TAKING THE FEAR OUT OF DATA ANALYSIS

COMPLETELY REVISED, SIGNIFICANTLY EXTENDED AND STILL FUN

ADAMANTIOS DIAMANTOPOULOS
BODO B. SCHLEGELMILCH
GEORGIOS HALKIAS

TAKING THE FEAR OUT OF DATA ANALYSIS

…to our families.

Seriousness is the only refuge of the shallow. –Oscar Wilde

TAKING THE FEAR OUT OF DATA ANALYSIS

COMPLETELY REVISED, SIGNIFICANTLY EXTENDED AND STILL FUN

SECOND EDITION

ADAMANTIOS DIAMANTOPOULOS
Professor of International Marketing, Department of Marketing and International Business, University of Vienna, Austria

BODO B. SCHLEGELMILCH
Professor of International Marketing Management, Department of Marketing, WU Vienna, Austria

GEORGIOS HALKIAS
Associate Professor of Marketing and Behavioral Research, Department of Marketing, Copenhagen Business School, Denmark

Edward Elgar PUBLISHING

Cheltenham, UK • Northampton, MA, USA

Published by
Edward Elgar Publishing Limited
The Lypiatts
15 Lansdown Road
Cheltenham
Glos GL50 2JA
UK

Edward Elgar Publishing, Inc.
William Pratt House
9 Dewey Court
Northampton
Massachusetts 01060
USA

A catalogue record for this book
is available from the British Library

Library of Congress Control Number: 2022950118

This book is available electronically in the Elgaronline
Business subject collection
http://dx.doi.org/10.4337/9781803929842

ISBN 978 1 80392 983 5 (cased)
ISBN 978 1 80392 984 2 (eBook)
ISBN 978 1 80392 985 9 (paperback)

CONTENTS IN BRIEF

FULL CONTENTS

TO THE READER

Effective data analysis requires the effective use of statistics. Unfortunately, statistics is boring. It is boring to learn, it is boring to teach, and it is usually boring people who actually *like* statistics. Indeed, the comment that a statistician is 'a person who didn't have enough charisma to be a cost accountant' says it all!

Statistics is also hard. It is hard to learn, it is hard to teach (properly), and it is even harder to remember what little you may have learned. In short, statistics is rarely fun. But it *can* be – as you will soon find out. Trust us.

INSTEAD OF A PREFACE

The first edition of this text goes back to 1997. Cars had just been invented, and computers were still coal fired – at least from the perspective of our new generation of students. There was no Facebook, no Amazon, and no iPhone. And, no, you could not watch videos on your mobile phone or surf the Internet on it. Times were clearly (very) different. Most people learning statistics were still blissfully unaware of some powerful techniques now described in this book, such as the much-feared multivariate analyses. It was also an innocent time! Questionnaires could still request the respondents' 'sex' in a dichotomous (male/female) fashion, and researchers did not even think about being politically incorrect for omitting dozens of other gender options. However, as unlikely as it sounds, even in these innocent times, the two original authors of this text, Adamantios Diamantopoulos and Bodo Schlegelmilch, in the profession widely referred to as the terrible twins (or more unkindly as the 'gruesome twosome'), still managed to get into trouble.

Upon seeing a couple of draft chapters of the first edition of this book, reviewers took exception to several of our jokes, found some of our examples politically incorrect, and queried the wisdom of not being 'sufficiently serious'. Having said that, one reviewer openly admitted that he/she might be getting to be an 'old sourpuss'(!), with which we politely agreed. Notwithstanding these obstacles, through a magic mix of constant encouragement and occasional threats of physical violence from the commissioning editor, we somehow managed to finish the book *and* get it published! Since then, astonishingly, the original version has been reprinted no fewer than six times, and we received numerous pieces of fan mail (well, we can at least recall two positive ones in the late 90s!).

So why a new, completely revised, and significantly extended version of the book? First, statistical software has progressed enormously and is much easier to handle; we now include up-to-date applications to illustrate the various techniques discussed. Second, partly as a result of the 'software revolution', sophisticated analytical techniques have become more accessible; we now include a discussion of the most important ones. And yes, we feel that we have a responsibility to our readers here: the mere mention of terms like 'non-hierarchical clustering' or 'orthogonal rotation' will substantially improve your chances on the job market and the dating scene! Finally, we now live in different times, so we had to (reluctantly) sanitize some non-PC jokes.

All these goodies – up-to-date software applications, new statistical tests, and (somewhat!) sanitized jokes – were largely made possible by taking a new and substantially younger co-author on board, Georgios Halkias. *He* is the one you should blame for any comments that may be (wrongly!) construed to be still not 100% PC (despite our best efforts). Any other mistakes or omissions you may find are clearly and squarely the responsibility of our wonderful support team, namely Martina Roth, Doris Lehdorfer, and Thomas Winter from the University of Vienna, as well as Hanife Özdemir, Erin Silangil, Sarina Mansour Fallah, Sanem Öztürk, Stephie Habicher, and Reiko Domai from WU Vienna – thank you all for doing a fantastic job! While the attribution of blame to others may contradict the spirit of this preface, there is an important reason for it. The two senior authors are striving for a second career in the clergy after retirement and, hence, we cannot possibly be responsible for anything bad!

And speaking of our future careers, if anyone knows of a mixed-gender monastery with basic amenities such as an infinity pool, a nice sea view, and 24/7 room service, where we do not have to get up in the middle of the night to pray, please let us know (and you'll get a generous 2% discount off the price of this book!).

We hope you enjoy reading the book as much as we enjoyed writing it!

Adamantios Diamantopoulos
Bodo B. Schlegelmilch
Georgios Halkias

DISCLAIMER

All examples, numerical figures, data, names, characters, places, products, and incidents mentioned in this book are fictitious and only reflect the authors' imagination. Any resemblance to actual people, elements, and events is coincidental. No such identification is intended or should anyhow be inferred.

ABOUT THE AUTHORS

Adamantios Diamantopoulos (BA, MSc, PhD, DLitt) is Chaired Professor of International Marketing and Head of the Marketing and International Business Department at the University of Vienna, Austria. He is also Visiting Professor at the University of Ljubljana, Slovenia, and Senior Fellow at the Dr. Theo and Friedl Schoeller Research Center for Business and Society, Germany. Previous full-time appointments include the Chair of Marketing and Business Research at Loughborough University and the Chair of International Marketing at the University of Wales, as well as positions at the University of Strathclyde and the University of Edinburgh. He has held several visiting professorships in the USA and Europe, including the Joseph A. Schumpeter Fellowship at Harvard University and the Nestlé Visiting Research Professorship of Consumer Marketing at Lund University, Sweden. He has taught at various university institutions in some 20 countries and has collaborated with several international companies. He is an elected Fellow of the European Marketing Academy and the British Academy of Management, and a recipient of the JIBS Silver Medal.

His main research interests are in international marketing and research methodology, and he is the author of some 200 publications in these areas with over 40,000 citations. His work has appeared, among others, in the *Journal of Marketing Research*, *Journal of International Business Studies*, *Journal of the Academy of Marketing Science*, *International Journal of Research in Marketing*, *Journal of Service Research*, *Journal of International Marketing*, *Journal of Retailing*, *MIS Quarterly*, *Organizational Research Methods*, *Psychological Methods*, *Information Systems Research*, *British Journal of Management*, and *International Journal of Forecasting*.

He has been the recipient of several best paper awards, including the Hans B. Thorelli Award for an article published in the *Journal of International Marketing* that has made significant and long-term contributions to international marketing theory or practice. He sits on the editorial review boards of several academic journals, and acts as a referee for various professional associations and funding bodies.

When not working, he likes skiing, riding big motorcycles (he has two of them), and playing the drums (but not well enough to give up the day job). For some reason, his wife and son both think he will never grow up.

Bodo B. Schlegelmilch heads the Institute for International Marketing Management at WU Vienna University of Economics and Business and is Chair of the Association of MBAs and Business Graduates Association (AMBA and BGA). For more than 10 years he served as founding Dean of the WU Executive Academy.

Initially educated in Germany, he obtained two doctorates (a PhD in International Marketing and a DLitt in Corporate Social Responsibility) from the University of Manchester (UK) and an honorary PhD from Thammasat University (Thailand). Starting at Deutsche Bank and Procter & Gamble in Germany, he continued his career at the University of Edinburgh and the University of California, Berkeley. Appointments as British Rail Chair of Marketing at the

University of Wales (UK) and Professor of International Business at Thunderbird School of Global Management (USA) followed.

Bodo serves on several business school advisory boards in Europe and Asia. He holds/held visiting appointments, for example, at the Universities of Minnesota (USA), Keio (Japan), Leeds (UK), Sun Yat-sen (China), and Cologne (Germany), as well as the Indian School of Business (India), and has taught in over 30 countries.

He has received numerous teaching and research awards, including the 'Significant Contributions to Global Marketing' award of the American Marketing Association, as well as fellowships from the Academy of International Business, the Academy of Marketing Science, and the Chartered Institute of Marketing. His research interests span from international marketing strategy to corporate social responsibility. He has published more than a dozen books in English, German, and Mandarin, and his work has appeared in journals such as the *Strategic Management Journal*, *Journal of International Business Studies*, *Journal of the Academy of Marketing Science*, and *Journal of World Business*. He was also the first European editor-in-chief of the *Journal of International Marketing*, published by the American Marketing Association.

When not researching, writing, or teaching, Bodo enjoys sailing, hanging out at nice beaches, eating Japanese meals, and writing long and complicated autobiographical statements. His wife knows he will never grow up.

Georgios Halkias is Associate Professor of Marketing at the Department of Marketing, Copenhagen Business School (Denmark) and a Visiting Professor at the University of Vienna (Austria). Prior to joining CBS, he was Associate Professor at the TUM School of Management, Technical University of Munich (Germany), while in the past he has held resident and/or visiting faculty positions at the University of York (UK), the University of Vienna and the WU Vienna University of Economics and Business (Austria), the University of Ljubljana (Slovenia), and the Athens University of Economics and Business (Greece). Georgios has also gained industry experience with multinational firms such as Société Générale and Procter & Gamble in Greece and the UK.

Georgios has received academic qualifications in three different countries. He holds a PhD Habil. from the University of Vienna (Austria), a PhD from the Athens University of Economics and Business (Greece), and an MSc from the University of Warwick, Warwick Business School (UK).

His main research interests lie in the areas of consumer psychology, branding, and research methods. His work has attracted several national/European Union funds and has been published in leading academic journals, including the *Journal of International Business Studies*, *International Journal of Research in Marketing*, *British Journal of Management*, *Journal of Advertising*, *Journal of Business Research*, *International Marketing Review*, and *International Journal of Advertising*. He has also made multiple contributions to academic books and international conferences.

Georgios acts as a reviewer for several top-tier journals and sits on the editorial review Board of the *Journal of International Marketing* (American Marketing Association). He has received various international distinctions and awards, including the Outstanding Reviewer Award 2019 and the Outstanding Paper Award 2020 from the *International Marketing Review* in the Emerald Literati Awards of Excellence. He has been repeatedly nominated for teaching awards in courses on quantitative research methods, consumer behavior, and branding, and is the recipient of the Best Teaching Award 2017 granted for outstanding teaching at the graduate level (University of Vienna).

When not engaged in highly scientific activities, Georgios practices the guitar because he wants to form a progressive metal band when he grows up. His wife desperately tries to make him realize that this ain't happening.

PRE-PUBLICATION REVIEWS FROM AROUND THE WORLD

'Written with wry wit and incredible clarity, the authors provide the reader a detailed understanding of seminal issues in data analysis. A masterful work that truly does "take the fear out of data analysis" – this book is a rare treat indeed.'

– David A. Griffith, Mays Business School, Texas A&M University, USA

'Written by a proficient team of authors, Taking the Fear Out of Data Analysis *is a fascinating … ah, forget the marketing blurb. This is a great text, you should read it! And your family too, provided that they want to learn about the basics of data analysis in the most entertaining way. There is no doubt that you will devour this book in no time and learn a lot about statistics on the way.'*

– Marko Sarstedt, Ludwig-Maximilians-University (LMU), Germany

'Awww … da da da da da da daayyyyy … boo boo boo brrrrrrr …'

– Penelope, 10-month-old daughter and Research Assistant of Georgios Halkias, Austria

'[H]idden behind some bizarrely memorable examples and illustrations is a very fine introduction to data analysis. This book will be of value to those approaching quantitative analysis for the first time, and should be on the reading list for project-based courses and for new research students.'

– Richard Speed, La Trobe University, Australia

'In the age of big data, at least a rudimentary understanding of data analysis is a must. Business students need to know about data analytics, but they are often intimidated by statistics and thus fail to appreciate the value of data-based and model-supported decision making. Even seasoned researchers are sometimes uncomfortable with conducting statistical analyses of their data because they lack the confidence to apply methods that are perceived as abstract and confusing. In this entertaining book, the authors gently guide the reader through the steps of the data analysis process and they brilliantly succeed in explaining difficult topics in an engaging, witty, and highly informative manner.'

– Hans Baumgartner, Smeal College of Business, Penn State University, USA

'Statistics. I know – you hate it. It's hard and confusing. Students of all levels find the topic hard. I tell them to get this book. And no! They cannot borrow mine, I don't want to lose it. Diamantopoulos, Schlegelmilch and Halkias knock another one out of the park with this excellent introduction to a great array of statistical issues. They start right at the beginning – which is always a good place to start if you're a beginner – and gently, often hilariously, and successfully guide the reader through the various learning moments that need to be negotiated if one is to become fearless in the face of columns of data. Priceless.'

– John Cadogan, School of Business and Economics, Loughborough University, UK

'What happens when three applied researchers write a book on data analysis? Well, as a minimum, you get a resource book written for the user, and not the statistician. In this

revised edition of the popular book Taking the Fear out of Data Analysis*, Diamantopoulos, Schlegelmilch and Halkias provide a highly practical and helpful book for the applied social scientist. Their writing (considerably more accessible than the spelling of their surnames) is plain, straightforward and fun to read. In an era when aspiring researchers will remain a one-legged scholar unless they master the foundational skills of data analysis and research methodology, this book is considered an essential read and frequent reference.'*

– S. Tamer Cavusgil, Georgia State University, USA

' Taking the Fear Out of Data Analysis *is one of the few books to provide a comprehensive, conceptually solid yet accessible overview of the theory and practice of data analysis. Building on their own extensive research experience, the authors use a pragmatic, elegantly ironic and competent style to "translate" the complex science of managing data and involve readers in an enjoyable learning journey. All the concepts and techniques are presented in a manner that is easy to read and understand and several (and mostly humorous) examples and illustrations have been integrated throughout the text. Definitely a must-have for anybody who is interested in discovering not only how to deal with data but also how pleasant it can be to learn it.'*

– Alessandro De Nisco, UNINT, Rome, Italy

'These guys never give me any tips and now they have threatened to buy their bento boxes elsewhere if I don't give them a review. So here it is: I like the book; it has a lot of words!'

– Sōta Tuna, Head Waiter, Harakiri Sushi Shop, Vienna, Austria

'This book is a real page-turner! Why? Because the book strikes an exceptionally good balance between fun (funny examples, witty remarks) and a sound introduction to statistics and data analysis. Thus, the book encourages its readers in an entertaining way to delve into the statistical concepts and methods covered. Despite its reassuring title, the book also does not conceal the mysterious "dark side of empirical data analysis", such as p-hacking or HARKing. But this, of course, makes perfect sense since you can only defeat your fear if you know the danger. The upshot is that this work definitely delivers on its promises.'

– Dirk Temme, Schumpeter School of Business and Economics, University of Wuppertal, Germany

'Read this book! It should not only be on the prescribed list of any student of the social sciences, it should also be compulsory reading for journalists and the media.'

– Leyland Pitt, Simon Fraser University, Canada

'It's been tried and tested and we can now reject the null hypothesis that "this book is no different to other data analysis books". It is, and in ways that make it an easier read and an easier ride for those starting out on their analysis journey. Like any great product, it lives up to its name and really does help to take some of the fear out of analysing research data. Congratulations to the authors for updating this now classic text.'

– Vince Mitchell, The University of Sydney Business School, Australia

'The authors gave me a choice of either giving a testimonial or reading the book cover to cover … you know the rest.'

– Anonymous PhD Student, Anonymous University, Anonymous Country

'The new edition of this book provides excellent guidance to data knowledge and competence using a problem-solving approach. With the digital becoming increasingly important, analytical skills should be key competencies in everybody's daily life. To achieve this goal, Taking the Fear Out of Data Analysis *is highly recommended.'*

– Zhongming Wang, Zhejiang University, China

' Taking the Fear Out of Data Analysis *is the best book for someone who has heard a lot of buzz about data analysis but doesn't have a firm grasp of the subject. The book is an eye-opening read for anyone who wishes to learn about data analysis: understanding of data, preparing the data for analysis and different analysis techniques. If I had to pick one book for an absolute newbie to the field of data analysis, it would be this one.'*

– Manish Gangwar, Associate Dean, Research and RCI Management and Executive Director, ISB Institute of Data Science, Telangana, India

'When I began my academic career as a Research Assistant, the first edition of this book enjoyed a prominent position on the shelves in the office I shared with three other early career researchers. We liked this book because it unpacked the complicated, procedure-heavy world of quantitative methods in a user-friendly way. It poked fun at the subject matter in a manner that was so disarming that even our office's hard-core qualitative researcher loved it. The significantly extended new edition is increasingly relevant as the world of quantitative methods has kept on expanding, in part due to an explosion in software programs that scholars can use seemingly without much understanding. Do not let the light-hearted nature of this book fool you. It is a statistics book that carefully leads readers through all the necessary stages of analysis. It effortlessly explains the analysis details and assumptions that PhD examiners, journal reviewers, and conference presentation audience members insist on raising. This excellent new edition is destined to be very well thumbed.'

– Matthew Robson, Cardiff Business School, UK

NOTE

Many testimonials were quite lengthy. While our tolerance for receiving praise tends to be alarmingly high, we felt it would be in the best interests of our readers to shorten the testimonials so that the part of the book talking about statistics would be *slightly* longer than the part devoted to testimonials. An extended list can be found on the publisher's website for this book.

INTRODUCTION TO *TAKING THE FEAR OUT OF DATA ANALYSIS*

The trouble with numbers is that they frighten a lot of people.
–Leslie W. Rodger, *Statistics for Marketing*

This book has been written for people who do not *like* data analysis, who do not *want* to become analysts or statisticians, but who *have* to learn about data analysis for whatever reason (e.g., to pass a course at college or university, complete a dissertation, get/keep a job, or to impress a new date). It has also been written for people who *think* they cannot understand data analysis and statistics, having had bad experiences with textbooks full of formulae and little substance, and teachers full of confidence and no humor. In fact, this book has been written for anybody suffering from the 'I hate numbers' syndrome – and we know there are plenty of you out there.

What we tried to do in this book is quite simple: take your hand and lead you through the entire data analysis process without boring you to death along the way. Our specific aims have been threefold:

1. To provide a comprehensive but digestible *introduction* into the strange world of data analysis, assuming no prior knowledge on your part.
2. To indicate the *linkages* among the various stages of the data analysis process and highlight the implications of good/bad early decisions on subsequent ones.
3. To demonstrate that learning about data analysis can be an *enjoyable* experience; hopefully, after you have finished reading this book, you will feel that way too.

Our philosophy behind the content and structure of the book is based on a few basic premises. First and foremost, our main concern is with *understanding* rather than memorizing. Thus, we urge you to channel your learning effort towards grasping and digesting the various concepts/techniques of data analysis rather than mechanically reproducing a bunch of formulae. Moreover, at this introductory level, we feel that your attention should be directed towards *key issues* and *major building blocks* rather than statistical refinements and details. Consistent with this view, we keep the number of formulae down to an absolute minimum and do not bother with providing mathematical proofs (the latter being a sure way of sending you to sleep!). We also firmly believe that a point is best driven home by means of an *example.* Consequently, we make liberal use of (mostly silly) examples and illustrations to show how the various concepts and techniques can be applied. Lastly, we see no harm in making you smile from time to time – a hefty dosage of humor is often the only way to keep one sane while learning/doing data analysis!

You can use this book in a number of ways, depending upon your background and objectives. For those of you with no prior experience of data analysis and little, if any, statistical knowledge, we strongly recommend that you go through the book chapter by chapter. In this way, you will be introduced to more complex material in a gradual manner and should not feel lost at any time. On the other hand, those of you with some previous exposure to research methods and/or statistics may opt for a more flexible approach. For example, you may wish to

concentrate on Chapter 4 (dealing with data preparation and coding) onwards and refer back to Chapters 1 to 3 (which are really background chapters) on an 'as required' basis. Finally, for those of you who are particularly interested in specific types of analysis, you should primarily focus on Chapters 7 to 14, which cover different analytical techniques (and presuppose familiarity with basic data analysis principles).

There are also some key chapters that everyone should read. These include Chapter 5 (on setting analysis objectives), Chapter 8 (on the nature of statistical estimation), Chapter 9 (on the principles of hypothesis testing), and Chapter 15 (on evaluating and presenting the analysis). Moreover, all readers would do well to heed the numerous **HINTS** and **WARNINGS** dispersed throughout the various chapters. These serve to emphasize key points, awareness of which can prevent problems and make life easier for you.

The Further Reading section at the end of each chapter should also be consulted. Rather than provide a long list of references, we have intentionally limited ourselves to a selection of a few key sources which we feel best amplifies and complements the material covered in the chapter. Each suggested reading has been briefly annotated, and there is no duplication of sources in the Further Reading sections of the various chapters. Having said that, many of the suggested readings may also be useful for issues discussed in the other chapters – so keep an open mind and be flexible.

The book is organized into three main parts, containing a total of fifteen chapters.

Part I. *Understanding Data,* provides the necessary background by looking at the nature of data, the sampling process, and the notion of measurement. These are essential building blocks underpinning data analysis and a prerequisite for understanding the application of statistical techniques.

Part II, *Preparing Data for Analysis,* focuses on the various tasks associated with converting raw data into a form that can be analyzed and on setting objectives for the analysis. Careful attention to the issues raised here will prevent a lot of problems at the actual analysis stage.

Part III. *Carrying out the Analysis,* examines the rationale behind different types of analysis and introduces a wide variety of analytical methods appropriate for different circumstances. Starting with simple approaches to describing and summarizing data, a number of techniques are considered, which, if properly applied, will enable you to get the most out of your data. A good taste of multivariate analysis is also provided in this edition, navigating the interested reader through more complex analytical techniques. Finally, several issues are raised relating to the evaluation and presentation of a data analysis project and the preparation of written and oral research reports.

By the time you finish reading this book, you should be in a position to know *what analysis* to apply, for *which purpose*, to *what kind of data.* The chapter-by-chapter overview that follows should help clarify what all this means.

Chapter 1 lays the groundwork for the rest of the book by introducing you to the **data matrix**, which is the raw material you will work with whenever you do data analysis. Here you will get acquainted with **units of analysis**, **variables**, and **values**, which are the essential ingredients of data. You will also learn about different **types of data** and about the distinction between data and **information**. By the time you finish this chapter, you should be able to talk about data as if you knew something about it!

Chapter 2 looks at **sampling** to demonstrate how units of analysis may come about in a particular project. You will understand the rationale for taking a **sample** rather than studying the whole **population** and the different **sampling methods** that you can use. You will also encounter the concept of **sampling error**, which, no doubt, will haunt you forever after! Finally, the numerous considerations for determining **sample size** will be piled upon you – if only to confuse you even further!

Chapter 3 describes what **measurement** is, why it is important and how it can be done. You will recognize the advantages and disadvantages of different **measurement scales** and get to know a variety of **scaling formats**. Following this, you will encounter a second kind of error, the notorious **measurement error**, which will also haunt you forever after (either with the sampling error or on its own). By the end of this chapter, you should have a firm grasp of the options available for measuring your variables and interpreting the resulting values, as well as for assessing the **validity** and **reliability** of measurement.

Chapter 4 moves into the practicalities of preparing data for analysis, focusing on the process of **data editing**. You will learn how to detect many errors you may make when processing data and how to deal with ambiguous, inconsistent, and missing data. You will be shown how to properly **code** your data, and, lastly, you will see how easy it is to perform **variable transformations** in order to take advantage of the full potential of your data set.

Chapter 5 emphasizes the need for having clear **analysis objectives** before embarking on the actual analysis (otherwise, you will not know how to start or when to stop). These will ensure that your analysis is relevant, comprehensive, and efficient. You will be introduced to different analytical perspectives, namely **description**, **estimation**, and **hypothesis-testing** (which will be dealt with in more detail in later chapters). The chapter will conclude with a discussion of the factors governing the choice of the **method of analysis** (which will undoubtedly become the subject of your worst nightmares for years to come!)

Chapter 6 focuses on **data description**, which is usually the first type of analysis that you will want to do. You will get to know the various forms of **frequency distributions** and the steps to take in grouping data. Your artistic talents will also be thoroughly stimulated by an examination of different types of **graphical representations**. At this stage, you will wonder what the purpose of life would be without percentiles, true class limits, and gives!

Chapter 7 soldiers on with data description and introduces you to different **summary measures** that can be used to capture **typical or average** responses as well as the extent of **variability** in your data. By using different summary measures in conjunction with one another, you will be able to identify the **shape** of a frequency distribution and compare it to known forms. In this context, the famous **normal distribution** will serve as a useful point of reference.

Chapter 8 deals with the process of **estimation** and shows how you can talk about the population when you only have a sample. First, you will become familiar with the concept of a **sampling distribution** and then learn how to set **confidence intervals** for different **population parameters** that you may wish to estimate. Being able to make inferences from a sample to a population will surely set you apart from the uninitiated punter!

Chapter 9 ventures into the mystical world of **hypothesis-testing** and provides you with an understanding of the basic principles associated with developing and testing specific research propositions. You will learn about different **types of hypotheses** and the rationale behind **sig-**

nificance testing, which, like estimation, also enables you to make inferences from a sample to a population. Concepts like 'null hypothesis', '*p*-values', and 'regions of rejection' will become second nature to you and a topic of conversation at every possible opportunity! Finally, in this chapter, you will be introduced to the concepts of **effect size** and **statistical power** and be alerted toward issues that go beyond statistical significance.

Chapter 10 deals with the simplest type of hypotheses, namely those involving a single variable and a single sample. Here you will be shown how to examine the **fit** of your frequency distributions against prior expectations or against a theoretical distribution. In addition, you will be able to determine whether your sample is likely to have been drawn from a population with known **parameter values**, including tests for **central location**, **proportions**, and **variability**, as well as whether your sample is, in fact, **random**. By the time you complete the chapter, you should feel confident enough to face more complex hypotheses (involving more than one sample or more than one variable).

Chapter 11 extends your journey into hypothesis-testing by concentrating on **comparisons**. First, you will learn how to compare two or more groups on the same variable – what is known as an **independent measures** (or samples) comparison. Next, you will address the issue of comparing two or more responses from the same group, involving a **related measures** (or samples) comparison. In both cases, you will be testing to see whether **significant differences** exist, that is, whether any observed differences on your sample results are likely to reflect 'true' differences in the population. By becoming an expert on comparisons, you will be able to impress your friends with profound statements of the sort: 'Basketball players are, on average, significantly taller than jockeys' and 'There is no significant difference between the proportions of male and female construction workers with a passion for ornithology.'

Chapter 12 shows you how to investigate **relationships** between two variables. Here you will be exposed to **measures of association** that can tell you whether two different variables are related to one another and which can be subjected to significance tests to see whether any observed relationship (based on your sample data) is also likely to hold in the population. Once you have established a **significant relationship**, your association measure will also enable you to assess its **strength** and **directionality** (i.e., positive or negative). This chapter will also introduce you to fundamental techniques that can not only help you identify a relationship between two variables, but also allow you to make specific **predictions** about the expected value of an outcome variable based on a given level of a predictor variable. The chapter concludes with an important note on the distinction between **correlation and causality** – make sure you don't miss this!

In Chapter 13, we show you that there is *much* more to data analysis – **multivariate analysis** procedures open up a whole lot of new opportunities for extracting information from your data and answering more complex research questions. We start by making a basic distinction between **dependence and interdependence methods** and proceed by giving you a good taste of the techniques included in the former. Here you will learn a great deal about **complex comparisons and relationships** that involve multiple variables at the same time. Advanced data analysis using multivariate procedures cannot be done justice in a single chapter – that is why we added another!

Chapter 14 deals with the remaining group of techniques in the multivariate universe, i.e., the interdependence methods. In this chapter, you will become acquainted with the most fundamental techniques of **identifying structures** in the data both in terms of **variables** and the **units of analysis** involved. If you reach this point of the book, you should feel rather comfortable in dealing with a wide range of complex analytical techniques.

Chapter 15 is an oasis of sanity at the end of your data analysis odyssey. Unbelievable as it may sound, in this chapter, you will *not* have to grasp new theoretical concepts, learn yet another technique, or interpret more statistical results. Responding to your pleas for mercy, the purpose of this final chapter is to make you sit back and think about the 'Now what?' question. Here we talk you through how to **present** your analysis to your audience(s): Is it too technical? Are the practical implications clear? Is the presentation attractive? Unless the presentation of your analysis (written and/or oral) is effective for the *specific* audience involved, all your efforts will have been in vain. Not only must you do the right thing and do it right: you must convince *others* that you have done so.

Have fun.

PART I
UNDERSTANDING DATA

1
What is data (and can you do it in your sleep)?

THE NATURE OF DATA

In today's digital economy, it is virtually impossible not to produce data. Automatic data capture through, for example, digital TVs, refrigerators or washing machines that read RFID tags embedded in your T-shirt or yogurt cup, traffic cameras (you should always look your best in public), and temperature sensors (China and some other countries want to ensure that new arrivals are healthy) constantly produce data. Tracking their shoppers' purchasing habits, supermarkets often learn earlier about their female customers' pregnancies than the respective fathers (of course, some fathers never learn about their fatherhood – but that is a different story). When driving a car, automatic measurement devices can let insurance companies know how safe or unsafe your driving style is (perhaps this explains why your insurance premium quadrupled last year). Taking your dog for a walk, holistic dog food bakeries and other stores can detect you via location-based services as potential customers and send you irresistible coupons on your smartphone, enticing you to visit. When you are surfing the web, cookies trace your behavior or misbehavior (couples beware when visiting dating sites). Even when you are sleeping, wearable devices can let your doctor, pharmacist, or manufacturers of anti-snoring aids know whether you urgently need their products or services.

Although we are swamped with data, sometimes market researchers still need more or different data. Most of us have already been accosted in the street by extremely 'friendly' (and sometimes obnoxious) characters who conduct personal interviews and ask to 'please answer a few questions' regarding our opinions on certain stores (e.g., 'Should El-Cheapo supermarket also offer Japanese bubble tea?'), product preferences (e.g., 'Do you prefer Superclean over Ultrasteril washing powder?'), voting intentions (e.g., 'If a general election were called tomorrow, would you vote for (a) The Conversation Party, (b) The Favor Party, or (c) The Anti-Everything Party?'), or opinions regarding the European Union (e.g., 'Do you feel that Greek producers of eucalyptus-flavored dog biscuits have benefited from EU membership?'). To top it all off, even the privacy of our own home is not enough to prevent the hungry information-seekers from reaching us. How many times have we had to miss a crucial part of our favorite Netflix series in order to answer the phone, only to find out that the caller is Belinda Pain, from Persistence Research, Inc., who was wondering whether you would participate in a survey regarding time-share holidays in Tirana?

Market researchers also like to make use of web-based survey instruments. Having just been to a hotel or restaurant or completed your weekly visit to a spiritual healer, you receive an email requesting you to fill in a questionnaire on various aspects of your experience. Countless university-based researchers have also discovered crowdsourcing systems (like Amazon's Mechanical Turk), where herds of volunteers complete questionnaires against micro-payments. Finally, very often, we have to provide details about ourselves (e.g., our age, income, place of residence) when applying for a driving license, passport, or bank account, or when filing a tax return (very painful) or booking a flight or holiday online.

While the approaches used to obtain data vary in each instance, the objective is always to learn about individuals with regard to certain characteristics of interest. Now, in statistical jargon (yes, you do have to learn some of it), the individuals are called units of analysis (or sometimes 'observations', 'cases', or 'subjects'), the characteristics studied are termed variables, and the responses linking the individuals to the characteristics are known as values. Together, units of analysis, variables, and values make up what we call 'data'. Thus when we refer to data, we implicitly address three distinct issues, notably (a) the respondents (as indicated by the units of analysis), (b) the topic of interest (as described by the set of variables, and (c) the responses of the latter in relation to the topic of interest (as reflected in the values of the variables). Variables can assume different values for different units of analysis; if this is not the case (i.e., when all units of analysis have the same value), then we are not dealing with a variable but with (surprise, surprise) a constant.

The examples at the beginning of this chapter will have made you realize that units of analysis do not have to be individuals (although in a great deal of social research, they are). They can be ordinary objects (e.g., nasal-hair removers, vodka brands, or horsewhips), time-periods (e.g., months, years, decades, centuries), events (e.g., strikes, accidents, shareholder meetings, visits to the dental hygienist), or other entities (e.g., firms, cities, nations, zoos). Neither do variables have to refer to human properties; they can be product features (e.g., speed, durability, color), organizational dimensions (e.g., centralization, formalization, span of control), or national characteristics (e.g., inflation rates, government spending, interest rates) – in fact, anything that can be used to characterize the particular unit of analysis.

When we have a number of units of analysis (e.g., 200 first-year parapsychology students), a number of variables (e.g., age, gender, parents' income, preferred method for reaching out to spirits), and a set of values linking the units of analysis to the variables, the result is a data set. This can be best visualized as a matrix, the rows of which represent the units of analysis, the columns the variables, and the matrix cells the relevant values; for our example, the idea is shown in Table 1.1. By the way, if you are afraid of numbers, you can give names to our tables. For example, Table 1.1 could become 'Sissi' or 'Rudoph'.

If there are n units of analysis (respondents, objects, events, etc.) and m variables, the data matrix looks like Table 1.2, where Rij is the response that unit i gives to variable j (in other words, Rij is the value for unit i on variable j). The subscripts i and j are simply used for counting purposes; in other words, i = 1,2, …, n and j = 1,2, …, m. Obviously, depending upon how many units of analysis and variables we have, n and m will vary; for example, in Table 1.1, n = 200 and m = 4.

Table 1.1 An example of a data matrix

Units of analysis	Variables			
	Age	Gender	Parents' income	Preferred method
Student 1	17 years	Male	€56,000	Chanting
Student 2	18 years	Non-binary	€92,000	A séance
Student 3	20 years	Female	€85,500	Ouija board
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
Student 200	19 years	Female	€77,000	Live chicken

Table 1.2 General form of a data matrix

Units of analysis	Variables												
	V1	V2	V3	•	•	•	•	Vj	•	•	•	•	Vm
01	R11	R12	R13	•	•	•	•	R1j	•	•	•	•	R1m
02	R21	R22	R23	•	•	•	•	R2j	•	•	•	•	R2m
03	R31	R32	R33	•	•	•	•	R3j	•	•	•	•	R3m
•	•	•	•					•					•
•	•	•	•					•					•
•	•	•	•					•					•
0i	Ri1	Ri2	Ri3	•	•	•	•	Rij	•	•	•	•	Rim
•	•	•	•					•					•
•	•	•	•					•					•
•	•	•	•					•					•
0n	Rn1	Rn2	Rn3	•	•	•	•	Rnj	•	•	•	•	Rnm

The main benefit of arranging data in matrix form is that the threefold nature of data (i.e., units of analysis, variables, and values) becomes immediately visible; in fact, most data sets can be represented in the form of Table 1.2. There are some exceptions (e.g., multivariate time-series data) but these need not concern us at this point.

The data matrix is the starting point for analysis, and its structure determines the kind of analysis that can be legitimately carried out. Specifically, the number of rows, n, indicates how many units are being studied; that is, the sample or population size (we will have much more to say about samples and populations in Chapter 2). The number of columns, m, on the other hand, indicates how many variables are used to characterize the units of analysis; depending upon the number of variables, k ($k \leq m$), that are simultaneously manipulated by applying statistical techniques, we talk about univariate ($k = 1$), bivariate ($k = 2$), and multivariate ($k > 2$) data analysis, respectively (we shall return to this issue in Chapter 5). Finally, the natures of the values, Rij, reflect the level of measurement of the variables and thus indicate what can and cannot be said about the units of analysis (measurement will be dealt with in some detail in Chapter 3, so don't worry if you feel totally confused at the moment!).

Table 1.3 Single- and multi-response questions

If you only had enough money to subscribe to just one of the following publications, which one would you choose?	
If money were not a problem, which of the following publications would you subscribe to?	
	(Please tick)
Wall Street Journal	[ ]
The Times	[ ]
Journal of the Mathematically Insane	[ ]
Mongolian Economic Review	[ ]
Unspeakable Acts	[ ]

WARNING 1.1 Questions and variables may not be the same: beware of multi-response questions.

Now, here's something you must always watch out for when you are dealing with a data set that is based on questioning respondents: a question and a variable are not necessarily the same thing. Very often, answering what appears to be a single question in a questionnaire may, in fact, require multiple responses and will thus result in a number of variables. Table 1.3 illustrates this point: irrespective of whether Question A or Question B is asked, the response alternatives are identical. However, the nature and number of the resulting variables are not. If Question A is asked, then only one option needs to be ticked to answer it. Consequently, to capture all possible responses, a single variable would be sufficient; this would be called something like 'most preferred publication' and would take five values, one for each publication involved (e.g., 1 = Wall Street Journal, 2 = The Times, …, 5 = Unspeakable Acts). If Question B is asked instead, a single variable is not sufficient to capture all possible responses, because a respondent may wish to subscribe to more than one publication (e.g., The Wall Street Journal and Unspeakable Acts) and thus legitimately tick multiple options. In this case, to ensure that all possible responses are captured, five variables would be needed (i.e., one for each publication). The reason for this is that Question B is not really a single question but a series of questions of the form: 'Would you subscribe to the Wall Street Journal?', 'Would you subscribe to The Times?', and so on. The response to each of these questions would be of the 'yes/no' variety, resulting in five variables, each taking two possible values (e.g., 1 = would subscribe, 2 = would not subscribe).

HINT 1.1 If a question does not specify to 'tick one option only', chances are it is a multi-response question and cannot be represented by a single variable.

TYPES OF DATA

Let us now move on to the different types of data that one may come across in a data analysis project. While there are many data classification schemes, at this stage we shall briefly look

at different types of data according to (a) their meaning, (b) their source, and (c) their time dimension; in Chapter 3, a fourth classification of data will be introduced based upon their measurement properties.

Focusing initially on different kinds of data according to their meaning, there are data that refer to facts; that is, characteristics or situations that exist or have existed in the past. Things such as age, gender, income, church membership, 1973 sales of Wartburg automobiles, and number of drunken Taiwanese visitors at last year's Oktoberfest in Munich are all examples of facts. Descriptions of individuals' present behavior (e.g., current shopping habits or using smartphones during sex) and past behavior (e.g., historical voting patterns in Azerbaijan) also fall into this category.

Secondly, there are data that refer to awareness or knowledge of some object or phenomenon. Typical examples here are brand awareness (e.g., 'Which of the following toothpaste brands do you recognize: (a) Draculadent, (b) Vampirmed, (c) Ghoulshine?'), knowledge of important events (e.g., 'When did the German chancellor first yodel in public?'), and mastery of a certain subject or topic (e.g., 'Who wrote Das Kapital? (a) Woody Allen, (b) John Grisham, (c) Stephen King, (d) Karl Marx, or (e) Karl Marx and Woody Allen together?').

Thirdly, there are data representing intentions that are acts that people have in mind to do (i.e., their anticipated or planned behavior). Such intentions can relate to future purchasing behavior (e.g., 'Having tried your free sample of Explosion laxative, do you intend to buy it in the future?'), social behavior (e.g., 'Mrs. Corleone, you will be sorry to hear that your son has failed all his final exams. Do you intend to (a) let him get away with it, (b) cut his pocket-money by 96%, or (c) confiscate his Ferrari?'), and personal behavior (e.g., 'Do you seek to enter the United States to engage in (a) export control violations, (b) subversive or terrorist activities, (c) any other unlawful activity, or (d) golf with the president?').

Fourthly, there are attitudes and opinions data, which indicate people's views, preferences, inclinations, or feelings toward some object or phenomenon. Examples of attitude/opinions data are product or service evaluations (e.g., 'Do you think that reporting by the New York Times (a) is brilliant, (b) is fake news, or (c) does not adequately reflect the views of the lesbian, gay, bisexual, and transgender community?'), political beliefs (e.g., 'In your view, does the Favor or the Conversation Party have the better policy for providing employment opportunities for single teenage mothers in the Outer Hebrides?'), and views on social issues (e.g., 'What is the single most important problem facing humanity today: (a) poverty, (b) drugs, (c) global warming, (d) the Welsh rugby team's recent bad streak?').

Lastly, there are data relating to the motives of individuals. Motives are internal forces (i.e., desires, wishes, needs, urges, impulses) that channel behavior in a particular way. Although motivations may be complex and difficult to articulate, data of this kind are quite important because they can tell us why people behave in the way they do. Examples include reasons given for doing or not doing a certain thing (e.g., 'I go to aerobics so that my fantastic body becomes even more irresistible' or 'I don't do any exercise whatsoever because I'm a lazy slob'), explanations for preferring something over something else (e.g., 'I prefer a wall between the US and Canada to a wall between Mexico and the US'), and rationales for holding certain views or opinions (e.g., 'I think that memorizing complex statistical formulae is a total waste of time because you can look them up in a book or use a computer to do the work instead').

Table 1.4 Cross-sectional and longitudinal data sets

Variables	Data set		
	A	B	C
February 2020 purchases of Glennfiddle scotch	15		12
February 2020 purchases of Johnny Stalker scotch	10		8
February 2020 purchases of Castledrain XXXX lager	5		7
March 2020 purchases of Glennfiddle scotch		18	14
March 2020 purchases of Johnny Stalker scotch		11	10
March 2020 purchases of Castledrain XXXX lager		8	6

Note: All purchases are measured in 1,000 liters!

Turning our attention to types of data according to their source, a broad distinction can be drawn between primary and secondary data. Primary data are data collected with a specific purpose in mind; that is, for a particular research project. The researcher usually gathers such data via surveys (conducted face to face, by telephone, or through the web), experiments (carried out in the laboratory or a 'natural' setting), or observation methods (using automatic data capturing or humans to record observed behavior). In contrast, secondary data are data that have not been gathered expressly for the immediate study at hand but for some other purpose; such data, however, might be of relevance for a particular research project (in other words, somebody else has done the work but you may be lucky enough to be able to use it without getting your own hands dirty). A wealth of secondary data can be found in published statistics (by government departments, trade associations, chambers of commerce, and research foundations), annual reports (published by business firms as well as non-profit organizations), and abstracting and index services (covering thousands of periodicals, academic journals, and newspapers). For those who can afford to pay for them (and beware, because they don't come cheap!), there are also syndicated services (providing regular detailed information on a particular country, industry, or product group) and database services (allowing fast access to digital information sources worldwide, or enabling 'electronic' transfer of data sets from one location to another).

The final classification of data to be considered has to do with the time dimension and distinguishes between data relating to a single point in time and data relating to a number of time-periods. Data of the former type are known as cross-sectional data, while the latter is commonly referred to as longitudinal data. From an analysis point of view, the distinction between the two is quite important because it determines whether inferences regarding change can be made. To illustrate this, consider for a moment the three data sets displayed in Table 1.4; the units of analysis in all cases consist of 600 insomniacs living in Auchtermuchty, Scotland. Data sets A and B are examples of cross-sectional data. Each provides a snapshot of the variables of interest at a particular point in time; in this instance, data set A informs us about whisky and lager purchases in February 2020, while data set B does the same thing for March 2020. One could compare purchases across product types within each data set and reach conclusions regarding the most popular drink in a particular time-period.

Data set C, on the other hand, is an example of longitudinal data and involves repeated measurements over time on the variables of interest; data set C informs us about whisky and lager purchases in February and March 2020. As a result, one can compare not only purchases across product types but also purchases of the same product over time; thus, in addition to being able to draw the kind of conclusions data sets A and B enable, conclusions regarding changes in the relative popularity of the three drinks are now possible.

Now, having just distinguished between cross-sectional and longitudinal data, we shall immediately confuse you by suggesting that one can do longitudinal analysis by using cross-sectional data! This is not as impossible as it first sounds. Take, for example, data sets A and B and combine them; that is, imagine they are parts of the same study. Piecing the two data sets together provides information on the same variables (here whisky and lager purchases) of different (but comparable) units of analysis (here two lots of 600 Auchtermuchty insomniacs) at different points in time (here February and March 2020, respectively). Data of this kind are known as trend data and enable inferences to be drawn regarding changes in aggregate behavior, attitudes, and so on. Election polls provide a good illustration in this context: '… a poll commissioned by the Daily Polygraph, in which the voting intentions of a nationally representative sample of 11,893 bird-watchers were obtained yesterday, shows the Anti-Everything Party standing at 23%, the Conversation Party at 37%, and the Pro-Birds Party at 40%. A similar poll, conducted two weeks ago by the Financial Crimes, had Anti-Everything neck and neck with the Conversationists at 44.3% and 44.1%, respectively, with the Pro-Birds trailing at an appalling 11.6%!'

WARNING 1.2	Unless you are certain that the same units of analysis have been measured over time on the same variable, then conclusions regarding change at the individual level are not possible.

Trend data should be distinguished from true longitudinal data (such as those provided by data set C), where the same units of analysis are studied at different points in time. The latter is sometimes also referred to as panel data, and in addition to capturing aggregate changes over time, they enable inferences to be drawn as to changes in individual behavior. A typical example of this kind of data is provided by consumer panels, in which a number of individuals or families (usually balanced on such variables as age, income, and geography) record their purchases of a number of products at regular intervals (e.g., monthly). Their records are subsequently used, among other things, to determine the degree of brand loyalty (i.e., the proportion of those buying brand X in period 1 who also bought brand X in period 2) and degree of brand switching (i.e., the proportion of those buying brand X in period 1 who bought some other brand in period 2). As you have probably fallen asleep by now, have a look at WARNING 1.2 to wake you up.

In general, given a set of variables, four basic kinds of studies can be distinguished according to (a) whether the variables concerned are measured once or repeatedly and (b) whether the same or different units are studied in each case (Table 1.5).

As you might have astutely suspected, there are more types of data sets than those we have discussed. The good news is that, for the most part, they represent different combinations or

Table 1.5 Basic types of studies

Units studied	Points in time for observations	
	One	Many
Same	Cross-sectional study	Panel study
Different	Cross-sectional replication	Trend study

variations of the basic types shown in Table 1.5. For example, the well-known experimental design of the 'before–after' variety with a control group is essentially a combination of two panels, one of which has been unlucky enough to be exposed to the experimental treatment (e.g., subjects were forced to watch 43 TV adverts of Loch-Ness-Café Gold Bland) and one of which has been spared the torture (i.e., subjects were left in peace). Similarly, an omnibus panel (no, this is not the rear section of a London double-decker bus) is essentially a series of cross-sectional studies, in which the same group of individuals (i.e., the panel members) is measured on different variables at different points in time (e.g., at one time, panel members may be asked to evaluate the aesthetic appeal of alternative packages for cat food and, at another time, to indicate their attitudes toward a new type of heavy-duty flea spray). Often, a sub-group of the total panel is selected for a particular purpose; for example, if one wanted to study consumer reactions to a new type of push-up bra, only those panel members that are female and over a certain age would (usually) be surveyed.

DATA AND INFORMATION

Before we move on to even more exciting stuff, we need to look briefly at the relationship between data and information. In everyday language, the two are usually taken to be synonymous; however, one can distinguish between them in at least two senses.

Firstly, one can look at information as the product of data; that is, information as data that has been digested and analyzed. In other words, information is the knowledge obtained and conclusions arrived at after appropriate analytical techniques have been applied to the data matrix in Table 1.2. Arguably, a mass of raw data as described at the beginning of this chapter (i.e., unsummarized/unstructured) is of little informational value in itself. Imagine, for example, that cookies traced the web-surfing habits of 20 million Russian consumers during the last three months and you have access to this data. Even if you are Ms. Bigbrains in person, you need to analyze these data in order to extract managerially useful information, such as on which sites to place banner ads for your new brand of egg-timers that play the national anthem when the eggs have been boiled to perfection.

A second distinction between data and information can be drawn on the basis of relevance; under this view, information is data relevant for a particular decision. To illustrate this, take again the example of deciding where to place banner ads for the new egg-timer. Assume that, in addition to the data on web-surfing habits of Russian consumers, you are also given the following: (a) price lists for banner ads of major websites, (b) a list of websites where your competitors place their banner ads, and (c) the shoe size of your boss's secretary. While most of us would agree that (a), (b), and (c) all constitute some kind of data, virtually no one (sober,

at least) would consider (c) as relevant information. Of course, for a different decision, (c) may be a perfectly legitimate informational input (e.g., if the objective is to order sports shoes for the office rugby team).

SUMMARY

Let us take a deep breath and try to summarize what we have learned so far. We first looked at the nature of data, distinguishing between units of analysis, variables, and values; we then put the three together to create a data matrix. Next, we warned against confusing questions with variables and discussed different kinds of data according to their meaning, source, and time dimension. Finally, we considered the link between data and information. It is now time to make a cup of coffee and mentally prepare for the things to come.

QUESTIONS AND PROBLEMS

1. What is the difference between a variable and a value?
2. Give two examples of different units of analysis and two examples of variables that could be used to characterize them.
3. What determines the size of a data matrix (i.e., the number of rows and columns)?
4. Give three examples of multi-response questions. How many variables would you need to capture the answers to them?
5. Distinguish between facts, awareness, intentions, opinions, and motives, and give an example of each. Why is it important to distinguish between these types of data?
6. What would you consider to be the main advantages (viz. disadvantages) of primary versus secondary data?
7. What type of data do you need in order to make inferences regarding change?
8. What is the key difference between trend data and panel data?
9. It has been argued that 'students often wonder what statistics is and why they should bother to study the subject'. What is your view on this?
10. What is your favorite dish? (We like to get to know our audience.)

FURTHER READING

Bryman, A. (2016). *Social Research Methods*, 5th edition. Oxford: Oxford University Press. An excellent introduction to all aspects of the research process.

Creswell, J. W. & Creswell, J. D. (2017). *Research Design: Qualitative, Quantitative and Mixed Method Approaches*. London: Sage Publications. All you need to know regarding thinking about, designing, and implementing different kinds of research projects.

Sheehan, K. B. & Pittman, M. (2016). *Amazon's Mechanical Turk for Academics: The HIT Handbook for Social Science Research*. Irvine, CA: Melvin & Leigh. A comprehensive guide to collecting data via Mechanical Turk.

