
- Standard error, of sample proportion, 244
- Standard error of the estimate, 497–500
- Standard normal distribution
  - defined, 198
  - inverse transformation, 207–209
  - transformation, 205–207
  - z values, 198–202
- Standard normal table, 198
- Standardizing, data, 87
- Starbucks Corp., 4, 151, 177, 231, 241, 246, 247, 339, 353–354, 356
- Statistical quality control; *see also* Control charts
  - acceptance sampling, 251
  - defined, 251
  - detection approach, 251
  - population variance used in, 376
  - sources of variation, 251–252
- Statistical software; *see* Excel
- Statistics; *see also* Sample statistics
  - descriptive, 5
  - importance, 4–5
  - inferential, 5
- Stem-and-leaf diagrams, 41–43
- Stocks; *see also* Returns
  - adjusted closing price, 665–666
  - Dow Jones Industrial Average, 9
- Stratified random sampling, 234–235

Studentized range table, 445  
 Student's  $t$  distribution; *see*  $t$  distribution  
 Study habits, hypothesis tests, 301, 322  
 Subjective probability, 111, 113  
 Sum of squares due to regression ( $SSR$ ), 501  
 Sum of squares due to treatments ( $SSTR$ ), 435  
 Sum of squares for factor  $A$  ( $SSA$ ), 452, 459  
 Sum of squares for factor  $B$  ( $SSB$ ), 452, 459  
 Sum of squares for interaction of factor  $A$  and factor  $B$  ( $SSAB$ ), 459–460  
 Surveys, 5, 231  
     of tween preferences, 3, 13  
 Symmetric distributions, 31, 66, 86, 196; *see also* Normal distribution  
 Systematic patterns, 626

**T**

$t$  distribution  
     characteristics, 278  
     defined, 277  
     degrees of freedom, 278  
     hypothesis testing, 319–322

$t$  test  
     assumptions, 688  
     two-tailed, 517–519

Test of independence, 410–414

Test statistics  
     for differences in means, 343  
     goodness-of-fit, 404–408  
     Jarque-Bera, 419  
     for Kruskal-Wallis test, 702  
     for mean difference, 353  
     for one-way ANOVA, 436  
     for population correlation coefficient, 480  
     of population mean, with known standard deviation, 308  
     for population proportion, 326  
     for population variance, 380  
     for population variance ratio, 389  
     for proportion differences, 362  
     for sign test, 712  
     for test of independence, 412  
     test of joint significance, 521  
     for test of linear restrictions, 528  
     for Wald-Wolfowitz runs test, 716  
     for Wilcoxon rank-sum test, 697  
     for Wilcoxon signed-rank test, 690

Thomsen, Jaqueline, 613

Time series data; *see also* Forecasting; Returns; Smoothing techniques  
     defined, 6, 7, 624  
     deflated, 676–678  
     nominal terms, 676  
     real terms, 676  
     systematic patterns, 626  
     unsystematic patterns, 626

Total probability rules, 131–134, 136

Total sum of squares ( $SST$ ), 437, 459, 500

Transformations; *see also* Logarithms; Polynomial regression models

    inverse, 207–209  
     normal, 205–207

Trend models  
     exponential, 634–637  
     linear, 633–634, 645  
     polynomial, 637–638  
     quadratic, 637–638

Trendlines, 486–488

Trends  
     extracting, 643–644  
     systematic patterns, 626

True zero point, 12

Tukey, John, 41, 442

Tukey's honestly significant differences (HSD) method, 442, 444–446, 455

Two-tailed hypothesis, 303–305, 314, 315–316

Two-tailed  $t$  test, 517–519

Two-way ANOVA  
     defined, 450  
     with interaction, 458–461  
     randomized block designs, 451  
     sample sizes, 451  
     uses of, 450  
     without interaction, 450–455

Type I error, 305–306

Type II error, 305–306

**U**

UCL; *see* Upper control limit

“Ultra-green” car, 269, 289

Unadjusted seasonal index, 642

Unbalanced data, 444

Unconditional probability, 120, 134

Under Armour Inc., 107, 126, 127, 128, 403, 411, 414

Unexplained variation, 500

Union, of events, 109, 110, 117

University of Pennsylvania Medical Center, 4

Unrestricted models, 527

Unsystematic patterns, 626

Unweighted aggregate  
     price indices, 670–671

Upper control limit (UCL), 252–253

U.S. Bureau of Labor Statistics, 6, 7, 677

U.S. Census Bureau, 7, 61

USA TODAY, 5

**V**

Vanguard, 59, 81, 375, 391, 506, 687, 709

Variability measures; *see* Dispersion measures

Variables; *see also* Dummy variables; Qualitative variables; Quantitative variables  
     continuous, 9  
     defined, 8  
     discrete, 8–9  
     response, 483, 492

Variance; *see also* Analysis of variance; Population variance; Sample variance  
defined, 78  
of discrete random variable, 158  
for frequency distribution, 89–91  
mean-variance analysis, 83–84  
of portfolio returns,  
163–164

Venn, John, 109

Venn diagrams, 109, 110

## W

Wald-Wolfowitz runs test, 715–718

*The Wall Street Journal*, 7

Websites, data sources, 7

Weighted aggregate price index, 671–673

Weighted mean, 66, 90

Width, confidence intervals, 273–274

Wilcoxon rank-sum test

for independent samples, 693, 695–697

for matched-pairs sample, 694–695

test statistic, 697

uses of, 346, 355

Wilcoxon signed-rank test

critical values, 690

for matched-pairs sample, 694–695

for population median, 688–692

software, 698

test statistic, 690

Within-treatments estimate of population variance,  
435, 436–437

Within-treatments variance,  
435

## X

$X^2$  distribution; *see* Chi-square distribution

$\bar{x}$  charts, 252

## Y

Yields; *see* Returns

## Z

$z$  table, 198

$z$  values

defined, 198

finding for given probability, 201–202

finding probabilities for, 198–200

inverse transformation, 207–209

normal transformation, 205–207

zillow.com, 7

$z$ -score, 87–88