2
Does *sampling* have a purpose other than providing employment for statisticians?

THE NATURE OF SAMPLING

Having taught you all we know about data sets, we shall ignore your cries for mercy and proceed to do exactly the same with regard to sampling. Crudely speaking, a **sample** is a part of something larger, called a **population** (or 'universe'); the latter is the totality of entities in which we have an interest – that is, the collection of individuals, objects, or events about which we want to make inferences. For example, in Table 1.4 earlier, the population implied was 'all insomniacs living in Auchtermuchty, Scotland'. Other examples of populations include 'all countries on the planet Earth', 'all furniture stores in the UK', 'all strikes at Italian knitting factories during 1975–79', and 'all vegetarian chiropodists based in Greater Manchester'. We can now define a sample as *a subset* of a given population. Going back again to Table 1.4, the sample consisted of 600 insomniacs living in Auchtermuchty, who were fortunate enough to have their purchasing behavior studied. The Scottish Tourist Board assured us that the total number of insomniacs living in Auchtermuchty is much higher! Possible samples for the other population examples mentioned above are 'member countries of NATO', '28 furniture stores located anywhere in Somerset and Devon', 'strikes at Italian knitting factories during the first six months of 1977', and '13 vegetarian chiropodists with city center practices in Manchester' (note that these are merely *possible* samples that could be drawn from the above populations, not necessarily *good* samples – an issue to be addressed later).

The rationale for sampling is really very simple: by checking out part of a whole, we can say something about the whole. However, you may ask, 'But why bother with a sample in the first place? Why not study the population instead?' Well, in some instances, the entire population is indeed studied as, for example, with the population **census**, which is conducted every few years by the government. However, in most cases, undertaking a census is not possible or desirable for reasons summarized in Table 2.1.

Despite the many advantages of sampling over a census, it is important to keep in mind that there is a price to pay, namely deciding whether the picture painted by the sample can be generalized to the population of interest. To illustrate this problem, we have applied our

Table 2.1 Reasons for preferring a sample over a census

Reason	Rationale	Example
Cost	A census is almost always more expensive than taking a sample; sometimes, the cost of a census is simply prohibitive.	Interviewing all car owners in all European Union member countries as to the color of windscreen washer liquid they prefer would cost millions of euros; the value of the information obtained to a car accessories manufacturer would not outweigh the cost of commissioning such a census.
Time	A census may take more time to conduct and analyze than is available for making the decision involved. In other words, doing a census may simply take *too* long.	The time needed to complete a census of all ski-enthusiasts in North America regarding their ski resort preferences would be unacceptably long if the research related to updating a *Snob Skiers Guide*, which *must* hit the bookshops by September (and it's already May!).
Destruction/ contamination of population members	A complete census may result in destruction or contamination of the entire population; 'destructive testing' procedures must, by necessity, be limited to a sample only.	Torture-testing each and every condom coming off the production line for durability, tensile strength, and friction resistance would result in no products reaching the shops (and a lot of pregnant people).
Decision importance	A minor decision would not normally justify the hassle of a census; a 'quick and dirty' sample may be all that is needed.	Deciding where to place a pink peach bubble tea dispenser in a tractor factory with 600 employees would not justify a census of 'perceived optimal locations'; on the other hand, whether to move the factory to a new location 50 miles away is an issue for which a census of the workforce's intentions to stay with the firm may be called for.
Confidentiality	A census is more likely to be noticed by interested parties than a sample study; with a census, the chances that competition will get wind of what's going on are much higher.	Test-marketing a new family-size package of beetroot-banana pickles by placing it in *all* major supermarket chain outlets (e.g., Aladi, Strangeway, Fresco, and Painsbury) is much more likely to invite competitive retaliation than if only a few carefully selected stores are used for test-marketing purposes.
Accuracy	Unbelievable as it may sound, a census may be *less* accurate than a sample; the latter's sampling error may be outweighed by the former's non-sampling error resulting in a greater *total* error in the case of a census.	Sending out 75 hastily recruited part-time interviewers to manually record the views of the entire student body of Edinburgh University on the introduction of a mandatory 'tartan dress code', and using another 20 semi-qualified, bored-to-death typists to input the data to a computer is likely to result in many non-sampling inaccuracies (e.g., interviewer variability and bias and/or transcription and typing mistakes).

magnificent artistic talents and creatively depicted the relationship between population and sample in graphical form; the relevant masterpiece is shown in Figure 2.1.

The first thing to realize is that, given a population containing a fixed number of elements (what is known among statistical geniuses as a **finite population**), a number of possible samples could be drawn. Just to put it into perspective (and have you faint in the process), a total of 1,099,511,600,000 different samples could potentially be drawn from the population

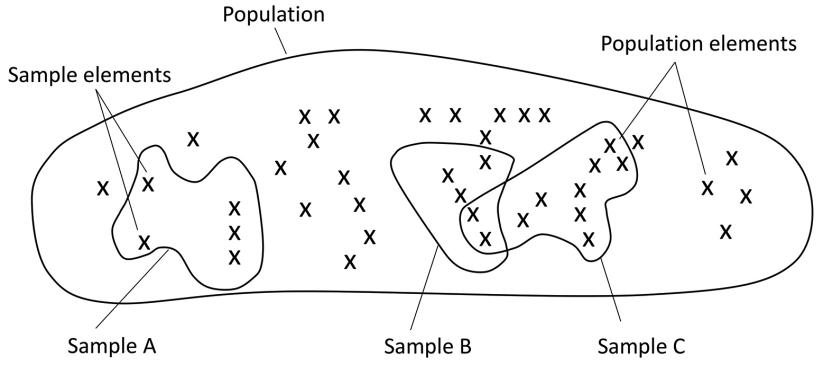


Figure 2.1 Population and sample

shown in Figure 2.1 (which only has 40 elements in the first place!). In general, if there are N elements in the population and the size of the sample is not fixed, one has a choice between 2^N possible samples (including taking no sample at all and conducting a census of the population). Of course, the choice of possible samples is drastically reduced if a certain **sample size** is pre-specified; that is, if the number of elements to be included in the sample is fixed beforehand (not an easy matter, either, as will be further discussed below). Going back to Figure 2.1, and assuming that we wish a sample of size 10 (such as sample C), our options are reduced to 'only' 847,660,530 possible samples! In general, given a desired sample size n (where $n \leq N$), then $\frac{N!}{n!(N-n)!}$ different samples could be drawn from a population with N elements. Thus, whichever way you look at it, one has a lot of choices when it comes to sampling.

SAMPLE SELECTION

Given the above, the key question becomes how to select the n **sample elements** (i.e., members of the sample) so that they are representative of the N **population elements** (i.e., members of the population). Unfortunately, the answer to this is just as difficult as selecting a 'good' husband/wife/partner (in other words, there is no easy answer!). However, one important consideration is the effect that *excluded* population elements are likely to have on the quality of the sample. Sampling, by definition, means that certain population elements will be excluded from the sample. This exclusion causes what is known as **sampling error**: the difference between a result based on a sample and that which would have been obtained if the entire population was studied (i.e., the 'true' value). Sampling error is generated whenever a sample is drawn by *whatever sampling procedure* and is a function of sample size (i.e., as the sample size increases, sampling error decreases).

To get a better feel of the concept of sampling error (while avoiding nasty technicalities), imagine that the population in Figure 2.1 consists of 30 female and 10 male students enrolled in an advanced entomology class at Mosquito State University, and you have just picked *at*

random a sample of five students (Sample A); you could have done this by drawing names out of a hat or throwing darts at the class roster (blindfolded, of course). Now, it just *could* be the case that as a result of pure chance, only male students were selected and, thus, Sample A would not be representative of the entire class (as female students – the majority in the population in this case – are not included in the sample). Remember that you did not *intentionally* try to influence the composition of the sample – it just so happened that no female students were chosen. However, the end result (i.e., no representation of female students) is exactly the same as if you had used a more 'subjective' procedure to select the sample, such as choosing only people of the same gender as you (assuming you are male) resulting in, say, Sample B.

So, is there a difference between Sample A (based upon random selection) and Sample B (based upon personal preference) if both contain sampling error and both are unrepresentative of the relevant population? Yes, there is. The difference is that with Sample A, we can *assess* the extent of sampling error, whereas, with Sample B, we cannot. More specifically, Sample A was obtained by means of a **probability sampling** procedure, whereby each element in the population had a known, non-zero probability of being included in the sample (i.e., the sample elements were selected by chance, and this chance was known for each element being selected). Consequently, we can apply the **laws of chance** (i.e., probability theory) and estimate how likely it is that it reflects the 'true' situation in the population. Sample B, on the other hand, was obtained utilizing a **non-probability sampling** procedure since the selection of sampling elements was left to the discretion of the researcher, and there was no explicit scientific model (such as probability theory) that could be used to assess the degree of sampling error.

Thus, the key difference between probabilistic and non-probabilistic sampling methods is *not* that the former will always produce a more **representative sample** than the latter. Rather, a *statistical evaluation* of sampling error can be undertaken with the former, thus enabling the researcher to assess *how likely* the sample is to be unrepresentative and by how much. Such an assessment is not possible with samples drawn by non-probabilistic methods (we will return to this topic in Chapter 8, where we will see how sample results can be used to estimate values in the population). It should also be borne in mind that **non-sampling errors** (e.g., measurement errors and non-response errors – see Chapters 3 and 4) also affect research results and, therefore, there is no guarantee that a probability sample will produce overall more accurate results than a non-probability sample.

WARNING 2.1 Probability sampling procedures do not ensure that the sample will be representative or the results accurate. What they do is to allow the assessment of sampling error.

From the above, you may wonder why we should bother with non-probability sampling methods, given that no calculation of sampling error is possible. The reason for this is that, in addition to statistical criteria, practical considerations also influence the choice of a sampling method. For example, applying probability sampling procedures tends to be more costly in time and/or money, requires a certain level of statistical training to use effectively, and assumes that a suitable **sampling frame** (i.e., a list of population elements) is available. Moreover, when one is not very concerned with the accuracy of **population estimates** (e.g., when only a

Table 2.2 Sampling methods

	Selection criteria	Example
Non-probability methods		
Convenience (chunk) sampling	Sample members are chosen based on their being readily available/accessible; thus, selection is made based on convenience.	A sociology professor interested in people's attitudes toward mixed saunas asks her introductory class to fill in a questionnaire on the subject.
Judgmental sampling	Sample members are chosen on the basis of the researcher's judgment as to what constitutes a representative sample for the population of interest; thus, potential sample members are screened judgmentally as to whether or not they should be included in the sample.	A marketing researcher wishing to test-market the new Starveline range of diet foods uses her knowledge and expertise to select a sample of stores in key UK locations (e.g., John O'Groats, Llannerch-y-medd and, Inverkirkaig) that are apparently reflecting national tastes.
Purposive sampling	Sample members are chosen with a specific purpose/objective in mind; the sample is thus intentionally selected to be non-representative.	A manufacturer of inflatable dog huts purposely over-samples a number of dog owners who are current non-users to gauge reactions to a new product line (the aim being to attract new customers while retaining existing ones).
Quota sampling	Sample members are chosen based on satisfying some pre-specified criteria thought to apply to the population; the researcher is free to choose which elements to include in the sample as long as they qualify on the pre-defined characteristics.	A government department trying to assess the effectiveness of its 'Get a Job or Get Lost' advertising campaign (designed to promote employment opportunities) surveys 1,000 adults, of which 120 are unemployed and the rest employed – the two groups (quotas) reflecting the current employment situation at the national level.
Multiplicity (snowball) sampling	Sample members are initially chosen either judgmentally or through a probability sampling method and are subsequently asked to identify others with the desired characteristics; thus, the final sample is constructed from referrals provided by the initial respondents.	A professional association (e.g., The National Society of Creative Accountants) wishing to up its membership contacts its existing members and solicits names of other individuals that may qualify for joining.
Probability methods		
Simple random sampling	Sample members are chosen randomly for inclusion in the sample, with each population element having an equal probability of being selected; each possible sample of n elements thus has a known (and equal) chance of being the one actually chosen.	Excessive Premium Insurers, an Australian insurance company with 850,000 motor insurance policies, randomly picks 2,000 policy-holders and approaches them via email to assess their level of satisfaction with the value-for-money provided.

	Selection criteria	Example
Systematic sampling	Sample members are chosen at regular intervals after a random start; the sampling interval is the ratio *N*/*n*, where *N* and *n* represent the population and desired sample size, respectively.	A dubious car dealer with 3,000 customers wants to survey 200 of them prior to offering a new 'all-inclusive' warranty scheme; she thus sets a sampling interval of 3000/200 = 15, picks at random a starting value between 1 and 15 inclusive – say, 8 – and subsequently selects from his customer records the 8th, 23rd, 38th, and so on victim (sorry – customer!).
Stratified sampling	Sample members are chosen randomly from different *segments* (strata) of an overall population; each stratum may be sampled in proportion to its size in the overall population (proportionate stratified sampling), or sample members of different strata may have disproportionate chances of being selected (disproportionate stratified sampling).	National Ripoff Bank splits its customer population into 'consumer' accounts (accounting for 80% of all accounts) and 'business accounts' (accounting for 20% of all accounts) and wants to conduct a survey of customer satisfaction among 500 account holders. If it wishes to retain the 80:20 consumer-to-business account ratio in the sample, it would select at random 400 consumer accounts and 100 business accounts. If, however, for one reason or another, it is felt that greater variability in perceptions is likely to exist among business account holders, then more of the latter would be included in the sample; for example, the sample composition decided upon may consist of 300 consumer accounts and 200 business accounts (picked at random within each stratum).
Cluster sampling	Sample members are chosen in groups (clusters) rather than individually; the clusters themselves are chosen randomly from a population split into groups.	A sales director wishing to conduct a study of salesforce motivation randomly picks three sales districts (out of a total of eight) and holds seven-hour, in-depth interviews with all 36 salespeople covering the chosen sales districts.
Multistage sampling	Final sample members are chosen by means of one of the other probability methods described above, but a number of stages precede the final selection.	Exaggeration Research Incorporated, an opinion poll organization, carrying out a nationwide poll on education standards, (a) selects at random 80 cities, then (b) selects randomly 30 blocks within each city, then (c) takes a systematic sample of households within each block previously chosen, and finally, (d) interviews the head of household (phew!).

'first-look', 'rough-and-ready', or 'broad-brush' view of the population is needed), when the population concerned is a **homogeneous population** (i.e., there is little variation among the population elements of the characteristic(s) of interest), and when the **expected cost of errors** in the obtained information is not very high (e.g., as in an unimportant decision), then using some kind of non-probability sampling procedure may be justified.

In terms of choice, there is a wide range of both probability and non-probability sampling methods available to suit all tastes, pockets, and desires for complicating things. Since there

are many books on sampling (written by much better authors than us), we shall refrain from getting into the intricacies of the different methods, but instead provide you with a quick summary of the most important ones (Table 2.2). Those of you with a desire for punishment can look at the Further Reading at the end of the chapter for more details.

One important point to note from Table 2.2 is that, in some instances, the sample elements (in the sense discussed previously) are not the same as the **sampling units**, the latter being the units actually chosen by the sampling procedure. More specifically, a sampling unit may contain one or more sample elements, as is the case, for example, with cluster sampling and multistage sampling. A major reason sampling units and sampling elements may not coincide is that there may not be a suitable sampling frame available to enable direct selection of the latter. For example, if one wished to conduct a survey of marketing managers in the funeral industry, it might be difficult to obtain an appropriate register listing these individuals; as a result, one might have to sample funeral homes instead as the basic sampling unit (e.g., using the relevant trade association directory).

SAMPLE SIZE DETERMINATION

Up to this point, very little attention has been paid to the question of sample size other than pointing out that it is inversely related to sampling error – a massively important point considering that sampling error is pretty much the source of all evil. As was the case with the choice of sampling method, the determination of sample size can be a rather complex issue, involving both statistical and practical considerations; let us highlight a few key points in what follows.

One of the key statistical considerations in sample size determination is the degree of **variability** in the population; the more heterogeneous the population, the larger the sample size needed to capture the diversity in the population. For example, if the population consisted of 300,000 identical pink 40-watt lightbulbs, a sample of one would be sufficient to describe the population perfectly; on the other hand, if the population concerned consisted of the 40 entomology students of Figure 2.1, a larger sample would be needed (if only to capture differences in age, gender, family background, etc.). Obviously, the degree of variability characterizing a population of interest requires the researchers' domain knowledge in order to be assessed and is inextricably tied to the specific context and research purpose at hand. (This highly eloquent sentence has been created by our co-author George – and we think it is appropriate to give him full credit for this!). For instance, a study focusing on whether consumers spend more money when highly attractive salespersons serve them would reasonably assume higher variability in the population as opposed to an eye-tracking experiment investigating visual attention to bold versus italic fonts in print advertisements.

A second statistical consideration is the desired degree of **precision** associated with population estimates based on a sample; the greater the precision required, the larger the sample size needed. For example, if one wanted to estimate the average height of the 40 students in the entomology class within ± 1 cm, a larger sample would be needed – all other things being equal – than if one wanted an estimate within ± 10 cm of the true average height.

A third statistical consideration relates to the desired degree of **confidence** associated with any estimates made. Sticking to our brilliant entomology example, if one wished to estimate the average height of the students in the class with a precision of, say, ± 5 cm *and* wanted to be 95% confident that this estimate would contain the true population value, a larger sample size would be required than if one wanted to have an estimate of the same precision but a lower confidence level (e.g., 90%). It does not take a PhD in Incredibly Advanced Mathematical Statistics to realize that there is a trade-off between precision and confidence; increasing confidence *without* sacrificing the level of precision requires a bigger sample size (and vice versa). Thus, in determining sample size, one has to balance these two considerations against each other.

A final statistical consideration concerns the extent to which the intended analysis will involve using sub-samples for **cross-classification** purposes and/or the use of statistical techniques, which assume a minimum sample size to produce meaningful results. For example, if you were only interested in an overall picture of the characteristics of the students in the entomology class, a smaller sample size would be required than if you had plans to cross-tabulate, say, three income categories (e.g., 'filthy rich', 'moderately comfortable', and 'totally broke') with two gender categories. In the latter case, you would be setting up a table with $3 \times 2 = 6$ cells, and you would obviously need to have sufficient observations in all of them if you want to draw inferences about the sub-populations involved. To drive this point home, if you take a sample of, say, five students, it would be *impossible* to fill all cells (since there are five people to fit into six feasible slots). Though it is difficult to generalize from project to project (particularly in the light of the practical limitations discussed below), as a rough rule of thumb (for which we accept no liability!), aim for about 100 observations in each category of the major/important sample breakdowns and around 20–50 in the minor breakdowns.

HINT 2.1	Think about the most important cross-classifications you want to produce with your data before deciding on the overall sample size.

Moving away from statistical criteria (hooray!) to more practical concerns, issues of resource availability in terms of time, money, and personnel available also have a (sometimes overriding) impact on the sample size. The poorer you are in terms of time or money, and the fewer the subordinates, friends, and/or relatives you can cajole to give you a hand, the less likely it is that you will be able to go for a large sample. In addition, expectations regarding **non-response** and/or **missing values** need to be considered in determining the sample size, as well as the **value of information** provided by different size samples in relation to their costs. Regarding the latter, the **Bayesian approach** to sample size determination can provide a formal procedure for selecting the sample size that maximizes the difference between the expected payoff of sample information and the estimated cost of sampling (see Further Reading). Also, rather than drawing a **fixed sample** (i.e., determining sample size *prior* to data collection), researchers sometimes prefer to employ a **sequential sample**. Under this approach, the researcher outlines a clear plan of stepwise data collection driven by the information obtained in each step. If the results are not conclusive after a certain sample is taken, more observations are made; if the increased sample size still does not furnish conclusive results, more population elements are

Table 2.3 The sampling process

Stage	Description	Potential errors
1. Define the population	Specification of the target population in terms of elements (e.g., teenagers), sampling units (e.g., households), extent (e.g., in Paris), and time (e.g., having lived in Paris for at least a year).	*Population specification error:* the target population is inappropriate for the problem at hand. For example, surveying only homemakers regarding their criteria for choosing a car would ignore that husbands/wives/partners sometimes also play a role in influencing the decision.
2. Specify the sampling frame	Specification of the listing, directory, or roster from which the sample will be chosen (e.g., the telephone directory); a sampling frame is essential if a probability sample is to be drawn.	*Frame error:* the chosen sampling frame may be inaccurate (i.e., capture other populations in addition to the one of interest) or incomplete (i.e., exclude some population members) or may include certain individual population elements more than once. For example, using a trade association list to draw a sample of motorcycle dealers will not cover those dealers that are not members of the association (because they are too 'dodgy' to be accepted or too mean to pay the relevant membership fee).
3. Select the sampling method	Specification of whether a probability or non-probability approach will be applied to draw the sample and exactly how the sample members will be selected.	*Selection error:* a non-representative sample is obtained as a result of a non-probability sampling procedure. For example, in a survey of attitudes toward animals in which a quota sample according to, say, age and location is used, bias will be introduced by interviewers who systematically avoid homes with cats, dogs, etc. (because they hate or are afraid of all animals, including goldfish).
4. Determine sample size	Specification of the number of sample elements to be included in the final sample as well as the number of any intermediate sampling units (e.g., when a multistage sampling procedure is being followed).	*Small numbers error:* the sample is too small to permit the desired statistical analysis and enable meaningful generalizations for the population. For example, making a random selection of eight individuals from the electoral roll and using their voting intentions to predict the outcome of the forthcoming general election would not impress many people.

Stage	Description	Potential errors
5. Draw the sample and collect the data	Specification of the operational procedures for the selection of sample members and carrying out the fieldwork (e.g., call every third name in the telephone directory; if not in, call back in half-hour intervals; if refuses to answer, pick next name, etc.).	*Sampling error:* the results based upon the sample will practically always differ from those that would have been obtained had the entire population been studied instead; recall that only probability methods allow for an assessment of sampling error.
		Non-response error: the researcher may fail to contact some members of the selected sample and/or some contacted members may fail to respond to all or some part of the research instrument. For example, an unscrupulous interviewer may make up the answers for sample members who happened to be out when they were called (thus avoiding the need for a call-back on another cold winter night), or respondents might simply refuse to cooperate (because of tiredness or sensitivity of the subject matter, or because they hate researchers).

added to the sample, and so on (until the information obtained becomes sufficient to permit a valid conclusion).

THE SAMPLING PROCESS

As a final point, hopefully not sending you into the depths of depression, Table 2.3 outlines the stages of the sampling process and, more importantly, indicates the kind of errors that can occur at each stage. The moral of the story is that the quality of the sample on which subsequent analysis is based is a *major* determinant of the quality of your conclusions. No matter how clever you are in using statistical techniques, you simply *cannot* produce meaningful, credible, and generalizable results from a poor sample.

SUMMARY

Let us summarize our journey into the wonderful world of sampling for the two readers who have not (yet) gone to sleep. We started by looking at the reasons for sampling and then spent some time exploring the link between a sample and a population. Next, we introduced the concept of sampling error and distinguished between probability and non-probability sampling methods. The influences bearing upon sample size determination were considered next and we concluded our journey with a look at the stages of the sampling process, and the things that can go wrong during it. Are you ready for more? If so, you are presumably in the top 1% of your peer group!

QUESTIONS AND PROBLEMS

1. Give five reasons for taking a sample instead of conducting a census; then, give five reasons for doing a census rather than taking a sample!
2. How would you define a 'population'? If you have a population with 10 elements, how many different samples of size 5 can you draw?
3. What is sampling error? What are non-sampling errors?
4. Do probability sampling methods *always* produce a more representative sample than non-probability methods?
5. What are the advantages and disadvantages of probability versus non-probability samples?
6. What are the key determinants of sample size?
7. Select one probability and one non-probability sampling method and give examples of research situations in which you would use them.
8. Is there a difference between a sampling element and a sampling unit?
9. Describe the stages of the sampling process and give one example of an error that can occur at each stage.
10. If you didn't have to learn about sampling, what would you be doing right now?

FURTHER READING

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. New York: Wiley. *The* classic on the subject, although very technical and heavy-going; an excellent source of reference, nevertheless.
Kalton, G. (2021). *Introduction to Survey Sampling*. London: Sage Publications.
Lohr, S. L. (2019). *Sampling: Design and Analysis*. Boca Raton, FL: CRC Press.

3
Why should you be concerned about different types of measurement?

THE NATURE OF MEASUREMENT

Having provided you with an unsurpassable discussion of sampling issues in the previous chapter, we need to complement this with an equally outstanding examination of measurement questions (yes, we know that modesty is one of our 328 virtues – shyness is another!). In terms of the basic structure of the data matrix, sampling considerations relate to the *rows* of the matrix (i.e., the *units of analysis*), whereas measurement considerations relate to the *columns* of the matrix (i.e., the *variables* and the assignment of *values* to them). Incidentally, if you don't remember what a data matrix looks like, then go back to Chapter 1 and have another look at Table 1.2.

Before we can look at how to measure something, we must think a little about what we want to measure; that is, we need to define the **concept** involved. Concepts express abstractions formed from observations from numerous particular happenings. (Is that a beautiful sentence or what?) For example, the concept 'bicycle' refers to the generalization of the characteristics that all bicycles have in common. In scientific research, we also often speak of **constructs**; these are concepts that have been consciously and deliberately *created* by scholars for particular scientific purposes (e.g., 'introversion', 'product life cycle', 'socialization'). For practical purposes, the terms 'concept' and 'construct' are often used interchangeably.

Defining a concept (or construct) is not always easy, particularly when the concept we are interested in does not have a physical referent. Compare, for example, the concepts 'bicycle' and 'brand loyalty'. The former is much easier to explicate as it is closely related to physical reality, and little disagreement would result if different people were asked to explain its meaning. In contrast, 'brand loyalty' is an abstraction that is much more difficult to define and, thus, measure.

In defining concepts, it is important to distinguish between two approaches. A **conceptual definition** defines a concept in terms of other concepts, the meaning of which is assumed to be more familiar to the reader. For example, brand loyalty could be conceptually defined as 'a consumer's preference for a particular brand in a particular product category over a (usually prolonged) period of time'. A conceptual (or 'constitutive') definition is thus roughly equiva-

lent to a dictionary definition. It aims to (a) capture the essence or key idea of the concept and (b) distinguish it from other similar but, nevertheless, distinct concepts. An **operational definition**, on the other hand, describes the meaning of a concept through specifying the *procedures* or operations necessary to measure it. For example, brand loyalty may be operationalized as 'a sequence of consecutive purchases (typically, three or four) of the same brand'. Thus, an operational definition aims to translate the concept into observable events by specifying what the investigator must *do* in order to measure the concept concerned.

Clearly, a conceptual definition logically precedes an operational definition and, thus, it should be used to guide the development of the latter. In this context, it is possible – and most often desirable – to have *multiple* operationalizations of the *same* concept. For example, brand loyalty could be captured by assessing to what extent consumers (a) intend to buy the brand in the near future, (b) would actively search for the particular brand in order to buy it, and (c) would buy other products of the same brand. Using multiple operationalizations offers a better way to capture the nuances of a complex construct, thus allowing for comprehensive measurement. In simple words, employing multiple operationalizations helps us make sure that we are tapping into all (or, at least, more) aspects of the concept we are trying to measure. Relying solely on single-item operationalizations limits the information available and may result in an incomplete representation of the concept concerned. This is why researchers typically prefer the use of **multi-item scales** to measure their constructs of interest.

HINT 3.1	Use conceptual definitions as guides for developing operational definitions. Generate multiple operational definitions and employ a number of them to comprehensively represent your concept.

It is important to appreciate that operationalization problems can be posed even by concepts that appear to be relatively clear and straightforward. Take, for example, the concept of 'firm size'. Although most of us would agree that Microsoft is a 'large' firm and the Chinese takeaway around the corner a 'small' business, if we wanted to take a sample of different-sized firms, how should we actually define size? Should we look at sales volume (and, if so, at physical units or revenues)? Should we look at employment (and how do we deal with part-time employees)? Or should we look at value of assets (gross, net, including intangibles such as 'brand equity', etc.)? What the complexities of this seemingly 'easy' operationalization task show is that, in most situations, there is no 'obvious' measure.

Operational definitions and measurement go hand in hand, the former specifying how the latter should be undertaken for the concept(s) involved. In general, the **process of measurement** can be thought of as the assignment of symbols to characteristics of persons, objects, states, or events according to certain rules. There are three key points associated with this definition. Firstly, it is not the persons, objects (etc.) themselves that are measured but rather their characteristics. As discussed in Chapter 1, certain variables describing our units of analysis are what are being studied. For example, we do not measure our friend Rudolph; we measure his age, height, opinions, capacity for drinking beer, sexual stamina (before and after beer drinking), or some other characteristic. Even when we count things, what we are, in fact, measuring is the characteristic of 'being present'.

WARNING 3.1 There are no 'obvious' ways of operationalizing even the simplest concept. Think twice (and then, think again!) before you decide on an operational definition.

Secondly, the assignment of symbols to characteristics is not done arbitrarily but according to pre-specified rules. **Measurement rules** ensure that the relations between the symbols assigned reflect the actual relations between the objects with respect to the variable concerned. To put it into statistical parlance (and give you an opportunity to practice your Greek), the assignment process is *isomorphic*; that is, there is a one-to-one correspondence between the symbol and the characteristic being measured (otherwise, the results of measurement would be useless, since knowledge of a particular symbol would tell us nothing about the person or object concerned).

Lastly, in most measurement situations, the symbols assigned are *numbers*; the use of numbers rather than other symbols (e.g., letters, colors, emojis, tattoos of the Chinese triad, etc.) provides a standardized means for communicating measurement procedures and results from researcher to researcher and user and also facilitates mathematical and statistical manipulation of the data. Thus, in practice, measurement can be seen as the use of numbers to represent the characteristics of persons, objects, events, and so on. Note that the meaning of the term 'number' in a measurement context is *not* the same as that understood in everyday life; that is, that numbers can be readily added, subtracted, multiplied, and divided. While this *may* be the case, in measurement we can have different kinds of numbers, the precise meaning of which depends on (a) the nature of the characteristic being measured and (b) the particular measurement rule used (indicating the basis upon which numbers are assigned to the characteristic).

Let us try and make this clear with an example. Suppose that you want to measure the weight of your beloved statistics Prof. Dr. Dr. Ignatz Zweistein (rumor has it that he is a distant relation of Einstein). Although the obvious thing would be to ask/bribe/force him to step on the scales and then take a reading (e.g., 105 kg), this is not the only measurement option available. Table 3.1 illustrates some alternatives.

WARNING 3.2 Numbers in measurement functions are symbols – as such, they can be used in different ways and, thus, their meaning is not invariant.

It is evident from Table 3.1 that, depending upon the measurement scheme used, the meanings of the numbers differ. This implies that the same number can tell us different things and, therefore, if we do not know the measurement rule behind the number, we cannot interpret it in a meaningful way.

MEASUREMENT SCALES

Different measurement rules result in different types of **measurement scales**. The latter can be visualized as a continuum upon which the measured units of analysis (e.g., objects, persons,

Table 3.1 Measuring the weight of Prof. Dr. Dr. Zweistein

Measurement scheme	Assignment of numbers
1. Create five categories	1 = very heavy for his age
	2 = rather heavy for his age
	3 = roughly average for his age
	4 = rather light for his age
	5 = very light for his age
2. Compare Prof. Dr. Dr. Zweistein to your other nine statistics professors	1 = heaviest
	2 = 2nd heaviest
	3 = 3rd heaviest
	•
	•
	•
	10 = 10th heaviest (i.e., lightest)
3. Define two classes	1 = fantastic weight
	2 = horrible weight

Table 3.2 Types of measurement scales

Properties	Scale type			
	Nominal	Ordinal	Interval	Ratio
Equivalence	Yes	Yes	Yes	Yes
Order	No	Yes	Yes	Yes
Equal intervals	No	No	Yes	Yes
Absolute zero	No	No	No	Yes
Typical usage	Store types Product categories Geographical locations Different species of bats	Education Preference order Choice rankings	Index numbers Temperature Calendar time Attitudes	Sales Costs Age Number of customers

etc.) can be located. The measurement schemes in Table 3.1, for example, use different kinds of scales on which to place Prof. Dr. Dr. Zweistein. A key distinction between different types of measurement scales is according to the **level of measurement** they provide; that is, the amount of information they convey about the measured objects and the permissible mathematical/statistical operations that can be applied to the resulting data. In this context, four major types of measurement scales can be distinguished (see Table 3.2).

As the name implies, a **nominal scale** is a scale 'in name only' and represents the simplest type of scaling. In nominal scaling, the numbers used have no mathematical properties in themselves and serve only as labels for identification and/or classification. For example, assigning a unique number to each athlete in a mud-wrestling event serves only to identify the individual athletes taking part (otherwise, how would you know who won?). This is the most

elementary nominal scale (sometimes referred to as a **label nominal scale**); there is a strict *one-to-one* correspondence between each number and each athlete and, as long as this correspondence is preserved, any set of numbers could be assigned.

A more common nominal scale is the **category nominal scale**, whereby the numbers assigned represent *mutually exclusive* and *collectively exhaustive* categories of persons, objects, and so on. For example, classifying individuals according to their nationality, eye color, or gender is done by means of category nominal scales. (Be careful and always ask for 'gender', not for 'sex'; if you unwittingly ask for 'sex', you may get 'three times a week' or 'not enough' as an answer.) As long as the same number is not assigned to different objects or different numbers to the same object, again, any set of numbers can be used. For example, in classifying individuals according to nationality, the following schemes would furnish identical results: [1 = Hungarian, 2 = Korean, 3 = German]; [1 = Korean, 2 = German, 3 = Hungarian]; [−7 = German, 23 = Korean, 247 = Hungarian]; [456 = German, 0 = Hungarian, 9 = Korean]. This illustrates that the only property conveyed by the numbers in a nominal scale is that of **equivalence** (i.e., identity) and, consequently, the only legitimate mathematical operation is *counting*; that is, the enumeration of persons, objects, and so on falling into each category. For example, suppose we classified 300 individuals according to their nationality using any of the scales above. In that case, we could conclude that we had so many Hungarians (e.g., 120), so many Koreans (e.g., 80), and so many Germans (e.g., 100). We could also say that 40% of our sample consisted of Hungarians and 60% of Koreans and Germans. Finally, we could state that the **mode** (i.e., the most frequently occurring value) was 1 (first measurement scheme), 3 (second measurement scheme), 247 (third measurement scheme), or 0 (fourth measurement scheme), representing in all cases the Hungarians. There is hardly any more analysis we can do with the results, such as calculating an arithmetic average (what does 'average nationality' mean?).

An **ordinal scale** establishes an *ordered relationship* between persons or objects. In ordinal scaling, numbers are used to indicate whether a person, object, and so on has more or less of a given characteristic than some other person or object. However, the numbers do not provide information as to *how much* more or less of the characteristic is possessed by the person or object concerned. Thus, with ordinal scales, we classify the units of analysis in meaningfully ordered/ranked categories without giving additional information about how much the units in these categories differ. That is, an ordinal scale can tell you which athlete won the gold, the silver, and the bronze medal in a sprint race but does not tell you anything about how much faster the gold medalist was compared to the other athletes.

In a similar sense, a consumer may be asked to try out and rank four brands of cornflakes (A, B, C, and D) according to 'digestibility' as follows: 1 = 'most digestible', 2 = 'second most digestible', 3 = 'third most digestible', 4 = 'least digestible'. Suppose that the following results are obtained: 1 = B, 2 = D, 3 = A, 4 = C. While we know that the order of preference is B, D, A, and C, we do not know anything about the *differences* in digestibility among the four brands (e.g., we cannot tell whether, say, the difference between B and D is less than, equal to, or greater than the difference between D and A). This means that as long as we preserve the ordering relationship, any set of numbers could have been used in our scale (e.g., 3 = 'most digestible', 48 = 'second most digestible', 49 = 'third most digestible', 1,245 = 'least digestible'). Notice that the assignment of low numbers to indicate better performance is a matter of

convention/convenience only, since when we rank things, we tend to think in terms of 'first class', 'second class', and so on. We can rank equally well by assigning low numbers to indicate poorer performance (e.g., as in 'first-class idiot'). In short, given that the key property conveyed by the numbers in an ordinal scale is that of **order**, we can transform the scale any way we wish as long as the basic ordering is maintained. Put into disgustingly complicated jargon, this means that any strictly increasing (i.e., positive monotonic) transformation is permissible on ordinal scales.

As you may have gathered, an ordinal scale subsumes the features of a nominal scale (since equivalent entities – i.e., persons, objects, etc. – receive the same rank) and, therefore, all the mathematical/statistical operations applicable to a nominal scale are also permissible on an ordinal scale. However, since ordinal scales incorporate order as an additional property, we can do more in terms of analysis with ordinal data than with nominal data (e.g., in addition to computing percentages and finding the mode, we could also calculate the **median**; that is, the value above which and below which half the scores lie). Having discussed nominal and ordinal scales, and before we proceed to the more elaborate levels of measurement, let us pose a practical question. Imagine that the manager of a toilet paper company is interested in the seasonality of its product (don't ask why) and thus wants to compare sales across the 12 months of the year. In this case, would you operationalize the variable 'month' as nominal or ordinal? If you have not already figured this out, make sure that you review the previous paragraphs and think carefully about whether the different levels of the 'month' variable can be objectively and meaningfully ordered.

An **interval scale** possesses all the characteristics of an ordinal scale (i.e., equivalence and order) and, in addition, is characterized by **equality of intervals** between adjacent scale values. In interval scaling, the numbers used permit inferences concerning the *extent of differences* that exist between the measured persons, objects, and so on with regard to a particular characteristic. The distances between the numbers correspond to the distances between the persons, objects, and so forth on the characteristic concerned. For example, when we measure temperature by the Fahrenheit scale, we can say that the difference between 70°F and 50°F is the same as the difference between 40°F and 20°F and that both are twice the difference between 10°F and 0°F. However, we *cannot* say that 50°F is 'five times as hot' as 10°F (i.e., suggest a 5:1 ratio). This is because the zero point on the Fahrenheit scale is arbitrary and does not reflect the 'true' zero of the underlying characteristic (i.e., absence of heat).

To make this point clear, let us convert the corresponding values into the Celsius scale (also an interval scale) using the famous formula: $C = (5F - 160)/9$ (very useful to know if you are a foreigner living in the United Kingdom or the United States). The corresponding temperatures for 50°F and 10°F are respectively 10°C and −12.2°C, which clearly do not stand in a 5:1 ratio. However, the ratios between the differences are still maintained (i.e., the difference between 21.1°C and 10°C is equal to the difference between 4.4°C and −6.7°C, and both are twice the difference between −12.2°C and −17.7°C). Since an interval scale has an **arbitrary zero point**, but the difference between any two adjacent scale points is equal to the difference between any other two adjacent scale points, we can assign any set of numbers to the scale so long as the intervality is preserved (as was done, for example, with the temperature scales above). In disgracefully incomprehensible technical terms, this means that any **positive linear**

transformation (i.e., of the form $y = a+bx$, $b>0$) will preserve the properties of the scale (note, in this context, that the conversion formula from Fahrenheit into Celsius is such a transformation). Our horizons for analysis are further expanded with interval scales, in that, in addition to calculating the mode and median, the **mean** (i.e., the arithmetic average) can also be computed from the data (as can various indicators of dispersion or variability – you will learn all about them in Chapter 7).

Finally, a **ratio scale** has all the features of an interval scale (i.e., equivalence, order, equality of intervals) plus an **absolute zero point** (also known as 'true' or 'natural' zero). In ratio scaling, the numbers assigned enable comparisons to be made between the measured persons, objects, and so on in terms of *absolute magnitude* on a given characteristic; equal ratios between the scale values correspond to equal ratios among the persons or objects concerned (which is not the case with an interval scale). For example, we can measure the speed of a motorcycle in kilometers or miles per hour. On either scale, a reading of zero actually corresponds to the absence of speed (i.e., the motorcycle is stationary). Moreover, we can say that driving at 240 kph is twice as fast as at 120 kph (a ratio of 2:1) or, converting into miles, that 150 mph is twice as fast as 75 mph (also a 2:1 ratio). The only legitimate transformation on a ratio scale is one that changes the unit of measurement. In emphatically unintelligible terminology, this means that the properties of the ratio scale are preserved up to a **positively proportionate transformation** (i.e., of the form $y = cx$, $c>0$); as you might have guessed, the conversion formula from kilometers to miles per hour is of this nature (since 1 mile = 1.609 kilometers, which is another useful formula for continental Europeans lost in the United Kingdom or the United States). In terms of analysis, ratio scales offer little additional advantage over interval scales as the vast majority of analytical methods are equally applicable to both scales. At the same time, while there *are* some statistical indicators that are only appropriate for ratio data, they are infrequently used in practice (and we will, therefore, unashamedly ignore them in the rest of this book).

From the above discussion (see also Table 3.2), it should be clear that the four types of measurement scales are *nested* within one another: as one progresses from a lower level of measurement to a higher one, the properties of the former are retained, and additional properties are gained (thus allowing for more refined measurement and more sophisticated analysis). This means that while one can always go back to a lower measurement level from a higher one (e.g., from ordinal to nominal), the reverse is *not* generally true: one cannot normally generate a higher level of measurement from a lower one (e.g., convert nominal data into interval). This asymmetry implies that the higher the level of measurement, the greater the resulting flexibility in terms of subsequent analysis. More specifically, with measures yielding **metric data** (i.e., interval or ratio variables), one can apply a set of analytical techniques known as **parametric statistics**, whereas with **non-metric**, **categorical data** (i.e., nominal and ordinal variables), only the less powerful **non-parametric** statistical techniques can be used (both types of techniques will be discussed from Chapter 7 onwards, so do not despair – yet!). Thus, given a choice, one should always opt for metric measures over non-metric ones.

HINT 3.2	Always aim for the highest measurement level possible. If feasible, obtain metric data from your measures.

Table 3.3 Measuring age

Please indicate your age by ticking the box applicable:	
Under 18 years	[ ]
18–24 "	[ ]
25–34 "	[ ]
35–49 "	[ ]
50–64 "	[ ]
65 plus "	[ ]

Some of you may feel uncomfortable with Hint 3.2 because some variables, by their very nature, offer little choice in terms of measurement level. In contrast to, say, weight or height, which are obviously **quantitative variables**, characteristics such as geographical location or race are obviously **qualitative variables** and can only be represented by nominal scales. Similarly, preferences for, say, six Uruguayan heavy-metal guitarists may only be amenable to a ranking procedure (e.g., 1st = Jose 'Animal' Martinez, 2nd = Juan 'Mad Reptile' Perez, …, etc.), precluding quantification of the distances/differences in creative talent between the 'artists'. However, while the nature of the characteristic under consideration is an important constraint on the discretion we have in measurement, in most circumstances we do have at least *some* choice. To illustrate this, have a look at Table 3.3, which shows a very common way of measuring the age of respondents in questionnaire studies (no doubt most of you will have come across this response format in one of the numerous silly forms you've had to complete at one time or another).

Being devastatingly perceptive, you will have noticed the wide variation in individual ages within each category (e.g., a 2-year-old still wetting his/her bed would be classified exactly the same as a 17-year-old about to enter university). You will also have noticed that there are unequal intervals across categories (e.g., compare the range of ages in the 18–24 category with that of the 35–49 category). Lastly, it would not have escaped your trained eye that there are open intervals (e.g., 65+), which means that we do not even know their true widths. All these problems are the direct result of the loss of information, which has occurred because the measurement opportunities have not been fully capitalized. Age is, by nature, a ratio variable, and, therefore, it could have been recorded by direct quantification (i.e., asking the respondent to state/write their age in years). What we, in fact, have ended up with is an ordinal measure of a ratio variable and, consequently, we cannot even calculate the average age of the respondents, let alone make statements such as 'Respondent X is twice as old as Respondent Y'. Put simply, we have missed an opportunity to obtain more meaningful information from the data.

Notice that, had we obtained the exact age of each respondent, we could (if we wished to) produce an age classification such as the one in Table 3.3 by appropriately grouping the individual ages (how this is best done will be discussed in Chapter 6). However, by virtue of our better measurement, we would also be in a position to comment on the average age of our sample, identify the oldest and youngest respondents, examine the proportion of respondents above/below any particular age, and so on.

Of course, if only to be awkward, you may argue that the real reason for using a measurement scheme such as the one in Table 3.3 is that respondents may be more willing to tick a box representing a range of ages than disclose their exact age (particularly in a face-to-face interview situation). This can certainly be true (as, for example, may be the case for a grandmother who is out with her toy-boy when intercepted by an interviewer) and can pose a response problem whenever 'sensitive' characteristics are to be measured (income, number of visits to disreputable night clubs in Amsterdam, tax bills, etc.). Under such circumstances, one may *consciously* opt for a lower level of measurement than would have been ideal. Nevertheless, one must think very carefully whether the expected gain in the response rate actually outweighs the loss of information due to poorer measurement. Moreover, in studies promising anonymous/confidential treatment of responses and/or conducted by means other than personal interviews (e.g., online surveys), there is much less reluctance to provide exact figures to open-ended questions, even for potentially sensitive variables. In short, Hint 3.2 stands firm, and you should keep it in mind.

Sometimes, you can improve upon the level of measurement by creating another variable that is related to the one on which you have data. To do this, you do not have to be the most creative person on earth (although this would help), but you definitely need an inquisitive mind and sufficient motivation (because there is additional work involved!). Imagine, for example, that you have just completed an international survey of mercenaries' attitudes toward disarmament in which, among other things, you have recorded the respondents' nationality on a nominal scale (e.g., 1 = Irish, 2 = Estonian, 3 = Peruvian, …, 47 = Greek, etc.). One thing you could do with these data is create another variable reflecting the level of economic development of each country concerned. This would involve allocating a new set of values to the various countries (e.g., 1 = post-industrial, 2 = industrial, 3 = developing, 4 = underdeveloped), and the resulting data could be treated as *ordinal* (reflecting successive levels of economic development). Of course, you would need to find out how, say, Estonia, should be classified, but an afternoon in your local library or a few quick Google searches (to consult appropriate secondary sources such as publications by the World Bank or the International Monetary Fund) should give you this information. Similar opportunities for generating new variables with a higher level of measurement exist for such characteristics as occupation (one can place individual occupations on an ordinal scale indicating occupational status or prestige, e.g., 1 = professional/managerial, 2 = skilled worker, 3 = unskilled worker) and education (one can fit individual qualifications according to educational level, e.g., 1 = postgraduate university education, 2 = undergraduate university education, …, 7 = part-time kindergarten only).

HINT 3.3	Investigate opportunities for improving upon the measurement level through the generation of new variables. Be creative!

A common problem encountered with scales measuring characteristics such as attitudes and opinions is that the level of measurement implied is not always unambiguously clear (indeed, sometimes it is unambiguously unclear). In most cases, this problem boils down to deciding whether the scale concerned yields only ordinal data or whether an assumption of intervality can be justified. Take, for instance, the scales majestically displayed in Table 3.4, all of which

Table 3.4 Dr. Igy's teaching performance

A. *Dr. Igy has been an excellent teacher.*

Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
[ ]	[ ]	[ ]	[ ]	[ ]

B. *Dr. Igy's examples have always been*

Excellent [ ]	[ ]	[ ]	[ ]	[ ]	[ ] Appalling

C. *When answering questions, Dr. Igy has been*

Very competent	Somewhat competent	Neither competent nor incompetent	Somewhat incompetent	Very incompetent
[ ]	[ ]	[ ]	[ ]	[ ]

D. *Pedagogically, Dr. Igy has been*

		+3		
		+2		
		+1		
		competent		
		−1		
		−2		
		−3		

attempt to measure students' perceptions of the previously mentioned statistics Prof. Dr. Dr. Ignatz Zweistein (you can now call him Dr. Igy because you already know each other).

The scales in Table 3.4 are all examples of the most widely used attitude-measurement techniques in social and marketing research (A is a **Likert scale**, B is an example of the **semantic differential**, C is an **itemized rating scale**, and D is a **Stapel scale**). Are these scales interval or ordinal? Unfortunately, to answer this question properly, we would have to spend at least three years delving into the relevant (and horrendously extensive) literature. Since we feel that you would not want us to do this, we shall settle this matter by giving you a choice. If you adopt the 'pragmatic' view followed by most social researchers, then you should treat the scales in Table 3.4 (or similar ones) *as if* they were interval. In this context, it is recommended to appropriately number the response alternatives on the scale so as to communicate to the respondent that the intervals between the scale points are intended to be of equal distance. On the other hand, if you adopt the 'purist' view most commonly followed by statisticians, then the scales of Table 3.4 should be treated as ordinal (unless you can *prove* otherwise).

Just so you don't get carried away, if you opt for the pragmatic view, you must be alert to the possibility of grossly unequal intervals and be cautious with the interpretation of data assumed to be interval. Moreover, you will *not* get away with it if you adopt the pragmatic view of scales that are patently ordinal. For example, if a respondent is asked to compare brand A of dog shampoo with brand B, on a scale reading 1 = worse, 2 = about the same, 3 = better, the resulting data cannot be reasonably treated as interval (so don't try tricks like that, or you'll be in trouble!).

At this point, we must deal briefly with a special kind of measurement situation, namely that involving **dichotomous variables**. A dichotomous (or 'binary') variable takes only two values (e.g., male/female, user/non-user, pass/fail) and is typically scored via **dummy-variable coding**; that is, by assigning the value of 1 to one category (usually to mark the presence of some characteristic) and 0 to the other category (to mark its absence). At first glance, dichotomous variables appear to reflect a nominal scale; for example, scaling gender with just two unordered categories (male/female) implies a nominal level of measurement. However, a dichotomous variable can also be viewed as reflecting both an order and equal intervals (see Table 3.2). Concerning the former, although a 'natural' rank order may not be inherent in the category definitions, either arrangement of the categories satisfies the mathematical properties of ordering. It makes no difference whatsoever which end of the ranking is considered 'high' or 'low' (thus '1 = female, 0 = male' and '1 = male, 0 = female' are equally acceptable). Moreover, the requirement of equally sized intervals is also satisfied because there is only one interval (which is naturally equal to itself). In short, dichotomous variables can be treated as nominal, ordinal, or even interval, depending upon the research situation – isn't this great?

WARNING 3.3 Think very carefully about whether the scale you are using can be assumed to be interval for analysis purposes. Think even more carefully about how you will interpret the results.

As an aside, note that using a dichotomous variable is not the only option for measuring gender. A more detailed gender classification can, for example, be found on Facebook, listing some 70+ gender types (including 'neither' and 'other').

SCALING FORMATS

Let us attempt to confuse you even further by having a quick look at the bewildering variety of scaling techniques at your disposal. Table 3.5 summarizes the main **scaling formats** you are most likely to encounter in data analysis projects.

As you can see, in many cases there is considerable scope for 'customizing' or 'tailoring' the final form of your scale by manipulating one or more of its basic features. To illustrate the opportunities (read: complexities) involved, consider the direct rating scale. Among the decisions you will have to make in using such a scale are (1) whether to provide an *odd versus even* number of response alternatives (i.e., whether to include a middle/neutral point), (2) whether to use a *balanced versus unbalanced* distribution of response alternatives (i.e., whether to provide an equal number of favorable/positive and unfavorable/negative scale points), (3) whether to use a *forced versus unforced* response format (i.e., whether to provide a 'no opinion', 'not applicable', or 'no knowledge' category), (4) whether to label *all versus some* of the scale points (i.e., whether to provide a verbal description of each response alternative or label only the extremes and/or middle point on the scale), and (5) *how many* response alternatives (i.e., scale points) to include. Different research problems, respondent groups, and characteristics to be measured all affect the precise form of the scaling format finally chosen. While space limitations preclude a discussion of the individual impact of these factors on (1)–(5) above, the

incredibly stimulating list of references we have generously provided (see Further Reading) contains comprehensive treatments of these issues.

MEASUREMENT ERROR

Before we say goodbye to measurement, we need to raise a very important issue that concerns the **quality of measurement**. We hate to break the news to you, but we can hardly have a perfect measurement as there is always a certain amount of deviation or error involved. In an ideal world (which is like an ideal boyfriend or girlfriend, i.e., it does not exist in reality), every time we measure something we could be sure that the score (i.e., value) obtained actually reflects the 'true' value of the underlying characteristic *and nothing else*. In other words, there would be a perfect one-to-one correspondence between the **observed score** (O) (i.e., our measurement) and the **true score** (T) (i.e., the reflection of the characteristic of interest) and, consequently, **measurement error** (E) would be zero.

Unfortunately, since we do not live in a perfect world (see above), measurement error is not normally zero and, therefore, we cannot take for granted that our measures, in fact, do a good job. Figure 3.1 shows our predicament by highlighting the different components of measurements and the sources of measurement error.

What Figure 3.1 tells us is that, given half a chance, measurement error will creep in whenever we are administering a measurement instrument. Respondent characteristics, both stable (e.g., respondent likes to disagree 'just for the hell of it') and temporary (e.g., respondent is in a bad mood), together with imperfections of the measurement instrument (e.g., lack of a 'don't know' alternative on a scale presented to an undecided respondent) and problems of the measurement situation (e.g., the respondent fancies the interviewer while their extremely jealous husband/wife is present), all combine to make our life difficult by introducing error in our measurements.

Two types of error can be distinguished: **systematic error** (S) occurs in a consistent manner every time a measurement is taken (e.g., a general tendency to respond negatively independently of one's true feelings, as in the case of the 'professional disagreer' above). Thus, systematic error results in either inflation (overestimation) or attenuation (underestimation) of the true score. **Random error** (R) (also known as 'variable' or 'unsystematic' error), on the other hand, does not manifest itself consistently every time a measurement is taken (e.g., a temporary characteristic such as mood may be reflected in overly favorable responses if the respondent is in a good mood and in excessively negative responses if the respondent is in a foul mood). Thus, random error tends to be self-compensating since it can occur in either direction.

The extent to which a particular measure is free from both systematic and random error indicates the **validity** of the measure; a perfectly valid measure implies $O = T$ (i.e., the measurement obtained reflects only the true score on the characteristic of interest). The extent to which a measure is free from random error indicates the **reliability** of the measure; a perfectly reliable measure implies $R = 0$ (i.e., no random error component).

Table 3.5 Different kinds of scaling formats

Format	Basic form	Description	Example
Checklist	Choose *m* out of *n*	Respondent is given *n* options and asked to select up to *m* ($m \leq n$).	*Which of the following items do you own?* 1. A car 2. A cattle prod 3. A yellow tea-cozy 4. A pair of purple wellies (i.e., rain boots for our US readers)
Determinant choice	Choose 1 out of *n*	Respondent is given *n* options and asked to select one of them.	*Which characteristics are important for the next president?* 1. Exceedingly good-looking 2. Exceedingly hedonistic 3. Exceedingly old
Dichotomy	Yes/no	Respondent is asked to respond with a 'yes/no' answer.	*Have you ever visited Siberia?* 1. Yes 2. No
Rank order	Rank *n*	Respondent is given *n* options and asked to rank them in ascending or descending order.	*Please rank the following chocolate bars according to taste (1 = best, 4 = worst)* Supernut ——————— Thunderslice ——————— Almondchunk ——————— Crumbleflake ———————
Constant sum	Allocate *n* to *m*	Respondent is given *m* attributes and asked to allocate *n* points (usually 100) to them.	*Please allocate 100 points to the following attributes according to their importance in your choice of pub.* Beer ————— Barmaid ————— Clientele ————— Opportunity for fights ————— Total: 100
Paired comparison	Choose X or Y	Respondent is given two objects (X, Y) and asked to select one according to the characteristic in question.	*Which of the following brands of soap would you choose if both cost £3.95 per bar?* 1. Moroccan Sunset 2. Andalusian Mist
Comparative rating	Rate X against Y on *n*	Respondent is given a rating scale with *n* points and asked to compare object X against Y by selecting one point on the scale.	*In terms of speed, how does the new Fjord Siesta compare with the new WV Molf?* 5. Much faster 4. Somewhat faster 3. About the same 2. Somewhat slower 1. Much slower

Format	Basic form	Description	Example
Direct rating	Rate X on *n*	Respondent is given a scale with *n* points and asked to rate object X by selecting one point on the scale.	*Having just driven the new Fjord Siesta, how would you rate its windscreen wiper speed?* 5. Extremely fast 4. Very fast 3. Average 2. Very slow 1. Extremely slow
Direct quantification	Give a figure	Respondent is asked to state (or write in) a fact that can be expressed as a number (integer or real).	*On average, how many liters of Czech lager do you drink per week?* liters

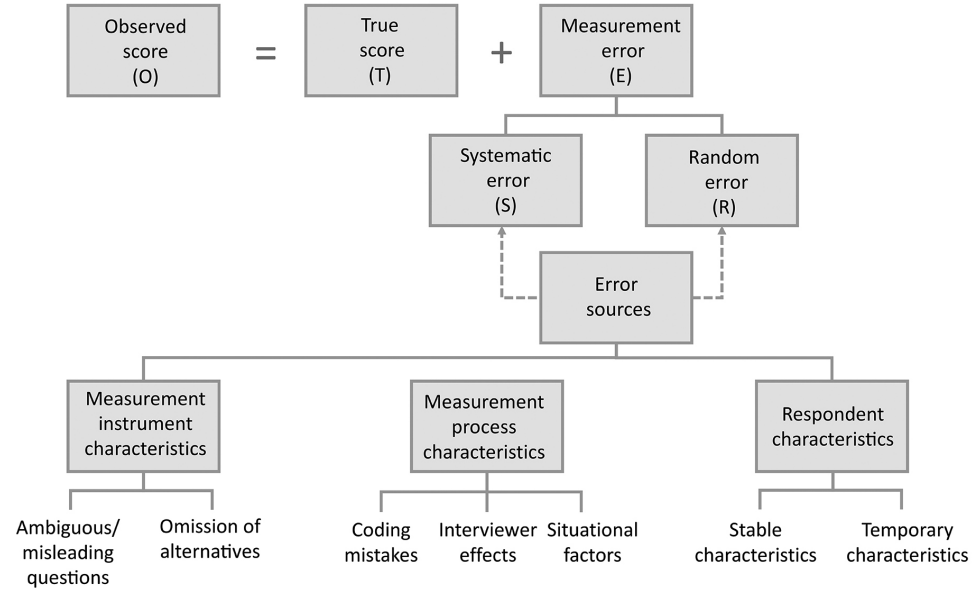


Figure 3.1 Measurement in the real world

Let us demonstrate the central ideas relating to measurement error using an unparalleled example. Imagine that you want to measure your friends' weight and, because of your living-on-the-edge nature, you play a prank on them by tampering with the bathroom scales used as a measurement instrument so that they show 3 additional kilos in every measurement. This will obviously decrease the validity of your measurement ($O > T$) by introducing a systematic deviation of +3 kg throughout (i.e., your scale reliably overestimates weight). In an alternative scenario, you prevent yourself from engaging in monkey business, but it just so happens that the bathroom scales you use are faulty. As a result, the measurements you obtain could be either inflating or deflating your friends' actual weight (e.g., +1 kg, −3 kg, +2.5 kg, −1.5 kg, etc.). In this case, validity is obviously damaged ($O \neq T$) because of measurement

error. However, because of the unsystematic nature of this error (sometimes overestimating and other times underestimating the actual weight), reliability would also suffer.

In principle, reliability is a necessary but not sufficient condition for validity (since, even if $R = 0$, the measurement obtained could still contain systematic error in addition to the true score, i.e., $O = T+S$). If a measure is not reliable, then it cannot be valid, but if it is reliable, it may or may not be valid; put differently, a measure that is valid is also reliable, but the reverse is not necessarily true. Confused? Don't worry, so are we! However, the famous dartboard analogy should help you set things straight (hopefully!). Every dartboard in Figure 3.2 corresponds to different measurements intended to capture the true score, the latter represented by the inner circle of the dartboard. In Panel A, things are not great because your measures fail to hit the target in a consistent way. Such a measurement scheme is reliable but not valid, therefore you can safely rely on it to consistently make the same mistake (ain't that great?). Panels B and C are even worse as now not only do the measurements fail to hit the target, but they also do so in an unsystematic, random fashion. Finally, in Panel D, you consistently hit the bullseye, thus your measure is both reliable and valid.

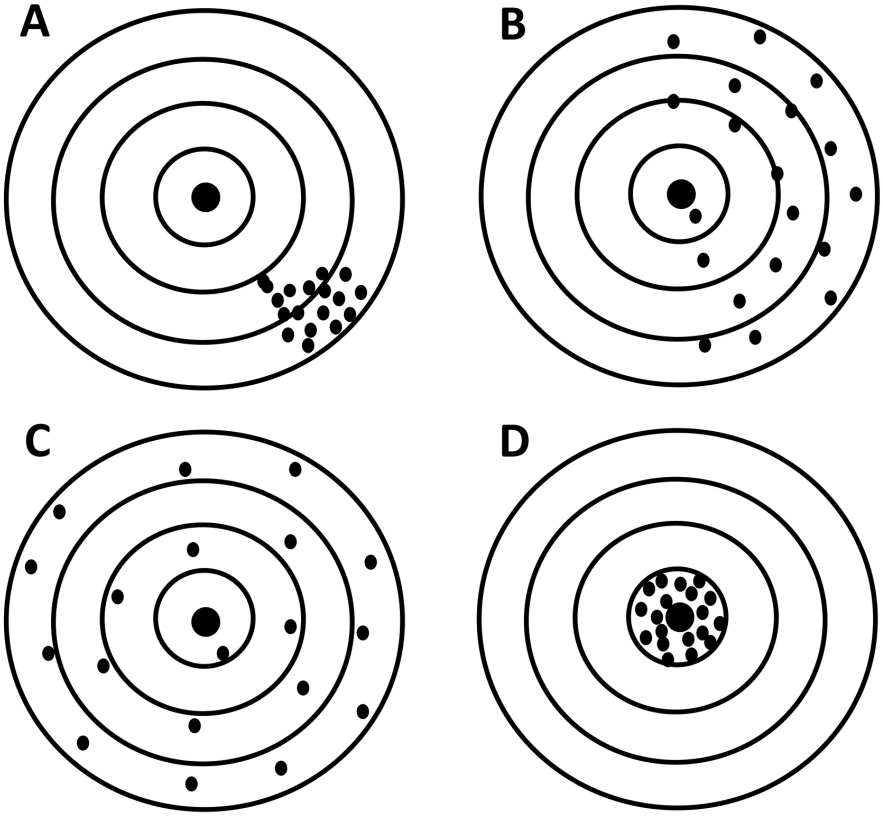


Figure 3.2 Reliability and validity

ASSESSING VALIDITY AND RELIABILITY

A rather annoying detail in any practical attempt to assess the validity or reliability of a measure is that we do not *know* the true score of the respondent on the characteristic concerned (if we did, there would be no point in measuring it, and we could all go home and sleep). Therefore, we must somehow try to gather some sort of evidence that will enable us to *infer* the extent to which our instrument is valid and reliable; yes, this does involve even more work, but that's science for you!

In validity assessment, the basic question that we try (usually unsuccessfully!) to answer is: 'Are we, in fact, measuring what we think we are measuring?'. Thus, our intention is to show that our measurement device does, in fact, measure what it is intended to measure. You may wonder why we are asking what appears to be a rather trivial question. For example, if we employ a ruler to measure our desk, it is pretty obvious that what we are measuring is size (i.e., length/width/height) and not sex appeal or religious beliefs. This may be so, but what if we are using a certain scale to measure 'citizen satisfaction' (a feeling) with 'police services' (an intangible and multifaceted object of evaluation)? Is our scale really a satisfaction scale or is it capturing other properties such as 'discontent', 'alienation', or 'law-abidingness' (got you there, eh?). Even with measures that are 'obviously' measuring 'obvious' things (such as with our ruler example above), problems of validity may still crop up; thus, a poorly calibrated ruler will provide very precise but wholly inaccurate (and, hence, invalid) measurements owing to the presence of systematic error. Thus, attention to establishing the validity of one's measures is good research practice and time well spent. The bad news is that validity assessment can be quite complex, as there are a number of different angles one can take (see Table 3.6). The *really* bad news is that it is the collective picture painted by evidence relating to the various kinds of validity that determines the overall validity of a measure. Using our standard excuse of 'space limitations', we will not get into the nitty-gritty of validity assessment procedures but direct you, once again, to the Further Reading section for some excellent guides to the steps involved. Shifting attention to the assessment of reliability, things are not as complicated as with validity assessment (but, then again, they aren't simple either!). The key question we are concerned with here is: 'Are we getting consistent results from our measures?' There are two types of consistency we are particularly interested in. The first is consistency over time; that is, the extent to which we get similar results from repeated applications of the same measurement instrument to the same set of respondents. This is known as the **stability** aspect of reliability. The second type of consistency is known as **equivalence** and indicates the extent to which the same set of respondents replies in a consistent manner on similar items; alternatively, equivalence can be seen as the extent to which different (but comparable) sets of respondents produce similar results on the same measurement instrument. Table 3.7 summarizes the key methods for assessing reliability in terms of stability and equivalence; detailed guidelines on how to go about it can be found (yes, you've guessed it) in the Further Reading section.

We do have some good news for you (it is about time, isn't it?). Some established statistical criteria for assessing reliability can easily be obtained from available software packages. The most common one is **Cronbach's alpha (α) coefficient**. Cronbach's α quantifies the internal consistency of a set of items that supposedly measure a single construct (*unidimensionality*

Table 3.6 Validity assessment approaches

Approach	Description	Procedure
1. Content validity	The extent to which a measure *appears* to measure the characteristic it is supposed to measure.	Subjective assessment of the appropriateness of the measure for the task at hand.
1.1 Face validity	The extent to which a measure seems to capture the characteristic of interest.	Agreement between expert and/or non-expert judges as to the suitability of the measure.
1.2 Sampling validity	The extent to which a 'content population' of situations/behaviors relating to the characteristic of interest (i.e., the characteristic's conceptual domain) is adequately represented by the measure concerned.	As above.
2. Criterion validity[a]	The extent to which a measure can be used to predict an individual's score on some other characteristic (the criterion).	Examination of the relationship between the measure and a criterion.
2.1 Concurrent validity	The extent to which a measure is related to another measure (the criterion) when both are measured at the *same* point in time.	Comparison of the scores obtained on the measure concerned and those obtained on the criterion.
2.2 Predictive validity	The extent to which *current* scores on a given measure can predict *future* scores of another measure (the criterion).	As above.
3. Construct validity	The extent to which a measure behaves in a theoretically sound manner.	Investigation of the relationships between the measure of interest and measures of other concepts/characteristics within a theoretical framework.
3.1 Convergent validity	The extent to which a measure is positively related to other measures of the same concept obtained by independent methods.	Examination of the relationships between measures of the same concept generated by different methods.
3.2 Discriminant validity	The extent to which a measure is not related to measures of different concepts with which no theoretical relationships are expected.	Examination of the relationships between measures of different concepts that are theoretically distinct.
3.3 Nomological validity	The extent to which a measure is related to measures of other concepts in a manner consistent with theoretical expectations.	Examination of the relationships between measures of different concepts that are theoretically related.

Note: [a] Also known as pragmatic o" empirical validity.

is thus an important underlying assumption regarding the meaningfulness of α) by means of a number that ranges between 0 and 1. The idea is that if a set of items reliably measures the same construct, then the items' values should be closely related. Very loosely put, Cronbach's α looks at the average inter-correlation among the items: the more items relate to one another, the higher the coefficient's value and the more reliable the multi-item measure (i.e., you can safely assume that summing or averaging the individual items into one composite scale will reliably capture the construct of interest). Note, however, that α also depends on the number of items in the scale, with more items resulting in higher values. Thus, although the literature

Table 3.7 Reliability assessment procedures

Approach	Description	Procedure
1. Test–retest reliability	Consistency of results of repeated applications to the same measure to the same respondents.	Assessment of the degree of correspondence between two (or more) applications of the same measure under similar conditions.
2. Alternative forms reliability	Consistency of results of applications of 'equivalent' forms of the measure to the same respondents.	Assessment of the degree of correspondence between two equivalent measures administered under similar conditions.
3. Split-sample reliability	Consistency of the results of applications of the same measure across randomly selected sub-samples of respondents.	Assessment of the degree of correspondence between random sub-samples of respondents (usually split 50:50) on the same measure.
4. Internal consistency reliability	Consistency of the results across individual items comprising a composite scale.	Assessment of the degree of consistency within a multi-item measure administered to a group of respondents.
5. Scorer reliability	Consistency of the results provided by different judges or scores when asked to categorize open-ended responses.	Assessment of the degree of agreement between independent categorizations of a set of items by multiple judges.

suggests that α values should be 0.70 and above for a multi-item scale to be reliable, you should not forget that, *ceteris paribus*, a longer scale will produce a higher alpha. Most statistical packages will calculate Cronbach's α coefficient for you and will also show you how it will change if an item is removed from the scale.

As if the notions of validity and reliability were not in themselves complicated enough to drive any sane person insane, two additional dimensions are sometimes referred to in discussions of measurement quality. **Sensitivity** refers to the extent to which a particular measure is able to capture variability in responses and is a particularly desirable property when *changes* in the characteristic of interest are measured. For example, a dichotomous scale such as '1 = agree, 0 = disagree' is unlikely to capture subtle attitude changes to the same extent as a five-point scale anchored at '5 = strongly agree, 1 = strongly disagree', and also containing a middle point. **Generalizability** refers to the extent to which a scale is applicable and interpretable in different research settings. For example, is a scale intended to measure 'hedonism' easy to employ with different data-collection methods (e.g., online questionnaire vs. telephone interviews) equally applicable to different respondent groups (e.g., males vs. females) and readily interpretable in different situations (e.g., in single-country vs. cross-cultural investigations)? In one sense, sensitivity and generalizability can be seen as sub-dimensions of validity and reliability. Thus, if a scale developed in setting X cannot be applied to setting Y (i.e., is not generalizable), this is because *for setting Y* the scale has low reliability and/or low validity. Equally, a scale exhibiting low sensitivity is unlikely to produce impressive results in reliability or validity assessments.

SUMMARY

In this mind-blowing chapter, we looked at a variety of issues relating to measurement and scaling. We began by considering the notion of measurement and the rationale behind the measurement process. We then distinguished between conceptual and operational definitions and followed this with a discussion of the different types of measurement scales. Next, we tried to make some sense out of the various scaling formats and called it a day

after taking a good look at measurement error, emphasizing the notions of measurement validity and reliability. If you need a drink, now is the time to get it.

QUESTIONS AND PROBLEMS

1. Why do we need conceptual definitions?
2. What is the role of numbers in measurement?
3. Describe the different levels of measurement and give one example of a measurement scale for each.
4. What is the difference between an interval and a ratio scale?
5. Why should you generally aim for the highest level of measurement possible?
6. What is so special about dichotomous variables?
7. What are the key factors determining the final format of a rating scale?
8. Which is more important in measurement: validity or reliability?
9. Give three examples of sources of measurement error; then, give three more!
10. Are we correct in assuming that a direct rating scale anchored at 1 = 'outstanding' and 5 = 'marvelous' will accurately capture your evaluation of this chapter?

FURTHER READING

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and Validity Assessment*. London: Sage Publications. A classic and very readable introduction to different approaches to evaluating the validity and reliability of measures.

DeVellis, R. F. (2016). *Scale Development: Theory and Applications*, 4th edition. London: Sage Publications. An easy-to-follow text on how to go about developing a scale from scratch.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling Procedures*. London: Sage Publications. Read this after DeVellis (2016) to get maximum benefit as it is a bit more advanced.

PART II
PREPARING DATA FOR ANALYSIS

4
Have you cleaned your data and found the *mistakes* you made?

THE ROLE OF DATA CLEANING

In Chapter 1, we pointed out that the raw material for analysis is always a data matrix, the rows of which reflect the units of analysis and the columns the variables on which the units of analysis are measured. In Chapters 2 and 3 we elaborated on the basic structure of the data matrix by discussing sampling issues (affecting its rows) and measurement issues (affecting its columns). Figure 4.1 summarizes what we have brilliantly covered so far and also indicates the next two stages involved, notably (a) preparing the data for analysis by means of editing and coding and (b) transforming the data into *results*.

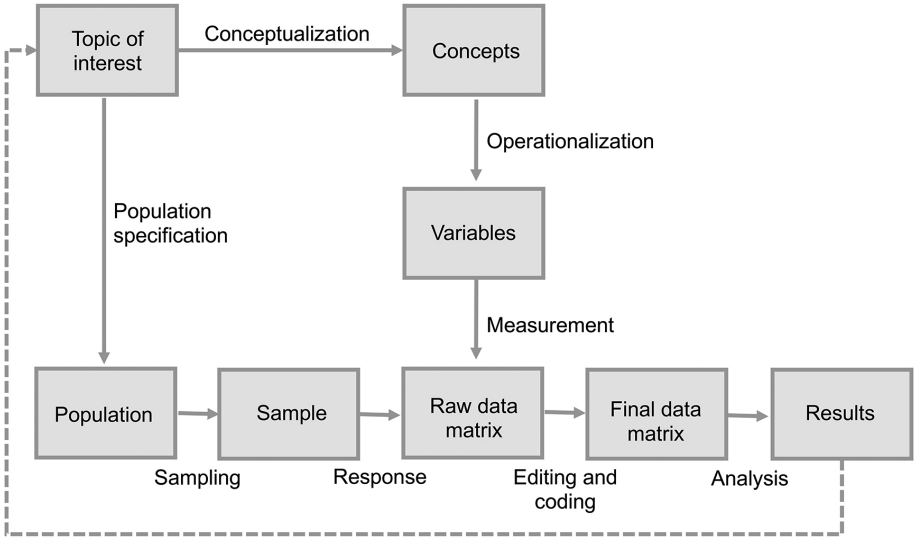


Figure 4.1 The long journey to producing results

In this chapter, we will deal with the preparation of data for analysis; that is, how one goes about 'cleaning' and coding data prior to analysis.

To avoid a possible misunderstanding right away, **data cleaning** does not mean you have to wash your completed questionnaires, take showers with your data matrix, or make sure that no obscene words slip into the responses to your open-ended questions! Instead, the objective is to find omissions, ambiguities, and mistakes in the responses. Thus, data cleaning refers to the preparatory work necessary to avoid errors in the data matrix during and immediately after collecting your data. Usually, this process is referred to as **editing** (see Figure 4.2) and will be discussed right away (we thought to emphasize 'right away', as we know how much you are looking forward to this!).

HINT 4.1	Always assign a respondent (questionnaire) number and use an interviewer code where a number of interviewers are involved in data collection.

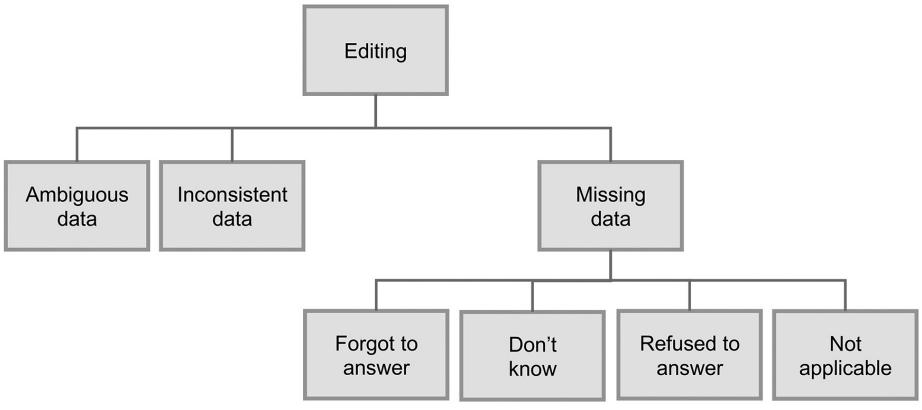


Figure 4.2 Main editing tasks

An important part of **data editing** involves keeping an eye on interviewers through monitoring and validation procedures, the basic motto being 'trust is good; control is better'! Thus, if you are in the lucky situation to have people doing the data collection for you, you may want to re-contact a sample of respondents to ensure that the interviews actually took place, that respondents were completely free to express their opinion, that there are no incomplete data in the questionnaire, that each respondent only participated once, and that the answers were not made up by the interviewer while enjoying a bi-weekly visit to the chiropodist. To apportion appropriate blame, it would be necessary to record the **interviewer number** on each questionnaire and assign a **respondent number**. This procedure should enable you to go back to the appropriate interviewer and/or respondent in case you have missing, confusing, or contradictory responses or simply if respondents' handwriting is unintelligible (this may happen if a respondent completes your open-ended question in Hindi and yours is a bit rusty these days). Online tools and computer-based surveying help you overcome most of these

shortcomings – within the confines of data protection laws – by informing respondents about unanswered questions and inappropriate responses (e.g., wrong language), incorporating unique respondent identifiers (through IP addresses, Internet cookies, etc.), and keeping interviewer–respondent interactions to a bare minimum. However, things might still go wrong, and you will have to deal with ambiguous, inconsistent, or missing data *without* (usually) being able to clarify the problem by contacting the relevant respondent. Consider the question, 'When did you purchase your last car?', which you asked in an online survey focusing on consumers' car-buying habits in Yorkshire, England. During data editing, you may find that the majority of respondents put the year (e.g., 2018) in the space provided. Some, however, may have answered the question in terms of months (e.g., 3 months ago), and one respondent from Huddersfield West answered, 'When my third wife walked out on me and took my old car.' This single question alone would require arduous editing of the responses (e.g., recoding, transformation) to be cleaned, and still, the problem with people like the bloke from Huddersfield West would remain. What, if anything, can we learn from this (apart from the fact that you have to be extra careful with giving a second car key to your partner)?

Firstly, you can reduce the amount of editing work due to **ambiguous answers** by not asking ambiguous *questions*. For example, putting 'months' in brackets would have helped with our Huddersfield West respondent.

Secondly, you need to guard against creating confusion by analyzing responses that may be ambiguous. For example, taking the average of values partly relating to months and partly to years is unlikely to result in anything useful.

Thirdly, you often need to make a pretty drastic decision about whether (a) the entire questionnaire ought to be thrown out (for example, if the respondent from Huddersfield West had answered the majority of questions in a similarly uninformative manner) or (b) a particular question should be ignored in further analysis (such as in our example question above).

Related to the problem of ambiguous answers is that of **logically inconsistent data**. An extreme case would be a male respondent who proceeds to answer a series of questions relating to problems encountered during his last pregnancy! A more common example is with respondents who might give a negative answer to the question, 'Are you aware of the soft-drink brand Peece?', yet move on to providing their ratings on how wonderfully Peece tastes. In such cases, you have to make certain that these data are not included in the analysis. However, you have to be careful with *seemingly* inconsistent data that are actually feasible. Consider a taste test between three new desserts: Slimmo-Stop, Obese-Plus, and Gut-Plug. A respondent may prefer Slimmo-Stop compared to Obese-Plus and Obese-Plus compared to Gut-Plug. However, the same respondent may still prefer the taste of Gut-Plug compared to that of Slimmo-Stop. Thus, in general, responses of the nature A > B, B > C, and C > A are sometimes feasible and are known as **intransitive responses**.

WARNING 4.1 Apparently, logically inconsistent data (intransitive responses) should not always be discarded as response errors, especially in taste comparisons and importance rankings; they are possible.

Perhaps the most common problem encountered when questionnaires are returned is that of **missing data**, which enables us to introduce yet another vital piece of jargon (useful as a conversation stopper when talking to your elderly Aunt Gwen from mid-Wales). The problem is known as **item non-response** and refers to specific questions that have been left unanswered. Note that, in online questionnaires, you can often avoid or reduce item non-response by ensuring that respondents cannot proceed to the next part of the questionnaire unless they have answered *all* questions on the screen. However, if you hate technology (like your Aunt Gwen does) and collected your data with, say, mail questionnaires, you must make up your mind *why* certain questions have not been answered and, if anything, what to do about it. Usually, there are four possibilities why a question has not been answered:

(a) The question did not apply to the respondent. This may occur when the respondent is a left-handed tea drinker whereas the branching instructions in the questionnaire only required right-handed coffee drinkers to answer a particular set of questions.
(b) The respondent refused to answer the question. For example, refusals are to be expected when clergymen are asked whether they harbor any 'impure desires' relating to professional rugby players.
(c) The respondent did not know the answer to the question. For example, most respondents would not know the cholesterol level of their dog.
(d) The respondent simply forgot to answer the question.

Remember that it is up to you to decide how best to deal with unanswered questions. In this context, you should consider whether it is possible (and, indeed, necessary) to distinguish between the above four reasons for item non-response. Suppose such distinctions can be achieved and are important for your intended analysis. In that case, you might want to use *different* **missing values** (also known as 'missing data codes') for 'not applicable' (NA), 'refused' (RF), 'did not know' (DK), and 'forgot to answer' (FA).

HINT 4.2 When it is not required to distinguish between different reasons for item non-response, define only one missing value. However, make absolutely sure that this value lies outside the range of possible answers.

Table 4.1 Measuring consumer reactions to a new chocolate spread

Having tasted our fantastic new chocolate spread, which one of the following words describes the taste of our excellent product most accurately?	
ecstatic	☐
outstanding	☐
exuberant	☐
thrilling	☐
mind-blowing	☐

Consider the example in Table 4.1 in which consumer perceptions to a new chocolate spread are measured. Say that some respondents ticked none of the response alternatives. If the question was supposed to be answered by all respondents, you may decide to interpret this as 'forgot to answer'. However, if the data you are editing are based on an interview with a respondent who, despite having received 27 complimentary packages, still could not bring himself to try your new chocolate spread, you may treat this as 'don't know'. Alternatively, a respondent may have scribbled on the questionnaire that she finds the chocolate spread truly repulsive and does not think the provided alternatives are fair; clearly a case of 'refused'! Finally, provided only teenagers were supposed to answer the question and you are editing the response of a grandfather who, correctly, skipped the question, you may decide to treat this as 'not applicable'.

In situations where it is impossible to distinguish between any of these alternatives (or simply where it doesn't really matter for the purpose of your research), you can assign a single missing value to all four cases (indicating 'no answer' for whatever reason). Regardless of the outcome of your deliberations, it is important that missing responses are ascribed with a value that lies outside the expected range of legitimate answers (we shall come back to this point later in this chapter).

Notwithstanding the reasons why some questions were unanswered, there are two main strategies for dealing with missing data. First, you may simply leave the answer blank or explicitly code 'empty cells' with a unique dedicated value. Statistical packages often include a data transformation function that automatically finds and replaces **user-** and **system-missing values** with whatever value you decide (we will say more about coding missing data later on in this chapter). You should, though, be aware that missing data will decrease your sample observations and might also create sample size imbalances (if, let's say, most missing data come from your male respondents). The same problem would occur if one decided to exclude altogether respondents who produced missing data.

Second, you may decide that it is safe enough to plug in an actual value where no answer is given. For example, missing data on a question might be replaced with the most frequent answer given by the other respondents or a value that is perceived to be 'very likely' or 'typical'. However, while this procedure has the advantage of not reducing the sample size as a result of item non-response, you have to be aware that, in essence, you are 'fabricating' your data! The literature has investigated a number of statistically justified ways of replacing missing data. Statisticians call the process of replacing missing values with substitute values **data**

imputation (just so you know, using a fancy term immediately increases the credibility of whatever it is that you are talking about). The most common techniques of data imputation include substitution with a randomly selected value with the overall mean value or by using other variables in the data set to build a model that can effectively predict the observed scores and, thus, be used as a basis for imputing missing values. Whether one should embark on an imputation endeavor and which approach to use is (once again) a complicated matter that pretty much depends on the specific situation at hand. The problem is exacerbated when a large proportion of answers to a particular question is missing. To labor the point: when 80% of the respondents do not answer a question, it is obviously unacceptable to replace the missing data with, let's say, the average of the remaining 20% who answered the question. So think *very* carefully before you engage in any strategy of 'replacing' missing data. Needless to say, drawing conclusions based on the responses of the 20% who *did* answer the question is not advisable either. You can never win, can you?

WARNING 4.2 Do not replace an excessive amount of missing data with averages. It is both misleading and unethical.

THE ROLE OF DATA CODING

Having carefully edited your data and corrected the many errors you have found, you should start thinking about **data coding**. This is something that most people leave to the last moment because, let's face it, coding questionnaires is not the most exhilarating task one can imagine (it only just beats sitting through a statistics lecture in entertainment value). However, coding is rather important in that mistakes made here might invalidate your results and might be difficult to detect at a later stage. In what follows, you will essentially learn two things: firstly, how to transform a bunch of completed (and edited) questionnaires into a coding system, which can be understood by any common statistical software package, and secondly, how to find mistakes invariably made during data input.

To prepare for the transformation of answers into a computer-readable format, it is useful to imagine the final product first. A woman must have a vision (and, of course, a man, too)! After your splendid transformation efforts, you should end up with a data matrix like the one shown in Table 4.2. This is actually how the '**Data View**' window in SPSS looks, and this is where you enter the responses to your questionnaires. If you have collected the data using an online tool, you will be able to automatically get a data matrix like the one in Table 4.2, which you can directly import to SPSS or similar programs.

Table 4.2 Example of a data matrix ('Data View' window in SPSS)

VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10	VAR11	VAR12
001	2	0	4	Totally useless	2	5	1	4	2,000	1	1
002	4	0	3	Should be a must	3	3	2	4	2,500	2	1
003	3	1	5	Too expensive	10	5	2	3	2,800	2	1
004	1	0	1	We should all walk	–999	2	4	5	3,100	2	0
005	2	1	2	I never knew how to spell this	5	2	4	5	1,320	0	1
006	2	1	2	My neighbor has one	8	3	3	3	–999	3	1
007	4	1	4	–999	12	1	–999	4	1,500	0	1

Looking at Table 4.2, it becomes apparent that a data matrix is rather useless unless we can establish a link to the answers in the questionnaire. For this, we strongly recommend that you set up a **code book** (sometimes also called a 'coding plan'). A code book describes how the responses relate to variables, labels the variables, shows whether a particular variable is **numeric** (i.e., a number) or an **alphanumeric** (i.e., a string of letters), gives the length of the variable specified (in terms of column width), and so forth. For example, the code book for the data matrix in Table 4.2 could look like the one shown in Table 4.3. While the 'Data View' in SPSS is used for data entry, the 'Variable View' window is designed for variable coding and serves as a code book. Note that every row in the '**Variable View**' represents a different variable and corresponds to a relevant column in the 'Data View' window.

Using the proper variable coding, the numbers in the data matrix (Table 4.2) can now be interpreted. We now know that the numbers in the first column merely identify a particular case; in our example, a questionnaire completed by a certain respondent. We can also see that the values in column 3 relate to the variable Gender and that 0 has been assigned to males and 1 to females. But before you become too bushy-tailed and think you have cracked this one, let's have a closer look at the elements typically involved in coding variables.

Variable name. The **variable name** is simply a name given to each of the variables in the data set so that the computer can reference them during the analysis. Note that some statistical packages do not really fancy variable names longer than eight digits. (We don't know why but it is a sad reality.) So if you would like to call a variable describing respondents' preferred winter holiday destination MOSTPREFERREDWINTERHOLIDAYDESTINATION, you may have to shorten it a bit (i.e., call it something like WINTDEST). Also, some programs do not permit you to use certain reserved words, usually those that have a particular meaning in 'computerspeak', such as AND, ALL, NOT, GE, LT, or WITH. Further, you need to make sure that variable names are *unique* (i.e., appear only once); you should not give the same variable

name to two (or more) variables as you are likely to create major havoc (computers are like humans – they can easily be confused!).

Table 4.3 Example of a code book ('Variable View' window in SPSS)

Name	Type	Label	Values	Missing
CASENO	Numeric	Case number	–	–
COUNTRY	Numeric	Country of origin	Togo = 1 USA = 2 Greenland = 3 Wonderland = 4	–999
GENDER	Numeric	Gender of respondent	Male = 0 Female = 1	–999
ATTENV1	Numeric	Attitude toward the environment	Strongly negative = 1 Somewhat negative = 2 Neutral = 3 Somewhat positive = 4 Strongly positive = 5	–999
COMMEN	String	Comments on catalytic converter	–	None
AGECAR	Numeric	Age of car (years)	–	–999
STATE2	Numeric	Attitude toward cars	Strongly disagree = 1 Somewhat disagree = 2 Neither = 3 Somewhat agree = 4 Strongly agree = 5	–999
STATE3	Numeric	Attitude toward walking	Strongly disagree = 1 Somewhat disagree = 2 Neither = 3 Somewhat agree = 4 Strongly agree = 5	–999
STATE4	Numeric	Attitude toward burial at sea	Strongly disagree = 1 Somewhat disagree = 2 Neither = 3 Somewhat agree = 4 Strongly agree = 5	–999
INCOME	Numeric	Monthly income (£)	–	–999
CHILD	Numeric	Number of children	–	None
CAR	Numeric	Car ownership	No = 0 Yes = 1	None

Novices to data analysis (and lazy people) are often tempted to take naughty shortcuts in variable names and simply consecutively number variables (for example, call them VAR1 to VAR723 – like in Table 4.2). We recommend that you refrain from this practice and use mnemonic variable names as in Table 4.3 (e.g., use COUNTRY instead of VAR2, GENDER instead of VAR3, etc.). Even if you have a nicely organized document including the detailed coding plan (which, in fact, you should always do), using easily recognizable variable names can be very helpful. For instance, during your analysis, you may not always see the complete variable

label (see below) and to remember just exactly what VAR437 stands for is far more difficult than working out that AGECAR stands for 'age of car'.

HINT 4.3	When defining variable names, use easily recalled mnemonics; that is, names that remind you of the meaning of your variable.

Variable type. We have already pointed out that there are two main types of variables, namely numeric and alphanumeric (the latter are also known as 'string' or 'alphabetic' variables). In Table 4.3, an example of the former type of variable is GENDER, as we have coded 'male = 0' and 'female = 1', instead of, say, 'male = M' and 'female = F'. On the other hand, the comments regarding the catalytic converter (COMMEN) have been coded as a string variable; that is, typed in as they appear in the open-ended question (see Table 4.3). However, we could quantify this variable if we wanted by creating specific response categories for these comments. For example, we could have grouped comments into 'positive', 'neutral', 'negative', and 'totally incomprehensible' and assigned a numeric value to each category (e.g., positive = 1, neutral = 0, negative = −2, totally incomprehensible = −5). This would result in a numeric nominal variable with four possible values.

Whether a string variable type is to be (re)coded into a numeric variable depends on (a) the scope for sensibly categorizing open-ended responses and (b) whether you would rather have a listing of the original comments or work with fewer response categories in your analysis. There is a whole literature on how best to code replies to open-ended questions, and several research methodology texts include extensive examples of coding instructions (see suggestions for Further Reading). For our purposes, it is sufficient to point out that open-ended questions, while providing useful information, do not lend themselves easily to quantitative data analysis. The number of open-ended questions should, therefore, be minimized. Where they appear unavoidable, the responses to such questions should be classified into exhaustive and mutually exclusive categories; the latter should then be given numerical values to facilitate subsequent computer analysis.

Variable label. Most software packages permit you to use **variable labels**, which can consist of a brief description of your variable, the full variable name, a short definition, or even the source of the original scale (if applicable). Variable labels are useful in that if you forget what a variable name stands for, a look into your code book will tell you what it is (e.g., in Table 4.3, CHILD stands for 'number of children' rather than, say, 'instances of childish behavior').

Value labels. **Value labels** are descriptions of the values for each variable. They are not always necessary, particularly when you have a ratio variable and it is quite obvious what the various values mean (e.g., when you record the respondent's income in euros or dollars). However, value labels need to be carefully defined for multiple-choice questions or scales in which you expect one particular answer out of a given number of alternatives. For example, going back to researching consumer perceptions of our new chocolate spread (Table 4.1), you will remember that we had the alternatives 'ecstatic', 'outstanding', 'exuberant', and so on. We could now code 'ecstatic' as 1, 'outstanding' as 2, 'exuberant' as 3, and so on. We could also code 'ecstatic' as 5 and 'outstanding' as 4, or, indeed, as 4,221. The particular number is irrelevant as long as we use different ones for the five alternatives (we are dealing with a nominal

variable, remember?). By the same token, it does not matter whether you use 0 for 'male' and 1 for 'female' as we did in Table 4.3, or vice versa. However, for some types of scale, the way you assign values *is* important. Consider the two statements in Table 4.4, which are both measured on a Likert scale (see Chapter 3 for this scaling format).

HINT 4.4	Code your answers numerically whenever possible.

Table 4.4 Attitudes toward driving and walking

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
STATE2 I like to use my car even for very short distances					
STATE3 Whenever possible, I leave my car and walk					

If we were to assign 5 for 'strongly agree' down to 1 for 'strongly disagree', a score of 5 would have a very different meaning in the two statements. Specifically, it would indicate a tendency for using cars in statement 2 and a tendency for *not* using cars in statement 3. Thus, if we were to code both statements in an identical fashion (as we did in our code book), it would make coding easier but we would have to remember the **polarity** of each statement (for example, when comparing the average scores of, say, males and females). Similarly, if we wanted to aggregate the scales (i.e., sum up or average the values for each respondent), we would be forced to recode one of the variables so that, for example, low values always expressed 'anti-car' attitudes and high values 'pro-car' attitudes. Consequently, we recommend (re)coding all questions relating to the same issue in such a way that high values express high intensity or more of the property being measured (e.g., liking of cars) and low values express low intensity or less of the property being measured (e.g., not liking of cars).

HINT 4.5	Whenever possible, use a consistent pattern for coding missing data.

Finally, it is important to have value labels for every possible type of response, including labels for missing data (see below). For data coded from open-ended questions, this often involves the specification of a category called 'other' to cope with responses that did not quite fit into the other categories you created.

Missing values. Through this field you can specify the values that represent missing data. In SPSS, for example, users can specify up to three unique missing value codes or specify a numeric range of values to treat as missing, plus one additional unique missing value code. Note that blank cells in the data matrix are treated as missing data, even if they are not explicitly assigned with the value code. Quite often, people use an extreme negative number as a missing value code (e.g., –999, like we did in Table 4.3). This not only ensures that missing data codes are 'out of range' but can also help in detecting missing values. For instance, if you visually inspect the data set (i.e., data matrix) developed based on your questionnaire for unusual cases, it is way easier to spot such long and extreme values.

In Table 4.3, in an attempt to keep our example simple, we have not distinguished between the four different types of missing data discussed earlier. If you want to consider different reasons for missing data, you can set more missing value codes or, even better, create a new nominal variable, which will exemplify why data are missing (see Table 4.5).

Table 4.5 Coding missing values

Name	Type	Label	Values	Missing
STATE4	Numeric	When my goldfish dies, I would like him to have a respectable burial at sea.	Strongly disagree = 1 Somewhat disagree = 2 Neither = 3 Somewhat agree = 4 Strongly agree = 5	–999
STATE4MS	Numeric	Reason for missing data in STATE4	Forgot to answer = 1 Don't know = 2 Refused = 3	None

HINT 4.6 Code variables relating to the same issue consistently; that is, low values should represent unfavorable/negative views and high values favorable/positive views.

Depending on the statistical program you use, it may be necessary to also define the level of measurement of your variables. For example, in SPSS, by default, variables with numeric responses are automatically designated as 'Scale' variables (i.e., assumed to be interval or ratio). If the numeric responses actually represent nominal or ordinal categories, you must change the specified measurement level to the appropriate setting. Note that it is essential to correctly define variables' measurement level as this can affect everything from producing graphs to choosing appropriate algorithms for statistical analysis. Incorrect specification of the measurement level can have potentially disastrous effects on your results (and peace of mind).

WARNING 4.3 Do not take chances by leaving your data matrix unchecked. Data coding and entry is bound to cause mistakes.

FINDING YOUR MISTAKES

Having coded and inputted your data, you might experience a big disappointment – you find mistakes in your data. Noticing mistakes in the input data can cause severe psychological consequences; one of the co-authors has been observed swearing in Greek and throwing himself on the floor in a tantrum. Even if you are very careful during your data input, this is virtually unavoidable, particularly when data entry is manually done; the lengthier and/or more complex your questionnaire, the more likely it is that mistakes will be made during data entry. But do not give up; there are some useful techniques that you can apply to find your mistakes (after the tantrum is over).

Let us first focus on those mistakes that are relatively easy to find to provide you with an instant feeling of success. Such mistakes include skipping a number or an entire row and typing the same number or row twice. To find mistakes of this kind, you should first check to see whether there is a discrepancy between the total number of cases in your data matrix (i.e., number of rows) and your sample size. You can do this by looking at a variable such as CASENO in Table 4.3 or, indeed, at any variable (the total number of legitimate answers plus the total number of missing values should always equal your sample size; if not, you have either omitted or duplicated one or more questionnaires). Next, you should calculate the minimum and maximum values (i.e., the range) for all your variables and compare them against your code book to see whether any of them are 'out of range' (e.g., if you are dealing with a five-point scale anchored at 1 and 5 and your computer tells you that its minimum is 0 and its maximum 7, you have inputted incorrect values somewhere).

If you want to be *really* thorough, you should try to examine conditional responses or responses that should have been filtered based on a previous answer. For example, going back to the variable coding in Table 4.3, maybe the survey was run online and included a filter question automatically preventing respondents who do not own a car from answering the question coded as 'AGECAR'. You should check that 'AGECAR' is indeed always coded as –999 when the value of 'CAR' is 0. If this is not the case, it is probably a good idea to start worrying about whether technical issues have compromised the data-collection process. A thorough examination of the data set would be necessary to identify if such problems represent isolated instances or systematic errors.

Sometimes the answer to one question is contingent on the answer to another, but there is no specific filter to 'enforce' no response. For instance, imagine that a respondent negatively answers the question, 'Have you heard of the brand of soft drinks, Almdudler?' (you might think this is yet another made-up brand but, as a matter of fact, it is a very popular brand in Austria), and later on they report a very positive attitude toward the brand (such as a rating of 5 on a five-point scale), even though a 'not applicable' option was available next to the rating scale. We can hardly trust respondents' attitudes toward something that they do not know. That said, we could recode this particular rating to be a missing value. However, such inconsistent responding casts doubt on whether the particular respondent completed the questionnaire in an appropriate way and is to be trusted altogether. Once again, response patterns need to be further examined to identify whether this inconsistency implies incidental or systematic bias.

Finally, there are mistakes that are *not* easy to detect. These refer to cases where values have been entered that are within the expected range but, nevertheless, wrong (e.g., entering a 3 for a response relating to a five-point Likert scale while the correct response was 4). With such errors, only a very time-consuming one-by-one comparison of the questionnaires (and/or coding sheets) with the data matrix in your computer can help. Nevertheless, you should at least select a few of your questionnaires at random and compare them with the corresponding entries in the data matrix; the more errors you detect, the less proud you should be about your coding and data entry performance. The moral of the story? Check your data – and then, check them again!

TRANSFORMING VARIABLES

Having gone through the painful processes of editing, coding, and finding input errors, you now deserve to have some fun. Changing some variables and/or creating new ones is the perfect excuse to set your creative side free! **Variable transformations** are often necessary to carry out a particular analysis and/or convenient to facilitate data reporting. Statistical software packages incorporate data transformation functions that allow you to form new variables by recoding or combining existing variables. Let us discuss the most common variable transformations. We know you can hardly wait!

Form new variables by recoding existing variables. A common variable transformation involves the recoding of **reverse-coded items** to ensure that values across the items in a questionnaire have consistent directionality or polarity (see the discussion of Table 4.4 above). Imagine, for example, that you have a series of statements capturing respondents' overall attitude toward a specific Albanian restaurant, all of which are answered on five-point interval scales anchored at 1 = strongly disagree and 5 = strongly agree. Now, say that some of the statements are positively framed (e.g., 'The food definitely tastes better than what I get at home') and others are negatively framed (e.g., 'Waiting times are way too long'). Obviously, agreement to these statements would suggest an inconsistency, as a favorable attitude toward the restaurant corresponds to high values in the positively framed and low values in the negatively framed questions. This difference in directionality makes the comparison and combination of responses problematic. For instance, how would you examine differences between an average score of 4.2 for taste and an average score of 1.9 for waiting time when prioritizing areas of improvement? The mathematical difference between these average scores is noticeable, although they both imply a roughly equally positive attitude to both characteristics of the restaurant. That said, any meaningful statistical comparison between the two scores requires **recoding** the variables, so that (low) high values consistently indicate (un)favorable attitudes.

Another common recoding practice involves **changing the measurement level** of a variable to enable subsequent analysis. For instance, you might be studying people's obsession with teddy bears and want to use some demographic data, say respondents' age, to identify differences and similarities between adult and non-adult individuals. In this case, you would have to recode a ratio variable (e.g., based on the answers to the question, 'What is your age?') into a categorical variable with two levels (e.g., 'Is respondent an adult?': 1 = yes; 0 = no). (For this example, we assume that age is an indication of 'adulthood'. However, as you know, you are only young once, but can stay immature forever.) In a similar sense, imagine that you are a brand manager and want to reward your loyal customers by offering them a discount voucher. You have collected detailed customers' brand loyalty scores on a five-point interval scale (1 = very low loyalty, 2 = low loyalty, 3 = average loyalty, 4 = high loyalty, 5 = very high loyalty) and now want to recode these responses in order to form a nominal variable that distinguishes between loyal (e.g., coded as 1) and non-loyal consumers (e.g., coded as 0). The criterion value used to distinguish between loyal and non-loyal customers can be the scale's midpoint (i.e., 3 = average loyalty) or the sample median (i.e., the value that separates the higher 50% from the lower 50% of observed scores – see Chapter 7).

WARNING 4.4 Do not permanently recode variables if the data transformation involves loss of information through a reduction in the level of measurement.

It is important to note that transforming variables by changing their measurement level is a one-way process and can only happen by *downgrading* more detailed levels of measurement to less detailed ones. That is, you can transform metric data (ratio and interval variables) into non-metric, categorical data (ordinal and nominal variables) but not the other way around (see Chapter 3). Note also that you can perform variable recoding either by transforming the original variable or by generating a new variable. You should always go with the latter option as the former implies that the character of the original variable will be altered irreversibly, and the more detailed information collected initially will no longer be available. You have been warned!

Form new variables by integrating existing variables. Another very common type of variable transformation involves forming new variables by combining multiple existing ones. In essence, the resulting new variable will be a mathematical function of the existing variables. Most often, this simply refers to counting, summing, or averaging individual variables. Imagine that a market research agency wants to investigate consumers' brand awareness in the product category of shampoos. To do so, it presents consumers with a series of shampoo brands and asks them to indicate whether they are aware of them or not. Each response is coded as a separate variable, as shown in Table 4.6.

Table 4.6 Brand awareness for different shampoo brands

Name	Type	Variable label	Value label	Missing value
BRAND1	Numeric	Do you know the shampoo brand Fluffy?	No = 0 Yes = 1	−999
BRAND2	Numeric	Do you know the shampoo brand Lavender?	No = 0 Yes = 1	−999
BRAND3	Numeric	Do you know the shampoo brand Durian?	No = 0 Yes = 1	−999
BRAND4	Numeric	Do you know the shampoo brand Petrol?	No = 0 Yes = 1	−999
BRAND5	Numeric	Do you know the shampoo brand Feta?	No = 0 Yes = 1	−999
BRAND6	Numeric	Do you know the shampoo brand Lilly?	No = 0 Yes = 1	−999
BRAND7	Numeric	Do you know the shampoo brand Mist?	No = 0 Yes = 1	−999

With these data, we can, of course, calculate how many consumers know each of the seven brands and use this percentage as a proxy for brand awareness. What is more, we can also create a category expertise index per consumer; the more brands a consumer is aware of, the more experienced he or she is with the specific product category. To do so, we need to compute a new variable that counts the number of brands each consumer is aware of. Given that the

existing variables are coded as 0 (no) and 1 (yes), we simply need to add them up (i.e., EXPERT = BRAND1 + BRAND2 + BRAND3 + BRAND4 + BRAND5 + BRAND6 + BRAND7). The new variable (EXPERT) would now be a discrete numerical variable ranging from 0 (no expertise) to 7 (high expertise). This new variable can now be used for further analysis to, let's say, explore whether consumers' expertise in the product category is associated with their spending pattern.

As another example, let us assume that you have gathered data on the different sources of income for a sample of business executives. At present, you have the five separate variables listed in Table 4.7, each of which records income in thousands of euros.

Table 4.7 Different sources of income

Variable name	Variable label
BLMAIL	Blackmail
EXTOR	Extortion
PROSTI	Prostitution
BRIBE	Bribery
LECTUR	Occasional lectures on business ethics at various universities

If you plan to analyze the overall income of this profession (in order, say, to compare it with the average annual income of a bagpiper in the 3rd Royal Scottish Highland Battalion), you need to create a new variable (call it TOTINC); this would be defined as the sum of the five original variables: TOTINC = BLMAIL + EXTOR + PROSTI + BRIBE + LECTUR. Note that TOTINC is an *additional* variable, with the five original variables still existing unchanged.

The final type of variable transformation is related to averaging and is a bit more delicate. As we discussed in Chapter 3, it is standard practice in research to use multiple questionnaire items in order to measure a particular construct. In these cases, researchers typically combine individual items into a **composite scale** by averaging responses to them. For example, you might be interested in measuring consumers' involvement in the product category of kitchen paper and thus ask respondents to indicate on three semantic differential items the extent to which kitchen paper is personally irrelevant/relevant, is unimportant/important, and means nothing/means a lot to them. To generate a composite scale of kitchen paper involvement, you can average respondents' ratings across these three items. In doing so, do not forget three important points. First, make sure that all of the items used to develop the composite scale have the same format. For example, they all need to be operationalized with five-point or seven-point scales. Second, make sure that all of the items have the same polarity. That is, in every item, higher (lower) values should indicate more (less) involvement. Third, make sure that you assess the validity and reliability of the resulting composite scale (see Chapter 3).

Finally, you should explore the scope for adding variables to your data set from secondary data. In industrial marketing research, for example, company information such as the number of employees, sales revenue, principal activity, and so on is often available from secondary sources (e.g., business directories), and there is no need to waste precious questionnaire space and/or risk alienating your respondent by requesting this information again. Nothing is stopping you from augmenting your data matrix by adding variables relating to your respondents

from whatever source you deem appropriate. Of course, you need to ensure that the data obtained from such sources are comparable to your own. For example, relying on a directory published in 1973 to obtain much-needed sales revenue information relating to firms surveyed during 2020 is not exactly a brilliant idea.

We conclude this exceedingly titillating chapter with the following words of infinite wisdom: *make sure that you have (several) back-up copies of all your important files, be it the data file, the command files you used to run your analysis, or the output files*. If you don't, you will live to regret it.

SUMMARY

Having demonstrated that there is more to editing, coding, and data input than meets the eye, how can we best summarize the key points? (This is a purely rhetorical question – after all, we are paid to write this book and should know the key points!) First, we focused on data cleaning or editing, which attempts to detect and correct data errors. Various editing tasks were discussed, including dealing with ambiguous data, logically inconsistent data, and item non-response. We then moved on to data coding and showed you the value of setting up a detailed coding scheme. We followed this with a discussion of strategies for finding mistakes resulting from data entry. Finally, we discussed different variable transformations and potential pitfalls associated with them. Throughout this chapter, we highlighted that statistical software packages can benefit you in working with data. Indeed, nowadays there is a great variety of accessible statistical software packages (including online applications) that enable users in (a) inputting and editing the data, (b) carrying out statistical analysis, and (c) reporting and visualizing the findings.

WARNING 4.5 Always make *several* back-up copies of your important files and do not keep all copies at the same place or on the same device.

QUESTIONS AND PROBLEMS

1. Distinguish between ambiguous answers and logically inconsistent answers and give two examples of each. What steps can you take to avoid such problems?
2. What are the four different forms of item non-response?
3. How would you code missing data when entering questionnaire results into a computer?
4. Construct an example to illustrate the difference between a variable name, a variable label, and a value label.
5. Under which circumstances would you change the polarity of attitude statements during the coding process?
6. What are alphanumeric (string) variables, and how would you go about coding them?
7. What are typical mistakes that can occur during data input and coding? How would you go about identifying them?
8. Under what circumstances would you consider transforming your variables? Give three examples of variable transformations.
9. What advice would you give to fellow sufferers to relieve boredom during coding?

FURTHER READING

Krippendorf, K. (2018). *Content Analysis: An Introduction to Its Methodology*. London: Sage Publications. Has an excellent section on how to code open-ended responses.

Little, R. J. A. & Rubin. D. B. (2019). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons. Everything you ever wanted to know about handling missing data.

Ruel, E., Wagner III, W. E., & Gillespie, B. J. (2015). *The Practice of Survey Research.* London: Sage Publications. Contains a comprehensive discussion on how to code survey data.

5
Why do you need to know your *objective* before you fail to achieve it?

THE NEED FOR ANALYSIS OBJECTIVES

So here you are, sitting in front of your computer, raring to go. Your data, meticulously edited and coded, sit restlessly in the depths of your computer's memory and eagerly await your instructions to reveal their secrets. Your problem at this stage is to decide exactly what instructions to give; in other words, what **analysis** to undertake.

One easy (but certainly not recommended) solution to your problem is to follow the 'pack the kitchen sink in' approach, also known as the 'run anything and everything' strategy. This approach simply involves asking your computer to perform all known statistical tests in the universe on all your variables – including relating everything to everything else (the rationale being that by producing 128 miles of results printout, something useful is *bound* to turn up!).

WARNING 5.1 Don't be tempted to use your computer's capability to run 'everything under the sun'. You will live to regret it.

Unfortunately, this approach has several drawbacks, the most important of which is that it doesn't work! While your computer will (usually!) do what it is instructed to do, the usefulness of the resulting information depends on the skill with which it is instructed to analyze the data. So, what's the alternative? Well, the alternative is to set some **analysis objectives** – preferably some good ones! Setting clear analysis objectives makes your data analysis effective and efficient, and ensures that you get the most out of your data.

If you think of analysis as the process by which one makes the data 'talk', setting analysis objectives is deciding exactly what kind of information one wants to get out of the data. Data can be made to 'talk' in different ways, depending upon how they are manipulated. The choice of the particular technique depends, above all, on the specific needs of the researcher; that is, the sort of answers that he/she seeks from the data (to be sure, the choice of analytical technique depends on other factors as well, as we will see later on).

Clear analysis objectives serve to direct and guide the analysis process and are absolutely essential for the success of the latter. Specifically, analysis objectives fulfill three roles. First,

they help ensure that only *relevant* analyses are undertaken; relevance refers to the extent to which any analysis performed contributes directly to answering the research questions of interest. For example, if you seek to determine what criteria company executives use in their choice of a massage parlor, a cross-tabulation analysis of age by income does not make you any the wiser about their choice criteria.

Second, analysis objectives provide a check on the *comprehensiveness* of the analysis. Comprehensiveness refers to the extent to which the set of analyses performed makes full use of the information potential in the data. Sticking with the same example, if you had data on both the income and the age of the respondent but only examined the impact of income (or age, but not both) on the choice criteria, then your analysis would have been relevant but not comprehensive.

Third, analysis objectives help avoid *redundancy* in the analysis; redundancy refers to the extent to which different analyses overlap – that is, how much they provide essentially the same information. In our massage parlor example, if you first compared the number of male executives considering criterion X (e.g., 'location') as important with that of females, and then contrasted the number of males and females *not* considering criterion X as important, then your additional contribution to knowledge would be precisely zero (because the content of the second analysis is already fully contained in the first).

In short, having analysis objectives helps ensure (a) that you only do relevant things, (b) that you do enough, and (c) that you don't do things twice.

SETTING ANALYSIS OBJECTIVES

The starting point for setting sound analysis objectives should always be the overall research objectives. Most of us do not wake up one Monday morning with an uncontrollable desire to analyze a data set we have never seen before and know nothing about. At some point in the past, we somehow got involved in a research project and got hold of some data, and now we actually have to do something with it (which is a hassle, but that's life for you). Obviously, there must have been a purpose to the research project (although, by now, you may have forgotten what it was!). The **research purpose** indicates why the study was carried out in the first place and what it hoped to achieve.

A careful re-examination of the overall aims of the research provides an excellent point of departure for developing analysis objectives. In this context, there has to be an explicit link between the analysis objectives and the overall research objectives, in the sense that the achievement of the former should contribute to the achievement of the latter. Clearly, the better specified the original research objectives, the easier it is to derive appropriate analysis objectives from them. For example, a study aimed at 'investigating the export behavior of nightgown manufacturers in the Swiss Canton Appenzell' provides hardly any clue about exactly what it hopes to achieve (other than that it has to do with exporting). Although more insight may be gained by looking at the specific variables that have been included in the study, one is still left unclear as to (a) why these specific variables were considered (and not others) and (b) what one is supposed to do with them (i.e., what kind of information one should be

looking for). The problem here is that the research purpose has been too broadly defined and not translated into concrete research objectives; thus, practically any set of analysis objectives would be consistent with the study's aims.

If, on the other hand, the same study had been defined as aiming to 'compare Swiss-owned and foreign-owned nightgown manufacturers in terms of (a) company size, (b) export experience, (c) extent of export operations, (d) resources allocated to exporting, and (e) export performance', then the development of appropriate analysis objectives would be far less problematic. To begin with, one would immediately know that the main focus should be on *comparisons* rather than an overall description of the nightgown manufacturing industry in Switzerland. Moreover, one would know what to look for in the variable list (i.e., specific indicators of size, export experience, export operations, and so on). Finally, one would be able to identify the limitations of the analysis (e.g., inadequate representation of 'export operations' due to, for example, the omission of the 'number of export destinations' from the list of variables).

WARNING 5.2 If you get involved in an ill-defined research project, chances are that you will encounter problems at the analysis stage. So, don't!

It is helpful to think of the process of setting analysis objectives as involving two sorts of decisions. The first relates to the **content** of an intended analysis and involves selecting a set of variables to be included in it; clearly, an analysis objective would be of little help if it did not reference the particular variable(s) to be analyzed. The second decision relates to the **focus** of the intended analysis and involves the specification of the analytical stance or orientation to be adopted (e.g., simple description vs. examination of relationships). Together, content and focus enable *operational* analysis objectives to be set such as: 'describe the sample/population in terms of variables X, Y, and Z'; 'break the total sample into two sub-samples according to variable A and compare them in terms of variables B and C'; 'examine the relationship between variables D and E'; and so on. Of course, you should always make sure that such specific objectives *do* indeed satisfy the criteria of relevance, comprehensiveness, and avoidance of redundancy mentioned earlier.

THE QUESTION OF FOCUS

While the content of any analysis is, by definition, project-specific (since it depends on the particular variables that have been included in the study), its focus can only take one of three basic forms: description, estimation, and hypothesis-testing.

With a **descriptive focus**, as the name implies, the aim is to paint a summary picture of the sample (or population) in terms of the variable(s) of interest. For example, assume that you have randomly selected 200 university students and conducted interviews on their clubbing habits. When you analyze the data, you may want to provide a sample breakdown in terms of, for example, gender, faculty, year of study, and clubbing frequency. Statements that 120 members of your sample are female, that two-thirds of all respondents are matriculated in the

Faculty of Theology, that the proportions of first-, second-, and third-year students are 20%, 30%, and 50%, respectively, and that, across the entire sample, the average club-visiting frequency is 4.6 times a week (during term time, of course, with a substantially higher frequency by students from the Faculty of Theology) are all examples of a descriptive focus. A number of statistical techniques can be used to undertake descriptive analysis; among the connoisseurs in the field, these are known as **descriptive statistics** and will be discussed in Chapters 6 and 7.

With an **estimation focus**, again as the name implies, the aim is to use the information one has on the *sample* to estimate the situation that is likely to exist in the *population* as a whole. For example, given that the average club-visiting frequency of 200 randomly selected students turns out to be 4.6 times a week, what is likely to be the average bar-going frequency across *all* students in the university concerned? Thus, estimation can be seen as the process of making an informed guess based on incomplete information. The guess is informed because we use the information we have on the sample (together with some statistical theory to be introduced in Chapter 7) to say something about the population from which it was drawn. Simultaneously, the very fact that we use a sample means that our information is incomplete because it is not based on the entire population but only on a fragment of it. Therefore, we should be fully aware that our information contains sampling error and, as such, it does not perfectly reflect the situation in the population (we talked about this in Chapter 2, remember?). However, to the extent that the sample has been drawn probabilistically (as is the case in the above example), we can *assess* the sampling error and incorporate it in our calculations of the population estimates, ending up with **confidence intervals** for the latter. How this is done is the subject of Chapter 8, so don't be impatient!

Finally, with a **hypothesis-testing focus**, the aim is to examine specific propositions concerning the variables of interest (e.g., relationships and/or differences between variables) and use sample evidence to draw conclusions about whether these propositions hold true for the population as a whole. For example, you may hypothesize that male and female students differ in terms of their clubbing frequency (i.e., that there is a relationship between gender and going clubbing). If, say, you found out that the 120 female students in your sample went to clubs 4.8 times a week on average, while the 80 males went only 4.3 times, could you conclude that female students, *in general*, tend to visit clubs more often than males? Well, your sample results suggest that this might be the case, but do not forget that this is an incomplete picture of the population that involves sampling error. How likely is it that the difference in the clubbing frequency observed in the sample actually reflects a 'true' difference in the population, rather than merely sampling error? To answer this and similar questions, we apply what are known as **significance tests**, which are statistical techniques designed to help us decide whether our sample results are likely to reflect the whole population of interest. Again, as was the case with estimation procedures, it is assumed that our data are based on a probabilistic sample so that assessment of sampling error is feasible. We will come back to the issue of hypothesis-testing in Chapter 9.

When the focus of analysis is on estimation or hypothesis-testing, we use our sample to make *inferences* about the population. This process is formally known as **statistical inference**, and the various techniques employed are commonly referred to as **inferential statistics**. Without inferential statistics, the only other way to make statements about the population is

to conduct a census; that is, obtain data on each and every population element. For reasons already discussed in Chapter 2, this is usually not feasible and, therefore, inferential statistics are indispensable in data analysis.

CHOOSING THE METHOD OF ANALYSIS

Although well-specified analysis objectives (in terms of content and focus) are essential prerequisites for successful analysis, there are additional factors to consider before selecting an appropriate analytical technique. Figure 5.1 provides a superb overview of the factors bearing on the choice of analytical technique. While their specific influence will become more apparent as we go through Chapters 6 to 9, we felt that you would be eternally grateful (and possibly include us in your last will and testament) if you could have a first taste in this section.

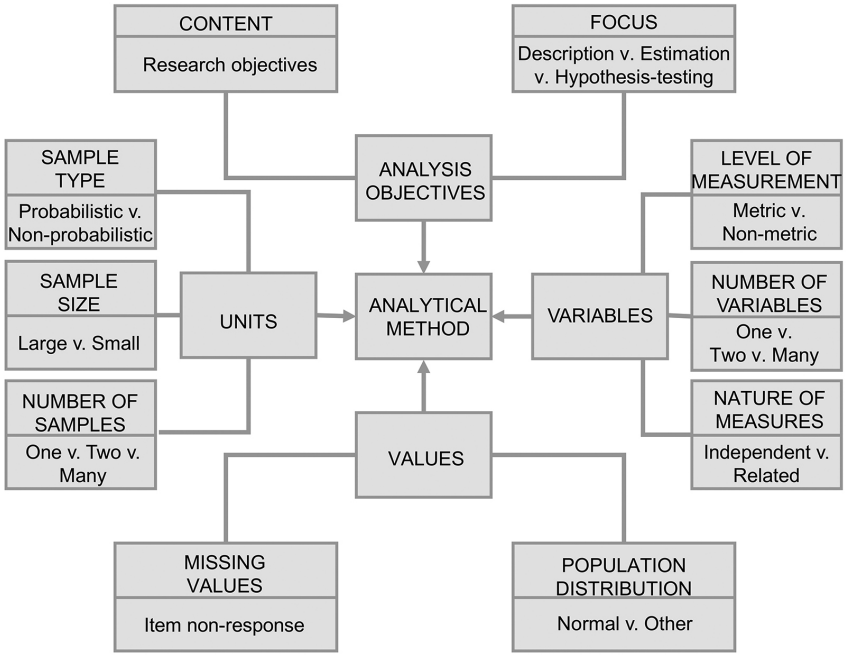


Figure 5.1 Factors influencing the choice of analytical technique

To begin with, the characteristics of the sample in terms of type and size will affect the choice of technique. With regard to *sample type*, as you should know by now, unless the sample has been drawn probabilistically, the use of inferential statistics is not legitimate, since the latter makes use of the concept of sampling error, which – as pointed out in Chapter 2 – cannot be assessed when non-probability sampling methods are employed. What this means is that you

will be treading on very thin ice if you try to estimate population parameters or test hypotheses with a non-random sample (e.g., a convenience or judgmental sample).

With regard to sample size, suffice it to say at the moment that some statistical procedures do not work well unless one has a 'sufficiently large' sample. In practice, for simple analyses, this translates to a sample size of at least 30 (this rule of thumb refers to a bare-minimum threshold that has some tortuous mathematical and empirical base underlying it). The procedures susceptible to small-sample problems are a set of inferential procedures known as **parametric statistics**. Parametric tests rely on specific assumptions about the distribution of parameters concerning the population from which the data are drawn and require reasonably large samples to produce valid estimates and hypothesis-testing results. With small samples, one may be limited to using only **non-parametric statistics** (sometimes also known as 'small-sample' statistics), which are not tied to stringent distributional assumptions of population parameters. Although they are generally less powerful techniques, they can yield trustworthy results (we briefly touched upon parametric and non-parametric statistics in Chapter 3 when we distinguished between metric and non-metric measures).

A final sample characteristic affecting the choice of analysis technique is the *number of sub-samples* that one may wish to consider in a particular analysis. This is particularly relevant when the aim is to undertake comparisons among different groups in one's data set (e.g., blond consumers vs. bald consumers, buyers vs. non-buyers, consumers owning tarantulas vs. consumers owning boa constrictors vs. consumers owning frogs). As will be discussed in Chapter 11, different statistical procedures are appropriate when only two groups are to be compared as opposed to when three or more groups enter the comparison.

HINT 5.1	Try to work with probabilistically drawn samples of sufficient size. You will have a greater choice of techniques (and fewer headaches) when it comes to analysis.

Another influence on the method of analysis comes from the *measurement characteristics* of the variables involved. As was pointed out in rather excruciating detail in Chapter 3, the higher the level of measurement, the more sophisticated the analysis that can be applied to the data; and, yes, parametric procedures only work with metric (i.e., interval and ratio) data (as another look at Chapter 3 should quickly confirm).

A related consideration is the *number of variables* that one wants to analyze simultaneously *and* the extent to which these differ in terms of their level of measurement. When one is interested in a single variable at a time (what we call **univariate analysis**), there is no real problem because there is only one level of measurement to consider. Things get a bit more complicated, however, when two variables are involved (the case of **bivariate analysis**). As can be seen from Table 5.1, six distinct combinations of levels of measurement are possible, each requiring a different statistical technique (interval and ratio measures are grouped together because, as noted in Chapter 3, practically all analysis methods suitable for ratio measures are also applicable to interval measures).

If you thought the case of bivariate analysis was bad enough, then **multivariate analysis** (i.e., taking three or more variables at a time) can only be described as positively nasty. Before

Table 5.1 Levels of measurement and bivariate analysis

	Variable 1		
Variable 2	Nominal	Ordinal	Interval/ratio
Nominal	I	II	III
Ordinal		IV	V
Interval/ratio			VI

you panic (and set this book on fire), however, let us assure you that we will be dealing with multivariate statistical procedures with love and tender care. Performing multivariate analysis (and living to talk about it) presupposes a sound background of basic statistical concepts and familiarity with both univariate and bivariate techniques. At this stage, you are likely to possess neither (otherwise, why did you buy/borrow/steal this book?), but make sure that you do obtain such familiarity before you dive into the abyssal fun of Chapter 13 onwards.

While the level of measurement and the number of variables involved impact the choice of technique, the latter is also influenced by the *nature* of the measures concerned; that is, whether they are independent or related. With **independent (between-subjects) measurements**, different units (i.e., sample elements) are compared on a given characteristic. For example, contrasting the sausage-eating capacity of Liverpool football supporters with that of Bayern München supporters is an analysis example involving independent measures (i.e., the measurement of the sausage-eating capacity of Liverpool supporters does not affect the measurement of the sausage-eating capacity of the Bayern München supporters and vice versa). With **related** (within-subjects or repeated) **measurements**, on the other hand, the same units are compared on different characteristics. For example, contrasting the sausage- and burger-eating capacities of a sample of Liverpool supporters is an example involving related measures (since the measurements obtained on each characteristic cannot be assumed to be independent from the measurements obtained on the other as *both* measures relate to the *same* units).

Typical analysis situations involving related measures arise from longitudinal research designs and experimental studies of the 'before–after' variety (see Chapter 1). However, even with cross-sectional data, related measures may be involved, for example, when comparing the opinions of the same group on two different attitude scales. Note that in the literature, the issue of measure independence is often referred to as **sample independence** (i.e., a distinction is drawn between independent and related samples rather than measures). We prefer to talk about measures rather than samples because it is not the sample elements themselves that are measured but rather their characteristics (see Chapter 3).

One practical way of checking the question of independence is to visualize the data matrix and ask the question: are we comparing different groups of respondents (i.e., units of analysis) on a given characteristic or different characteristics of the same respondents? In the first instance, as Figure 5.2 shows, we are partitioning the data matrix by rows (i.e., splitting it into different sub-samples reflecting independent observations). In the latter instance, we are partitioning the data matrix by columns (i.e., into different variables relating to the same set of observations). In the atrociously simplified example of Figure 5.2, the two partitioning approaches result in the following very different analyses: first, a comparison of supermarkets

(A) and department stores (B) in terms of their *total* aftershave sales and, second, a comparison of Macho Man (A) and Super Wimp (B) sales irrespective of store type. In the first case, we are dealing with independent measures as we are comparing two distinct sub-samples (i.e., four supermarkets and three department stores). In the second case, we are comparing two variables across all seven stores. Different statistical techniques are appropriate for each type of comparison, and you will learn all about them in Chapter 11.

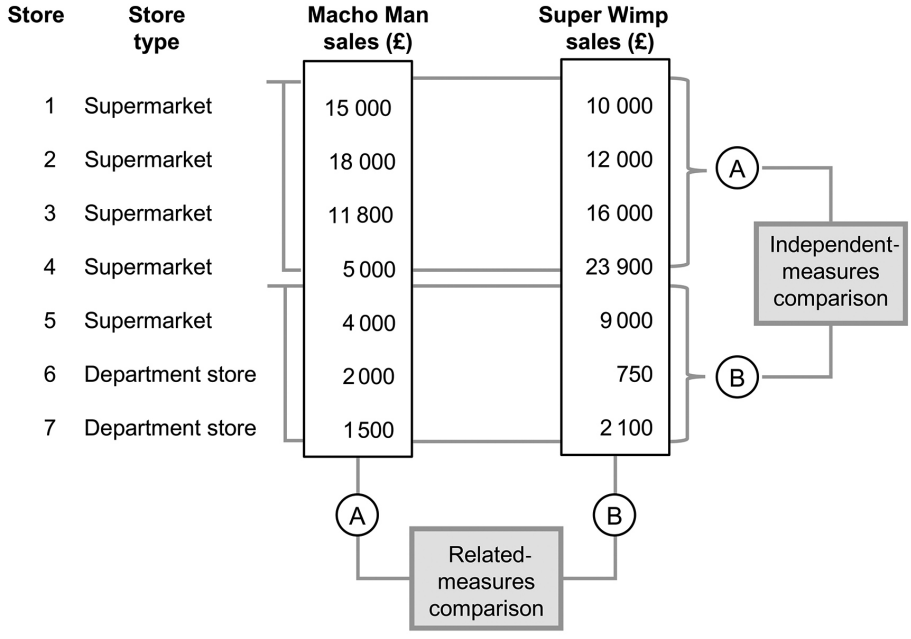


Figure 5.2 Independent- versus related-measures analysis

The final set of influences affecting the choice of analysis technique has to do with the responses themselves; that is, the actual data values obtained. A key issue here is that of item non-response; that is, missing values on one or more variables entering the analysis (see Chapter 4 to refresh your memory as to why missing values may turn up and what you can do about them other than throwing a tantrum). A direct consequence of missing values is the reduction in the **effective sample size**; that is, the number of observations or data points actually available for analysis. This, in turn, might affect what kind of statistical procedures can legitimately be employed (see the previous discussion on the influence of sample size).

To illustrate the potential severity of the missing-value problem, consider a situation where you have a sample of 100 left-handed roller-skating enthusiasts (the fact that they are left-handed is, of course, entirely irrelevant, but we want to make sure that you pay attention). Also, imagine that you have asked them two questions, namely: (1) 'How many pairs of roller

skates do you own?' and (2) 'How many hours per week do you spend roller-skating?' Now, assume that, for one reason or another, you end up having 10% item non-response on each question. You may think that, given this non-response pattern, a total of 90 cases will be available for analysis. However, this is not necessarily the case since it is possible that a *different* 10% of respondents did not answer Question 2 from the 10% that did not answer Question 1. This means that, in any analysis including both questions, your effective sample size may be as low as 80 cases (depending upon the degree of overlap among non-respondents on the two questions). To make matters worse, even when you take one variable at a time, you still have the problem of **comparability of responses** (since the 90 respondents who answered Question 1 might include some or all of the 10 respondents who did not answer Question 2 and vice versa). Extend this line of thinking to a situation of many variables (i.e., a real-life data set), and you can imagine the kind of horrors that missing values can cause!

A second consideration concerns the assumptions made regarding the distribution of values *in the population*. Many inferential statistical procedures assume that the values of the variable(s) of interest in the underlying population are 'normally' distributed; that is, they form a symmetrical, bell-shaped curve. (This curve is known as the **normal distribution**, and you will make its acquaintance in later chapters.) 'Violations of the normality assumption' (an expression widely used among statisticians) can render a great deal of parametric statistical techniques inoperative. However, 'departures from normality' (another favorite expression among the statistical 'in'-crowd) have to be rather severe since (mercifully) many statistical tests are quite *robust* (i.e., they do not immediately break down in tears with the slightest violation in their assumptions). Nevertheless, if the assumption of normality cannot be reasonably entertained (and there are tests one can use to check for this, as we shall see in Chapter 10), then one may have to make do with non-parametric statistical tests. The latter make no assumption about the distribution of values in the population and are, therefore, also sometimes referred to as 'distribution-free' statistics. Having said all this, if the normality assumption cannot be satisfied but the sample is large enough, one may still be able to use some parametric statistical procedures (the reasons for this are quite complex to explain here, so we will not even try).

From the above, it should be evident that a recurrent theme in the choice of analytical technique is deciding between parametric and non-parametric methods of analysis. Figure 5.3 should make this task easier for you by providing a summary of the major considerations. (Dear, dear, are we spoiling you or what?)

As a final point, it would be highly immoral not to mention that, in addition to the various influences displayed in Figure 5.3, different statistical procedures have some 'unique' assumptions and requirements that must also be satisfied in order for them to yield valid results. (Yes, we know that this makes things even more complicated, but it is not our fault – blame the statisticians if you must.) However, we thought it would be better to mention such additional assumptions on an 'as-needed' basis (i.e., when discussing specific procedures) rather than overloading an already demanding chapter.

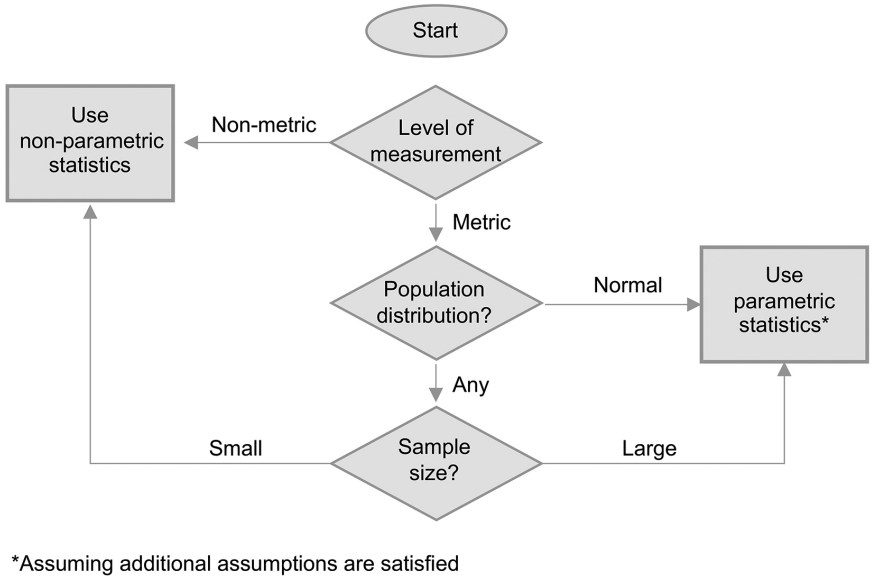


Figure 5.3 Choosing between parametric and non-parametric procedures

SUMMARY

The purpose of this chapter was to provide an overview of the kinds of issues that must be considered *before* the analysis can be carried out. We started by stressing the importance of setting analysis objectives and followed this up by showing you how the overall research aims can be used to this effect. Next, we distinguished among different foci of analysis as reflected in description, estimation, and hypothesis-testing, respectively. Last, but not least, we took a hard look at the various factors that influence the choice of analytical procedures, highlighting, in particular, the conditions favoring the use of parametric versus non-parametric statistics. We are now ready to actually *do* some analysis.

QUESTIONS AND PROBLEMS

1. What are the main reasons for setting clear analysis objectives?
2. Describe the connection between analysis objectives and the research purpose. Use an example to illustrate your answer.
3. Why is it useful to distinguish between content and focus issues when setting analysis objectives?
4. Construct an example to illustrate the differences between description, estimation, and hypothesis-testing.
5. In simple terms, what is the purpose of inferential statistics?
6. Which key factors, other than the objectives of analysis, determine the choice of sta-

tistical technique?
7. When would you use parametric versus non-parametric statistics?
8. What are the implications of item non-response for data analysis?
9. Draw up an example to illustrate the difference between independent and related measures.
10. Wouldn't life be simpler without Figure 5.1?

FURTHER READING

Sorry guys, there is nothing useful we know of to refer you to for your bedtime reading. For some reason, and despite their importance, the issues we covered in this chapter have not received explicit and integrated treatment in mainstream data analysis textbooks.

PART III
CARRYING OUT THE ANALYSIS

6
Why not take it easy initially and *describe* your data?

PURPOSES OF DATA DESCRIPTION

Imagine that you have just completed a nationwide survey of Mongolian attitudes toward year-round swimming. Your research instrument, a 24-page mail questionnaire, requested information on some 120 different variables ranging from sociodemographic characteristics (e.g., age, education, and marital status) to behavioral aspects (e.g., number of times that respondents went swimming in the sea in December). Out of 2,000 questionnaires initially sent out, you managed to get back 310 completed (as well as 1,444 letters calling you a lunatic, 231 threatening phone calls from outraged respondents, 12 personal visits from respondents who 'did not see the funny side', and three letter bombs). Following your recovery in hospital, you have coded your data and are now facing something like 37,200 entries in your data matrix (310 cases multiplied by 120 variables). Now, no matter how brilliant you are or how long you spend staring at the data matrix, you will be unable to make any sense out of the endless list of seemingly random numbers.

Data description is a typical first step in any data analysis project. In addition to being an important, self-standing activity when a descriptive focus characterizes the analysis objectives (see Chapter 5), descriptive analysis provides a very useful initial examination of the data even when the ultimate concern of the investigator is inferential in nature (i.e., involving estimation and/or hypothesis-testing). Specifically, the purpose of descriptive analysis is to:

(a) Provide preliminary insights as to the nature of the responses obtained, as reflected in the distribution of values for each variable of interest.
(b) Help detect errors in the coding process (see also Chapter 4).
(c) Provide a means for presenting the data in a digestible manner through the use of tables and graphs.
(d) Provide summary measures of 'typical' or 'average' responses as well as the extent of variation in responses for a given variable (to be discussed in Chapter 7).
(e) Provide an early opportunity for checking whether the distributional assumptions of subsequent statistical tests are likely to be satisfied (see also Chapters 5 and 7).

In short, data description is like a first date: it enables you to get to know your data before you try something more adventurous.

FREQUENCY DISTRIBUTIONS

The starting point in descriptive analysis is the construction of a **frequency distribution** for each variable of interest. This simply shows in absolute or relative (e.g., percentage) terms how often the different values of the variable are actually encountered in one's sample. In other words, a frequency distribution indicates how 'popular' the different values of the variable are among the units of analysis. A few examples of frequency distributions using your data on Mongolian swimming habits are shown in Table 6.1. Both **absolute frequencies** (i.e., simple counts) and **relative frequencies** (i.e., percentages) are given, and in an attempt to spoil you even more, the variables summarized cover different levels of measurement (and if you can't tell *at a glance* which variable gives you what level of measurement, you'd better hurry back to Chapter 3 and brush up!).

HINT 6.1	Always begin your analysis by doing some data description. It will make your life easier when you get around to doing estimation and/or hypothesis-testing.

There are a few things worth noting about frequency tables. First, a frequency (whether absolute or relative) can *never* be negative, as a value cannot be encountered fewer than zero times. Second, the sum of absolute frequencies *must* equal the total number of observations (i.e., the sample size), while the sum of relative frequencies *must* equal 100% (if expressed in percentages) or 1 (if expressed in proportions). Nonetheless, if your data set has missing values, the total frequency will obviously not be equal to your sample size, nor will the relative frequency add up to 100%. Note how the variable 'Education' in Table 6.1b is organized into 'Valid' and 'Missing' sections. In the 'Missing' section, we can see that the numerical code '–999' is explicitly used to designate missing values (10 cases or 3.2%). We also notice that there are eight additional cases (2.6%) coded as 'System'. These are missing values that are not explicitly coded and simply refer to blank cells in your data set (you might want to double-check why this happens). As a result of these 18 missing values, the 'Valid' section of the frequency table has shrunk to a total of 292 observations. The column 'Percent' shows the relative frequencies (percentage of observations) out of all observations in the data set, while 'Valid Percent' displays the percentage of observations in each category based on the total number of *non-missing* responses only. Finally, as shown in Table 6.1, frequency tables can be set up for any variable, irrespective of the level of measurement.

WARNING 6.1	Use your frequency tables to check the coding of your data. Finding your mistakes early in the analysis will save you a lot of trouble later on.

The above features can be used to your advantage because they can alert you to possible errors in your data (yes, finding errors is depressing, but not half as depressing as detecting errors *after* you have completed your analysis). Thus, if you find that the sum of absolute frequencies for a given variable is greater than your sample size or that you obtain frequencies for 'out-of-range' values, something has gone wrong somewhere. For example, if you find that

138 of your respondents were 86 years old, and your sample consisted of 75 primary school children, chances are that some entries have been coded incorrectly (probably because you did not follow our brilliant advice on editing and coding in Chapter 4). Your frequency table will be able to tell you *where* these errors are likely to have happened so that you can go back to your original data and sort things out before carrying on with the analysis.

It is also possible to construct **cumulative frequency distributions**. These show how many observations take values that are 'greater than' or 'less than' a specified value. Going back to Table 6.1, you may want, for example, to find out how many respondents had three children at most or how many had more than five children. To answer these and similar kinds of questions, you can use the absolute (i.e., counts) and/or relative frequencies (i.e., percentages) to compute the corresponding cumulative frequencies. This simply involves adding the frequencies associated with a particular value to the sum of the frequencies corresponding to all preceding values. The derived cumulative relative frequency distribution for a variable is provided in the last column of Tables 6.1a–d. Obviously, you can also derive the cumulative frequency distribution in absolute terms, but you know how to count; computers will not do everything for you!

Table 6.1a Examples of frequency distributions: marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Single	112	36.1	36.1	36.1
	Married	87	28.1	28.1	64.2
	Widowed	48	15.5	15.5	79.7
	Divorced	63	20.3	20.3	100.0
	Total	310	100.0	100.0	

Table 6.1b Examples of frequency distributions: education

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Postgraduate degree	57	18.4	19.5	19.5
	Undergraduate degree	100	32.3	34.2	53.8
	High school diploma	77	24.8	26.4	80.1
	Primary school certificate	47	15.2	16.1	96.2
	Kindergarten only	11	3.5	3.8	100.0
	Total	292	94.2	100.0	
Missing	–999	8	2.6		
	System	10	3.2		
	Total	18	5.8		
Total		310	100.0		

Table 6.1c Examples of frequency distributions: 'I love to go swimming in December when I get a chance'

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Strongly disagree	45	14.5	14.5	14.5
	Disagree	34	11.0	11.0	25.5
	Neither disagree nor agree	66	21.3	21.3	46.8
	Agree	36	11.6	11.6	58.4
	Strongly agree	129	41.6	41.6	100.0
	Total	310	100.0	100.0	

Table 6.1d Examples of frequency distributions: number of children

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	95	30.6	30.6	30.6
	1	60	19.4	19.4	50.0
	2	30	9.7	9.7	59.7
	3	105	33.9	33.9	93.5
	4	10	3.2	3.2	96.8
	5	9	2.9	2.9	99.7
	more than 5	1	.3	.3	100.0
	Total	310	100.0	100.0	

Again, cumulative frequencies (whether absolute or relative) must be non-negative. Moreover, the total absolute cumulative frequency must be equal to the sample size, while the last relative cumulative frequency must be equal to 100%. Note that both the absolute and relative frequencies will obviously fall short in the presence of missing data. For example, in the second panel of Table 6.1b, the total frequency is less than the sample size. Also, the total percent in the second column now adds up to 94.2% since 5.8% of the values are missing. Finally, note that the cumulative frequency distribution is typically constructed on the basis of *valid* values only.

Cumulative frequency distributions such as the ones shown in Table 6.1 are of the 'up to' or 'less than' variety in that they tell us the percentage of observations *not exceeding* a certain value. For example, the last column for the variable 'number of children' in Table 6.1d tells us that 59.7% of respondents had two children or fewer. However, it is very easy to calculate the percentage of observations that *exceed* a given value (e.g., to find out how many respondents had more than two children). This is quickly done by subtracting the relative cumulative frequencies from 100%. In our example, 100% − 59.7% = 40.3% of all respondents. Given our outstanding preparatory work in the previous chapters, it should now be crystal clear to you that cumulative frequencies are not useful for nominal variables. This is because their values don't have an inherent order. For example, with the variable 'marital status' in Table 6.1a, it doesn't really make sense to say that '79.7% of our respondents are at least widowed …'. Thus,

looking at cumulative frequencies is only informative with ordinal-, interval-, and ratio-level variables (and if you can't remember what these are, you should urgently revisit Chapter 3).

A particularly useful application of cumulative frequencies is the calculation of what are known as **percentiles**. These divide a set of observations into 100 equal portions; thus, there are 99 percentiles (P_1, P_2, P_3, …, P_{99}). The *p*th percentile P_p is the point on the measurement scale such that *p* percent of the observations have values less than P_p and (100–*p*) of the observations have values greater than P_p. For example, the 25th percentile P_{25} (also known among statistical boffins as the bottom or **first quartile**, Q_1) is the value below which 25% of the observations fall; the 50th percentile P_{50} (the **second quartile**, Q_2) is the value below which 50% of the observations fall; and the 75th percentile P_{75} (the **top quartile**, Q_3) is the value below which 75% of the observations fall. By looking at percentiles, we can get a feel for how our sample is spread along the variable of interest. Thus knowing that P_5=10.7 and P_{15}=20.5 tells us that (a) 5% of our sample have values less than 10.7, (b) 15% of the sample have values less than 20.5, and (c) 10% of the observations lie between 10.7 and 20.5. Similarly, knowing that P_{90}=26.4 indicates that only 10% of the observations have values greater than 26.4, while 90% lie below.

As an example, say that we want to identify the 80th percentile (P_{80}) for the 'number of children' variable in Table 6.1d. In other words, we want to find out the value below which 80% of the observations fall (this corresponds to (80/100) × 310 = 248 observations). Looking at the relative cumulative frequencies, we can see that 59.7% of respondents had two children or fewer (i.e., 95 + 60 + 30 = 185 respondents). To get to 80%, we must account for some of the respondents in the next category to make up for the shortfall. So, the value of the 80th percentile should be somewhere between two and three children; in fact, SPSS would tell us that 80% of the respondents had fewer than 3.18 children. Obviously, in this particular example, the exact value for the 80th percentile is not very illuminating, as '3.18 children' does not make much sense. This is because, being a **discrete variable**, the 'number of children' can only be measured in whole units; in this context, we might just as well have read directly from Table 6.1d the percentages of respondents with a maximum of two children and/or three children (which come to 59.7% and 93.5%, respectively). However, had we been dealing with a **continuous variable** for which we can assume (at least theoretically) any value within an interval (e.g., age, weight, height, income), then all percentiles would be meaningful (as all values, fractional or not, would be legitimate).

GROUPED FREQUENCY DISTRIBUTIONS

Sometimes it is not practical to construct frequency distributions based on the original values of a given variable. With a continuous variable or with a discrete variable that takes on many individual values, deriving a frequency distribution along the lines described previously is unlikely to be very informative. Consider, for example, the variable 'age' in the context of the Mongolian swimming survey. Being a continuous variable, in the extreme case, the age of each of the 310 respondents is likely to be different from that of everyone else's (unless two or more respondents happened to be born at exactly the same time!). A frequency distribution constructed under this (admittedly extreme) scenario would convey nil additional information

over the original data as there would be 310 individual frequencies – with each frequency being equal to 1! Even if we had measured the 'true' age only approximately (e.g., by asking respondents to indicate their 'age at nearest birthday'), there are still likely to be a very large number of values to deal with. For example, even if the survey was confined to adults (i.e., 18-year-olds and over) and the oldest respondent was only 87 years young, no fewer than 69 different yet perfectly legitimate responses (i.e., ages) would be contained in this range (including the youngest and oldest respondents). Again, a frequency distribution with 69 categories is not going to be much help for interpretation purposes.

As you may have guessed, the solution to problems posited by too many values is to group the latter into a smaller number of **classes**. Each class or **class interval** is defined by setting two **class limits** (an upper and a lower) within which a subset of the original values is subsumed. Having grouped the original values into appropriate class intervals (we will see how this is done shortly), one can proceed to derive a **grouped frequency distribution**, which shows how often the different classes of the variable concerned are encountered in the sample. Table 6.2 shows the grouped frequency distribution for our age example.

WARNING 6.2 When setting class intervals, make sure that they are not overlapping and that all cases can be assigned to a class.

Table 6.2 An example of a grouped frequency distribution

Age of respondent	Absolute frequency	Relative frequency	Cumulative frequency (%)
18–24	67	19.1	19.1
25–34	89	25.4	44.5
35–44	106	30.3	74.8
45–54	33	9.4	84.2
55–64	28	8.0	92.2
65–74	23	6.6	98.8
75 and older	4	1.2	100.00

Note that class limits are set in a *non-overlapping fashion* (i.e., *not* 18–25, 25–35, etc., but 18–24, 25–34, etc.); had overlapping limits been set, then a respondent who was exactly 25 years old could have been allocated to two different classes. This would violate a key classification principle, notably that of having **mutually exclusive** categories. Another key classification principle is that of **collectively exhaustive** categories, which says that you cannot have any leftovers at the end – every observation has to fit somewhere.

Class limits reflect the nature and extent of rounding off of the figures. The ages in the present example are all rounded off to the *nearest* year (since it was the age at the nearest birthday that was measured). Thus, a respondent between 24 and 24.5 years old would be placed in the first class interval, while a respondent between 24.5 and 25 would be counted in the second class interval. This implies that the 'true' boundary between the first and second class intervals is 24.5. Similar 'true' or 'exact' boundaries can be determined between the second and third intervals (34.5), the third and fourth (44.5), and so on. These are known as **true class limits**

(or simply 'class boundaries') to distinguish them from the (non-overlapping) **stated class limits** typically used in displays of grouped frequency distributions. In general, when data are recorded to the nearest unit (as is mostly done), the lower true limit of any given class lies one half-unit below the lower stated limit, and the upper true limit lies one half-unit above the upper stated limit. Thus, the true class limits of the first three intervals in Table 6.2 are 17.5–24.5, 24.5–34.5, and 34.5–44.5, respectively.

By using the true class limits, we can establish the **class width** (which shows the number of measurement units included in the class, i.e., the size of the class interval) and the **class midpoint** (which can be thought of as the 'representative' value of the class). Specifically, the class width is defined as

$$w = U - L$$

(where U and L are the *true* upper and lower limits, respectively), while the class midpoint (M) is defined as

$$M = L + 0.5w = U - 0.5w = \frac{U + L}{2}$$

In the example shown in Table 6.2, the width of the first class interval comes to 24.5 − 17.5 = 7 and the class midpoint is, therefore, 17.5 + (0.5 × 7) = 21. Similarly, the width of the second class is 34.5 − 24.5 = 10 and the class midpoint is 24.5 + (0.5 × 10) = 29.5. Just as a piece of useless information, if the width of the interval is odd (i.e., the interval contains an odd number of units), the class midpoint will be a whole number; if the interval width is even, the class midpoint will be a fraction.

Clearly, the distinction between stated and true class limits only makes sense for continuous variables. With *discrete* variables, the stated and true class limits are identical as only whole units are permissible as values. For example, if 'number of children' is grouped into three classes, 0–2, 3–4, and 5 or more, there are no legitimate values between the upper stated limit of the first class and the lower stated limit of the second class: someone can have *either* two *or* three children with no other possibility in between (see also the previous discussion on percentiles).

There are a few words of wisdom to bear in mind when setting up grouped frequency distributions. First, it goes without saying that there is inevitably a loss of information by grouping values together. Thus, from Table 6.2, it is impossible to identify the individual ages of the 89 respondents who are between 25 and 34 years old; some of them may be 25, others 27, and so on. We could take the class midpoint as a 'representative' value (which is mostly done in practice). Still, by doing this, we implicitly assume that responses are evenly distributed within the class interval (or that all responses are the same and equal to the midpoint value). If either is roughly the case, then there is no harm in using the class midpoint to represent a given class. On the other hand, if responses are heavily concentrated in the upper or lower parts of the interval, then using the class midpoint may result in a distorted picture of the data. Of course, we could go back to the original (i.e., ungrouped) data and see how responses are actually

distributed within the interval(s) concerned. However, this assumes that we *can* go back to the original data; that is, that access is not a problem (e.g., that we have not torn/burned our questionnaires and/or coding sheets while throwing a tantrum).

Second, for a given set of data, the loss of information becomes larger as the class width increases and the number of classes decreases. Had we, for example, grouped all ages between 18 and 34 in a single interval, there would be 156 respondents whose ages might vary by as much as 17 years from each other; at the same time, the number of classes would have been reduced from seven to six. So, the dilemma one faces is that there is substantial information loss with too few intervals, while with too many intervals, the purpose of summarizing the data is defeated.

Fortunately for you, there are some general guidelines to help you construct grouped frequency distributions that provide an effective summary of your data without incurring an unacceptable loss of information; these are gracefully outlined in Table 6.3.

Table 6.3 Guidelines for grouping data

Categories should be mutually exclusive.
Responses within categories should be similar.
Substantive differences in responses should exist between categories.
Categories should be exhaustive.
The number of categories should be neither too large nor too small (say, between 6 and 15 categories).
Preferably, categories should be of equal width.
Class intervals of 5, 10, or some multiple of 10 units tend to be easier to comprehend.
Open-ended categories are to be avoided if possible.

There are also a couple of weird formulae that may help you decide on the number of class intervals and their widths and, thus, demonstrate your unparalleled statistical grouping skills to friends and family alike. These formulae come in particularly handy when dealing with a continuous variable and a large sample (e.g., if you recorded the exact weight of 1,200 stray cats during your last holiday in Greece).

Deciding on the 'optimal' number of categories (c) can be accomplished by following Sturges's famous rule:

$$c=1+3.22(\log_{10}n), \text{ where } n= \text{ sample size}$$

Although this is not a hard and fast rule to be followed uncritically in all situations (see additional considerations in Table 6.3), it can be quite helpful. For example, the application of Sturges's rule suggests that we may have used too few class intervals in grouping the respondents' ages in Table 6.2, resulting perhaps in too crude a classification (about nine intervals would have been 'optimal' according to this rule – check it yourself if you don't believe us).

Having decided on the number of intervals, it is a simple matter to approximate the class width (w) you should be aiming for. This is given by:

$$w=\frac{max-min}{c}$$

(where *max* and *min* are the highest and lowest values, respectively). In our example, the youngest respondent was 18 years old and the oldest 87, so $w = (87 - 18)/9 = 7.66 \approx 8$ (based on nine intervals). Again, the formula for w should not be followed blindly as the resulting class width may be inconvenient to work with or not customarily used for the variable under consideration. For example, ending up with a class width of 18 to record income levels (e.g., 1,000–18,000, 19,000–36,000, 37,000–54,000, etc.) is not nearly as appetizing as, say, a class width of 20 (i.e., 1,000–20,000, 21,000–40,000, 41,000–60,000, etc.). As the more observant among you (or those still awake) may have noticed, the above class width has been calculated using the true class limits, as income is a continuous variable.

As Table 6.3 recommends, ideally all class intervals should be of the same width. This makes life much easier both for comparing the frequencies of the various classes and for graphical display purposes (to be discussed shortly). However, suppose the nature of the data is such that there are a large number of values relatively close to each other, coupled with a few values very far apart, so that one may be forced to use **unequal class intervals** (i.e., narrower intervals for the former set of values and wider intervals for the latter set). For example, in Table 6.2, the first (18–24) and second (25–34) intervals are of unequal width, with the former spanning a narrower range of ages than the latter.

Sometimes **open-ended intervals** are used at either or both ends of a grouped frequency distribution. However, as with unequal class intervals, they should be avoided if at all possible. These are typically employed when there are a few extremely small or extremely large values in the data in order to avoid having one or more class intervals with zero frequencies. A rather nasty problem with open-ended intervals is that their widths are not readily apparent. For example, looking at the (grouped) age distribution in Table 6.2, there is no way of knowing how close or far apart the four 'veteran' year-round swimmers (75+) are in terms of their age. The inability to determine class width leads to a second headache associated with open-ended intervals, namely identifying a 'representative' value (as a 'proper' class midpoint does not exist). One must either go back to the original (ungrouped) data (assuming this is feasible) or take a chance by guessing a midpoint (e.g., by assuming that the class width is the same as for the other intervals; however, if unequal class widths have been used for the latter, things become very messy indeed as there is no obvious choice).

HINT 6.2	As the grouping of data into categories results, inevitably, in some loss of information, calculation of summary measures based on ungrouped data is preferable.

By now you must be getting the rather uncomfortable feeling that, while grouped frequency distributions are helpful for *summarizing* a mass of data, there is a price to pay when it

comes to making *calculations* based on such data. Questions of distributions of observations within class intervals, unequal class widths, and the problems of open-ended intervals can all introduce error in subsequent analysis, such as the computation of summary measures (e.g., averages) for the variables concerned. While it is possible to calculate such summary measures based on grouped data (and there are plenty of horrible-looking formulae that you can find, if you must, in the Further Reading section), our position on this issue is simple: by all means, construct grouped frequency distributions to present your data in a digestible fashion. Even better, make good use of graphs and diagrams (advice on which will follow shortly) to improve the effectiveness of your presentation. However, always use the original (ungrouped) data when computing measures of average and/or dispersion so as to avoid potential inaccuracies introduced by the grouping process (we will discuss a variety of summary measures in Chapter 7).

GRAPHICAL REPRESENTATION OF FREQUENCY DISTRIBUTIONS

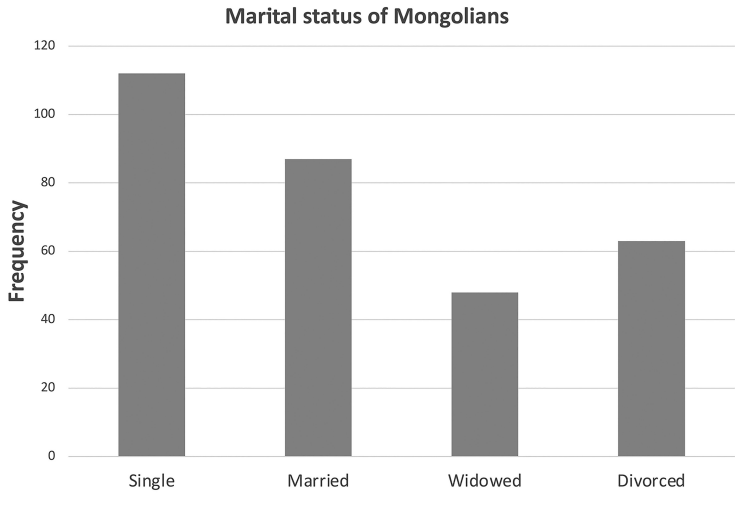
It is often said that 'a picture is worth a thousand words', which is undoubtedly the case with frequency distributions. Rather than relying solely on frequency tables to describe your data, you may decide to go for a **graphical representation**, if only to impress the reader with your amazing ability to draw pretty pictures or show off the extortionately expensive graphics package you recently bought for your computer. On a more serious note, using graphical representations is almost always a good idea, not least because the average casual reader pays more attention to graphs than to continuous text or tables (see also Chapter 15 on presenting the analysis).

WARNING 6.3 The heights of bars in a histogram are not proportional to frequencies unless all class intervals are of exactly the same width.

While the tabular form of representing frequency distributions can be used with any variable irrespective of its level of measurement (see the various examples in Table 6.1), things get slightly more complicated if you want to be 'artistic'. This is because some types of graphs are better suited to one level of measurement than others.

If you are dealing with a nominal or ordinal variable, you cannot go wrong with a **bar chart** or **pie chart**. In the former, values are represented by vertical or horizontal bars, the height of which is proportional to each value's absolute or relative frequency. In a pie chart, a circle is divided into slices representing the various values, with the size (i.e., area) of each slice being proportional to the value's *relative* frequency (pie charts are not really used to display absolute frequencies). Figure 6.1 shows the information on 'marital status' and 'education' from Tables 6.1a and b in bar chart and pie chart form, respectively.

Now that you have learned that 'bar' charts have nothing to do with your usual drinking hole and 'pie' charts are unconnected to pies (although most of them are also round), we can proceed to the graphical representation of interval and ratio data. This is altogether much more fun! With interval and ratio data, the choice of graph depends partly on whether the



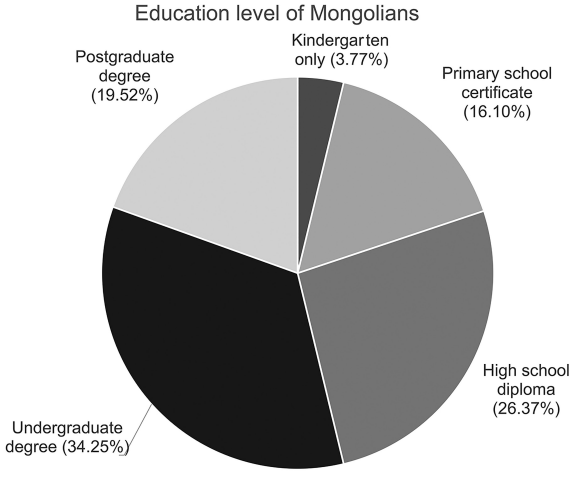


Figure 6.1 Examples of bar and pie charts

variable concerned is discrete or continuous and partly on whether the data is grouped or not. With a discrete variable that takes only a few values, you can use the same type of display as for nominal and ordinal variables (i.e., bar charts and/or pie charts); the five-point Likert scale used to register attitudes toward swimming in Table 6.1c is a prime example here. With discrete variables, which take a lot of different values, and with continuous variables, which are in the form of grouped frequency distributions, one has the choice between a **histogram** and a **frequency polygon**. In a histogram, absolute or relative frequencies are represented by areas in the form of bars. The total area under the histogram is 1.0 (or 100%), so the proportion of

the area over a range of values shows the proportion of observations falling into that range. The bars are erected at either the true limits of each class interval (the 'purist' approach) or the stated limits (the 'practical' approach); in both cases, adjacent bars 'touch' each other, unlike in bar charts where the bars are set apart. Also, unlike bar charts, the height of each bar is *not* proportional to a value's frequency; rather, it is the *area* of the histogram bar that reflects the latter. Consequently, unless all class intervals are of equal width (in which case heights and areas are in constant proportions), it is not meaningful to have a frequency scale on the vertical axis (incidentally, statistics texts often display histograms with a vertical (frequency) scale, but few make explicit the assumption concerning equality of intervals – so beware).

Now, if you connect the bars of the histogram at the midpoints of the class intervals, you will magically end up with the second type of graph, namely the frequency polygon. Figure 6.2 shows the histogram and frequency polygon for the (grouped) age data in Table 6.2; the figures at the top of each histogram bar indicate the percentage of cases falling within each age class (see Table 6.2).

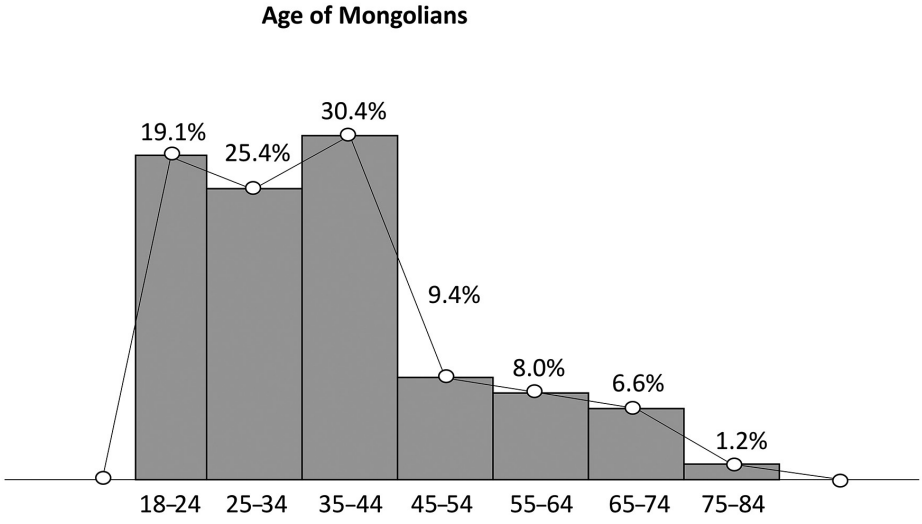


Figure 6.2 Examples of a histogram and a frequency polygon

Note that the open-ended interval at the top end of the age distribution in Figure 6.2 has been replaced by a closed interval (75–84) to enable the construction of the complete histogram (otherwise, the last class could not have been represented by a histogram bar). Note also that the frequency polygon is closed, by convention, by connecting the midpoint of the first (last) interval with a point on the horizontal axis one-half a class interval below (above) the lower (upper) true limit of the first (last) class (if this sounds a bit complicated, have another look at Figure 6.2 and things should become clear).

Histograms and frequency polygons are based on slightly different assumptions regarding the distribution of observations within each class interval. In a histogram, frequencies are

represented as being equally distributed over the range of a given interval, and changes from interval to interval are shown in a stepwise fashion. In contrast, with a frequency polygon, we assume that all cases within an interval are concentrated at the midpoint of the interval and, thus, changes from interval to interval are depicted as continuous. In fact, if we imagine a situation whereby (equal) class intervals become smaller and smaller, and the number of observations (i.e., sample size) becomes larger and larger, we would eventually end up with an infinitely small class interval and an infinitely large number of cases (see Figure 6.3(a)); that is, a truly **continuous frequency distribution**. The frequency polygon of the latter would then majestically show up as a smooth curve (see Figure 6.3(b)). This type of representation is often used to depict a frequency distribution in general terms graphically, and we will be making use of it in subsequent chapters.

HINT 6.3	If you want to graphically compare frequency distributions on the same scale, a plot of their frequency polygons will give a tidier picture than a plot of their histograms.

One advantage of the frequency polygon over the histogram is that when one wants to plot two (or more) distributions that overlap on the same base line (i.e., horizontal axis), the histogram can give a rather messy and confusing picture. In contrast, the frequency polygon, being simpler, usually enables a better comparison. An illustration indicating the final examination scores of two groups of meteorology students is shown in Figure 6.4; note that since the data are recorded in equal intervals, a frequency scale is given on the vertical axis of the histograms and frequency polygons to aid interpretation.

Histograms provide you with a wonderful opportunity to explore and understand your sample in relation to the variables of interest. In fact, the histogram shows the shape of the distribution of values in a variable. It thus can reveal the central location (i.e., 'What do the values center around?') as well as the dispersion (i.e., 'How spread out are the values?') of your data. In other words, they tell us which values are 'mainstream' and which ones are less common. For example, in Figure 6.2, ages within the range of 24 and 44 years seem to be quite common, whereas ages between 75 and 84 are rather rare. Moreover, histograms show whether values are symmetrically distributed around the most common value(s) or whether they tend to cluster toward the lowest/highest value. Most of the values in the distribution of Mongolians' age in Figure 6.2 are clustered toward the left side, indicating that 74.8% of the Mongolians surveyed are between 18 and 44 years old. Finally, histograms are useful for identifying potential **outlier values** (i.e., extremely low or high values). Overall, the properties of the histogram render it one of the most valuable tools for visually inspecting data, which very often plays a central role in deciding the most suitable approach for analysis.

By now, you must be getting extremely anxious to learn what kind of graph one may use to depict *cumulative* frequency distributions; after all, the various types of graphs we have discussed so far only apply to absolute or relative frequencies. One way of doing this is by means of the **cumulative frequency polygon** (what a mouthful), which differs from the 'normal' frequency polygon in two respects. First, instead of plotting points corresponding to absolute frequencies, we plot points corresponding to cumulative frequencies. (You would never have

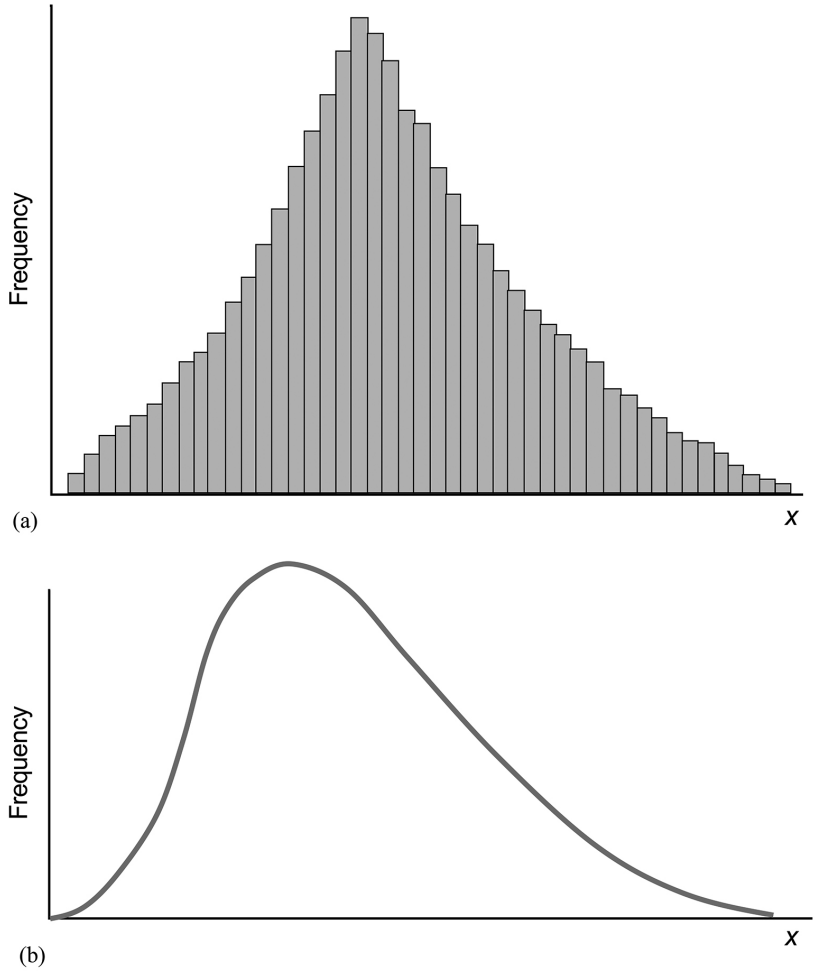


Figure 6.3 Graphical representation of a continuous frequency distribution. (a) histogram; (b) frequency polygon

guessed, would you?) Second, instead of plotting points above the midpoint of each class interval, we plot points above the upper true limit of each class (so that the graph shows the number of observations falling above/below particular values).

Figure 6.5(a) shows the cumulative frequency polygon for the 'number of children' data in Table 6.1d. Note that since the 'number of children' is a discrete variable, the stated and true class limits are identical (remember?) and, given that the data have not been grouped, each class interval has a width of one unit. Note also that the general trend of the cumulative frequency polygon is progressively rising; there are no inversions or setbacks. This is because

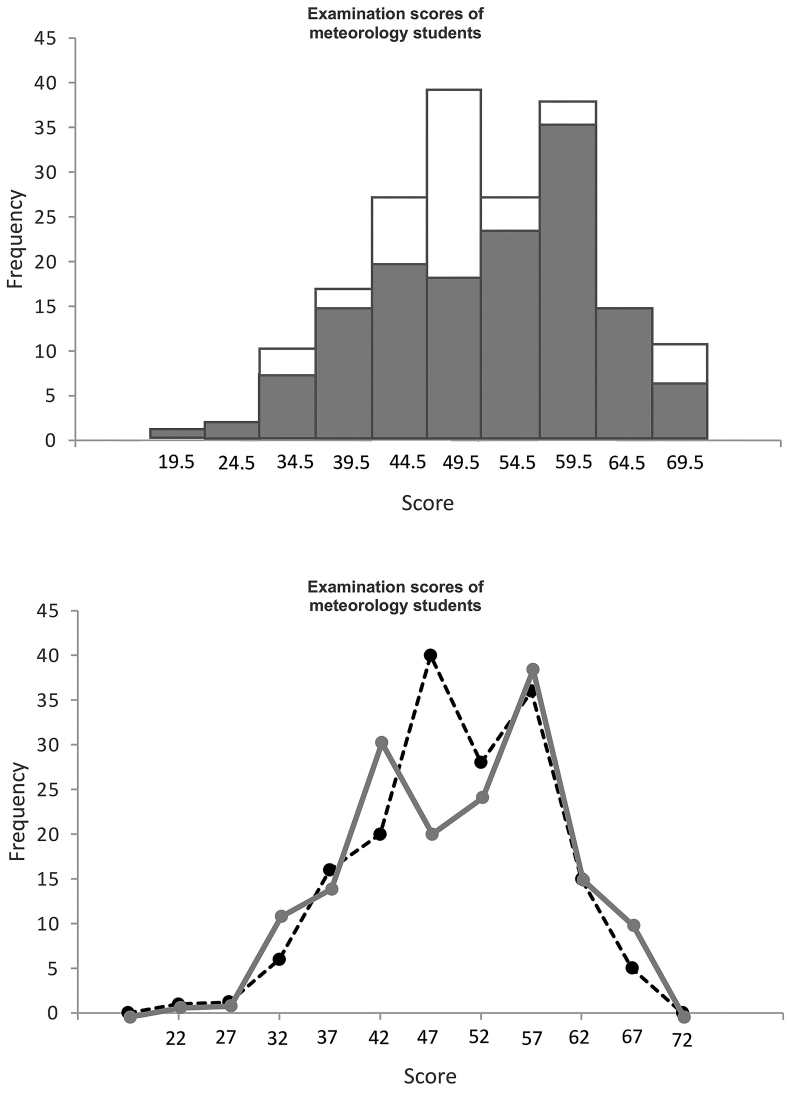


Figure 6.4 Graphical comparison of frequency distributions

all (non-cumulative) frequencies are positive except the odd zero frequency (e.g., as in the case of more than five children).

A more useful type of graph (yes, we know you have had enough of graphs – this *really* is the last one) is the **ogive**, which plots the cumulative relative frequencies against the variable's values (see Figure 6.5(b)). You can use the ogive to (a) name your dog (hey, Ogive, here boy!), (b) read off directly the percentage of observations less than any specified value, and (c) determine the value below which a specified percentage of observations lies. Thus in Figure 6.5(b), if you were interested in finding out the percentage of respondents who had two children at most, you would follow arrow A; this would tell you that almost 60% of all respondents had

between zero and two children. On the other hand, if you wanted to know how many children 90% of the respondents had, following arrow B would give you the answer 'fewer than three children'. Thus, the ogive can be used as an alternative (and quick) way to look at percentiles.

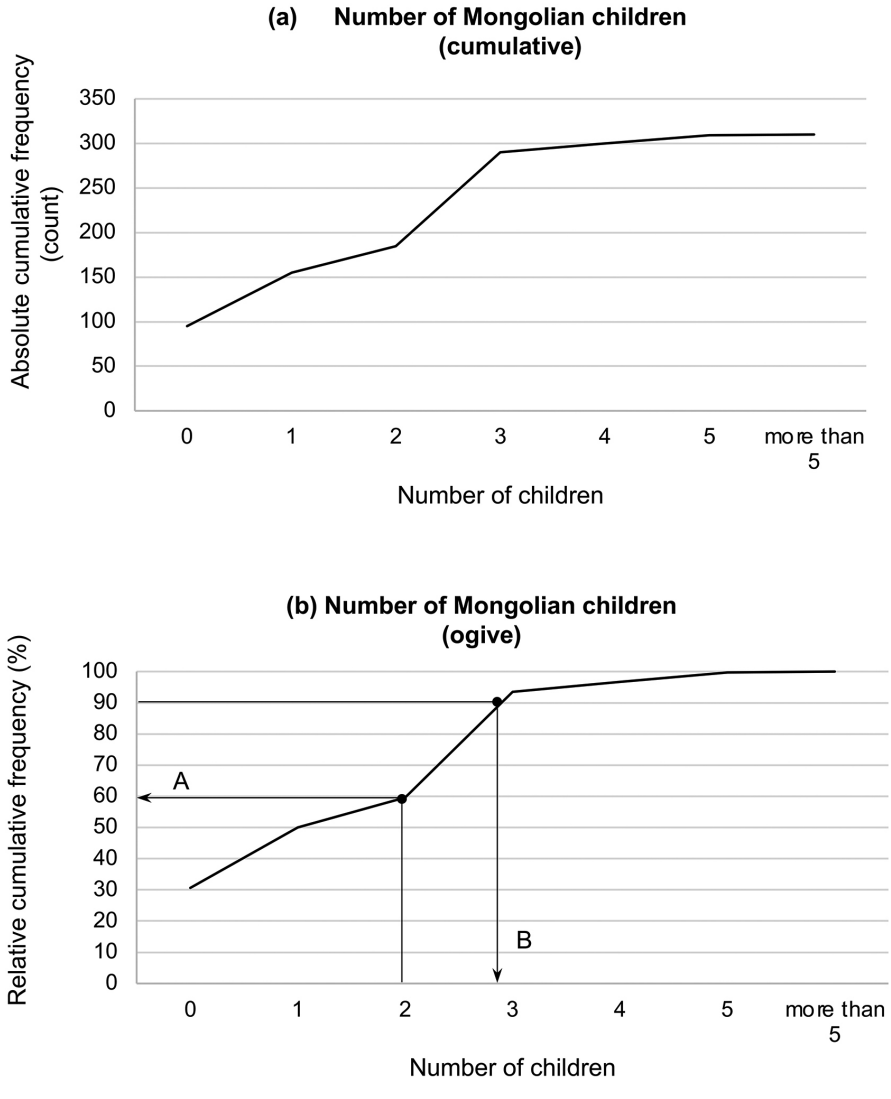


Figure 6.5 Cumulative frequency polygon (a) and ogive (b)

Most statistical analysis packages have facilities for producing the various kinds of frequency distributions and graphical displays we have discussed in this chapter (and usually many more). This means that the only thing you have to do is input your data and the computer does

the rest. Table 6.4 shows the kind of output you can expect from the various statistical packages available out there. Here we are using SPSS on data relating to attitudes toward winter swimming from a sub-sample of 100 Mongolians (see five-point Likert scale in Table 6.1c).

Table 6.4a Example of frequencies output: statistics

Attitude toward winter swimming		
N	Valid	98
	Missing	2
Percentiles	25	3.000
	50	4.000
	75	5.000

Table 6.4b Example of frequencies output: attitude toward winter swimming

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Strongly disagree	7	7.0	7.1	7.1
	Disagree	6	6.0	6.1	13.3
	Neither agree nor disagree	19	19.0	19.4	32.7
	Agree	36	36.0	36.7	69.4
	Strongly agree	30	30.0	30.6	100.0
	Total	98	98.0	100.0	
Missing	−999.0	2	2.0		
Total		100	100.0		

Finally, when using the graphics options of statistical software or employing a specialized visualization package, it is worth bearing in mind some general guidelines regarding graph construction (as sole reliance on 'default' settings may result in boring or even weird-looking graphs!). A set of such guidelines is generously provided in Table 6.5.

Table 6.5 Guidelines for graph construction

Every graph should have a title (usually below the diagram).
The general arrangement of a graph should proceed from left to right.
Where possible, quantities should be represented by linear magnitudes as areas or volumes are more likely to be misinterpreted.
The horizontal scale for a graph should read from left to right and the vertical scale from bottom to top. Low numbers on the horizontal scale should be on the left and low numbers on the vertical scale should be toward the bottom.
Both the horizontal and the vertical axis should be clearly labeled. In graphing frequency distributions, it is customary to let the horizontal axis represent values and the vertical axis frequencies.
Preferably, the zero point should be included in the graph (e.g., for a frequency scale on the vertical axis); if this is not practicable, a break in the axis should be used.
While the distance on each axis chosen to serve as a unit is arbitrary, it can affect the appearance of a graph; a rough 3:5 height to length ratio seems to work well in most cases.
The coordinate axes for a graph should be sharply distinguished from any other lines included.
When representing percentages, it is helpful to emphasize in some distinctive way the 100% line (or any other basis of comparison).
The graph should not be overly cluttered with too much information and the key information should be made as clear as possible.

SUMMARY

In this chapter, we took the first steps in data analysis by attempting to describe a data set. We started by looking at the concept of a frequency distribution and the various forms it can take. We then focused on the process of grouping data and the potential problems that may be encountered along the way. Finally, we let our artistic talents run wild by examining different types of graphical displays, putting a bit more emphasis on the pivotal role of histograms. We are now ready to start some serious work, namely to derive summary measures that can be used to represent the properties of a set of data succinctly.

QUESTIONS AND PROBLEMS

1. What is the role of data description?
2. What is the difference between an absolute and a relative frequency distribution?
3. Why is it not possible to calculate a cumulative frequency distribution for nominal-level variables?
4. What are percentiles?
5. What is the difference between a discrete and a continuous variable?
6. What is the difference between ‘true’ and ‘stated’ class limits?
7. Which factors should be considered when constructing grouped frequency distributions?
8. Why should unequal class intervals and open-ended intervals be avoided, if possible?
9. When would you use a frequency polygon over a histogram?
10. Do you think Ogive is a good name for a dog?

FURTHER READING

Cleff, T. (2014). *Exploratory Data Analysis in Business and Economics*. Cham: Springer International Publishing. A guide to different perspectives for exploring the structure of data. It complements the material covered in this as well as the next chapter.

Few, S. (2012). *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Burlingame, CA: Analytics Press. Enough said!

Johnson, R. B. & Vogt, W. P. (2015). *Dictionary of Statistics and Methodology: A Non-Technical Guide for the Social Sciences*, 5th edition. London: Sage Publications. It's about time we sent you to one of these, as the amount of terminology you must digest is increasing all the time; this book should help you maintain your sanity!

7
Can you use few numbers in place of many to *summarize* your data?

CHARACTERIZING FREQUENCY DISTRIBUTIONS

In the previous chapter, you learned more than you ever wanted to know about frequency distributions. Here we will build upon your knowledge by showing you how frequency distributions can be conveniently described and compared to one another with reference to some key properties they possess. These properties are important because they allow us to calculate some **summary measures** capturing the essential characteristics of different distributions. By using summary measures, we can condense the information contained in the individual values, thus making the interpretation of the data much more manageable. Moreover, instead of graphically displaying distributions (as was done in Figure 6.4 in Chapter 6), we can compare their respective summary measures. Finally, as we will see in Chapter 8, we can use summary measures derived from our sample data to make inferences about the population from which the data have come.

Consider the three histograms in Figure 7.1 (if you don't remember what a histogram is, you probably did not pay much attention to Chapter 6, and this is not a nice thing to do). These represent the frequency distributions of IQ scores of economics students enrolled at three universities, A, B, and C. Although they are identical in shape, they are markedly different in terms of the central values about which the observations in each distribution appear to concentrate. In other words, the **central location** (or 'central tendency') of each distribution is different. Measures of central location are also called **averages** and reflect 'middle' points in the sense that they are near the center of the distribution of values. In our example, distribution A has a lower average than distribution B, and B has a lower average than C. The familiar **arithmetic average** or **mean** (i.e., the value obtained by adding together a set of values and dividing by their number) is one such measure; however, there are several other types of averages, as we will see shortly.

Now consider Figure 7.2, showing the distributions of the IQ scores of biochemistry students in the same three universities. Although they all have the same average, in university A the IQ scores are more closely concentrated around the average than in university B; the same applies to the IQ scores in B in relation to those at university C. Thus, distributions A, B, and C differ in terms of their **variability** (or 'dispersion'); that is, the degree of clustering around a central

value. In general, if all the observations are close to the central value (i.e., the average), their variability will be less than if they tend to depart markedly from the central value.

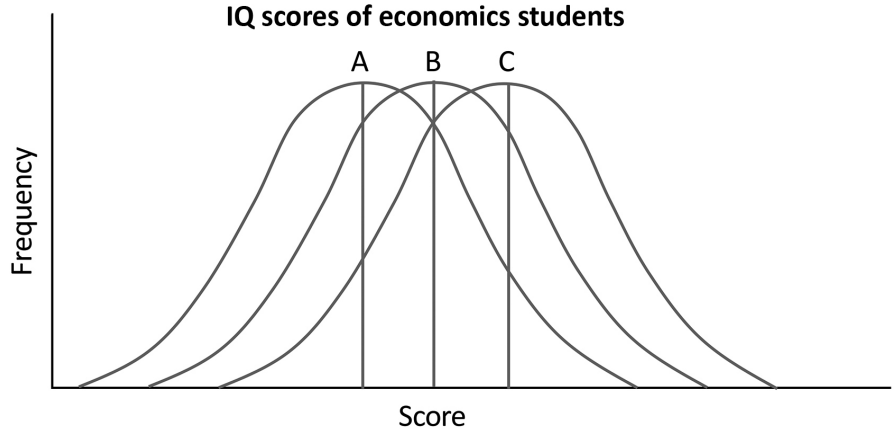


Figure 7.1 Three frequency distributions identical in shape but with different averages

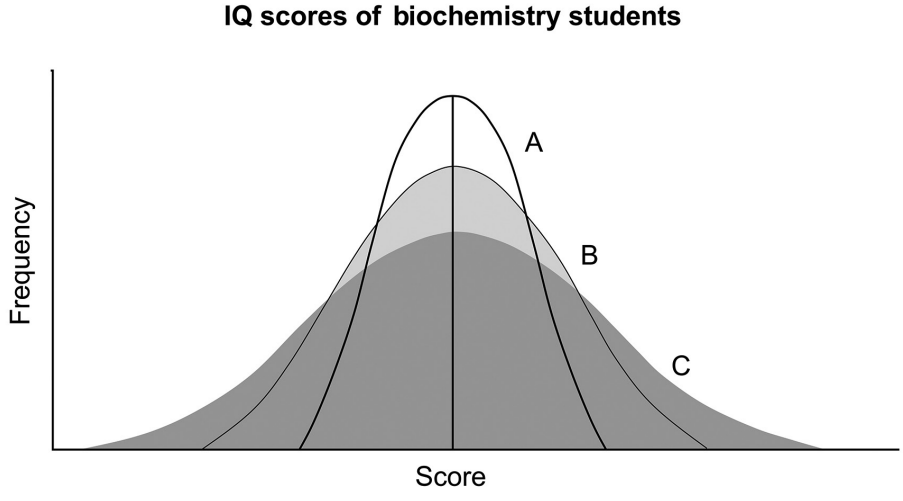


Figure 7.2 Three frequency distributions with the same average but differing in variability

A third way in which frequency distributions can differ is in terms of their **skewness**, which reflects the 'symmetry' (or lack thereof) in the distribution. Figure 7.3 illustrates this mysterious property by comparing the distributions of IQ scores of engineering students in universities A, B, and C. Note that distribution B is **symmetrical**, for if we were to 'fold' it over about the average, we would find that it had the same shape on both sides. In contrast,

distributions A and C are **asymmetrical** in relation to their central location, the shape to the left of the average being different from the shape to the right. More specifically, distribution A is **positively skewed** as the larger frequencies tend to be concentrated toward the low end of the variable and the smaller frequencies toward the high end (i.e., the 'tail' of the distribution extends to the right). The opposite holds true for distribution C, which is **negatively skewed** (the larger frequencies are toward the high end of the variable and the smaller frequencies toward the low end; that is, the 'tail' is to the left).

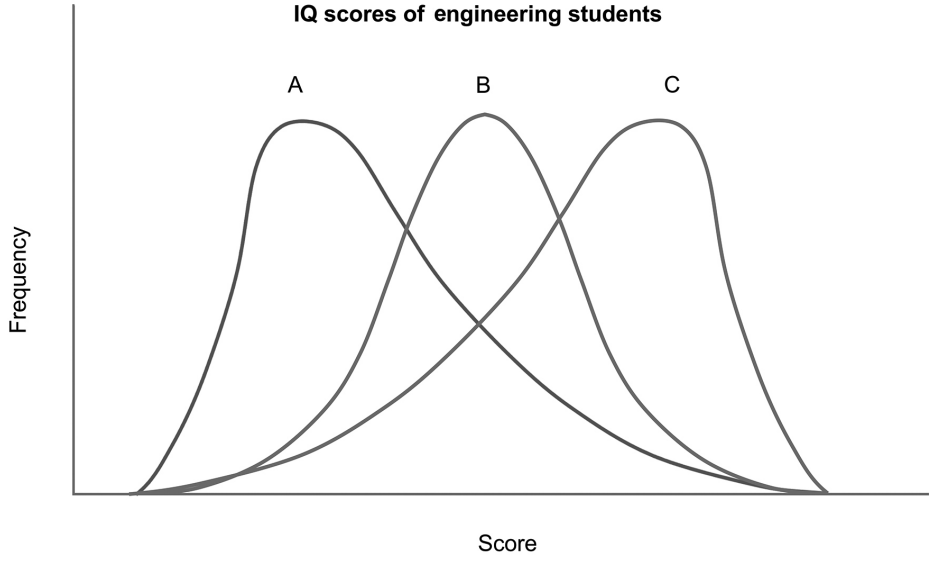


Figure 7.3 Three frequency distributions differing in skewness

Finally, look at Figure 7.4, which shows the distribution of IQ scores of physics students in the three universities. These distributions differ in terms of their **kurtosis**, which describes the 'flatness' or 'peakedness' of one distribution in relation to another. If one distribution is more peaked than another, it is said to be more **leptokurtic** (compare A to B), and if it is less peaked with values more frequently being clustered in the tails of the distribution, it is said to be more **platykurtic** (compare B to C). It is conventional to speak of a distribution as leptokurtic (platykurtic) if it is more (less) peaked than a particular type of distribution known as the 'normal distribution' (to be discussed later in this chapter). The latter is referred to as **mesokurtic**, which means that it falls somewhere between leptokurtic and platykurtic shapes. Note that, although people may often refer to it as the amount of 'peakedness' of a distribution, kurtosis is essentially related to the tails and not the peak of the distribution. As such, a high value for kurtosis (i.e., platykurtic distribution) means that extreme values tend to be more frequent, whereas a low kurtosis value (i.e., leptokurtic distribution) indicates that extreme values are not very common.

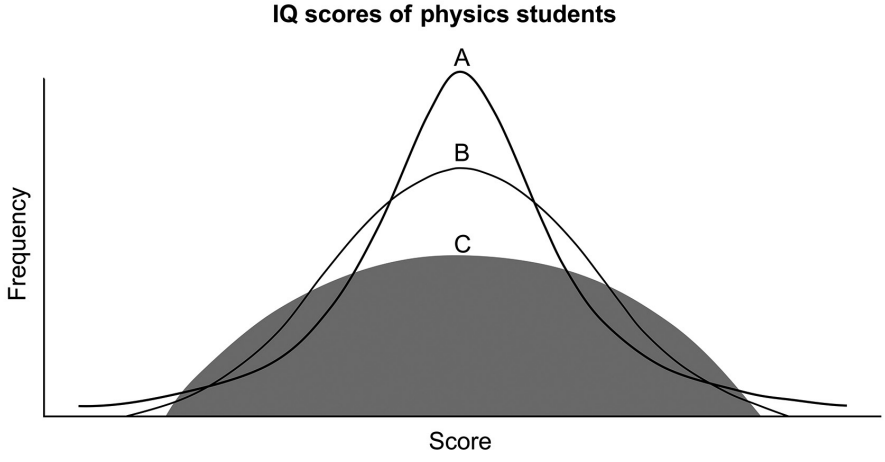


Figure 7.4 Three frequency distributions differing in kurtosis

Taken collectively, central location, variability, skewness, and kurtosis can provide a very detailed and yet efficient description of the nature of a frequency distribution. Statisticians have spent their lives developing a whole range of measures that can be used to represent these properties (well, somebody has to do it). Here we will concentrate on the most widely used summary measures and refer you to the Further Reading section for the more exotic ones (and, believe us, there are plenty of them!). Note that, in what follows, we will be solely concerned with **sample statistics**; that is, with measures of central location, variability, and so on based on data obtained from a sample; issues relating to corresponding **population parameters** will be considered in detail in Chapter 8. Moreover, for reasons already explained in Chapter 6, only summary measures based on original (i.e., ungrouped) data will be discussed; methods of calculating equivalent statistics based on data in the form of grouped frequency distributions can be found in the Further Reading section (but read Hint 6.2 again in Chapter 6).

When looking at specific measures of average, dispersion, and so on, it is useful to bear in mind some general criteria for evaluating alternative measures. While the answer to the question 'What makes a good descriptive statistic?' is just as difficult to give as to 'What makes a good husband or wife?', Table 7.1 lists a number of desirable properties of summary measures, the role of which should become clearer as you fight your way through this and the next chapter (if you need a break, you can always compile a list of 'desirable properties' for your ideal partner – but don't show anyone!).

MEASURING CENTRAL LOCATION

Our main concern here is with what we may call 'typical' or 'average'; that is, with computing a *single* value that is in some way representative of the entire set of observations for the variable concerned. Indeed, the average value is usually close to the point of greatest concentration of

Table 7.1 Criteria for evaluating summary measures

A summary measure should be *appropriate for the level of measurement* of the variable in question; e.g., if it involves summation or averaging of values, it should only be applied to metric data.
It should ideally be *based on all the observations*; i.e., its calculation should make use of the entire set of individual values.
It should be *simple to comprehend, easy to calculate, and expressed in algebraic terms.*
It should be *unique rather than multivalued*; i.e., given a set of data, it should take on *one* value only.
It should be *resistant to outliers*; i.e., not severely affected by extreme values.
It should ideally be *equal to actual data values* and should be *expressed in the original measurement units*; i.e., those of the variable in question.
It should be *consistent*; i.e., the sample statistic in question should approximate the true value of the corresponding population parameter as the sample size approaches infinity.
It should be *unbiased*; i.e., if several samples are taken, and the sample statistic in question is calculated, then the average of these sample statistics should equal the population parameter.
It should be *efficient*; i.e., the statistic in question should exhibit minimum variation across different samples.

the measurements and may therefore be thought to typify the whole set. We shall now look at the three most important central location measures, namely the mode, the median, and the mean.

The mode

Undoubtedly, the simplest measure of central location is the *mode*, which is defined as the most frequently occurring value. As it is the value corresponding to a frequency and does not involve any algebraic manipulation of the individual observations, the mode can be applied to any variable irrespective of its level of measurement. However, the mode is particularly useful for summarizing nominal data, as other measures of central location (to be discussed shortly) assume at least ordinal-level data. Note that the mode is not itself a frequency but the value associated with the highest frequency; this is often an issue of confusion, so be careful.

Table 7.2 shows both the ease with which the mode can be determined and the limitations of the mode as a measure of average; the example shown relates to a variable indicating the favorite color for socks among a small sample of casino owners.

In Panel A, the mode is four, reflecting a clear preference for shocking pink socks among casino owners (these guys have taste, don't you doubt it!). Here the value of four occurs 11 times, much more frequently than any other value, so there are no problems. Things get a bit more complex in Panel B, where all the values occur equally frequently. For all sock colors, the corresponding frequency is four, so identifying *a* modal value is not possible (incidentally, a distribution such as the one in Panel B, in which all values occur with equal frequency, is notoriously known as a **uniform distribution**). In Panel C, we have a different problem as eight casino owners prefer dawn scarlet socks and another eight shocking pink socks (all other colors are preferred by fewer people). Thus, this distribution has two modes (two and four) and is commonly referred to as a **bimodal distribution**; other than reporting both modes, there is little one can do here. Clearly, for distributions with two or more modes (i.e., for

multimodal distributions), it would be misleading to talk about the mode as a measure of 'central' location.

WARNING 7.1 The mode is an actual value and not a frequency of occurrence.

The possibility of having multiple modes exposes another disadvantage of this measure, namely its instability under sampling: taking several samples from a bimodal distribution with population modes M1 and M2 is likely to result in some samples having M1 as their mode and other samples M2. Thus, the mode could fluctuate considerably from sample to sample (which is a pain in the neck, to say the least). To depress you even more, have a careful look at Panels A and D in Table 7.2. Both panels have a mode of four (i.e., shocking pink is the most preferred color in both cases). However, thinking of these two cases as being equivalent would not be representative of reality, as in Panel A there is a very clear preference for shocking pink socks, whereas in Panel D shocking pink and dawn scarlet are fighting tooth and nail to win the hearts of our fellow casino owners. This is because the mode only cares about what the most frequent value is without bothering with the rest of the values. The fact that the mode provides no information whatsoever about the relative importance of the central value exposes another disadvantage of this measure, namely its potential lack of comparability across samples.

WARNING 7.2 The mode can be considered as a measure of central location only for distributions that taper off systematically toward their extremes.

Note that with metric (i.e., interval and ratio) data, it is sometimes possible to have two values with 'the highest' and equally occurring frequencies and still be able to calculate a single mode (and all this without violating Warning 7.2). Consider the following set of values, indicating the number of (pairs of) socks bought by our 20 casino owners in the past week:

8, 9, 10, 10, 11, 11, 11, 12, 12, 13, 13, 13, 13, 14, 14, 14, 14, 15, 15, 16

Here, the values 13 and 14 both occur with a frequency of four, which is greater than the frequency of occurrence of the remaining values. As the two 'modal' values are adjacent, the mode may be arbitrarily taken to be the arithmetic average of the two values (i.e., in this case $(13 + 14)/2 = 13.5$). It goes without saying that this 'averaging' procedure is only legitimate if *adjacent* values of a *metric* variable are involved. Where two non-adjacent values occur with equal frequencies, which are higher than the remaining frequencies, the distribution must be treated as bimodal and Warning 7.2 applies. Similarly, with non-metric (nominal and ordinal) data, averaging of values (whether adjacent or not) is *not* legitimate for reasons explained in detail in Chapter 3 (think about the outcome of averaging sock colors, and you will immediately see why).

Note that, even with **unimodal distributions**, the mode may not be very informative of the structure of the data because the 'most frequently' occurring value may not occur very often.

Table 7.2 The mode in different situations

Variable		Frequency	Mode
Favorite sock color			
Virgin white	= 1	1	
Dawn scarlet	= 2	3	
Sickly green	= 3	3	
Shocking pink	= 4	11	4
Deathly black	= 5	2	
Favorite sock color			
Virgin white	= 1	4	
Dawn scarlet	= 2	4	
Sickly green	= 3	4	Uniform distribution
Shocking pink	= 4	4	
Deathly black	= 5	4	
Favorite sock color			
Virgin white	= 1	2	
Dawn scarlet	= 2	8	
Sickly green	= 3	1	2 and 4
Shocking pink	= 4	8	
Deathly black	= 5	1	
Favorite sock color			
Virgin white	= 1	2	
Dawn scarlet	= 2	7	
Sickly green	= 3	2	
Shocking pink	= 4	8	4
Deathly black	= 5	1	

For example, in the following set of values, the mode is eight, but you could not possibly designate it as the *typical* value (with a straight face, that is!).

1, 2, 3, 6, 7, 8, 8, 9, 12, 24, 25, 26, 38, 44, 52, 53, 54, 112, 313, 414

All in all, the mode does have its share of problems (and who doesn't, after all?) and would not score highly on most of the evaluation criteria listed in Table 7.1. However, as it is the only measure of average that is available for nominal data, we should not complain too loudly. Nevertheless, use it with care.

WARNING 7.3 Make sure that you have at least ordinal-level data before you even think about calculating a median!

The median

The **median** is another measure of central location and is defined as the value above and below which one-half of the observations fall. In other words, the median is the value of the 'middle' case when all individual observations have been arranged in rank order. Needless to say, calculating a median for nominal data is totally meaningless as nominal values cannot be ordered. (See Chapter 3 if you are not clear on this point.)

WARNING 7.4 Do not confuse the position of the median with the value of the median. The former indicates where the median is located in relation to the individual (ordered) observations, while the latter indicates what the median value is.

When the number of observations (i.e., sample size) is odd, one value is always in the middle, and this is the median. For example, consider the following ordered set of seven values:

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

The median is the fourth value; that is, 3. When the number of observations is even, then there are two middle values and it is customary to think about the median as being halfway between the two middle values. Consider, for example, the following ordered set of eight values:

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

Here the median lies between the fourth and fifth values; that is, between 4 and 3. What happens now depends on the level of measurement of the variable in question. If the variable concerned is an interval or ratio, then it is usual to take the arithmetic average of the two middle values as the median, in this case (4 + 3)/2 = 3.5. On the other hand, if the variable concerned is only ordinal, averaging of values is not legitimate and we can only say that the median lies between the fourth and fifth values (i.e., between 3 and 4) in the ordered set of observations (although, in practice, 'illegal' averaging of the two middle values with ordinal data is commonplace).

In general, the location of the median, given an ordered set of values, is found by computing $(n+1)/2$, where n = sample size. Thus, in our first example above, the median is positioned at the $(7+1)/2 = 4$th value, while in our second example the median is located at $(8+1)/2 = 4.5$th value (i.e., midway between the fourth and fifth values). Note that the formula $(n+1)/2$ does not give the *value* of the median, only its *location*.

Sometimes people have difficulty determining the median from a frequency distribution because they confuse the location and actual value of the median. Table 7.3 recasts our two examples above in frequency distribution form (the variable involved being a five-point

Table 7.3 The median in different situations

Variable	Frequency	Median
Quality rating of The Dodgy Chicken		
Excellent = 5	2	
Good = 4	1	
Passable = 3	1	3
Poor = 2	2	
Atrocious = 1	1	
Quality rating of The Dodgy Chicken		
Excellent = 5	3	
Good = 4	1	
Passable = 3	1	Between 3 and 4
Poor = 2	2	
Atrocious = 1	1	
Quality rating of The Dodgy Chicken		
Excellent = 5	1	
Good = 4	1	
Passable = 3	3	?a
Poor = 2	1	
Atrocious = 1	1	
Quality rating of The Dodgy Chicken		
Excellent = 5	2	
Good = 4	2	
Passable = 3	2	?a
Poor = 2	1	
Atrocious = 1	1	

?a = *You* tell us!

ordinal scale describing the perceived quality of food at The Dodgy Chicken, a notorious local takeaway). In Panel A, you should immediately see that the median rating is 'passable' (the value of the fourth case), while in Panel B, the median rating is between 'passable' and 'good' (the values of the fourth and fifth cases, respectively). If you cannot see these medians *at a glance*, then please go over the examples again. Then check that you've *really* got the hang of it by going to Panels C and D and determining the medians for the data shown.

The median has two advantages as a measure of central location. First, unlike the mode, it is based on the entire distribution of a variable (as one needs to rank the observations in order to locate the median). Second, the median is not affected by extreme values; this makes it an attractive measure even for metric data, particularly when **outliers** (i.e., extreme or atypical values) are involved. For example, the medians for the following three sets of data (A, B, and C) are identical (we leave it up to you to decide what the median value is!). That said, you had better be extra cautious when you look at medians from different samples or different variables

because, even though medians are quite robust against outliers, they can be very sensitive to additional values. Compare data set A to D or simply delete the last two cases from data set C, and you will catch the drift. Finally, note that calculating the median does require looking at the whole distribution of values, but it hardly captures the variability in the data set. As mentioned above, the median value in data sets A, B, and C are identical, but the variability or dispersion is materially different.

A: 5, 6, 9, 10, 15, 16, 20
B: 5, 6, 9, 10, 15, 16, 200
C: 1, 1, 1, 10, 20, 400, 6,000
D: 5, 6, 9, 10, 15, 16, 20, 21

The mean

The **mean** is your familiar arithmetic average and is defined as the sum of a set of values divided by their number; as its computation involves algebraic manipulation of the individual data values, the mean is an appropriate measure of central location for metric data only. As an example, consider the following data indicating the weight in kilograms of six sumo wrestlers (after they had a light breakfast):

175, 233.7, 199.6, 304.7, 254.9, 388

The mean weight of these healthy lads comes to

(175+233.7+199.6+304.7+254.9+388)/6 = 1,555.9/6 = 259.3 kg

In general, given n observations with values $X1$, $X2$, $X3$, …, Xn, the mean is found by applying the following ridiculously simple formula:

$$\bar{x} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} = \frac{\sum X_i}{n}$$

The mean has a number of properties worth noting. First, it makes full use of the data available in that its calculation is based on all the individual data values. The flip side of this is that it can be greatly affected by one or two extreme values (outliers); as such, the mean is said to be a **non-resistant measure**. For example, while the median in both sets of data below is 13, the mean is 14.1 in A and 24.1 in B.

A: 8, 9, 12, 13, 18, 19, 20
B: 8, 9, 12, 13, 18, 19, 90

Another rather annoying characteristic of the mean is that it can take fractional values, even when the variable involved is discrete (and thus should only take integer values – see Chapter 6). For example, if data set A above represented the number of children of seven families in Rabbitsville, Pennsylvania, the average number of children comes to 14.1 – which raises the interesting biological question as to what 0.1 of a child actually means!

A related issue is that the mean can take a value that does not reflect any of the individual values of the variable in question; for example, if the Domingo family has two cars, the Pavarotti family 18, and the Caruso family 16, the mean number of cars comes to (2+18+16)/3=12, a value that is not representative of any of the families.

But not all is bad with the mean – quite the contrary. First, the mean is unique in the sense that the total sum of (signed) deviations around it is zero (i.e., $\Sigma(X_i - \bar{x}) = 0$). Second, and related to the first point, the sum of negative deviations from the mean is always equal to the sum of positive deviations from the mean (use data sets A and B above to verify that this is so). This gives the mean a special interpretation as a balance point (or 'center of gravity') for the distribution of individual values, as negative deviations are exactly offset by positive deviations. Third, the sum of **squared deviations** around the mean, $\Sigma(X_i - \bar{x})^2$, is smaller than the sum of squared deviations around any other value; this property is used in the calculation of measures of dispersion, as we will shortly see. Without wishing to go into details, you should note that this **least-squares** criterion is very important and underpins many statistical techniques. Finally, the mean is much more stable than the median or mode over repeated sampling. (It exhibits less variation from sample to sample compared to other measures of central location.)

Having looked at the characteristics of three different central location measures, it is useful to briefly consider them in conjunction with each other. This is both because no individual measure is ideal on all the criteria listed earlier in Table 7.1 and because different measures can furnish complementary perspectives on the same set of data (assuming, of course, that the level of measurement for the variable concerned allows the calculation of all three measures).

Take a look at the fictitious distribution shown in Figure 7.5. Its mode is the point on the horizontal axis, which corresponds to the 'peak' of the curve (this represents the most frequently occurring value). Its median is a point on the horizontal axis where the ordinate divides the area under the curve into two equal parts; half the area of the curve falls to the right of the median value and half to the left. Lastly, its mean is also a point on the horizontal axis, reflecting the center of gravity of this distribution; if we were to cut out this distribution and try to balance it on the edge of a ruler, the point of balance would be the mean (now, don't get carried away and deface your book just to try this out!).

If you are lucky enough to have a unimodal, symmetrical distribution, then the mode, the median, and the mean will all coincide at the center of the distribution (see Figure 7.6(a)). In contrast, with a positively skewed distribution, the mean will have the largest value and the mode the smallest value, with the median in between (see Figure 7.6(b)). With a negatively skewed distribution, on the other hand, the mode has the largest value and the mean the smallest; again, the median will be between the mean and mode (see Figure 7.6(c)).

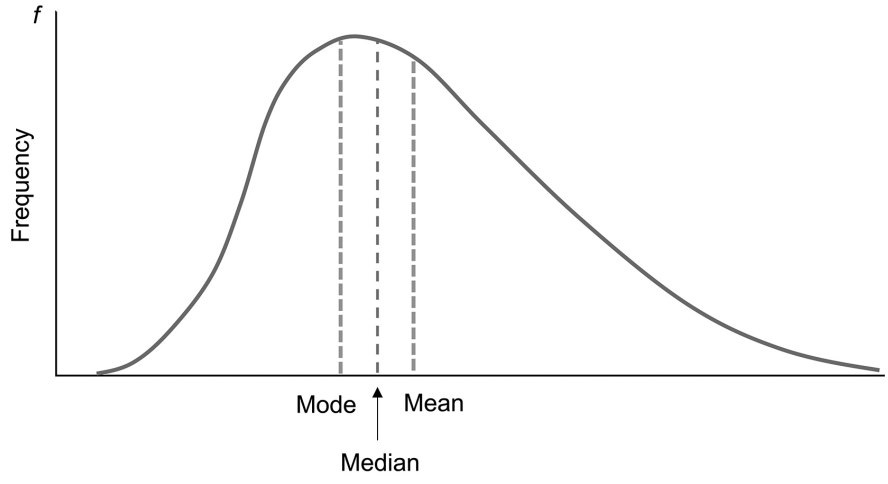


Figure 7.5 The mean, median, and mode in a distribution

HINT 7.1	Use several central location measures in conjunction with one another to provide a comprehensive picture of your data.

The fact that the *relative* magnitudes of the mode, median, and mean differ according to the shape of the distribution suggests that we can use them to get some rough indication of skewness (there are also formal indices to assess skewness, as we will see later on). In this context, we usually rely on the relationship between the median and the mean alone, as the mode is the least reliable of all measures of central tendency (particularly in small samples). Thus, if the mean is larger than the median, the distribution is positively skewed; if the mean is smaller than the median, the distribution is negatively skewed. Information regarding the skewness of a distribution is more important than you may initially think. This is because the shape of a distribution often determines the analytical techniques that can be legitimately applied to the data and guides the interpretation of the results. Just to give a simple example, at the end of every semester, we check the distribution of students' grades in our (advanced) course Data Analysis for Gamblers in order to identify potential skewness. If the distribution of grades is positively skewed, the majority of students score below the mean, whereas if the distribution is negatively skewed, most students score above the mean, in which case we should perhaps make the test a bit harder (just kidding, this never happens; in reality, we throw all exam scripts in the air and look at their distribution when they land. This widely used practice speeds up grading tremendously!).

Okay, so what measure of average should be used when? First, make sure that you *do* have a choice; that is, that your level of measurement allows you to pick among the three measures! Assuming this is so, do *not* use the mean (at least not on its own) if your distribution is markedly skewed, particularly when one or more very extreme values are at one side of the

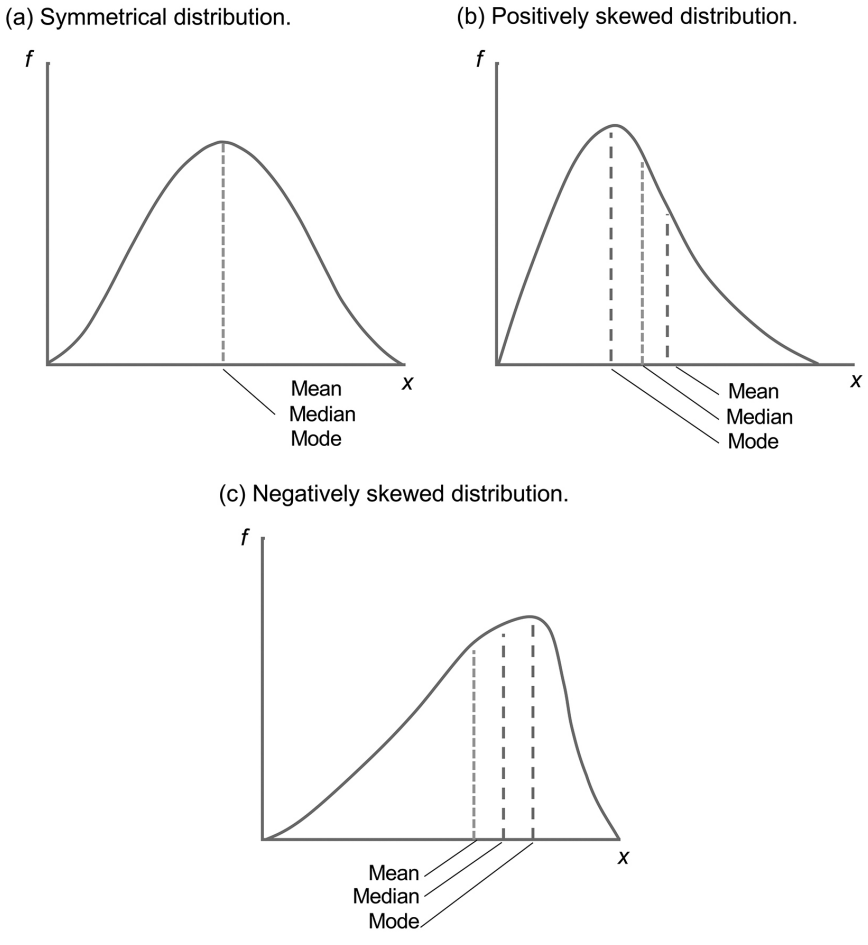


Figure 7.6 Central location measures in different distributions

distribution; opt for the median instead. With roughly symmetrical distributions, the mean is probably the best measure to use, although (to be on the safe side) it would not do any harm to report the median as well. Finally, as far as the poor mode is concerned, it should not be reported on its own but only used to 'spice up' a picture painted by the median and/or mean (unless, of course, you are dealing with a nominal variable and the mode is your only choice).

At this point, you should note that there are many more central location measures available with such colorful names as the 'midextreme', the 'midhinge' (nothing to do with door support), the 'trimmed mean', the 'windsorized mean', the 'geometric mean', and the 'harmonic mean', to name but a few. The references in the Further Reading section should be seriously consulted by those of you who wish to impress other people at parties by subtly displaying your knowledge of different measures of average.

Having an effective way of measuring central location is an important component in describing different sets of data but not always enough to make relevant decisions. Imagine, for example, that you want to buy a new robotic vacuum cleaner and go online to check customers' quality ratings across products in order to help you make your decision (see Figure 7.7). Assuming that customers' ratings are largely symmetrically distributed for each product, you can easily rely on average ratings to decide what the best product is, right? Well, not quite. Would you randomly choose between Robosuck and Autoblow just because the two products have the same average (Panel A), or would you readily opt for Robosuck just because its mean quality rating is higher than that of Funneltron (Panel B)? Figure 7.7 indicates that examining measures of central location in isolation can be problematic. The following section explains why this is the case in more detail.

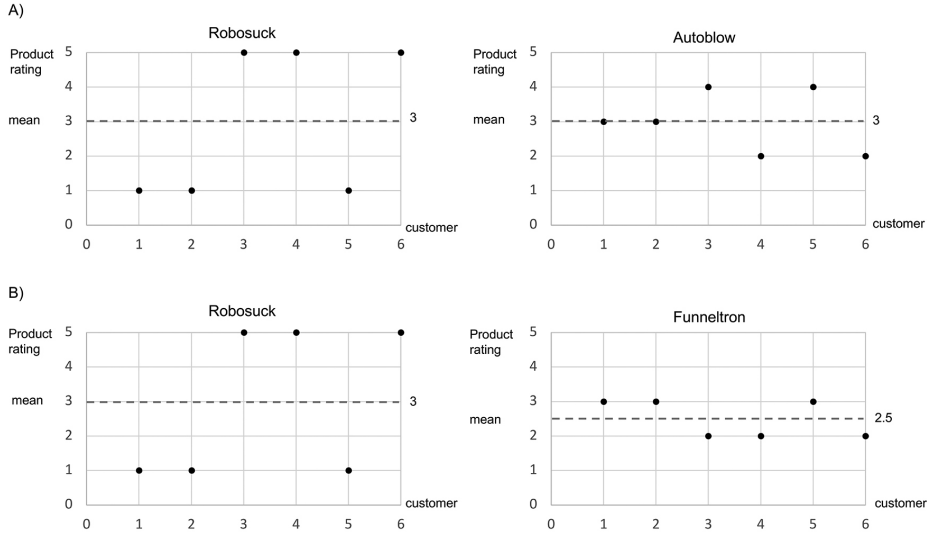


Figure 7.7 Comparing product quality ratings

MEASURING VARIABILITY

While measures of central location identify 'average' values in a set of data, they tell us absolutely nothing about the extent to which the individual values are similar to or different from one another. However, in many cases, information concerning variability among a set of values is more valuable than information about the 'average'. Consider the following two data sets, indicating the number of lost working days (due to hangovers) by a sample of seven French wine tasters:

October 2019: 4, 5, 5, 5, 5, 5, 6
November 2019: 0, 1, 3, 5, 5, 9, 12

In both months, mean = median = mode = 5 days (i.e., all their averages are the same). Despite this, however, we can see that the pattern of days lost in November is substantially different compared to that in October. In October, all wine tasters lost between four and six days due to hangovers (an unfortunate professional hazard). In November, the number of days lost ranged from zero (by the 'employee of the month') to 12 days (no doubt by the 'bad employee of the month'). Clearly, the extent of variation in November is much greater than in October.

Measures of variability provide a complementary picture to central location measures by summarizing the degree of **dispersion** (or 'scatter' or 'spread') in a variable. They all equal zero when there is no variation and increase in value with greater dispersion in the data. We will look at the following variability measures in this section: the index of diversity for nominal variables, the range and interquartile range for ordinal variables, and the variance and standard deviation for interval and ratio variables.

The index of diversity

The **index of diversity** (D) is a measure of variability for nominal data and is based on the frequencies associated with the variable in question. It involves (a) determining the relative frequencies (i.e., proportions) in each category of the variable, (b) squaring them, (c) summing the squares, and (d) subtracting the resulting sum from 1 (phew!). For example, using the data on sock color preferences in Table 7.2 (Panel A) gives us

$$D = 1 - [(1/20)2+(3/20)2+(3/20)2+(11/20)2+(2/20)2] = 1-0.360 = 0.640$$

This measure shows the degree of *concentration* of the cases in a few large categories (as squaring relative frequencies affects the large frequencies much more than smaller ones) and tends to be zero if almost all cases fall into the same category. It is at a maximum when each category occurs just once; however, its maximum value depends partly on the number of categories. Thus, D cannot be used to compare nominal variables with different numbers of categories. However, given a number of categories c, one can calculate the maximum value of the index, which is $(c - 1)/c$, and then compare it to the actual value obtained from the data. In our example, the maximum diversity is $(5 - l)/5 = 0.80$ (since there are five categories in our variable – see Table 7.2). Given that the actual value of D is 0.640 for our data, we can conclude that there is quite a bit of variation among casino owners in terms of sock color preferences.

Before we move on, you should note that measures of variability for nominal data are not used very often (nobody really knows why). Indeed, most standard texts in statistics do not even mention them and, what is infinitely worse, many computer packages do not compute them! Yes, that's right – *you* have to do the work! While you can find out more about them by following the excellent sources in the Further Reading section, please do not hold us responsible for bumping into such weird measures as the dreaded 'standardized entropy index' or the much-feared 'fragmentation statistic'.

The range and interquartile range

The range and interquartile range are measures of variability particularly useful for ordinal variables (although also applicable to metric variables). The **range** is simply defined as the difference between the highest and lowest values in the data; when there is no variation in the variable, the range is obviously zero. While it is extremely easy to calculate, the range does not use all the information in the data and, thus, is very much affected by extreme values. For example, the range for both data sets below comes to 19, although arguably the fluctuation in values in A is greater than in B.

A: 1, 2, 5, 9, 11, 20
B: 1, 1, 1, 1, 1, 20

For large samples, the range is an unstable measure of variability as its fluctuation from sample to sample increases as the sample size increases. Further, there is a higher chance of obtaining extreme values with large samples than with small samples; what this implies is that, for the same variable, ranges calculated on different sample sizes are not directly comparable.

To deal with the above problems, it makes sense to opt for a modified range, which is established by eliminating a certain percentage of the extreme values at the two ends of the distribution (thus providing a more stable and reliable measure of variability). One such modified range (there are others) is the **interquartile range** (yes, it does sound like a testing ground for ballistic missiles!). It is calculated by (a) finding the values corresponding to the 75th and 25th percentiles (i.e., the upper and lower quartiles, Q3, Q1) and (b) subtracting the latter from the former (i.e., forming the difference Q3 − Q1). Any decent computer package should be able to provide you with Q3 and Q1 (unless you prefer going back to Chapter 6 and following our superb guidelines for manually calculating percentiles). As a quick example, in the following data set, the range comes to 39, while the interquartile range comes to 18 (and if you cannot tell how we have arrived at these figures, you should brush up on your percentiles).

5, 10, 25, 28, 44

The interquartile range, along with the upper and lower quartiles (Q3 and Q1), can give you a good idea about the dispersion in a data set. The **boxplot** in Figure 7.8 (also known as a 'box-whisker-plot') shows the distribution of the age of casino owners split by quartiles (note that the median is the second quartile Q2 or, put differently, the 50th percentile). The smaller the body of the interquartile range (the 'box') and its legs (the 'whiskers'), the less the variability in the data. The interquartile range is also very helpful in identifying outliers (extreme scores) or potential mistakes that might have occurred during data input. For example, the boxplot in Figure 7.8 shows that Casino Owner Number 34 is 75 years old and considered to be an outlier in relation to the age of the rest of the casino owners, while there is something strange with Casino Owner Number 20, who seems to be 148 years old! Unless this casino owner is some sort of supernatural creature from Mars doing business in Texas, this has to be a typing error. Needless to say that with ordinal or interval variables with fixed levels (e.g., five-

or seven-point scales), spotting potential errors is more straightforward as any value exceeding the legitimate range would be unacceptable.

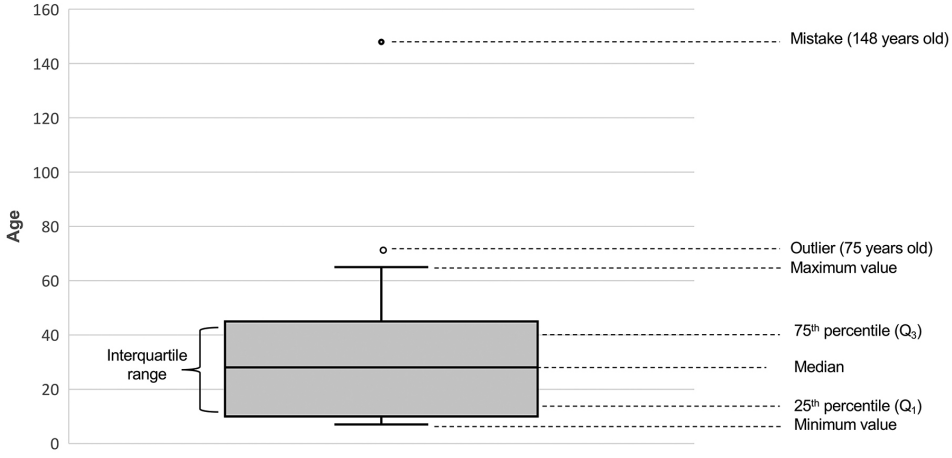


Figure 7.8 Boxplot showing the quartiles and interquartile range

Generally speaking, numerical outliers are considered to be values that are larger than Q3+1.5×interquartile range or smaller than Q1−1.5×interquartile range. In the example of Figure 7.8, Q1 is 26 and Q3 is 44; thus, the interquartile range is 18 (Q3 − Q1). Any value larger than 71 (i.e., 44+1.5×18) or smaller than 17 (i.e., 44−1.5×18) will be flagged as an extreme value. Note that this is a preliminary examination of what could be perceived as an extreme value and that there are way more sophisticated methods of identifying outliers that (thankfully for you and us!) will not be discussed in depth in this book.

There are several other measures of variability for ordinal data, and, if you will not be satisfied until you learn everything there is to know about the 'coefficient of quartile variation' or 'Leik's ordinal consensus measure', the Further Reading section is eagerly awaiting you.

The variance and standard deviation

These are by far the most widely used and highly regarded measures of variability – you cannot open any statistical text and find no reference to the variance and standard deviation (it's like opening *Cosmopolitan* magazine and finding no reference to sex). They are widely used not only as descriptive measures but also in connection with inferential statistics (see Chapters 8 and 9). So, what are these apparently amazing dispersion measures?

The **variance** involves (a) subtracting the mean from each individual value – that is, forming all deviations from the mean, (b) squaring these deviations, (c) summing them, and (d) taking

their average. Thus, the variance is the average squared deviation from the mean. Using the horrible notation introduced previously, the variance is defined as

$$\frac{\sum(X_i-\bar{x})^2}{n}$$

where *Xi* are the individual values, $\bar{x}$ is the mean, and *n* is the number of observations. Unfortunately, when calculated using sample data, the above formula provides a **biased estimate** of the population variance; specifically, it shows a systematic tendency to underestimate the population variance (why this happens need not concern you here). As this is clearly undesirable (see Table 7.1), we use the following slight modification to calculate the sample variance *s*2; this gives an **unbiased estimate** of the population variance, so everybody is happy:

$$s^2=\frac{\sum(X_i-\bar{x})^2}{n-1}$$

As an example, let us calculate the variance for the following small set of data, indicating the number of domestic cats reported lost in 12 police stations in Wigan, Lancashire:

37, 59, 71, 75, 78, 78, 81, 86, 88, 92, 95, 96

The mean number of cats reported lost is

(37+59+71+75+78+78+81+86+88+92+95+96)/12 = 78

so the variance is

[(37−78)2+(59−78)2+(71−78)2+…+(96−78)2]/(12−1) = 3,082/11 = 280

One clear advantage of the variance is that it takes into account each and every piece of data available, as all the individual values are used in its calculation. Second, it expresses variation in individual values in relation to a measure of central location, namely the mean; when the values are scattered widely about the mean, the variance is large, and when the values are concentrated near the mean, the variance is small. Third, the variance emphasizes large deviations from the mean as these, when squared, increase much more than small deviations; however, while large deviations from the mean indicate greater variability, squaring the deviations implies that extreme values can have a disproportionate impact, making the variance a non-resistant statistic.

Also on the downside, the variance is unfortunately expressed in *squared* units of the variable in question. In the above example, the variance of 280 does not refer to 'number of cats' but 'number of cats squared'! (Don't ever try to square a cat in reality; this would definitely be animal cruelty.) As this makes interpretation rather messy, it is usual to report the square

root of the variance: this is the famous **standard deviation**. Thus, in our example, the variability across police stations would be described in standard deviation terms as being 16.7 cats $(\sqrt{280})$. In general, the sample standard deviation is defined as

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (X_i - \bar{x})^2}{n-1}}$$

where all the symbols are as before.

While describing the dispersion of a variable using the standard deviation is preferable to using the variance because the original measurement units are retained, it may still be difficult to interpret standard deviation values directly. For example, is a standard deviation of £100 large or small? It may be considered large if we are measuring the weekly amount spent on sweets but small if we are looking at the incomes of statistics professors. In cases such as this, it may be more helpful to look at the standard deviation *in relation* to the mean for the variables in question; in this way, we can make comparisons across different variables. A measure that accomplishes this beautifully is the **coefficient of variation**, *CV*, which is defined as the standard deviation divided by the mean, or

$$CV = \frac{s}{\bar{x}}$$

Thus, if the mean weekly amount spent on sweets is £125 and the mean weekly professorial salary for statisticians is £1,300, then the respective coefficients of variation come to 100/125 = 0.80 and 100/1,300 = 0.077, respectively; thus, the amounts spent on sweets are relatively much more variable than the incomes. Note that the coefficient of variation is independent of the unit of measurement (i.e., it is 'unitless'). This is because both the standard deviation and the mean are in the same units, which cancel out when their ratio is formed. This makes it particularly useful in case we want to make comparisons across distributions of values that correspond to variables with different measurement scales (e.g., variables measured with five- vs. nine-point interval scales).

The standard deviation is also used in conjunction with the mean to compute what is known as **standard scores** (or '*z*-scores'). Sometimes the information provided by a raw value has little meaning unless seen in relation to the other values. For example, if you score 63% on a Creative Accounting exam, whether you did really well or badly would depend on how the rest of the class did. If the mean grade was 12%, then you are flying (and all your colleagues will hate you!). If, on the other hand, the mean grade was 92%, there is little cause for celebration. A standard score allows us to place an individual case in context; it shows us how many standard deviations above or below the mean a certain observation lies. To calculate a standard score from raw scores (what we eloquently call a process of **standardization**), we (a) subtract

the mean from the value of the individual observation (i.e., find the deviation from the mean) and (b) divide this difference by the standard deviation; thus

$$z_i = \frac{X_i - \bar{x}}{s}$$

Standard scores have a mean of zero and a standard deviation of one; however, the shape of the distribution of standard scores is identical to the distribution of raw values. That is because standardization is simply a process of re-scaling the data to facilitate interpretation and by no means influences how the values are actually distributed (this is important, as we will see in a later section). A positive standard score indicates that the observation in question has a value greater than the mean, while a negative score indicates the opposite. Thus, if you converted your Creative Accounting grade to a standard score and found $z = 1.25$, you could brag to your friends that you scored 1.25 deviations above the mean grade awarded to the class as a whole; conversely, if you were unlucky enough to find that $z = -1.25$, you can be ready to change the subject if a conversation concerning grades ever comes up.

Another important use of standard scores is that they enable comparisons to be drawn between different distributions. For example, if you wanted to compare your performance in the Creative Accounting exam with that in the Unfair Taxation exam, standard scores should do the trick. If you scored 63% in the former and 70% in the latter, does that mean that Unfair Taxation is your better subject? Not necessarily – it depends on the mean score and the standard deviation of scores in each of the two exams. If the mean grade for the class for Creative Accounting was 52% with a standard deviation of 8%, while the respective figures for Unfair Taxation were 75% and 4%, your standard scores would be

Creative Accounting:	(63−52)/8 = 1.375
Unfair Taxation:	(70−75)/4 = −1.25

Thus, in relation to the other students taking the two exams, you did much better in Creative Accounting than in Unfair Taxation.

Standard scores are also helpful if one wants to combine several variables to create a **composite scale** (or 'index'; see Chapter 4 on variable transformations to refresh your memory on this). If the variables in question have very different variances or are measured in different units, then, obviously, combining (e.g., adding) their raw scores is inappropriate. In such cases, the variables should first be standardized and the standard scores used to form the desired index.

There is yet another important use of standard scores; we will see what that is after a brief excursion into the wonderful world of skewness and kurtosis.

MEASURING SKEWNESS AND KURTOSIS

Although for most practical purposes one can get a reasonable idea of skewness and kurtosis by looking at a histogram or frequency polygon and examining the relative magnitudes of the mean and median, it is sometimes desirable to compute some formal measures. A measure of skewness or kurtosis will be close to zero if the distribution of values is symmetric and mesokurtic (i.e., like the nice bell-shaped curve of the normal distribution discussed later in this chapter). Note that with sample data it is extremely unlikely that *exactly* zero values will be obtained because of sampling fluctuations. Positive values for skewness indicate a positive skew, and positive values for kurtosis reflect a more leptokurtic distribution.

Note that people often confuse the notion of kurtosis with that of variance, erroneously treating flat or peaked distributions as indicating more or less variance in the data. This is not the case and, without getting into a complex discussion about it, the best way to illustrate it is through Figure 7.9, which shows the frequency distribution of IQ scores of 150 students who finished reading this entire book and another 150 who did not. The two distributions have the same variance, yet their histograms obviously have very different shapes, the left being rather leptokurtic while the right is completely flat (platykurtic).

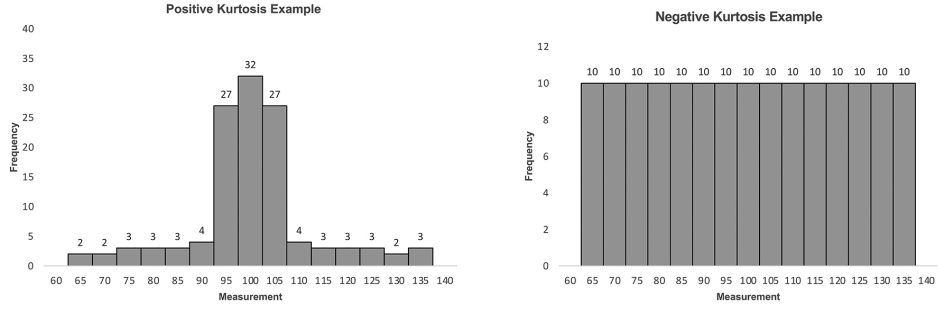


Figure 7.9 Example of leptokurtic (positive kurtosis) and platykurtic (negative kurtosis) distributions

As the formulae for skewness and kurtosis are rather scary, and given that decent computer packages can provide a **coefficient of skewness** and a **coefficient of kurtosis** upon request, we will restrain ourselves and will not reproduce the relevant formulae here. Instead, we will present you with an example of SPSS-generated output showing a variety of summary measures (Table 7.4); the variable in question shows the age (in years) of a sample of 80 snake charmers.

We can see that the mean and median ages are practically the same, but the modal age is lower; this suggests that the distribution may be somewhat positively skewed. This is indeed the case, as indicated by the positive coefficient for skewness; moreover, the negative coefficient of kurtosis suggests that the distribution of ages is rather platykurtic. What is perceived to be too much of a skewness or kurtosis in a distribution is (as always) not very straightforward

Table 7.4 An example of computer-generated descriptive output

Statistics		
Age		
N	Valid	73
	Missing	7
Mean		40.890
Median		41.000
Mode		30.000
Std. Deviation		9.379
Variance		87.960
Skewness		.359
Std. Error of Skewness		.078
Kurtosis		−.279
Std. Error of Kurtosis		.304
Range		41.000

and depends on the researcher's judgment about the specific situation at hand. However, a common guideline in statistical literature is that values within +/− 2 are acceptable.

CHEBYSHEV'S THEOREM AND THE NORMAL DISTRIBUTION

After this brief encounter with skewness and its mate kurtosis, we can now return to more important issues. The first thing to consider is a wonderful theorem developed by a clever Russian named Chebyshev, well over 100 years ago. This enables us to calculate for *any* set of (metric) data the *minimum* proportion of values that can be expected to lie within a specified number of standard deviations from the mean. According to this theorem, at least 75% of the individual values in a set of data can be expected to fall within two standard deviations from the mean, at least 88.9% will fall within three standard deviations from the mean, and at least 96% of all values will be within five standard deviations from the mean. More generally, **Chebyshev's theorem** tells us that given a set of observations, the probability is at least (1−1/k2) that an individual observation will take a value within k standard deviations from the mean (where $k > 1$).

To illustrate the application of the above theorem, let us use the snake-charmer data in Table 7.4. We know that the mean age (rounded) is 41 and the standard deviation about nine years (also rounded). If we set $k = 2$, then the probability is at least (1−1/22) = 0.75 that the age of a snake charmer will be within two standard deviations (i.e., within 2 × 9 = 18 years) from the mean (which is 41). In other words, we can expect that at least 75% of the snake charmers will be between 41−18 = 23 and 41+18 = 59 years old. Isn't this great? We can, of course, select any value for k and compute the relevant intervals for our variable. Note that, given the same mean and a specified value for k, the intervals will be narrower the lower the standard deviation (try the example as above but using a standard deviation of six years, and you will

see what we mean). This is because, for a set of data with a small standard deviation, a larger proportion of the values will be concentrated near the mean than for a data set with a large standard deviation.

The key point to take home from all this is that by using the standard deviation in conjunction with the mean, we are able to make some *probabilistic* statements (i.e., 'informed guesses') about the position of individual values relative to the mean. Note that we are able to make such statements without making any assumptions about the specific form or shape of the distribution; indeed, Pafnuty Chebyshev's theorem (no, we did not just make up his first name) can be applied to any distribution of values (and, indeed, this is one of its advantages). The question now is: can we do any better if we know something about the form of the distribution concerned?

The answer is yes. Consider the distribution shown in Figure 7.10. This is a symmetric, bell-shaped, mesokurtic (neither flat nor peaked) distribution known as the **normal curve**. (Yes, this is the normal distribution – say hi to it!) Its mean, median, and mode all coincide, and each half of the distribution is a mirror image of the other half; moreover, the tails of the curve tend toward but never actually touch the horizontal axis (this 'you can reach but never touch' quality makes the normal distribution *asymptotic*). A key characteristic of this distribution is that approximately 68% of the individual values fall within one standard deviation from the mean, approximately 95% of individual values fall within two standard deviations from the mean, and approximately 99.7% of all values fall within three standard deviations from the mean.

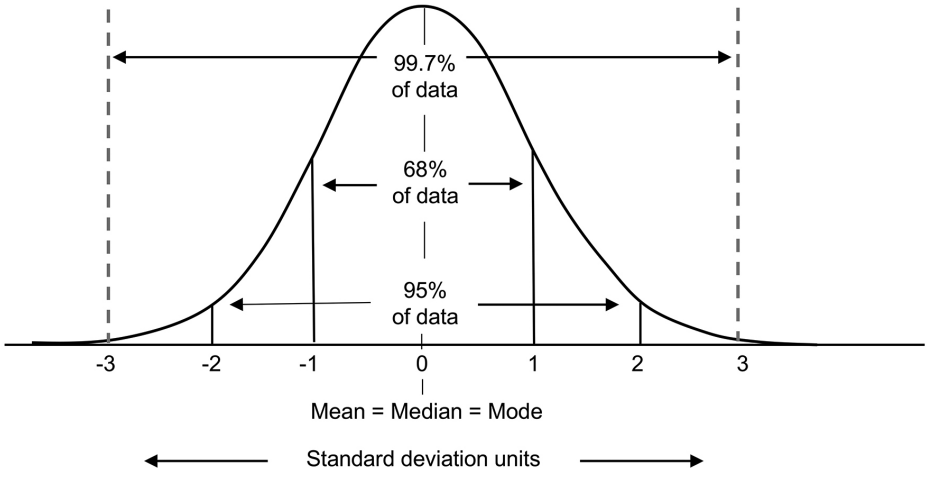


Figure 7.10 The normal distribution

Let us now use our snake-charmer data and calculate the proportion of ages within two standard deviations from the mean, assuming a normal distribution for the age variable. As before,

the mean is 41 and the standard deviation nine years. Thus the desired interval is defined by 41−2(9) = 23 and 41+2(9) = 59, which, assuming a normal distribution, indicates that approximately 95% of the snake charmers are between 23 and 59 years old. If we compare this interval with that established using Chebyshev's theorem, we see that, while they are identical in size, we now include 95% as opposed to 'at least 75%' of the individual cases (see above). In other words, whereas before we could only say that there is at least a 0.75 probability that an individual value will lie within two standard deviations from the mean, we are now able to attach a 0.95 probability to the same estimate. We are able to do this because we know the *form* of the distribution involved (in this case, the normal), and we can use this information to our advantage in order to make more precise probabilistic statements (i.e., more informed guesses). Aren't you highly impressed?

The normal distribution is just one of several distributions with known properties. Other well-known distributions include the **binomial distribution**, the **hypergeometric distribution**, the **Poisson distribution**, the **chi-square distribution**, the **t-distribution**, and the **F-distribution**, to name but a few. In each of these distributions, the probabilities associated with all possible values are known, and for this reason they are often referred to collectively as **probability distributions**. While a detailed discussion of different probability distributions is way beyond the scope of this text (we do not want to make enemies of you!), there are some basic things you ought to know.

First, we can distinguish between probability distributions that have been developed for discrete variables and those for continuous variables (see Chapter 4 if you have forgotten the difference between the two). The former are known as **discrete probability distributions** and include the binomial, hypergeometric, and Poisson distributions. The latter are referred to as **continuous probability distributions** and include the normal, *t*-, chi-square, and *F*-distributions.

Second, some of these distributions are symmetrical about their mean (e.g., the normal and *t*-distributions), whereas others are not (e.g., the *F*- and chi-square distributions). This is an important distinguishing characteristic, the implications of which will become evident in Chapter 11.

Third, the exact form of some probability distributions (e.g., the chi-square distribution) is dependent upon what is known as **degrees of freedom**, which is a rather complex concept to explain in simple terms (so we will not even try!). Suffice it to say that distributions dependent upon the degrees of freedom are really families of distributions (the exact shape of a distribution changes according to the number of degrees of freedom).

Probability distributions are also known as **theoretical distributions** because they are used as models to study real-life variables that may be approximately distributed as the theoretical distributions. Put differently, we often use theoretical distributions to study **empirical distributions**; that is, those resulting from actual data. We also use theoretical distributions in our efforts to estimate population parameters (see Chapter 8) and to test hypotheses (see Chapter 9). Thus, the 'idealized' distribution types that statisticians have developed over the years are very important for data analysis purposes: they provide the foundations for inferential analysis and enable us to go beyond a mere description of our sample data.

By far the most important theoretical distribution is the normal distribution, which is the backbone of much statistical analysis. There are several reasons for this. A large number of natural phenomena (e.g., height and weight) are approximately normally distributed; errors of measurement and prediction errors are also usually assumed to be normally distributed; certain sample statistics such as sample means are also approximately normally distributed in the population, assuming they come from a reasonably large sample size; a number of other probability distributions (e.g., binomial and Poisson) can be effectively approximated by the normal distribution; the normal distribution is easy to work with because of its symmetrical nature and the fact that it is completely described by its mean and standard deviation; the list goes on.

One particular version of the normal distribution deserves special attention. This is the **standard normal distribution** and is based on the concept of standard scores (see the discussion of *z*-scores earlier). Also sometimes referred to as the 'unit normal distribution', this is the distribution of normally distributed standard scores. Thus, the standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1. We now need your undivided attention for a second. Any normal distribution with a given mean and a given standard deviation can be transformed into the standard normal distribution simply by converting the raw scores into standard scores (see earlier). This means that if you ever bump into *any* normal distribution in a standardized form, you will immediately be able to know a bunch of important stuff about it, such as that the value of the mean is 0; the standard deviation is 68% of the scores will be within ±1, 95% of the scores will be within ±2, and 99.5% of the scores will be within ±3 (i.e., one, two, and three standard deviations, respectively); half of the scores will be below/above 0 (i.e., the median); and the most frequent score equals 0 (i.e., the mode).

A direct implication of the above is that we can use the standard normal distribution as a general model and generate **statistical tables** showing the proportion of cases falling within a certain distance from the mean. In this context, the area under any normal curve is proportional to the frequency of values, so the proportion of values below (above) the mean is represented by the area under the curve that lies below (above) the mean. To determine this proportion (without the help of statistical tables), one would have to use integral calculus – which is as pleasurable an experience as eating live eels (if you have ever seen the mathematical formula describing the normal distribution, you'll know what we mean!). What's even worse, one would have to engage in the dubious pleasure of integration for all possible distances from the mean (in order to find the proportion of cases falling a certain distance above/below the mean or, which amounts to the same thing, find the probability that a case will fall within a certain distance from the mean). Naturally, this process would have to be repeated every time a different distribution was under study, as a different mean or standard deviation results in a different normal distribution! Indeed, there are an infinite number of possible combinations of means and standard deviations, and consequently the number of different normal distributions is also infinite.

This is where the standard normal distribution comes to our rescue. Luckily, there are published tables for the standard normal distribution (developed by geniuses in integral calculus that certainly had plenty of time at their disposal and sadly no access to technology), showing (a) the values of standard scores, (b) the proportion of the area falling between the mean and

a given standard score, (c) the proportion of the area falling beyond a given standard score, and (d) the height of the curve (i.e., the ordinate) for the given standard score. Different statistics texts incorporate different combinations of the above tables in their appendices (see Further Reading section), and, yes, you can find these tables online too. Note that as the total area under the standard normal curve is equal to 1 (i.e., 100%) and as the curve is symmetric, the tables usually cover only one-half of the distribution (to the right of the mean); thus, only positive *z*-scores are covered. Figure 7.11 reproduces the standard normal distribution and the areas under the curve for $z = 0, \pm1, \pm2$, and ±3 (recall from our earlier discussion that a standard score reflects the distance from the mean in standard deviation terms, e.g., $z = 2$ indicates that the raw value lies two standard deviations above the mean. The mean is, of course, at $z = 0$).

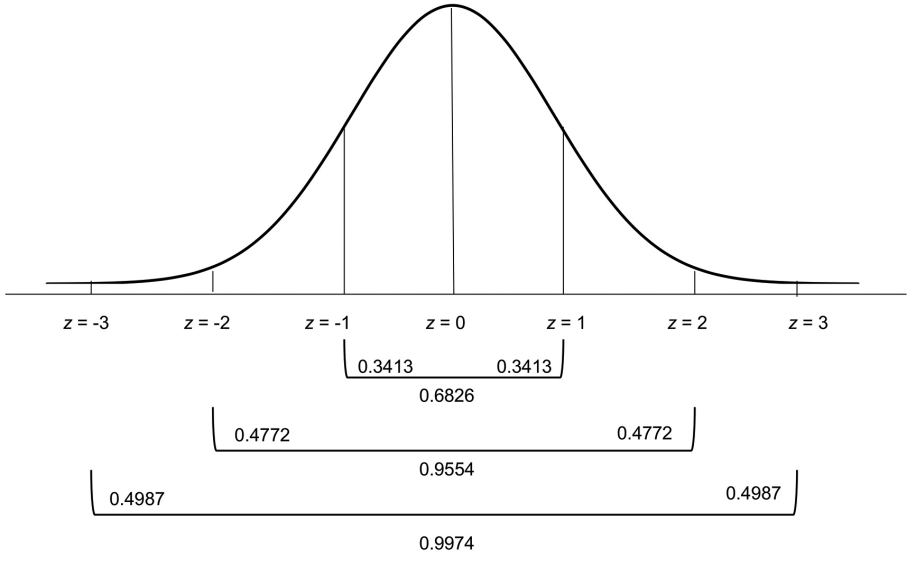


Figure 7.11 The standard normal distribution

Enough of theory. Let us see what the standard normal distribution can *do* for you. Just for the sake of it, we will assume that you have a normal distribution with a mean of 20 and a standard deviation of five, and you want answers to the following questions (you can ask similar questions with *any* normal distribution and the specific figures involved are for illustrative purposes only):

1. What is the proportion of scores lying *within* 0.5 standard deviations from the mean?
2. What is the proportion of scores *below* 1.5 standard deviations from the mean or *above* 2.0 standard deviations from the mean?
3. What is the proportion of scores *between* 1.0 and 1.5 standard deviations *above* the mean?
4. What is the proportion of scores lying *between* 0.5 standard deviations below the mean and 2.0 standard deviations *above* the mean?

5. Where in the distribution is a raw score of eight located?
6. What proportion of scores exceeds a raw score of 22?
7. What is the 70th percentile of this distribution?
8. What is the percentile rank of a raw score of 10?
9. What percentage of raw scores are between 6 and 12?
10. Did aliens really visit an Apple store in New York?

Yes, you can get the answers to all of these questions (well, the first nine anyway) simply by making use of the tables of the standard normal distribution; here, we are using the 'classic' tables of Fisher and Yates (see Further Reading section), but you should be able to replicate the results using any standard normal distribution tables (if not, *you* are doing something wrong, the tables are *always* right!). Okay, here goes:

1. This is a real giveaway; all you need to do is go to your statistical tables and look up the area between $z = 0$ (the mean) and $z = -0.50$ and then between $z = 0$ and $z = 0.50$ and add the two areas together; this comes to 0.1915 + 0.1915 = 0.383. Thus, approximately 38.3% of the observations will lie within ±0.5 standard deviations from the mean. See Figure 7.12(a).
2. This is a bit trickier but dead easy too. First, determine the area between $z = 0$ and $z = -1.5$; a quick peek at the tables tells you that this comes to 0.4332. Given that the area to the left of the mean is 0.50 (remember, the total area under the standard normal curve is equal to 1), the area below $z = -1.5$ is 0.50 − 0.4332 = 0.0668; make a note of this. Now, do exactly the same for $z = 2.0$. The area between $z = 0$ and $z = 2$ is 0.4772, which means that the area above $z = 2.0$ is 0.50 − 0.4772 = 0.0228; make a note of this too. Now, add your two values together (0.0668 + 0.0228 = 0.0896), and you have your answer: approximately 9% of the observations have scores 1.5 standard deviations below the mean or two standard deviations above the mean. See Figure 7.12(b).
3. This is truly a piece of cake since you only need to look at one side of the distribution. First, find the area between $z = 0$ and $z = 1.0$; this comes to 0.3413. Now find the area between $z = 0$ and $z = 1.5$; this comes to 0.4332. As the area between $z = 0$ and $z = 1.0$ is *enclosed* within the area between $z = 0$ and $z = 1.5$, all you need to do is subtract the former from the latter (0.4332 − 0.3413 = 0.919), and you have your answer. Thus, approximately 9.2% of observations have scores between 1.0 and 1.5 standard deviations above the mean. See Figure 7.12(c).
4. This is similar to answer 3 above (i.e., you are looking for the area between two *z*-scores), the only difference being that *both* sides of the distribution are involved. Proceed exactly as above and establish the areas between $z = 0$ and $z = -0.50$ and between $z = 0$ and $z = 2.00$; these come to 0.1915 and 0.4772, respectively. As each of these areas falls on a *different* side of the mean, you need to add them to get the total area; this gives 0.1915 + 0.4772 = 0.6687, which indicates that approximately 66.9% of the observations fall in the interval between 0.5 standard deviations below the mean and two standard deviations above the mean. See Figure 7.12(d).
5. You've done this one before. Just transform this raw score to a standard score and then check your standard normal distribution tables. Given that your original distribution has a mean of 20 and a standard deviation of five, a raw score of eight corresponds to a standard score of (8 − 20)/5=−2.40. Thus, this raw score lies 2.4 standard deviations below the mean. See Figure 7.12(e).

6. Elementary, my dear Watson. First, you must transform your raw score of 22 to a standard score; this should come to (22 − 20)/5 = 0.40. Now find the area between $z = 0$ and $z = 0.40$; this is 0.1554. Now since you want to find the area beyond $z = 0.40$, you need to subtract 0.1554 from 0.50 (see also answer 2 above). This gives 0.50 − 0.1554 = 0.3446, which indicates that approximately 34.5% of observations have raw scores greater than 22. See Figure 7.12(f).
7. This needs a bit of thought. What you want to do is get the standard score below which 0.70 of the total area lies and then transform it back to a raw score. Now, you know that the half of the standard normal distribution to the left of the mean has an area equal to 0.50, so you need another 0.20 to the right of the mean. So, what you must find is a *z*-score such that the area between it and $z = 0$ is equal to 0.20. Rushing to your standard tables, you should be able to find that the *z*-score you want has a value of 0.52. Great! Now, translate this back to a raw score. How? What do you mean, *how*? If $z=(X_i-\bar{x})/s$ gives you the standard score (see previous section), then if you know *z*, all you have to do is solve for *Xi*; specifically

$$X_i=\bar{x}+sz$$

where $\bar{x}$ = mean and *s* = standard deviation. In your case $z = 0.52$, $\bar{x} = 20$ and $s = 5$. Thus the raw score is 20 + 0.52(5) = 22.6, and this is the 70th percentile of the distribution in question. See Figure 7.12(g).
8. Compared to answer 7 above, this is a doddle. First, get your raw score translated into a standard score; you should get $z = -2.00$. Then find the area between $z = 0$ and $z = -2.00$; this is 0.4772. Now subtract this from 0.50 (see also answer 2 above) and you will get the area below $z = -2.00$ (which is what you want since you are after a percentile rank). Thus, the percentile rank of a raw score of 10 is 0.50 − 0.4772 = 0.0228, which indicates that approximately 2.3% of the observations have raw scores less than 10. See Figure 7.12(h).
9. This is essentially the same problem as the one in answer 3 above, the only difference being that you must first transform your raw scores to standard scores. If you do this, you will get a standard score of $z = -2.8$ for a raw score of six and a standard score of $z = -1.6$ for a raw score of 12. The area between $z = 0$ and $z = -2.8$ is 0.4974 and the area between $z = 0$ and $z = -1.6$ is 0.4452; you now have to subtract the latter from the former and you have the required area: that is, 0.4974 − 0.4452 = 0.05222. Thus approximately 5.2% of the observations have raw scores between 6 and 12. See Figure 7.12(i).

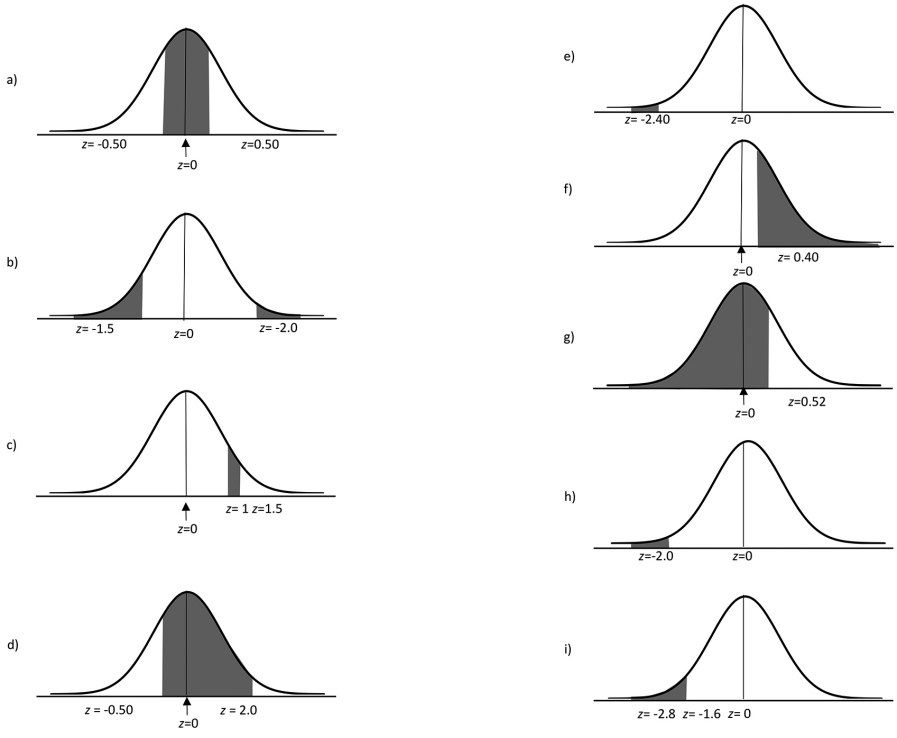


Figure 7.12 Using the standard normal distribution

Hopefully, we have not completely turned you off statistics with these beautiful examples! It should now be clear to you how one can use the known properties of the standard normal distribution to find the probability of occurrence associated with different outcomes (i.e., values). At this point, we should decrease your excitement by saying that you will most likely never have to look at statistical tables or perform calculations similar to the ones above as statistical packages automatically do this for you. Nonetheless, it is really important to have an understanding of the processes that take place in the background.

Other probability distributions are used in a similar manner as the standard normal distribution, although their applications are rather more limited. As already mentioned, we will be making much use of different probability distributions in the chapters to come in order to set up confidence intervals (see Chapter 8) and test various kinds of hypotheses (see Chapters 9–12). In fact, directly or indirectly, probability distributions will be with us for the remainder of this book, and thus you should go through the ideas introduced in this section once again to make sure that all is clear (better still, make up some additional examples and spend your weekend playing around some more with standard normal tables).

SUMMARY

In this chapter, we looked at different summary measures to characterize and compare frequency distributions. First, we showed that different distributions may vary in several respects, namely central location, variability, skewness, and kurtosis. Next, we discussed a variety of summary measures for describing these characteristics and highlighted their advantages and limitations. We then examined how central tendency can be looked at in conjunction with variability and learned about standard scores and the famous Chebyshev theorem. Lastly, we introduced the concept of a probability distribution and spent quite a bit of time discussing the normal distribution and the use of statistical tables. We have now concluded our journey into descriptive analysis and are ready to face the challenge of inferential statistics. Enjoy.

QUESTIONS AND PROBLEMS

1. In which ways can frequency distributions differ from one another?
2. What measures of central location could you use for (a) a nominal, (b) an ordinal, and (c) an interval variable?
3. What are the potential problems with using the mode as a measure of central location?
4. Why is it important to distinguish between the location of the median and the value of the median?
5. How could you use the mode, median, and mean to draw inferences about the skewness of a distribution?
6. Name one measure of variability for each different level of measurement.
7. Under what circumstances would you use the coefficient of variation?
8. What is a standard score? What are its uses?
9. Why is the standard normal distribution so important in data analysis?
10. Do you think it is politically correct to refer to the body shape of your partner as platykurtic?

FURTHER READING

Fisher, R. A. & Yates, F. (1974). *Statistical Tables for Biological, Agricultural and Medical Research*, 6th edition. London: Longman. Although its entertainment value is equal to emptying a dishwasher, this is the collection of statistical tables to have and treasure (however, any set of tables will do).

Huff, D. (2020). *How to Lie with Statistics*, short edition. It's short, it's witty, it's brilliant! If you want to learn how to mislead with your data and get away with it, this book will tell you.

Powell, F. C. (2009). *Statistical Tables for the Social, Biological, and Physical Sciences.* Cambridge: Cambridge University Press. This one is for the 0.01% of our readership who prefer statistical tables in print. The rest of you can surely find this riveting stuff online.

Weisberg, H. F. (1992). *Central Tendency and Variability*. London: Sage. An excellent little guide to different summary measures (including some very strange ones); although not a spring chicken anymore, descriptive statistics do not age!

8
What about using *estimation* to see what the population looks like?

THE NATURE OF ESTIMATION

In the previous two chapters, we saw how one can use different statistical techniques to describe, display, and summarize data. While this is an important endeavor in its own right when the focus of the analysis is purely descriptive (see Chapter 6), we often want to go beyond describing our sample data and say something about the population from which our sample was drawn; in other words, we want to make *inferences* about the population on the basis of what we observe in the sample. For example, we may have conducted a survey of 450 randomly selected penguin breeders and found that 63% absolutely hate going to the dentist. Now, although we may have a random sample, can we say that the corresponding percentage in the penguin breeder profession *as a whole* is 63%? Could it be higher? Or lower? By how much?

WARNING 8.1 Do not confuse population parameters with sample statistics.

It is questions like these that we try to address every time we engage in **estimation**. Specifically, estimation can be defined as the process of using a particular **sample statistic** (e.g., mean, standard deviation, proportion) to estimate the corresponding **population parameter**. In this context, parameters are fixed values that relate to the population and are generally unknown. Statistics, on the other hand, vary from one sample to the next and their values can be computed. Clearly, the notion of the term 'statistics' as used here is different from the everyday usage of the term; that is, as in published statistics or statistics as a subject area. Under the current perspective, a statistic is any quantity or summary measure (e.g., a mean) calculated from sample data, whereas a parameter is any quantity or summary measure calculated from population data. This is an extremely important distinction to keep in mind for reasons that will become evident shortly.

Let us go back to our example and explore the concept of estimation a little further. In the first instance, we may be content to take the sample proportion (63% or 0.63) as being our 'best' guess for the population proportion. In doing so, we would be using our sample statistic as a **point estimate** of the relevant population parameter. In general, a point estimate is a *single* value that is obtained from sample data and is used as the 'best guess' of the corresponding population value. Thus, in our example, the point estimate of the population proportion is

0.63. Now, the problem with a point estimate is that, since it is based on sample data, it reflects not only the relevant population parameter but also sampling error. As discussed in Chapter 2, sampling error reflects the difference between a result based on a sample and that obtained if the entire population in question (here all penguin breeders) had been studied. Thus, if π represents the (unknown) population proportion and P our sample proportion, then

$$\pi - P = e$$

where e = sampling error.

Three things should become immediately obvious. First, in any one sample, sampling error can be positive or negative. What this means is that a population parameter may be overestimated or underestimated by a sample statistic used as a point estimate (in our example, P = 0.63 so, depending on whether $\pi > 0.63$ or $\pi < 0.63$, then $e > 0$ or $e < 0$).

Second, the magnitude of sampling error will vary from sample to sample; if we were to take several samples and compute the sample proportions, some of them would be likely to be better point estimates of the population proportion than others (if we took, say, 20 different samples of 450 penguin breeders and calculated P, could we really expect that P would be *exactly* equal to 0.63 *every single time*?).

Third, the variation in sampling error across samples is solely due to the variation in the sample statistic (here the sample proportion) as the value of the population parameter is always fixed. Although π is unknown, being a population parameter, it cannot vary in value as its computation includes all the elements in the population. In contrast, P can vary as it is based only on a fragment of the population; that is, those population elements that happen to be included in our particular sample.

If we pull the above observations together, we can see that a point estimate of the population parameter will be inevitably influenced by sampling error and the specific characteristics of the particular sample used. What seems to make more sense is to try to establish a *range* for our estimate by taking into account the likely variation in sampling error. This is exactly what the purpose of *interval* estimation is: to predict that the population parameter in question is somewhere within a given interval on either side of a point estimate. In our example, an **interval estimate** would take the form of a statement such as 'the proportion of *all* penguin breeders who hate going to the dentist is somewhere between 0.59 and 0.66'.

This is all very well, you may say, but how do we go about constructing such an interval? Given that we only know the value of the sample statistic for our particular sample and given that the population parameter is unknown, we are clearly unable to specify the magnitude of sampling error involved. However, if we took repeated samples of the same size and calculated the sample proportions, then we should gain some appreciation of sampling error by observing how these repeated measurements varied from each other; if the sample proportions fluctuated little from sample to sample, then we would gather that the sampling error was less in magnitude than if they fluctuated a lot. Now, imagine that we took *all* possible samples of a given size and calculated the sample proportion. We would eventually end up with a *distribution* of all possible sample proportions, representing all samples of a given size. We call this distribution the **sampling distribution of a proportion**. In general, the sampling distribution

of some statistic (e.g., mean, standard deviation, proportion) can be defined as the distribution of all possible values that can be assumed by this statistic as computed from samples of the same size drawn at random from the same population.

Now, as with all distributions for metric data (and recall from Chapter 3 that dichotomous variables can be treated as such), we could use the standard deviation to measure the variability of the sampling distribution (see Chapter 8). We call this standard deviation the **standard error of the proportion** and denote it with *s*p (to distinguish it from the sample standard deviation, *s*).

Thus, the standard error of the proportion is a measure of the fluctuation in proportions from sample to sample, and we can, therefore, use it to describe the variation in sampling error. Before we go any further, you should be absolutely clear that the standard error of the proportion is a standard deviation of a particular distribution and that this distribution is neither the distribution of individual values in the sample nor the distribution of values in the population, but the distribution of sample proportions calculated from all samples of a given size selected randomly from the population – what we call the sampling distribution of a proportion. The same applies to *any* standard error (e.g., the standard error of the mean): a standard error is *always* the standard deviation of the sampling distribution for the statistic under study.

WARNING 8.2 Do not confuse the standard deviation of the sampling distribution (i.e., the standard error) with either the sample standard deviation or the population standard deviation.

'Hang on a second!', you may shout – what's the point of attempting to construct a sampling distribution so that we can measure the fluctuation in sampling error if we need to take *all* possible samples of a given size from the population? Wouldn't it be just as quick (or even quicker) to conduct a census; that is, forget about sampling altogether and measure the whole population instead? This is absolutely true *if* we actually had to go through the procedure outlined above – that is, if we had to take all possible samples of a given size from the population, calculate the statistic of interest for each one of them, and use them to construct the sampling empirically. However, in practice, we do not have to do this because we can use **theoretical sampling distributions**. These probability distributions have known properties (see Chapter 7) and can be used in conjunction with a *single* sample to generate estimates of sampling error. Thus, there are theoretical sampling distributions for the sample proportion, the sample mean, and the sample variance, to name but a few. Such sampling distributions have been derived mathematically by incredibly bright statisticians (for goodness's sake, don't ask *how*!) but can be taken 'off the shelf' by mere mortals such as ourselves every time we want to estimate a certain population parameter (for example, assuming a normally distributed population, the sampling distribution of the mean is a normal distribution, while the sampling distribution of the variance approximates a chi-square distribution). Thus, all we have to do is use the value of our sample statistic together with the standard error of the appropriate (theoretical) sampling distribution and generate an interval estimate as follows:

$$\text{Population parameter} = \text{sample statistic} \pm k(\text{standard error})$$

where k = number of desired standard errors for the estimate.

Thus, if we wanted to estimate the population proportion, we would (a) calculate our sample proportion, P, based on our sample data, (b) look at the theoretical sampling distribution for sample proportions and calculate its standard error s_p (there are standard formulae for this), and (c) set up an interval by adding/subtracting so many standard errors, k, from the sample proportion; thus our estimate would take the following form:

$$\pi = P \pm ks_p$$

So far, so good – but we still have not shown you an actual example of how to get an interval estimate for the dreaded proportion of penguin breeders who despise going to the dentist. What is s_p for the specific sample of 450 penguin breeders under consideration? How small/large should k be? Who will win the pole vault in the next Olympics? Patience, dear friends – all will soon be revealed.

SETTING CONFIDENCE INTERVALS

Consider the following question: is an interval estimate that places the population proportion between 0.60 and 0.70 necessarily better than another estimate, which places the population proportion between 0.40 and 0.80? *Of course*, you may think – the former interval is substantially narrower than the latter, so our estimate of the population parameter is more precise. That may be so, but can we be as *confident* that the population parameter will, in fact, be captured by the first interval as we can that it will be in the second? In other words, is the risk of stating that the parameter is somewhere within the interval *when in fact it is not* the same for both cases? Probably not.

In attempting any kind of interval estimation, the first thing we have to decide is the **confidence level** for our estimate; that is, we have to decide how often we want to be correct that our interval will, in fact, contain the population parameter in question. Different samples are likely to yield different values for the sample proportion P (see discussion of sampling error earlier) and, therefore, will produce different interval estimates of the population proportion π. Some of these intervals will contain π but others will not (see Figure 8.1). As the theoretical sampling distribution for a proportion has known properties, it will enable us to set k (i.e., the number of standard errors) in such a way that any specified percentage of these intervals will contain the desired parameter π (see also Chapter 7). Therefore, if k is set so that, say, 95% of all possible intervals under repeated sampling would contain π, then we have a 0.95 probability of selecting one of the samples that will produce an interval containing π. This interval, computed from the specific sample at hand, is known as a **95% confidence interval**, and its upper and lower limits are denoted as **95% confidence limits**. As already mentioned, we can never be certain (100% confident) about our estimations when using a sample. And, needless to say, the less confident we are, the more room for error we allow. Thus, operating with a confidence level of 95% allows 5% of incorrect estimations, while a confidence level of 99% implies that only 1 out of 100 intervals will fail to contain the population parameter. The allowed 'room for error' is

typically referred to as α (alpha) level, and, as we will see in the following chapters, it is a rather crucial benchmark value.

In simple terms, a confidence interval reflects a range of values that we are confident (but not certain) contains the population parameter. Clearly, the wider the confidence interval, the more confident we can be that it will contain our (unknown) parameter, but our estimate of the latter will not be very precise. Conversely, the narrower the confidence interval, the less confident we can be that it includes the parameter in question – but *if* it does, then our estimate will be more precise. This brings us to the important distinction between estimation *accuracy* and estimation *precision*. **Estimation accuracy** refers to whether an interval actually captures the true population parameter or not and is reflected in the confidence level we set (e.g., 95%). However, for a *given* confidence level, **estimation precision** might differ (i.e., confidence intervals may be wider or narrower depending on how much sampling error is involved). Ideally, of course, we prefer a narrow interval (precision) with a high degree of confidence (accuracy), and this can be achieved by employing bigger sample sizes (and more reliable measurement instruments).

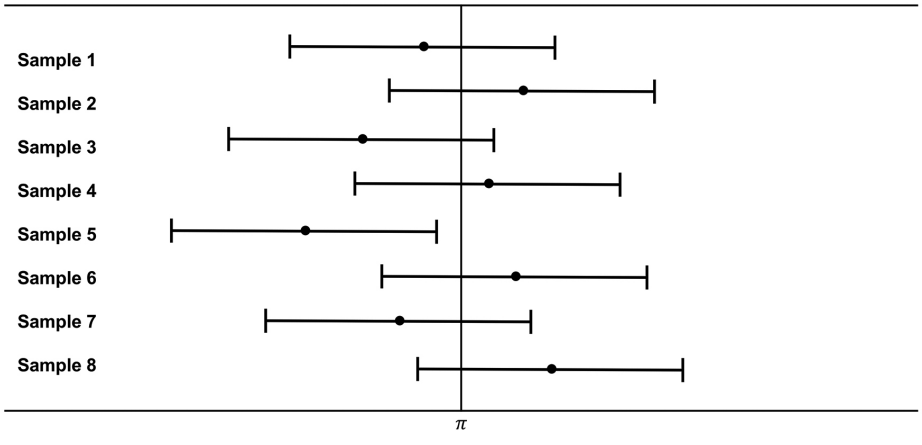


Figure 8.1 Confidence intervals for several samples around parameter value π

In general, a 100(1−α)% confidence interval for any population parameter Θ takes the following form:

$$\text{Sample statistic} - k_c(\text{standard error}) \leq \Theta \leq \text{sample statistic} + k_c(\text{standard error})$$

where k_c is known as the **critical value** corresponding to the **confidence level** of (1−α). This critical value indicates the number of standard errors needed to create the desired confidence interval around Θ. As α increases, the confidence level decreases (e.g., if α goes from 0.05 to 0.10, the confidence level reduces from 0.95 to 0.90) and, therefore, k_c also decreases; this results in a narrower interval for Θ (but accompanied by a reduction in confidence – see the

discussion about accuracy and precision above). Conventional values for α are 0.01, 0.05, and 0.10, resulting in 99%, 95%, and 90% confidence intervals, respectively; however, there is nothing sacred about these values, and if you want to set up an 83% or 74% confidence interval, nobody can stop you.

People often get (very!) confused when trying to interpret a confidence interval. A typical mistake is to make statements such as 'the probability is 0.95 that the parameter in question is included in the interval'. This is clearly *not* correct because the probability that the parameter is within a *particular* interval is either 1 or 0; the parameter in question is either included in the interval or it is not (see Figure 8.1). The correct interpretation of a confidence interval emphasizes the *procedure* used to generate it, not the specific interval obtained. More specifically, we can only say that the procedure is such that 95% of the intervals obtained will include the true (fixed) parameter in the long run.

Having said all this, making a statement such as 'we are 95% confident that the particular interval we have constructed contains the population parameter' is actually OK, as long as we are clear that what we *really* mean is that '95% of all possible intervals constructed using this procedure will include the parameter concerned'. And how do you know that your interval does not belong to the 5% of intervals that fail to capture the true population parameter? Well, you don't, and most likely you will never know (unless you get access to the whole population). What you *do* know, though, is that your particular interval has a 95% chance of being one of those that *do* contain the population parameter, so the odds are clearly in your favor.

So how about putting all this seriously theoretical stuff into practice? By now, you must be *dying* to estimate the proportion of penguin breeders who hate going to the dentist (and throw in a confidence interval for good measure). Let's do it.

ESTIMATING THE POPULATION PROPORTION

You will be pleased to know that the steps we shall follow below are not specific to the population proportion (or to penguin breeders) but reflect a general sequence that can be followed whenever you feel like estimating *any* population parameter (e.g., you may wake up in the middle of the night feeling an uncontrollable urge to engage in estimation!). Thus, if you understand the basic logic and decisions involved in estimating the population proportion, you will be in a position to apply your knowledge to estimate means, variances, and the like (and thus discover a whole new way of spending your weekends!).

To refresh your memory, the specific estimation problem we face is that of estimating the proportion of penguin breeders who hate going to the dentist. Thus, our unknown population parameter is π and we would like to develop, say, a 95% confidence interval for it. This would reflect the uncertainty associated with the fact that we are using a sample instead of the whole population. We know from our (random) sample of 450 penguin breeders that 63% of them hate visits to the dentist, so $n = 450$ and $P = 0.63$.

What we must do now is find the appropriate sampling distribution for our sample statistic (i.e., the proportion). This can be quickly achieved by consulting any standard statistics textbook (see Further Reading section for some readable examples) or by accosting your local

friendly statistician. It just so happens that, with large samples, the sampling distribution of the proportion approximates a normal distribution (hurrah, we know this one!), with mean equal to the population proportion π and standard deviation equal to:

$$\frac{\sqrt{P(1-P)}}{n}$$

The latter is, therefore, our beloved standard error *s*p (see earlier) and comes to $\sqrt{\left[0.63(1-0.63)/450\right]}=\sqrt{\left[(0.63)(0.37)/450\right]}=0.023$.

HINT 8.1 When interpreting a confidence interval, always remember that the particular interval is only one out of many possible ones based on different samples; hopefully, it is one of those intervals that does include the population parameter!

All we need to do now is set our critical value *k*c and then we have all the information required to construct our confidence interval shown below (note that this is merely the application of the general formula for confidence intervals with respect to the population proportion):

$$P-k_c s_p \leq \pi \leq P+k_c s_p$$

Since we want a 95% confidence interval and we know that the sampling distribution involved approximates a normal distribution, our problem is to determine a value for *k*c such that the interval will take up 95% of the area under the normal curve. Yes, we *have* done this before in Chapter 7. And yes, you are right that we can get the value for *k*c simply by going to the standard normal distribution tables, which tell us, at a glance, how many standard deviations either side of the mean enclose 95% of the total area. And yes, you are right again if you think that *k*c can be directly determined by looking at the *z*-values (i.e., standard scores) defining this area.

Indeed, if we rush to our statistical tables, we should find that 95% of the area under the standard normal distribution lies between $z = -1.96$ and $z = 1.96$ (see Figure 8.2(a)). Thus, if we set $k\text{c} = 1.96$, we've cracked it. Applying the formula for a confidence interval to our data gives us the following result:

$$0.63-1.96(0.023)\leq\pi\leq 0.63+1.96(0.023) \text{ or } 0.584\leq\pi\leq 0.675$$

Thus, we can conclude with 95% confidence that the proportion of penguin breeders who hate going to the dentist is within 58.4% and 67.5%. End of story.

So, what if we want to set a 99% confidence interval instead? Do we have to go through the entire set of calculations and bore ourselves to death again? Not at all. Everything stays the same, the only bit of information that needs changing being *k*c (in order to correspond to our new confidence coefficient). Back to our trusted standard normal distribution tables, we look for *z*-scores on either side of the mean, such as that 99% of the total area under the curve is

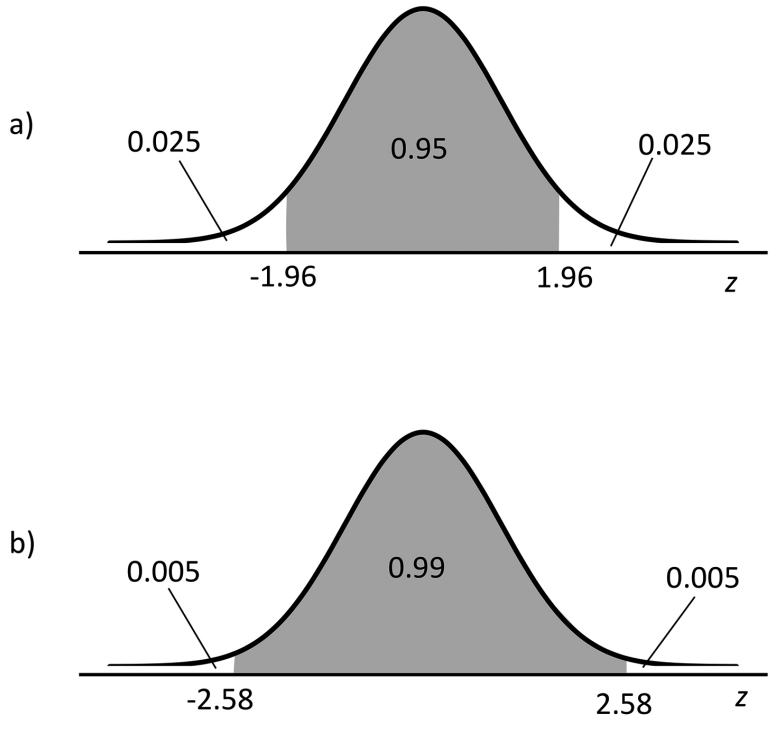


Figure 8.2 95% and 99% confidence intervals: normal distribution

covered. These come to $z = -2.58$ and $z = 2.58$, so setting $kc = 2.58$ gives us the following 99% confidence interval (see also Figure 8.2(b)):

$$0.63-2.58(0.023)\leq\pi\leq0.63+2.58(0.023) \text{ or } 0.570\leq\pi\leq0.689$$

Thus, we can be 99% confident that between 57% and 68.9% of penguin breeders hate going to the dentist. Note that this interval is wider than the corresponding 95% interval, reflecting the trade-off between the increase in confidence and loss in precision discussed in the previous section. Needless to say, we could follow the above procedure to set intervals at *any* confidence level – see how handy the standard normal tables are?

If only to get statistical purists off our back, we need to quickly mention that estimating the population proportion using the normal distribution as we have done above is only legitimate when we have a reasonably sized sample and when P is too close neither to 0 nor to 1. A rule of thumb often used to check these requirements is that both nP and $n(1-P)$ are greater than 5 (in our case, they come to 283.5 and 166.5, so no problems). If nP and/or $n(1-P)$ are less than 5, then the appropriate sampling distribution is not the normal but the **binomial distribution**, which is the probability distribution for a dichotomous variable (we will come across the binomial distribution again in Chapter 11). However, as the sample size increases, the bino-

mial distribution approximates the normal (a phenomenon that statisticians proudly describe as the 'normal approximation to the binomial'); it is for this reason that, with large samples, it is OK to go ahead and use the normal distribution as the appropriate theoretical sampling distribution for estimating a proportion.

We should also mention that if the sample constitutes more than 5% of the population, we should use what is called a **finite population correction** in our calculation of standard error. This correction is defined as

$$\sqrt{\frac{N-n}{N-1}},$$

where N = population size and n = sample size. Thus, the standard error of the population proportion corrected by this factor is

$$s_p = \sqrt{\left(\frac{P(1-P)}{n}\right)} \times \sqrt{\left(\frac{N-n}{N-1}\right)}$$

Note that, in most practical applications, the samples taken are usually a small fraction of the entire population and, thus, the finite population correction can safely be ignored. Having said that, a sample of 450 randomly selected penguin breeders may well be above 5% of the entire population of penguin breeders; after all, breeding penguins is not exactly the most widespread profession, is it?

ESTIMATING THE POPULATION MEAN

The procedure one uses to estimate a population mean μ is identical to the one followed above to estimate the population proportion. The only difference comes in the calculation of standard error, as means and proportions are different statistics and, not surprisingly, require different formulae.

If the population is normally distributed (and there are tests to establish this, as we will see in Chapter 11), the **sampling distribution of the mean** is also normally distributed with a mean equal to the population mean μ and a standard deviation equal to $\sigma/\sqrt{n}$, where σ is the *population* standard deviation and n is the sample size. What's even better, according to the notorious **Central Limit Theorem**, even if the actual values in the population are *not* normally distributed, the sampling distribution of the mean (i.e., the distribution of means computed from all possible samples of a given size drawn from the population) will be approximately normally distributed (see also Chapter 11). So, if the sample is large enough (most experts suggest $n > 30$), the sampling distribution of the mean can be approximated by a normal distribution with mean μ, and standard deviation $s/\sqrt{n}$, where s is the *sample's* standard deviation.

Thus, with a decent sample, the **standard error of the mean** is $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ and the sampling distribution is normal, regardless of the form of the distribution in the population.

As an example, say that, in addition to estimating the proportion of penguin breeders who hate going to the dentist, we also wanted to estimate the number of hours they spent at the golf course each week and set a 90% confidence interval. Say that from our sample of 450 penguin breeders, we found that the mean number of hours per week spent at the golf course was 9.5 with a standard deviation of 6.8. All we need to do is to fill in the numbers in the following (by now surely familiar) formula:

$$\bar{x} - k_c s_{\bar{x}} \leq \mu \leq \bar{x} + k_c s_{\bar{x}}$$

Note that we have simply substituted the symbols for the sample mean, population mean, and standard error of the mean in the general formula for a confidence interval that we introduced earlier in this chapter. There is nothing new here. Given that our sample size is large (well over 30), we can use the normal distribution as the appropriate theoretical sampling distribution for the mean and calculate the standard error as

$$s_{\bar{x}} = s/\sqrt{n}$$

This comes to $6.8/\sqrt{450}$ = 0.32. Given that our confidence coefficient is 90% and the sampling distribution involved is normal, we can go to the standard normal tables (exactly as we did before when we estimated a proportion) and find the *z*-scores between which 90% of the total area under the curve is enclosed; these come to $z = -1.64$ and $z = 1.64$ (see Figure 8.3). We now set the critical value to kc = 1.64, and we have our desired confidence interval:

$$9.5 - 1.64(0.32) \leq \mu \leq 9.5 + 1.64(0.32) \text{ or } 8.98 \leq \mu \leq 10.02$$

Thus, we can be 90% confident that the mean number of hours (per week) spent at the golf course by penguin breeders is between 8.98 and 10.2 hours.

What happens if we do not know the population standard deviation *and* our sample is small? Well, not all is lost. If the distribution in the population is approximately normal, then, instead of the normal distribution, we should use the **t-distribution** (with $n-1$ degrees of freedom) to set kc (the formula for calculating the standard error $s_{\bar{x}}$ stays the same). Remember the degrees of freedom? Well, we won't hold it against you if you don't because we hardly mentioned it in previous chapters. Suffice to say that the degrees of freedom is an important concept that, among other things, determines the *exact* shape of a distribution.

The *t*-distribution is a symmetrical distribution like the normal distribution and almost as beautiful, but a bit more platykurtic; that is, less peaked at the center and higher in the tails. As the sample size increases to infinity, the *t*-distribution approaches the normal (at about

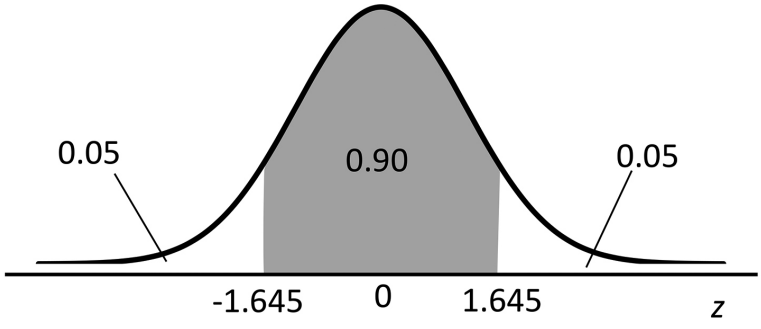


Figure 8.3 90% confidence interval: normal distribution

n = 120, the two distributions are practically identical). The t-distribution is really a 'family of distributions' since there is a different t-distribution for each different number of degrees of freedom; the latter are given by $n-1$, where n is the sample size. There are statistical tables available for the t-distribution (for different degrees of freedom) showing the area under the curve corresponding to different **t-values** (or 't-statistics'); these are equivalent to z-scores in the standard normal distribution, and their interpretation is very similar. However, published tables for the t-distribution are nowhere near as detailed as those for the standard normal distribution, as t-values depend on the degrees of freedom; typically, only the t-values enclosing 90%, 95%, and 99% of the total area under the curve are reported for a range of degrees of freedom. Note also that, since the t-distribution approaches the normal distribution as the sample size increases, for practical purposes the t-distribution is often substituted by the standard normal distribution when $n > 30$. With 30 degrees of freedom, the t-values enclosing 95% of the area come to −2.04 and 2.04, while with 120 degrees of freedom the corresponding values are −1.98 and 1.98. Given that the equivalent z-scores are −1.96 and 1.96 (see Figure 8.2(a)), the loss in precision is trivial if the standard normal distribution is used instead of the t-distribution once the sample size exceeds 30.

Here is a quick example of estimating the population mean using the t-distribution with a small sample. Suppose that we have asked 20 randomly selected primary school children to tell us the number of fancy-dress parties they have attended during the past month, and we want to estimate the mean 'fancy-dress party attendance' with 95% confidence (clearly a profound social issue!). Assume that we have calculated the mean and standard deviation from our sample data, and we got 14.6 and 3.9, respectively. Now, we cannot use the normal distribution as our sampling distribution because not only do we not know the standard deviation of the population (see earlier), but our sample is small as well. However, assuming that the population is normally distributed (or roughly so), we can use the t-distribution (with 20 − 1 =

19 degrees of freedom) to find the critical value *k*c. As before, we calculate the standard error of the mean as

$$s_{\bar{x}} = s/\sqrt{n}$$

This gives us $3.9/\sqrt{20} = 0.87$. Next, we go to the tables of the *t*-distribution (for 19 degrees of freedom) and look up the *t*-values such that 95% of the area under the curve is enclosed between them; these come to $t = -2.093$ and $t = 2.093$ (see Figure 8.4). So, we set $k\text{c} = 2.093$, and we get

$$14.6 - 2.093(0.87) \leq \mu \leq 14.6 + 2.093(0.87) \text{ or } 12.78 \leq \mu \leq 16.42$$

Therefore, we can be 95% confident that, on average, primary school children attended between 13 and 16 fancy-dress parties over the past month.

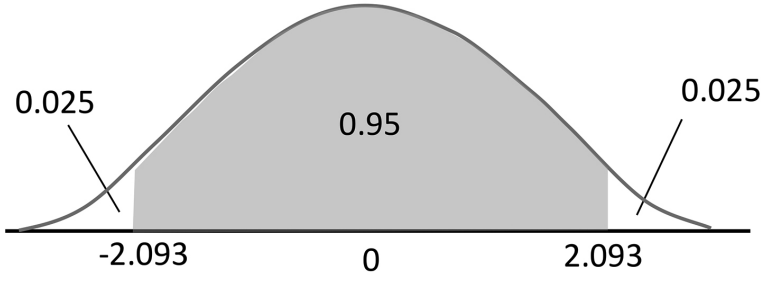


Figure 8.4 95% confidence interval: t-distribution (df = 19)

Now, if only to play devil's advocate, you might ask: 'What is the sampling distribution of the mean if (a) we do not know the standard deviation of the population, (b) the sample size is small, *and* (c) the population distribution is not normal?' The 'formal' answer to this question is that, under these conditions, we cannot specify a probability distribution for the fluctuation of the mean from sample to sample to enable accurate estimation to take place (i.e., we are really and truly stuck!). The 'informal' answer is that you should have got a decent-sized sample in the first place, and then you would not be asking such awkward questions!

Note that the comments made in the previous section concerning the application of the finite population correction $\sqrt{(N-n)/(N-1)}$ when the sample is more than 5% of the population also apply here. We must adjust the standard error of the mean in a similar fashion as we did for the standard error of the proportion; the correction factor formula stays the same.

Table 8.1 Steps in estimation

1.	Decide on which population parameter, Θ, you want to estimate; this could be a proportion, a mean, a variance, etc.
2.	Decide on the confidence and α level (1−α) for your estimate of Θ and set your confidence level 100(1−α)%.
3.	Calculate the corresponding statistic (i.e., sample mean, proportion, etc.) from your sample data; needless to say, this has to be a *random* sample, otherwise estimation goes out of the window as sampling error cannot be assessed (see Chapter 2).
4.	Find the appropriate sampling distribution of the statistic and calculate its standard deviation; this is your standard error.
5.	By looking at the tables of the sampling distribution, determine the critical value, *k*c, that corresponds to your confidence coefficient (1−α).
6.	Construct your 100(1−α)% confidence interval by multiplying the critical value, *k*c, with the standard error and subtract/add this product to the sample statistic.
7.	Eat a big chocolate cake to reward yourself.

ESTIMATING OTHER POPULATION PARAMETERS

While the estimations of the population mean and the population proportion are, by far, the most frequent estimation applications in practice, one can follow similar procedures to estimate other parameters, such as the variance of a population, the difference between means in two populations, the difference between two population proportions, or the ratio of two variances. We see little point in going through what are essentially identical steps to demonstrate the setting of confidence intervals for such parameters (it would be incredibly boring and add little to what you already know). Instead, we have compiled in Table 8.1 a neat little summary of the general steps one has to go through in order to estimate *any* parameter, and we refer you to the Further Reading section for more specific details (i.e., appropriate sampling distributions, computation formulae for standard errors, and relevant statistical tables).

HINT 8.2 While (sample) size is not everything, in estimation it helps!

Computer packages are obviously of considerable assistance when you go through the steps outlined in Table 8.1 as they patiently provide you with the standard error of a requested statistic, go through the distribution tables, and even construct the confidence intervals for you. While such facilities mean that you have to spend only a minimum amount of time constructing confidence intervals for your estimates, they do not reduce the burden of *interpreting* confidence intervals – *you* still have to do this, and frequent reference to Hint 8.1 should ensure that you do it correctly. Computer science has been tremendously helpful in relieving us from all computational issues and calculations, but it cannot possibly do anything to ensure that we understand what we are doing.

Before we leave this chapter, we should draw your attention to an interesting feature of the formulae for standard errors, namely that the sample size is always included in the denominator. Thus, the larger the sample size, the smaller the standard error. Given that the standard error is always multiplied by the critical value, *k*c, when setting a given confidence interval

(see earlier sections), the smaller the standard error, the narrower the interval (all other things being equal, of course). Thus, at a given level of confidence, increasing the sample size results in a narrower (i.e., more precise) interval estimate of the population parameter. This indicates the importance of sample size, and it is worth bearing in mind when determining the sample size required for a study. Studies employing smaller samples are, inevitably, prone to more sampling error and yield less precise estimates at a given level of confidence.

Finally, we should mention that all the procedures and formulae presented in this chapter assume that a **simple random sample** has been drawn (see Table 2.2 in Chapter 2 for a description of different sampling methods). The calculation of standard errors and confidence intervals under alternative sampling methods (e.g., stratified or cluster sampling) is more complex, and we refer you to the Further Reading section and the sources recommended in Chapter 2 for more details.

SUMMARY

In this thoroughly titillating chapter, we looked at how we can say something about the population when we only look at a sample. We began by distinguishing between point estimates and interval estimates and highlighted the importance of understanding the fluctuation of sampling error. Next, we introduced the concept of a sampling distribution and indicated how its standard deviation can be used to represent the variation in sampling error for the statistic concerned (i.e., the standard error of the estimate). We then considered issues associated with the construction and interpretation of confidence intervals and provided illustrations involving different population parameters and different sampling distributions. Estimation was our first close encounter with the inferential branch of statistical analysis; we will be using this knowledge in the next chapter, in which statistical inference will be approached from a different angle, namely that of hypothesis-testing.

QUESTIONS AND PROBLEMS

1. What is the general objective of statistical estimation?
2. What is the difference between a population parameter and a sample statistic?
3. What is the difference between a point estimate and an interval estimate?
4. What is a sampling distribution? Why is it useful?
5. Explain the difference between standard error, standard deviation of the sample, and standard deviation of the population.
6. What does the term ‘confidence limits’ mean? How would you go about setting a confidence interval for the population mean?
7. What do precision and accuracy mean in the context of parameter estimation?
8. Explain what the ‘critical value’ indicates. Give an example to illustrate your answer.
9. Under what circumstances would you choose the *t*-distribution in preference to the standard normal distribution?
10. What role does the size of the sample play in estimation?
11. What would be your best estimate of the average mental age of the authors?

FURTHER READING

Berenson, M. L., Levine, D. M., Szabat, K. A., & Stephan, D. F. (2019). *Basic Business Statistics: Concepts and Applications*, 14th edition. London: Pearson. Chapters 7 and 8 will tell you all you want to know about sampling distributions and estimation.

Churchill, G. A. (2018). *Marketing Research: Methodological Foundations*, 12th edition. Nashville, TN: Earlie Light Books. Chapters 11 and 12 contain good discussions of estimation issues and highlight the role that sample size plays when setting confidence intervals.

McClave, J. T., Benson, P. G., & Sincich, T. (2018). *Statistics for Business and Economics*, 13th edition. London: Pearson. If you are still yearning for more punishment, check Chapters 5, 6, and 8.

9
How about sitting back and *hypothesizing*?

THE NATURE AND ROLE OF HYPOTHESES

In the last chapter, we introduced basic techniques for estimating population parameters from sample statistics. A complementary approach to making inferences about the population is via **hypothesis-testing**. While in estimating population parameters we use sample data to make 'informed guesses' about the true value of a certain population characteristic, in testing hypotheses we use samples to examine whether a particular proposition concerning the population is likely to hold or not. Thus, essentially, hypotheses capture specific expectations about what holds true (or not) at the population level. For example, an advertising executive may hypothesize that placing perfume ads in *Heavy Metal Weekly* will attract more readers than placing them in *Shipbuilders' Digest*. A personnel manager may hypothesize that employees who have undergone a 'find your elusive inner-self' training program (offered by Weird Solutions Inc.) will be more productive than employees who have not. Or an economist may hypothesize that the average personal income not declared to the tax authorities is more than €4,000 per annum. In short, a hypothesis is a statement regarding a population (or populations) that may or may not be true.

Let us assume that, on the basis of past evidence, theory, or seven years of conceptual deliberation, we hypothesize that the Dutch are more likely to have visited a professional ear-wax remover than the Vietnamese; this is simply a *conjecture* on our part, which may or may not be true. Now, if we had unlimited resources in terms of time and money, we could undertake a census in the Netherlands and a census in Vietnam and determine whether our hypothesis is correct or not. However, as most of us do not have unlimited resources (otherwise, we would be spending our time and incredible wealth on better things), what we have to do is make inferences as to the correctness of our hypothesis from *sample* information. If we took random samples of Dutch and Vietnamese consumers and found, say, that 57% of Dutch consumers had visited a professional ear-wax remover as compared to 28% for the Vietnamese, our conjecture would find support in the data. If, on the other hand, we found that the proportions were 15% and 36%, respectively, our conjecture would be refuted by the data (indeed, it would appear that the opposite would hold). Finally, if we found that, in both samples, exactly 34% had visited a professional ear-wax remover, again, our conjecture would not find support (as both country samples show identical visit rates).

This is all very nice, but what if the figures do not turn out as clear-cut as that? Can we really say that we found support for our hypothesis if the percentages turn out to be, say, 29% for the Dutch and 28% for the Vietnamese? And, equally, should we immediately conclude that the 'true' situation in the population is exactly the opposite from what we predicted if the proportion for the Dutch sample is 36.4% and that for the Vietnamese sample 36.5%? If you feel uncomfortable about drawing such conclusions, well, you should! The problem, as you should know by now, is the dreaded sampling error. As was the case with estimation, data from a sample are just that: data from a sample. Any quantity (such as the percentages calculated here) based on sample data is subject to sampling error, which is not constant but variable (see Chapter 8). Unless we somehow incorporate sampling error into our deliberations, it becomes very difficult to determine whether our hypotheses are supported by the data or not. Thus, conceptually, the problem we are facing is very similar to that faced when estimating any population parameter: we must consider sampling error whenever we make inferences about the population. Before we show you how this is done, it is important to look at the notion of a hypothesis from several different angles.

To begin with, every time we set a hypothesis, we are really setting *three* of them. No, we have not been drinking – any hypothesis implies two other hypotheses, as we shall immediately demonstrate beyond reasonable doubt. Take our previous hypothesis (let's call it *H1*), which states:

> *H1*: A greater proportion of Dutch consumers compared to Vietnamese consumers visited a professional ear-wax remover in the last year

This hypothesis indicates what we think the situation is in the population. By implication, we do not think that

> *H2*: A smaller proportion of Dutch consumers compared to Vietnamese consumers visited a professional ear-wax remover in the last year.

or that

> *H3*: The proportions of Dutch and Vietnamese consumers who visited a professional ear-wax remover in the last year are the same.

Any *one* of *H1*, *H2*, or *H3* could reflect the actual situation in the population. Obviously, we do not *know* which one (otherwise, why bother with hypothesis-testing?), but we *think* that *H1* is correct. Note that nothing would change if our initial hypothesis was *H2* (or *H3* for that matter); the other two hypotheses would be 'automatically' generated. Why this happens is quite simple: the three hypotheses constitute *mutually exclusive* and *collectively exhaustive* descriptions of all possible situations in the population. *Either* the Dutch are more likely to visit a professional ear-wax remover, *or* the Vietnamese are more likely to visit a professional ear-wax remover, *or* the Dutch and the Vietnamese are equally likely to visit a professional

ear-wax remover; there are no other possibilities (if you can come up with additional possibilities then you deserve a Nobel prize!).

From the above, it should be clear that whenever we examine a hypothesis, we are really comparing our hypothesis against two other **competing hypotheses**; consequently, support for our hypothesis implies rejection of the other two hypotheses. However, in practice, what we usually do is test one hypothesis against a *combination* of the other two (it simply makes life easier). Thus, if our interest is in *H1*, we can combine *H2* and *H3* into a new hypothesis (call it *H4*) as follows:

H4: The proportion of Dutch consumers who visited a professional ear-wax remover in the last year is smaller than or equal to that of Vietnamese consumers.

What if our initial hypothesis is *H2* instead of *H1*? No problem. We formulate another hypothesis (say, *H5*) in a similar way by combining *H1* and *H3*:

H5: The proportion of Dutch consumers who visited a professional ear-wax remover in the last year is larger than or equal to that of Vietnamese consumers.

Note that both *H4* and *H5* include the possibility of no difference between the two groups. Hypotheses of this kind are known as **null hypotheses**; whenever a hypothesis includes an equal sign, it is a null hypothesis. If you go back to *H3*, you will see that it too is a null hypothesis; the competing hypothesis it implies is simply the combination of *H1* and *H2* (call it *H6*):

H6: The proportions of Dutch and Vietnamese consumers who visited a professional ear-wax remover in the last year are not equal.

HINT 9.1	Look for any mention of equality when reading a hypothesis; if you find one, then you are dealing with a null hypothesis; if not, then you are facing an alternative hypothesis.

If you look again at hypotheses *H1* through to *H6*, you will see that the null hypotheses are accompanied by other hypotheses, which are not of the null type; we call such hypotheses **alternative hypotheses** (sometimes also known as 'research hypotheses'). An alternative hypothesis is the complement of the null hypothesis in that it postulates some difference or inequality; as such, it can *never* include a statement of equality. Professional statisticians tend to have fits and throw severe tantrums if you confuse the null and alternative hypotheses, so please keep Hint 9.1 in mind.

Hypotheses *H3*, *H4*, and *H5* are all null hypotheses, while hypotheses *H6*, *H1*, and *H2* are their respective alternative hypotheses. Altogether, as Table 9.1 shows, we can form three pairs of hypotheses concerning the proportions of Dutch ($\pi 1$) and Vietnamese ($\pi 2$) consumers who visited a professional ear-wax remover; note also that these three pairs cover all possible hypotheses that can be formulated concerning the population proportions $\pi 1$ and $\pi 2$.

Table 9.1 Examples of null and alternative hypotheses

	Case A	Case B	Case C
Null hypothesis	$\pi 1 \leq \pi 2$	$\pi 1 \geq \pi 2$	$\pi 1 = \pi 2$
Alternative hypothesis	$\pi 1 > \pi 2$	$\pi 1 < \pi 2$	$\pi 1 \neq \pi 2$

Why should we bother distinguishing between null and alternative hypotheses? Does it really matter that some hypotheses are null hypotheses and others alternative hypotheses? The answer is an unqualified yes: it *does* matter, and it matters *a lot*. This matters since we can never *directly* test an alternative hypothesis – as we shall see shortly, our hypothesis-testing procedures can only deal with null hypotheses; it is only when a null hypothesis is *rejected* as being untenable that we obtain *indirect* support for the corresponding alternative hypothesis. As we cannot test *H1*, *H2*, and *H6* directly, what we do is test their respective *null* hypotheses (i.e., *H4*, *H5*, and *H3*); this is why we always form pairs of hypotheses (such as the ones shown in Table 9.1). Indeed, we have to unashamedly admit that we can never prove anything. The best we can do is to show that an idea is untenable because it is associated with an unsatisfactorily small probability. Expressed differently, statistical tests are designed to *disprove* hypotheses, and the particular hypothesis we are trying to disprove is always the null hypothesis. There are good reasons why we have to follow this 'roundabout' procedure in hypothesis-testing, which is more formally known as **Null Hypothesis Significance Testing (NHST)**. While a formal explanation is beyond the scope of this text, by the time you finish this chapter you should get a rough idea (see also Further Reading section).

WARNING 9.1 Only null hypotheses can be tested – if they are rejected, this is taken to signify support for the alternative hypothesis. We can never test an alternative hypothesis directly, nor can we ever prove a null hypothesis.

Bearing the above in mind, if you compare the three alternative hypotheses in Table 9.1, you will notice that all three postulate the *existence* of differences in the proportions of Dutch and Vietnamese consumers who visited a professional ear-wax remover in the last year. However, in Cases A and B, the *direction* of the difference is also specified (e.g., $\pi 1$ is bigger/smaller than $\pi 2$), while in Case C it is not (i.e., $\pi 1$ is different to $\pi 2$). Alternative hypotheses, which in addition to the existence of differences also indicate the direction of the expected differences, are known as **directional hypotheses**. In contrast, hypotheses that only postulate a difference without any *a priori* expectations as to the direction of the differences are called **non-directional or exploratory hypotheses**.

The formulation of directional hypotheses presupposes greater prior knowledge about the issue at hand; such knowledge may be available from past theoretical work and/or empirical evidence. Exploratory hypotheses, on the other hand, are more 'cautious' or 'conservative' and reflect either a lack of prior knowledge or conflicting past findings; thus, one is uncertain about the direction in which a difference will be manifested. For example, if you know that there had been four previous studies in the past 20 years and all had concluded that Dutch consumers are more likely to have visited a professional ear-wax remover than Vietnamese consumers, then

it would not be inappropriate to formulate a directional hypothesis along the lines of Case A in Table 9.1. On the other hand, if two of the previous studies had shown a greater propensity of visits to professional ear-wax removers among the Dutch, one study had found no difference, and yet another indicated that the Vietnamese are more likely to visit a professional ear-wax remover than the Dutch, then an exploratory hypothesis (i.e., Case C) would be called for. An exploratory hypothesis would also be appropriate if no previous studies had been undertaken to enable you to formulate some directional expectations about visits to professional ear-wax removers in the two countries, and you had little theory to rely on. Note that the distinction between directional and exploratory hypotheses has important implications for the actual testing of such hypotheses, as we will see in the next section.

So far, we have been discussing null and alternative hypotheses relating to how two groups (in this instance, Dutch and Vietnamese consumers) may differ in terms of some characteristic (in our example, visits to professional ear-wax removers). Hypotheses concerning differences between two (or more) groups are only one of many types of hypotheses. Another type involves hypotheses concerning differences between measurements. Such hypotheses are typically formulated in the context of an experimental design, in which the same group is compared on the same variable before and after administration of the experimental treatment. For example, one may hypothesize that drinking vodka increases typing speed; to test this hypothesis, one may compare the typing speed of a group of secretaries before and after they drink a nice glass of vodka on the rocks. Similar comparisons take place in longitudinal studies, where a sample is repeatedly measured on a characteristic of interest across different points in time. As another example, not involving experimentation, consider a hypothesis suggesting that the leader of the Exceedingly Conservative Party has a more positive image among voters in Great Yarmouth than the leader of the Excessively Liberal Party. This could be tested by taking a random sample of voters and asking them to state their opinions for *both* leaders – again, the comparison would involve the same group but different measurements. Hypotheses involving comparisons between groups and differences between measurements will be examined in detail in Chapters 11 and 13.

Another type of hypothesis-testing involves the examination of *relationships* (or 'associations') between variables. For example, a sociology professor may hypothesize that the higher one's income, the greater one's consumption of strawberry ice cream; here, a positive relationship between income and consumption is postulated, indicating a directional hypothesis. A directional hypothesis would also be indicated if a negative link had been specified between income and strawberry ice-cream consumption. Had the direction of the relationship not been specified (i.e., income and strawberry ice-cream consumption are related, but we do not know *how*), then the hypothesis would have been an exploratory one.

Note that, strictly speaking, hypotheses concerning differences between groups could also be interpreted as implying a relationship; for example, postulating differences between Peruvian and Australian consumers in terms of bicycle ownership can be interpreted as hypothesizing a relationship between nationality and bicycle ownership. Be that as it may, we usually talk about relationships when 'positive' and 'negative' connotations are meaningful and readily interpretable. Thus, while a positive relationship between income and ice-cream consumption states unambiguously that the more money one has, the more ice cream one is likely to eat,

a positive relationship between nationality and bicycle ownership is not as clear. The reason for this is that nationality is a *nominal* variable and, as such, it contains no meaning of order (see Chapter 3). Therefore, the values that this variable can assume are arbitrarily determined and, as a result, any 'relationship' between it and another variable is dependent upon how the variable levels have been coded. Thus, a 'positive' link between nationality and bicycle ownership with Peruvian = 1 and Australian = 0 automatically becomes a 'negative' link if we code Peruvian = 0 and Australian = 1! However, the substantive nature of the link is clearly the same in both instances. In general, when a nominal variable is involved, it is much better to state hypotheses in terms of differences and reserve the term 'relationship' for hypotheses involving variables measured at (at least) ordinal level. Hypotheses involving relationships will be examined at length in Chapters 12 and 13.

WARNING 9.2 Inadequate attention to the basic logic and steps associated with hypothesis-testing is bound to result in confusion, frustration, and erroneous interpretations when specific tests are faced.

In addition to hypotheses regarding differences (between groups or measurements) and hypotheses concerning relationships, there are *single-variable* hypotheses. Such hypotheses can relate to postulated values of particular population parameters (e.g., that a mean, median, or standard deviation is equal to a certain value), the shape of the distribution (e.g., that a variable follows a normal distribution), and the nature of the observations (e.g., that a set of observations is indeed random). Hypotheses of this kind will be the subject of Chapter 10. For your convenience, Figure 9.1 provides a 'road map' for the various types of hypotheses covered in the rest of this book.

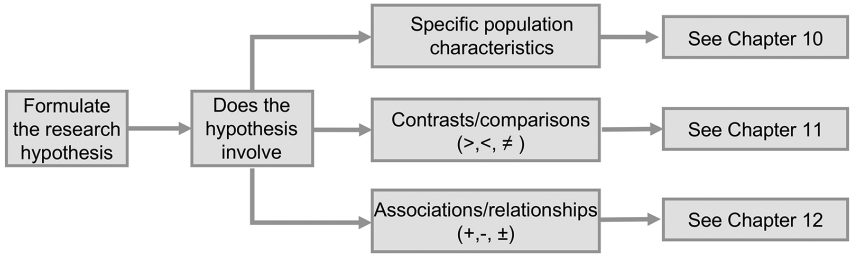


Figure 9.1 Types of hypotheses

We are now ready to tackle the more 'technical' part of hypothesis-testing. In what follows, we shall go through the hypothesis-testing procedure step by step, highlighting the main issues along the way. The principles of hypothesis-testing are the same irrespective of (a) the type of hypothesis under consideration and (b) the specific statistical test involved. As long as you grasp the key concepts discussed in the next section, you should be able to deal with any hypothesis and have no problem whatsoever understanding the specific techniques presented in Chapters 10 to 12. Conversely, failure to devote sufficient time to understanding the basic

principles underlying the testing of hypotheses means that you are likely to end up in tears and blame us endlessly for tempting you to buy this book.

HINT 9.2	Hypotheses involving a nominal variable are best stated as differences between groups. Hypotheses involving variables at higher levels of measurement are best stated as relationships.

A GENERAL APPROACH TO HYPOTHESIS-TESTING

As Table 9.2 shows, there are five distinct steps associated with formulating and testing a hypothesis. While some of these steps will normally be carried out by your computer, it is imperative that you understand what is happening at each step and why. Consequently, we will take a hands-on approach here and illustrate the general principles involved by the use of a specific example without the use of a computer (shock, horror!); yes, this means that there will be some calculation involved, but it's for your own good!

Table 9.2 Steps in hypothesis-testing

1.	Formulate the null and alternative hypotheses.
2.	Specify the significance level.
3.	Select an appropriate statistical test.
4.	Identify the probability distribution of the test statistic and define the region of rejection.
5.	Compute the value of the test statistic from the data and decide whether to reject or not reject the null hypothesis.

Step 1: Formulation of null and alternative hypotheses

The first step in Table 9.2 is rather obvious; however, it is also extremely important because it largely determines what happens next. As already mentioned, the null hypothesis should contain a statement of equality (i.e., either =, ≥, or ≤). By convention, a null hypothesis is denoted as *H0* (read: H-nought) and is always given the benefit of the doubt; that is, it is *assumed* to be true unless it is rejected as a result of the testing procedure. Note that inability to reject the null hypothesis does not *prove* that *H0* is actually true. It may be true, but our tests are only capable of disproving (but not confirming) a (null) hypothesis. In this context, evidence supporting a null hypothesis is never conclusive, whereas contradictory evidence is sufficient to confidently reject a null hypothesis.

To illustrate this logic, consider a claim by a friend of yours that drinking a cup of coffee with three spoons of pepper will cure any hangover within five minutes. Next time you have a bit to drink, you decide to test his hypothesis and drink this revolting concoction; within five minutes, you feel fine! Now, would you, on the basis of a single trial, conclude with certainty that it was your friend's amazing remedy that cured your hangover? Probably not. Nevertheless, you may be encouraged enough to repeat 'the treatment' when you get your

next hangover. Imagine now that after successfully treating 128 major hangovers within the prescribed time, this time your hangover does *not* go away within five minutes (but takes three agonizing days to clear!). This one incident is enough to invalidate the hypothesis that the remedy is effective within five minutes; it took one piece of contradictory evidence to reject this hypothesis while 128 instances of results in favor of the hypothesis were not sufficient to conclusively prove it.

If, as a result of testing, the null hypothesis is rejected, this is interpreted as signifying support for the alternative hypothesis, which, again by convention, we denote as *H1* (read: H-one). Remember that since *H0* and *H1* are complementary, the alternative hypothesis should always include a statement of inequality (i.e., ≠, >, or <). Depending upon whether the alternative hypothesis is exploratory or directional in nature, the form of the null hypothesis will also vary (see Table 9.1 earlier); as we will see, this has important implications for Steps 4 and 5.

Okay, we've conceptualized enough. Let us illustrate some of the points discussed above by means of an example. Suppose that we wish to test a hypothesis concerning the average number of monthly visits to the hairdresser by teenagers. Having researched the voluminous literature on teenagers' habits, we have unearthed 148 empirical studies in 73 different countries, indicating that, on average, teenagers go to the hairdresser between 2 and 13 times each month. None of the studies, however, have been conducted in the United Kingdom, and we see this as a unique opportunity to make our mark in the scientific community by providing much-needed UK-based evidence. Bearing in mind the findings of previous research, we speculate that UK teenagers will visit the hairdresser *more than* three times a month. This is our alternative hypothesis (*H1*), which means that our null hypothesis (*H0*) is that the average number of visits is three times or fewer (in practical applications, you may generally find it easier to state the alternative hypothesis first – as was done here – and then 'derive' the relevant null hypothesis from it). Denoting the population mean with the beautiful Greek letter μ (as we did in Chapter 8), we can formally state our hypotheses as follows:

$H0: \mu \leq 3$
$H1: \mu > 3$

Now, what we want is to be able to reject *H0* in favor of *H1* (which will make us famous when we publish our findings in the prestigious *Journal of Hairdressing Research*). We thus set out to see whether we should or should not reject *H0* (always remembering that failure to reject does not mean that we have proved *H0* – see Warning 9.1 above). We first have to collect some data, so we take a random sample of 200 teenagers across the UK (recall from Chapters 2, 5, and 8 that we *must* have a probability sample in order to engage in statistical inference). Next, we interview our sample, and we find that the mean number of visits to the hairdresser is 3.5 with a standard deviation of 2.4. Where do we go from here?

Step 2: Specification of significance level

Having formulated the null and alternative hypotheses, you should take a break. Don't exhaust yourself – research should be fun! After the break, the next step is to specify the circumstances under which we will reject *H0* and not reject *H0*. Bearing in mind that we do not know for sure whether *H0* is true or false in the population (otherwise, there is no point in hypothesis-testing), a rejected *H0* may be true or false; similarly, a non-rejected *H0* may also be true or false. Thus, there are four possible outcomes whenever we test any hypothesis, namely:

1. Not rejecting *H0* when *H0* is true (i.e., failing to reject a true null hypothesis).
2. Rejecting *H0* when *H0* is true (i.e., rejecting a true null hypothesis).
3. Not rejecting *H0* when *H0* is false (i.e., failing to reject a false null hypothesis).
4. Rejecting *H0* when *H0* is false (i.e., rejecting a false null hypothesis).

As Table 9.3 indicates, outcomes 1 and 4 are desirable, representing correct decisions on our part. On the other hand, outcomes 2 and 3 are undesirable since they indicate that we got it wrong. Our decisions are erroneous since they do not correspond to the true situation in the population. However, the two types of errors are very different. What we term **Type I error** occurs if we reject the null hypothesis when we should not (i.e., when the null hypothesis is, in fact, true). In contrast, **Type II error** occurs if we do not reject the null hypothesis when we should reject it (i.e., when the null hypothesis is false). Put differently, you commit a Type I error when you find an effect that, in reality, does not exist (your result is a '**false positive**'), whereas you commit a Type II error when you do not find an effect that, in reality, exists (your result is a '**false negative**').

If only to improve our knowledge of the Greek alphabet, let us call the probability of committing a Type I error 'α' and the probability of committing a Type II error 'β'; as *H0* can be *either* true *or* false (but not both), the probabilities of making a correct decision are 1−α and 1−β, respectively. Our problem is that since we do not know in advance whether *H0* is true or false, we need to guard against the possibilities of making *both* a Type I *and* a Type II error. Thus, ideally, we would like to keep α and β as low as possible so as to maximize the probabilities of reaching a correct decision. Unfortunately, there is a snag: for a fixed sample size, decreasing α results in an increase in β, and decreasing β results in an increased α. Only by increasing the sample size will both errors decrease (the exact relationship between α and β is quite complex and discussing it here will probably confuse rather than enlighten; for more details, see the Further Reading section). This emphasizes, yet again, how important sample size is.

Table 9.3 Errors in hypothesis-testing

Decision made	Situation in the population	
	H0 is true	*H0* is false
H0 is not rejected	Correct decision (1−α)	Type II error (β)
H0 is rejected	Type I error (α)	Correct decision (1−β)

Bearing the above in mind, we are faced with a dilemma: should we minimize the risk of committing a Type I error (i.e., keep α as small as possible), or should we be more concerned with making a Type II error (and thus opt for a low level of β?). The way that we think about resolving this dilemma is best illustrated by the famous 'judicial analogy' shown in Table 9.4. Here, the two possible true conditions are paired with the two possible verdicts. Ideally, we want to find the defendant innocent when he/she is, in fact, innocent and guilty when he/she is, in fact, guilty; thus, the top-left and bottom-right cells of Table 9.4 indicate correct decisions. What about the implications of getting it wrong? Are the two 'error cells' (i.e., bottom left and top right) equally undesirable? In other words, can we view the condemnation of the innocent in exactly the same way as the release of the guilty? Most people would not: finding an innocent person guilty when he/she, in fact, is innocent is seen as a much graver error than failing to convict a guilty person. Indeed, the judicial principle of 'innocent until proven guilty' that characterizes modern law is based upon the recognition that the two types of error in Table 9.4 are not of equal importance; making a Type I error is seen as a much more serious failure of the judicial process than making a Type II error.

Recalling that in hypothesis-testing, the null hypothesis is always given the benefit of the doubt (see step 1 earlier), we have an 'innocent until proven guilty' situation. The implication is that we should be particularly careful not to reject *H0* unless we have very strong evidence against it. This, in turn, suggests that we need to minimize the risk of wrongly rejecting *H0*; that is, committing a Type I error, α. We denote α as our **significance level** and use it to indicate the maximum risk we are willing to take in finding a false positive; that is, rejecting a true null hypothesis: the less risk we are willing to assume, the lower the α. Typical values for α are 0.10, 0.05, 0.01, and 0.001; however, these are largely arbitrary and reflect tradition as much as anything else.

Table 9.4 Judicial analogy of Type I and Type II errors

	True situation: defendant is	
Verdict	Innocent	Guilty
Innocent	Correct decision (1−α)	Wrong decision (β)
Guilty	Wrong decision (α)	Correct decision (1−β)

As was the case with interpreting a confidence interval (see Chapter 8), confusion often reigns in the interpretation of a significance level. Among the most common misinterpretations is that the significance levels show the probability that the results occurred by chance; that one minus the significance level shows the confidence we can have in the alternative hypothesis; and that the significance level shows the probability that the null hypothesis is true. None of these interpretations is even remotely right. You should always associate a significance level with *a probability of making a mistake* and, at that, a particular kind of mistake: finding a false positive or, more formally stated, rejecting the null hypothesis when you shouldn't reject it (because it is true). Thus, when we select the 5% significance level (i.e., set $\alpha = 0.05$) to conduct a hypothesis test, what we are saying is that we will conduct our test in such a way that only five times out of 100 the result will be a false positive. Adhering to formal statistical articulation, this is to say that we will reject the null hypothesis when, in fact, it is true only five times out of

100. In our example of teenager visits to the hairdresser, if the true number of visits is three or fewer (as our null hypothesis specifies), we would wrongly conclude that the number of visits is *not* three or fewer only five times out of 100; in other words, we would make a wrong ruling against a true null hypothesis relatively rarely.

HINT 9.3	Always remember that the significance level is a probability of making a mistake: rejection of a true null hypothesis.

Having specified a significance level, the way we use it is simple. If the result of a statistical test has a probability of occurrence less than or equal to α, then we reject *H0* in favor of *H1*, and we declare the test result as *significant*. If, on the other hand, the probability associated with the test result is greater than α, then we cannot reject *H0* and we denote the test result as *non-significant*. Thus, by using a particular significance level, we can 'partition' all possible test results into (a) those that lead us to reject the null hypothesis (and, thus, indirectly support the alternative hypothesis) and (b) those that prevent us from rejecting the null hypothesis.

Note that in hypothesis-testing, we run a statistical test assuming that the null hypothesis is true and use it in conjunction with a significance level to decide whether or not to reject a null hypothesis. This is the reason such statistical tests are often referred to as **significance tests** or simply as **significance testing**.

Let us go back to our example of teenager visits to the hairdresser and agree on $\alpha = 0.05$ as our significance level. Note that we should really decide on the size of α *before* we collect any data (otherwise, by 'tweaking' α after we know what the sample data look like, we can influence whether *H0* is rejected or not, and this is a very questionable practice – see also step 4 below). Recall that the mean number of monthly visits to the hairdresser came to 3.5. Our problem is to determine the probability of obtaining such a result under the assumption that the null hypothesis is true (i.e., that the actual number of visits to the hairdresser is three or fewer). If this probability is lower than or equal to 0.05 (our significance level), we could reject the null hypothesis and rule in favor of our alternative hypothesis (i.e., that the number of visits is more than three). If, on the other hand, the probability is greater than 0.05, we cannot reject the null hypothesis and, thus, we have no evidence in support of the alternative hypothesis. In either case, to determine this probability, we need some sort of statistical test; the selection of an appropriate statistical test is the next and an extremely important step in hypothesis-testing. Thus, make sure you are well rested before you go on. The next step demands your undivided attention! Consider asking one of your friends or classmates to give you a massage before you continue.

Step 3: Selection of an appropriate statistical test

A statistical test is merely a technique that can be used to test a particular hypothesis. There is little doubt that the selection of an appropriate statistical test is the most difficult and frustrating step in hypothesis-testing. The reason is simple: there are so many of them (imagine the unemployment rate among statisticians if there were one test you could apply to all possible hypotheses!). Choosing among the multitude of tests currently available for testing every conceivable type of hypothesis is not exactly an enjoyable activity and can put you off data analysis

for life! Unfortunately, it is a necessary evil because if you select the wrong test for your type of hypothesis and/or data, the rest of the hypothesis-testing procedure goes out of the window (i.e., your results, significant or not, will be meaningless).

Before you get totally depressed, let us cheer you up by mentioning that, in Chapters 10 to 12, we have already selected for you those tests that you are most likely to need. Moreover, not only do we show you *how* to use *which* test for *what* type of hypothesis, but we also highlight the key assumptions and limitations underlying individual tests so that you avoid the pitfalls involved (what more could you possibly ask for?). Consequently, in this section, we discuss only the broad considerations involved in choosing an appropriate test for your needs and provide a specific illustration applied to our unforgettable example of teenager visits to the hairdresser.

Several factors enter into the choice of a statistical test. Many of these have already been mentioned in Chapter 5 in the context of the broader discussion of the influences affecting the choice of the method of analysis (so it would not harm you to go back and have another look at Chapter 5 and, in particular, Figure 5.1). Nevertheless, it is useful to look at choice criteria as relating specifically to techniques appropriate for hypothesis-testing purposes.

First, the *type* of hypothesis to be tested will require a different test each time; for example, different tests are appropriate for hypotheses concerning differences between groups compared with those for hypotheses concerning relationships between variables. Moreover, even within a broad class of hypotheses, different tests may be required; for example, one test might be more appropriate for testing differences between two means than, say, between three means or two medians. Finally, even for exactly the same hypothesis, a different test may be needed depending on your sample size; for example, testing a hypothesis about a proportion with a large sample is done with a different test from that used with a small sample.

Second, the *distributional assumptions* made regarding the population(s) from which the sample(s) was drawn will affect the choice of test. A typical assumption, in this context, is that the sample data have been drawn from a normally distributed population; this assumption permeates a great many statistical tests, including tests for hypotheses involving means, variances, proportions, and correlations. Another assumption that often crops up in hypotheses involving comparisons is that the samples come from populations with equal variances; tests of mean differences across two or more groups usually have this requirement. As already mentioned in Chapter 5, statistical tests that make (often stringent) assumptions about the nature of the populations from which the sample data are drawn are known as **parametric tests**; tests that do not make such assumptions are known as **non-parametric tests** (or 'distribution-free' tests). Although minor violations of the distributional assumptions of many parametric tests do not totally invalidate their results if the sample is large enough, small samples coupled with substantial deviations from distributional assumptions and/or measurement requirements (see below) are a sure recipe for disaster.

Third, the *level of measurement* of the variable(s) involved in the hypothesis under consideration is also relevant. Some tests are appropriate for nominal data (e.g., when frequencies are compared across two groups), other tests are suitable for ordinal data (e.g., when two rank orders are related to one another), and still others should really only be used on interval/ratio data (e.g., when differences between means are examined). A point made in Chapter 5

and worth repeating here is that parametric tests should only be applied to metric data (i.e., at least interval) and that parametric tests work best with 'large' sample sizes (i.e., at least 30 observations per variable/group); a trip back to Figure 5.2 in Chapter 5 at this point is highly recommended.

Given the restrictive nature of parametric statistical tests, you may wonder why we bother with them. If non-parametric tests can do the job, what is the benefit of using their parametric brethren? The answer is provided by the fourth (and, you will be glad to know, last) consideration affecting the selection of a statistical test, namely its *power*. The **statistical power** of a test is defined as the probability of rejecting the null hypothesis when it should be rejected (i.e., when it is in fact false). Reference to Table 9.3 shows that power is the complement of making a Type II error; that is, it is equal to $1-\beta$. In our judicial analogy in Table 9.4, power is illustrated in the bottom-right cell where the guilty defendant is indeed found guilty by the court. It is in terms of power that parametric tests have the upper hand over non-parametric tests. Specifically, given a sample size and a null hypothesis that could be tested by both a parametric and a non-parametric test, the former test will be more powerful in terms of its ability to detect and reject a wrong null hypothesis (i.e., detect a true effect). This is not surprising since the more restrictive the circumstances in which a test is expected to apply (i.e., the more assumptions that must be satisfied), the more sensitive the test will be *if* these circumstances apply (i.e., if the assumptions are met). Bearing in mind that the power of the test increases with increases in sample size (see earlier), a parametric test will be more *efficient* than a corresponding non-parametric test as a lower sample size will be required by the former to achieve the same power as the latter. What all this boils down to is that if the assumptions of a parametric test are satisfied (and some of them can actually be tested, as we will see in the next few chapters), opting for a non-parametric test makes little sense as one would be losing out on power. Moreover, many parametric tests are relatively **robust**; that is, relatively insensitive to mild-to-moderate departures in their assumptions (particularly with a large sample), which further encourages their use. Note that the degree of robustness is *test-specific* – that is, not all parametric tests are equally robust. Having said all that, if serious violations of the underlying assumptions are evident, turning a blind eye and persisting with the parametric test is a cardinal sin (punishable by having to memorize and then recite backward the 138-volume *Encyclopaedia of Advanced Statistical Procedures*!).

WARNING 9.3 Do not let your lust for (statistical) power drive your test choice. A powerful test with violated assumptions is not powerful – it is simply invalid.

Let us go back to our example and try to apply some of the riveting issues discussed above. Here's what we have: we want to test a hypothesis about a mean, we have a nice large sample ($n = 200$), our data are ratios (number of monthly visits to the hairdresser), and, obviously, the more powerful the test we can come up with, the better. Given these specifications, the appropriate test to use is the ***z*-test** for a population mean; this is a parametric test that assumes either that the population is normally distributed and the population variance is known (the latter being as realistic as saying 'It never rains in Ireland') or that the sample is large and the popu-

lation variance can be approximated by the sample variance (i.e., $n > 30$). The z-test makes use of a distribution that you are already familiar with, namely the standard normal distribution; indeed, as we will see in the next section, its test statistic is directly interpretable as a standard score (see Chapters 7 and 8 to refresh your memory on z-scores). Note that the z-test is not the only test that can be used to test a hypothesis concerning a population mean (another one, based on the t-distribution, will be introduced in Chapter 10); nevertheless, it is the best choice given the characteristics of our example data. Here's how it works.

Step 4: Identification of the probability distribution of the test statistic and definition of the region of rejection

Each test generates what is known as a **test statistic**, which is a measure for expressing the results of the test. The test statistic is a single numerical value that summarizes your sample data using a mathematical formula (an equation), which reflects your hypothesized effect. Needless to say that the formula of a test statistic differs depending on the research hypothesis at hand and the appropriate statistical test involved (otherwise life would just be too easy, wouldn't it?). Table 9.5 provides examples of different tests along with their test statistics.

Table 9.5 Examples of statistical tests with their test statistics

z-test (one-sample)	$z = \frac{\bar{\chi} - \mu_0}{\left(\sigma / \sqrt{n}\right)}$
t-test (two-sample)	$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
ANOVA (analysis of variance)	$F = \frac{\sum_{j=1}^{k} n_j \left(\bar{y}_j - \bar{y}\right)^2 / (k-1)}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} n_j \left(\bar{y}_j - \bar{y}\right)^2 / (k-1)}$
$\chi 2$ (chi-square) test	$\chi^2 = \sum^{\kappa} \frac{\left(O_{ij} - E_{ij}\right)^2}{E}$

The most perceptive of you might have noticed that every test statistic in Table 9.5 is a ratio. That's right! You might not (nor should you) be able to remember the exact formula for every statistical test (besides, you've got statistical software packages and the Internet for this), but

you should be able to remember that virtually all test statistics quantify a hypothesized effect in relation to the sampling error involved.

In any one application, the numerical value of the test statistic will obviously reflect the particular sample data at hand. However, the test statistic itself is a variable quantity that can assume many different values – indeed, the probability distribution of a test statistic indicates all values that the statistic can assume when *H0* is true. Thus, the distribution of the test statistic allows us to assess sampling error. Clearly, for a test statistic to be of any use, its probability distribution should be calculable; otherwise, we could not 'partition' our test results into those that would qualify as being significant and those that would be non-significant (see step 2).

In many statistical tests, the probability distribution of the test statistic takes known forms (e.g., a normal distribution). In other tests, the probability distribution of the test statistic *approximates* known forms. Finally, in those instances where the probability distribution needs to be re-calculated every time the test is applied, computer packages will do the hard work for you, so don't panic! In Chapters 10 to 12, you will get to know intimately several test statistics as you go through the various exciting tests we have carefully selected for you. Here, we will simply illustrate with our example the role that the test statistic plays in the hypothesis-testing process.

The test statistic associated with a *z*-test for a population mean is known as the ***z*-statistic**, which is simply a *z*-score defined as follows:

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ = sample mean, μ0 = hypothesized value of the population mean, *s* = sample standard deviation, and *n* = sample size (strictly speaking, the population standard deviation (σ) should be in the denominator as in Table 9.5; however, since we do not know it, we 'cheat' a bit by using *s* instead; this is OK as our sample is large.

The *z*-test statistic follows the standard normal distribution, which we discussed extensively in Chapter 7 and used in Chapter 8; recall that the properties of the standard normal distribution are known and that there are statistical tables available for using this distribution in practice. Thus, all *we* have to do is decide on the range of values of *z* that would lead to the rejection of the null hypothesis; in other words, establish the **rejection region**. All other values would indicate that the null hypothesis should be 'accepted' and, thus, define an **acceptance region** (but read Warning 9.1 for a cautious interpretation of 'acceptance' as non-rejection). The value of *z* that separates the rejection and acceptance regions is called the **critical value**. Assuming we have established the region of rejection, then if the computed value of the test statistic based on the sample data falls within the rejection region, we reject *H0* and take this as evidence in support of *H1*. If, on the other hand, the computed value of the test statistic falls within the acceptance region, then we have no grounds for rejecting *H0* (and thus find no support for *H1*).

How do we decide on the rejection region? Well, it is here that our choice of significance level comes into play. Specifically, we use the significance level we set in step 2 to identify the

'unlikely' values of z, given that our null hypothesis is true. In our example, we decided that $\alpha = 0.05$, so we need to partition the distribution of the test statistic into an area covering 5% of the values (the rejection region) and another area covering the remaining 95% of the values (the acceptance region). Bearing in mind that for our specific null hypothesis, rejection implies that the mean has a *greater* value than that hypothesized, it becomes evident that the region of rejection should be located at the *right* tail of the distribution (in other words, it is large values of z that should be included in the rejection region). With these considerations in mind, we look into our standard normal tables and find that, at $z = 1.645$, we can define the region of rejection (see Figure 9.2(a)). This is our critical value since 5% of the values of the test statistic exceed 1.645 and 95% are below 1.645; or, what amounts to the same thing, the probability that the test statistic exceeds 1.645 given that the null hypothesis is true is 0.05 (our significance level α), and the probability that the test statistic does not exceed 1.645 is 0.95 (which is equal to $1-\alpha$ in Table 9.3 earlier).

In the light of the above, we can restate our decision rule for rejecting *H0* as follows: if, assuming *H0* is true, the probability of obtaining a value of the test statistic as extreme as or more extreme than the critical value is less than or equal to α, we reject *H0*. Otherwise, we do not reject *H0*. Since the significance level is instrumental in determining the region of rejection, the latter is also widely referred to as the **significant region**. Note that any change in the significance level (e.g., from 0.05 to 0.10) will affect the definition of the rejection region and, thus, will have an impact on whether the null hypothesis will be rejected or not. Therefore, it is not legitimate to change the level of significance retrospectively (e.g., if you see that, given your data, the null hypothesis cannot be rejected at, for example, the 5% significance level, but it would, for example, at the 10% level, it is cheating if you report significant results at the more 'liberal' level).

For reasons already discussed, the region of rejection in our example was established in the right-hand tail of the distribution of the test statistic. If our null hypothesis had been of the form $H0{:}\mu \geq \mu 0$, then the region of rejection would have been established in the left-hand tail of the distribution (see Figure 9.2(b)); however, the rationale for and determination of the relevant region would be identical to those just described (the only difference being that the 'extreme' values of the test statistic in the rejection region would be small rather than large).

WARNING 9.4 Redefining the region of rejection in light of the sample data is not a legitimate practice.

What if, instead of a directional hypothesis, we have an exploratory hypothesis, leading to a null hypothesis of the form $H0{:}\mu = \mu 0$? Clearly, in this instance, the extreme values of the test statistic calling for the rejection of the null hypothesis could be large or small – so where's the region of rejection? In *both* tails, of course! What we do is the following little trick: we take our significance level and chop it in half. Then if, say, $\alpha = 0.05$, we define *two* regions of rejection each, capturing 0.025 of the area under the curve (i.e., each rejection area is equal to $\alpha/2$). This is shown in Figure 9.2(c). Note that we now have *two* critical values ($z = -1.96$ and $z = 1.96$, respectively, as our standard normal tables show), which are symmetrically placed and enclose between them the acceptance region (i.e., $1-\alpha$).

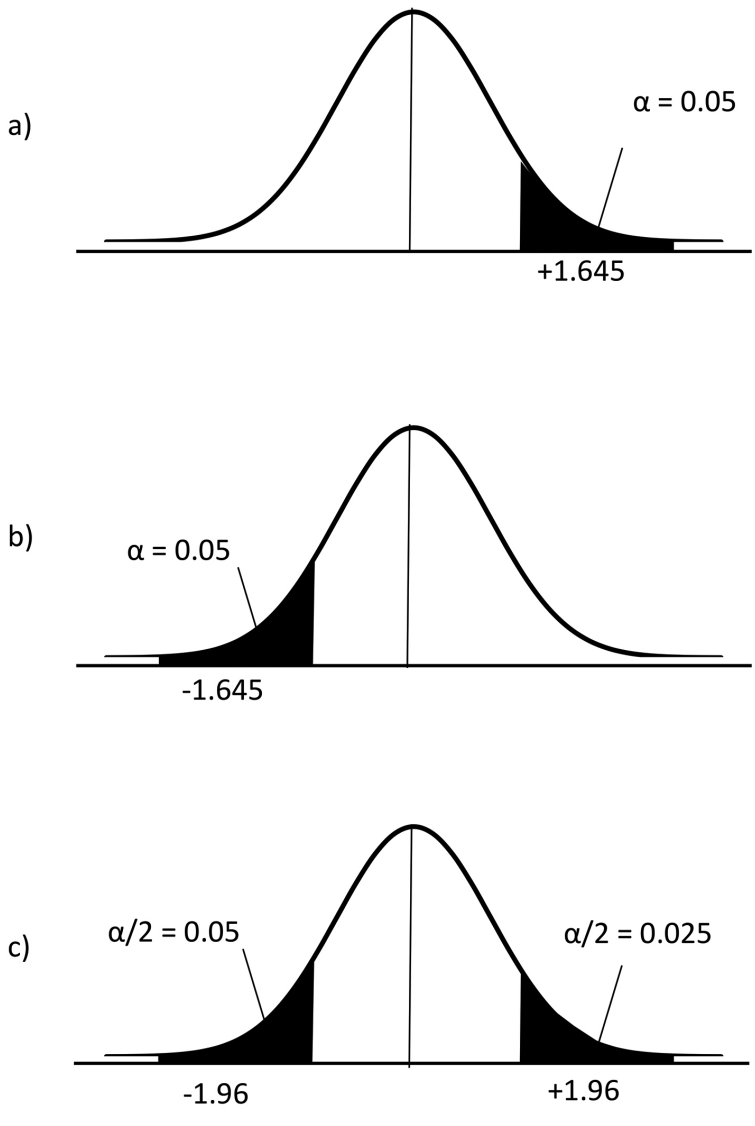


Figure 9.2 The region of rejection

By convention, whenever a statistical test is used with the region of rejection defined in one tail of the test distribution only, we speak of a **one-tailed test**; if both tails of the distribution are used in defining the region of rejection, we speak of a **two-tailed test**. A one-tailed test is appropriate when a *directional* alternative hypothesis is specified (which, as already mentioned in step 1, implies considerable prior knowledge about the nature of the hypothesized phenom-

enon), while a two-tailed test is the one to use if an *exploratory* hypothesis is all that can be reasonably specified (because of the absence of or conflicting prior knowledge).

HINT 9.4	If your alternative hypothesis is directional (i.e., includes a > or < sign), then use a one-tailed test. If it is only exploratory (i.e., includes a ≠ sign), use a two-tailed test.

From the above, you should be able to work out why the way in which the null and alternative hypotheses are stated is important (see step 1). If the alternative hypothesis is directional in nature, then only *one* tail of the distribution will be used to define the region of rejection, and this one tail will contain all the extreme values for rejecting the null hypothesis. If, on the other hand, the alternative hypothesis is exploratory in nature, *both* tails will have a region of rejection, each containing half the extreme values for rejecting the null hypothesis. Consequently, *at the same level of significance*, a one-tailed test will result in the rejection of the null hypothesis more often than a two-tailed test; that is, it is *always easier to reject the null hypothesis with a one-tailed test*. This becomes immediately obvious if you compare the critical values in Figure 9.2(a) and (c); although the significance level is 0.05 in both cases, the critical value of z in Figure 9.2(a) is 1.645, while in Figure 9.2(c) it is 1.96. Thus if z turns out to be, for example, 1.853, the null hypothesis would be rejected in the former case but not in the latter (a similar conclusion is reached if Figures 9.2(b) and (c) are compared). The corollary of this is that if a two-tailed test gives a significant result, then its corresponding one-tailed test will also produce a significant result (but the reverse does not always hold). Note that changing a hypothesis from exploratory to directional *after* the data have been collected (so as to justify the use of a one-tailed test) is not permissible; instead of being theory-driven, you become data-driven, and this defeats the very purpose of hypothesis-testing. Also important: if someone tries to fool you by suggesting a three-tailed test, there is something drastically wrong! As such a test does not exist, we recommend severely slapping the person's wrist and breaking off all contact for at least a day and a half.

WARNING 9.5 Reformulating a hypothesis from exploratory to directional after the data have been collected is a no-no.

Step 5: Computation of the test statistic and rejection or non-rejection of the null hypothesis

Having defined the region of rejection in the light of your hypothesis and chosen the significance level, all that remains is to compute the value of the test statistic using the information from your sample and compare it to the critical value(s). In our wonderful example, the computed value of the test statistic is as follows (see step 1 to remind yourself where these figures came from):

$$z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{3.5-3}{2.4/\sqrt{200}}=2.94$$

Going back to Figure 9.2(a), we can see that the critical value (at the 5% significance level) is 1.645. As our computed test statistic exceeds the critical value, it falls within the region of rejection, and therefore we reject *H0* in favor of *H1*. Thus, we conclude that the average number of monthly visits to the hairdresser by teenagers is more than three. Fame and fortune, at last! A publication in the prestigious *Journal of Hairdressing Research* is within reach!

As previously described, the test statistic is characterized by a known (or approximate) distribution of values. Therefore, you can also use the probability distribution of the test statistic values (i.e., their sampling distribution) to identify how likely or unlikely different test statistics are to occur. In times past, one would have to look at long statistical tables to find the probability associated with different test statistics. Thankfully, computer science and human laziness have joined forces, and today any decent statistical software package will not only calculate the test statistic for you but also provide you with its corresponding probability of occurrence. The probability associated with a test statistic is known as the **p-value** and shows how likely it is to get a test statistic at least as big as the one observed if no effect exists in the population (i.e., if the null hypothesis is true). In our example, the *p*-value associated with a *z*-value of 2.94 is 0.0016. This shows us that, if the null hypothesis were true, we would have observed a test statistic of 2.94 (or more extreme) less than twice in 1,000 times.

The *p*-value provides a bit more information on how far down in the significant region our result lies. Articles in scientific journals often discuss their results in terms of *p*-values: they are full of statements such as $p < 0.05$, $p < 0.01$, $p < 0.001$, and so on. Unlike critical values, which are test-specific (e.g., a critical value of a *z*-test and a critical value of a chi-square test will be very different for a given α), *p*-values represent a 'common currency' across which the results of different tests can be compared. In this context, it is better practice to report the actual *p*-value (e.g., $p = 0.004$, $p = 0.013$, etc.) rather than use the conventional cut-off points (i.e., $p < 0.01$, $p < 0.05$, etc.). This allows readers to form their own judgment about the significance of the results, as different people may feel that different levels of significance are appropriate; reporting *p*-values keeps everybody happy by enabling everyone to decide individually on the strength of the evidence against the null hypothesis. Needless to say, you *must* accompany *p*-values with some information indicating whether they relate to a one- or two-tailed test (see step 4 above); this is usually accomplished by such statements as '$p < 0.05$, sum of both tails' or 'one-tailed $p < 0.001$'. Failure to do so is bound to result in confusion.

WARNING 9.6 Whenever reporting *p*-values, do not forget to indicate whether they relate to a one- or two-tailed test.

The discussion of *p*-values concludes our presentation of the hypothesis-testing procedure. As we mentioned at the beginning of this chapter, some of the above steps will be executed by your statistical package. While you cannot really expect your computer to set hypotheses on your behalf (step 1), decide on a significance level (step 2), and choose an appropriate test (step 3), once you have made these decisions, it will automatically know which test statistic to use and how to present significant results. In the next three chapters, you will have the once-in-a-lifetime opportunity to see several examples of computer-generated output for

a variety of statistical tests carefully selected by yours truly. But we're not finished yet – not quite.

HYPOTHESIS-TESTING AND CONFIDENCE INTERVALS

Consider a situation where you want to test an exploratory hypothesis concerning, say, a mean. For the sake of convenience, let us use our existing example of teenager visits to the hairdresser but reformulate our initial hypotheses (see step 1 in the previous section) as follows:

$H0$:$\mu = 3$
$H1$:$\mu \neq 3$

Is there a way of testing *H0* at a significance level α without applying a significance test? Yes, there is – and you already know how to do it (even if you don't know that you know). Instead of following the hypothesis-testing procedure just described, let us construct a 100(1−α)% confidence interval for μ. Then if 3 is included in this interval, we do not reject *H0*; if, on the other hand, 3 is not included in the interval, we reject *H0* in favor of *H1*. So, if we set α = 0.05, we need to set a 95% interval for μ using our sample data. Following the procedures so eloquently presented in Chapter 8, the 95% confidence interval in our case is as follows (and all the strange symbols are as defined in Chapter 8):

$$\bar{x} = k_c s_{\bar{x}} \leq \mu \leq \bar{x} - k_c(s/\sqrt{n}) \leq \mu \leq \bar{x} + kc(s/\sqrt{n})$$

Placing the appropriate values into this equation results in the following calculations:

$$= 3.5 - 1.96(2.4/\sqrt{200}) \leq \mu \leq 3.5 + 1.96(2.4/\sqrt{200})$$
$$= 3.17 \leq \mu \leq 3.383$$

Since the 95% confidence interval for μ does not include 3, we reject the null hypothesis and conclude that the population mean is not 3; indeed, from the upper and lower levels of the confidence interval it can be seen that the true mean is likely to be greater than 3.

HINT 9.5	For hypotheses involving population parameters, significance testing and confidence interval estimation are equivalent procedures.

Would we have arrived at the same conclusion had we applied a two-tailed *z*-test for a population mean? Indeed, we would. The *z*-value from our data comes to 2.94 (see previous section), while the critical values for a two-tailed *z*-test and a 5% significance level come to −1.96 and 1.96, respectively (see Figure 9.2(c)). As the calculated value of the test statistic lies outside the critical values, we reject *H0*; moreover, since the relevant region of rejection turns out to be in the right tail of the distribution, we can conclude that the true population mean is likely to be higher than the hypothesized mean.

In general, hypotheses involving population parameters can be tested either by following the hypothesis-testing procedure outlined previously or by setting a confidence interval and observing whether the hypothesized value of the population parameter falls within or outside this interval. Given a significance level α, the corresponding confidence interval is 100(l–α); recalling that 1–α corresponds to the confidence coefficient associated with interval estimation (see Chapter 8), we can see that the significance level is the complement of the confidence coefficient and vice versa. In this context, the confidence interval approach to hypothesis-testing focuses on the region of acceptance, while the significance approach to hypothesis-testing focuses on the region(s) of rejection; both produce identical results.

Despite the fact that using different inferential rules in order to conclude statistical significance leads to the same outcome, within a given context one approach might be more informative than the other. Imagine, for instance, that you measure your friends' intelligence and find that with a 99% confidence level (i.e., $\alpha = 0.01$), their average score is lower than the population average of 100, and this is statistically significant ($p < 0.01$). In light of this finding, you might feel deeply concerned about your life choices and start thinking about what this says about yourself. However, if you were to look at the confidence interval around their average score and found that it ranged between 91 and 99, you would probably not be tempted to dive into the abyss of self-reflection. Table 9.6 brings everything nicely together, summarizing the inferential rules applied to decide whether your results are statistically significant or not.

Table 9.6 Inferential rules for statistical significance

Test statistic	*p*-value	Confidence interval	Decision
valueobserved ≥ |valuecritical|	$p \leq \alpha$	The (1–α)% confidence interval does NOT include the *H0* value.	Reject *H0* (Support for *H1*)
valueobserved < |valuecritical|	$p > \alpha$	The (1–α)% confidence interval DOES include the *H0* value.	Do not reject *H0* (No support for *H1*)

STATISTICAL AND SUBSTANTIVE SIGNIFICANCE

Before we move on to the final section of this incredibly exciting chapter, a few words of caution are in order. Newcomers to hypothesis-testing often make the mistake of confusing statistical significance with *importance*. Just the fact that your test produced a statistically significant value does not imply that you have an important or even remotely interesting finding. This misunderstanding has led many people (unfortunately, even professional researchers) to misuse the *p*-value as an indication of the magnitude or importance of the result. The blunt truth is that the *p*-value gives virtually *no information* about whether the results really matter! You should always have in mind that statistical significance (and, thus, a *p*-value) is a statement linking a sample to a population. Remember that you are only looking at a specific sample (out of numerous possible samples one could draw from a population) to say something about the whole population. Establishing statistical significance means that, with a given level of confidence, you can use your sample to *infer* what holds in the population. Whether your finding is important or not is a whole different story.

There are many reasons statistical significance does not equal actual importance or **substantive significance**. One has to do with the particular hypothesis you set out to test. For example, if you compare the average weight of 20 elephants with the average weight of 20 ducks, surprise, surprise, you are going to get a significant result (probably at $p < 0.00000001$!); however, reporting a significant result for a hypothesis so blatant is not going to impress many people. Bottom line: if your hypotheses are trivial, significant findings are not going to cut much ice.

Second, the fact that you have observed a significant result (even with a reasonable hypothesis) tells you nothing about the *magnitude* of the effect involved. For example, observing a significant difference between the average annual income of Greek and German cost accountants does not necessarily mean that the difference is of *practical* significance (the difference may be very small – after all, if the former earn €2 a year more than the latter, does this make them richer?). Recall that the power of a statistical test increases with the sample size; thus, even a tiny difference may turn out to be statistically significant because huge samples are involved. Indeed, with very large sample sizes statistical tests become extremely powerful and will declare even minor deviations from the null hypothesis significant. This is precisely the reason it is always good practice (if not necessary) to complement the results of significance testing with some indication of the magnitude of the observed effect, the so-called **effect size**. You can think of effect sizes as belonging to two main categories that actually correspond to the research hypothesis being tested: (a) effect sizes that express how big a *difference* is in a given comparison (e.g., Cohen's *d*) and (b) effect sizes that express how *strong* an association between variables is (e.g., Pearson's *r*). The nice thing about effect size measures is that they are standardized, and thus we can use them to draw direct and meaningful comparisons across the findings of studies conducted in different contexts and/or using different research settings and operationalizations. For most measures, there are also statistical formulae to convert one form of effect size (e.g., Pearson's *r*) to another (e.g., Cohen's *d*). The literature has provided us with rules of thumb about what constitutes a small, medium, and large effect. Table 9.7 provides an example for Cohen's *d* and Pearson's *r*. It should be emphatically noted, though, that these rules of thumb are only rough guidelines, and one should always consider what is small or big within the context of one's research field. For example, what business researchers may consider a small or negligible effect can be considered quite important in medical studies.

Table 9.7 Commonly used effect sizes

Pearson's *r*	Cohen's *d*	
$r = 0.10$	$d = 0.20$	small effect
$r = 0.30$	$d = 0.50$	medium effect
$r = 0.50$	$d = 0.80$	large effect

Third, do not be tempted to dismiss your findings because they turn out to be non-significant. Granted, a null hypothesis is often formulated with the express purpose of rejecting it as the real interest lies in the alternative hypothesis (i.e., you want to find differences, establish relationships, etc.). However, failure to reject the null hypothesis can be just as interesting/important/exciting as rejecting it, or even more so. For example, if you are testing the well-established theory that statistics professors are much more introverted than salespeople and are unable to

reject the null hypothesis, your finding would contain a high degree of surprise. Consequently, it would probably generate more discussion among the scientific community than if you had simply supported conventional wisdom one more time. After all, finding that statistics professors are *not* more introverted than salespeople (when everybody thought they were) is bound to hit the front pages! Thus, a non-significant finding is not necessarily a bad finding and, conversely, a significant finding is not necessarily a good finding.

Fourth, and related to the above point, sometimes the aim is to 'fail' intentionally; that is, *not* to reject the null hypothesis. For example, if you are applying a statistical test to check the distribution of your variable against the normal distribution, you want to get non-significant results because this would indicate that making the assumption of normality would indeed be justified (we will see an example of such a test in Chapter 10). Under these circumstances, you are deliberately interested in obtaining *null results*; that is, findings that would not lead to rejection of the null hypothesis.

Fifth, do not be tempted to run test after test on the same set of data in a mad search for significance. In today's statistical jargon, 'forcing' the analysis to fit your expectations or theory is referred to as **p-hacking** and involves applying different tests until you find support for your hypothesis. This would be perfectly fine in the context of exploratory data analysis, but it is completely unacceptable for hypothesis-testing purposes, which requires a transparent and pre-specified plan about how the researcher will test a given hypothesis (see Table 9.2 earlier).

Sixth, do not confuse the process of *discovery* with the process of *verification*. You shouldn't explore the data for significant effects and then try to develop a theory that conforms to what you have found. This shameful practice is known as **HARKing** (Hypothesizing After the Results are Known) and neglects the distinction between exploratory (discovery) and confirmatory (verification) analysis, often leading to non-replicable findings and poorly developed theories that were driven by the data of an idiosyncratic sample. That said, it is perfectly legitimate, as a result of exploratory analysis, to come up with more specific (i.e., directional) hypotheses concerning the topic at hand. However, it is *not* legitimate to test these hypotheses on the same set of data that suggested them in the first place! Consequently, you should get hold of another set of data in order to verify that your hypothesis developed previously is, in fact, reasonable. Yes, this means extra work, but who told you that the way to scientific stardom is paved with roses?

Taken together, the above comments suggest that statistical significance should never be allowed to obscure or overrule substantive significance. At the end of the day, it is the substance of your findings in terms of their implications for theory, practice, or policy that matters, not merely whether these findings happen to be significant or not. Significance tests and effect size measures are tools that enable us to investigate issues of theoretical and/or practical relevance and used alone have no intrinsic value whatsoever. You should always remember this.

STATISTICAL POWER REVISITED

We have already mentioned the power of a test in the context of parametric versus non-parametric methods. However, statistical power is too important a concept not to have a dedicated section in this chapter.

In step 2 above, we explained that in hypothesis-testing we do not know in advance if an effect exists, so there are two plausible scenarios: (1) an effect does not exist in the population (the null hypothesis is true) and (2) an effect exists in the population (the null hypothesis is false). We also said that, even though we do not actually know which scenario is true, the null hypothesis is given the benefit of the doubt, and thus the scenario with no effect in the population takes priority. That said, we inevitably give priority to minimizing Type I error (i.e., the probability of a false positive) by keeping the α level as low as possible. That's all fine, but do understand that giving priority to Scenario 1 is not to say that we should altogether neglect Scenario 2. We still don't know what holds true in the population. (Well, if we knew, we wouldn't be testing.) Thus, we should also be concerned about Type II error (the probability of a false negative or β).

WARNING 9.7 Never allow significance testing to become an end in itself.

This is where statistical power comes into play. Statistical power focuses on the probability that a test will find an effect, assuming one exists in the population (see the lower-right cell in Table 9.3). In simple words, operating with a statistical power of 80% (thus a Type II error or β of 20%) implies that if an effect does exist, 80 times out of 100 you will manage to identify it and only 20 times out of 100 you will miss it. In fact, the literature recommends 0.80 (and thus $\beta = 0.20$) as the desired minimum threshold level for power. Assuming a significance level of $\alpha = 0.05$, this recommendation considers the cost of Type I error to be four times more severe than the cost of a Type II error (i.e., $\beta/\alpha = 4$ or $\beta = 4\alpha$).

Statistical power depends on sample size, effect size, and the significance level (α level). It increases, other things being equal, with a bigger sample, bigger effect sizes, or a more relaxed significance level (e.g., an α of 10% as opposed to 5%). Importantly, given three of these elements, one can derive the fourth. For instance, you can use your test's results (effect size) and specifications (sample size and α level) to calculate how powerful your study was. Even better, you can set the desired α level (e.g., 5%), use existing literature and expert knowledge to *a priori* approximate how big a potential effect size might be or what would be the minimum effect of interest in the context of your research (e.g., Cohen's $d = 0.20$), and then calculate the sample size required to achieve a desired level of statistical power (e.g., 80%). In the unlikely case that you are interested in learning even more about power, it is worth having a look at G*Power, which is free-to-use statistical software to perform power analysis and sample size determination (http://www.gpower.hhu.de).

BEYOND STATISTICAL SIGNIFICANCE: *READ AT OWN RISK*

Let us now bring everything together and confuse you even more. Imagine that there is a new study claiming that wearing sunglasses makes people smarter and that the effect size corresponds to, say, Cohen's $d = 0.20$. Given that the study was published in a very credible scientific journal (*Journal of Optician Wealth*), you decide to use it as the foundation of your PhD project. Thus, you assume that there is indeed an estimated effect size of 0.20 in the population. At the same time, though, you know that because of sampling error, similar studies will not produce exactly the same result. If you were to repeat the original study, again and again, you would notice that sometimes the result overestimates and other times underestimates the proposed effect. That is, some studies might find a smaller effect size (thus be closer to *H0*), while others might find a bigger effect size (more strongly contradict *H0* in favor of *Ha*, where *Ha* is the alternative hypothesis). To better understand this, you need to think of two sampling distributions simultaneously, one under *H0* (i.e., sampling distribution assuming there is no effect in wearing sunglasses) and one under *Ha* (i.e., sampling distribution assuming there is an effect of $d = 0.20$). Now, suppose that the original study was based on a sample size of $n = 50$ and a 95% confidence level ($\alpha = 0.05$). The relevant specifications are represented in Panel A in Figure 9.3.

Panel A shows that the Type II error associated with the original study is 59% (region C). So, if we replicate this study again and again, 59% of the time we would not obtain significant results, and only about 4 out of 10 (i.e., 41%) of the studies would produce significant results. In simple words, the original study doesn't seem to perform better than flipping a coin! In fact, it is even worse, as the chances that the effect will occur in the future are lower than it not occurring. This is not very good news about the veracity of the effect or about the future of your PhD research. 'But the results of the original study are statistically significant. Why wouldn't I trust them?', you might mumble. Here's the thing: a study with statistically significant results is not necessarily informative. A small sample generates a large sampling error, and therefore it could be that the original study has overestimated the effect in the first place, which in reality may be smaller ($d < 0.20$) and even negligible. Maybe the original study was a false positive, meaning that it belongs to the 5% of cases that are expected to find an effect, even when no such effect exists in the population (Type I error).

Now, what if the original study had exactly the same specifications but used a bigger sample size? Panel B shows that by increasing the sample to $n = 90$, statistical power increases to 60%. Now, your trust in the original study would increase. The Type II error decreases ($\beta = 40\%$) and statistical power (region D) indicates that in the long run, 6 out of 10 similar studies will corroborate the original result. You are now starting to feel that the original study is less likely to be a false positive because more studies than not would support the proposed effect. If the original study had an even bigger sample size, things would be considerably better. With $n = 153$, the original study would reach a desirable level of statistical power (80%). Panel C shows that, with this sample size, the original study would have more effectively controlled for both types of error, and its results would be more likely to be reproduced in the long run, thus having greater evidential value.

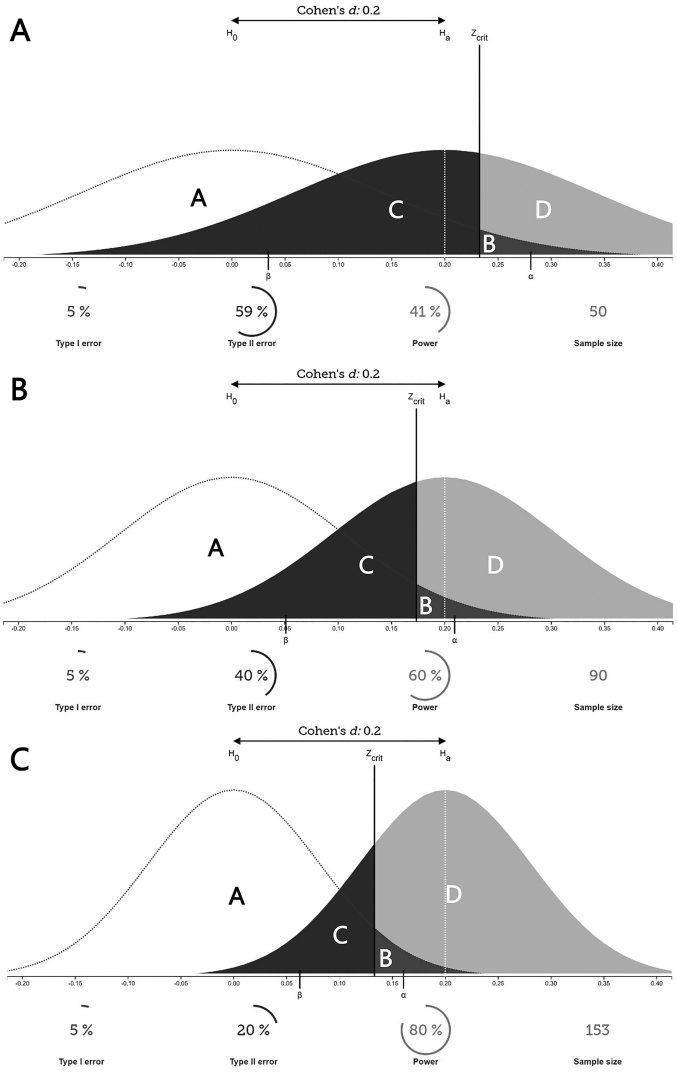


Figure 9.3 Partitioning the regions of statistical inference

Let us now try a different scenario. Imagine that you work as a microbiologist and you are commissioned to develop a new test that detects whether the concentration of procrastin (a protein that threatens one's mental state) in someone's blood is so high as to require treatment (e.g., a low-voltage intervention with a cattle prod). The old test is quite accurate but, unfortunately, it is very time consuming and not cost-effective at all. So, to test the effectiveness of the new test you developed, you recruit 50 patients who have already been diagnosed with severe procrastinitis of $d = 0.20$ (a d close to zero would indicate a healthy individual) and run a study by setting the significance level to $\alpha = 5\%$ (see Panel A of Figure 9.3). You obtain a statistically

significant result with $p < \alpha$ (i.e., your p-value lies somewhere in region B). Feeling elated that you have succeeded, you shut down your lab and start preparing a report for the Ministry of Health. Your main argument is that your new test should replace the old one, and you are presenting the relevant statistical evidence about this claim. Well, think again before you submit your report and demand an exorbitant payment for your 'new and improved' test. The test you developed has a statistical power of 41% and generates false negatives 59% of the time. This implies that about 6 out of 10 people who suffer from procrastinitis will *not* be detected and thus will not be eligible for the relevant (albeit somewhat unpleasant) treatment. Working your way down to Panel B and Panel C in Figure 9.3, you will see that with a bigger original sample, the insights obtained from your study about the new test would be much better. For instance, Panel C indicates that, in the long run, you would correctly detect 80% of people suffering from procrastinitis and wrongly identify that a healthy individual needs treatment 5% of the time.

Overall, given the α level and the effect size, as the sample size increases, estimation precision increases, and the sampling distributions under *H0* and *Hα* get narrower, indicating more statistical power. Hopefully, you now realize that running a *single* study and finding a $p < \alpha$ doesn't really mean that much in itself, especially when the statistical power of the study is not high. The results of underpowered studies are not trustworthy, often misleading, and the reason several well-known effects in the sciences have failed to pass the test of time, as they could not be successfully reproduced by subsequent studies.

As an aside, the graphs shown in Figure 9.3 were created with Kristoffer Magnusson's awesome online interactive visualization tool, which can be accessed through https://rpsychologist.com/d3/NHST. Visiting his webpage and playing around with different visualizations can help you develop a more intuitive understanding of significance testing and statistical power. But don't procrastinate!

SUMMARY

The purpose of this chapter was to introduce you to the second approach to statistical inference, namely hypothesis-testing. First, we looked at hypotheses from a number of different perspectives and highlighted the indirect nature of obtaining evidence in support of a hypothesis. We then moved on to consider the practicalities of hypothesis-testing by going through a five-step procedure, which can be used as a blueprint for testing any hypothesis. Following this, we examined the relationship between hypothesis-testing and interval estimation, discussed the notions of effect size and statistical power, and concluded with some words of wisdom concerning the role and interpretation of statistical significance. If you have grasped the issues in this chapter, then you should be able to take Chapter 10 onwards in your stride. If you are still a bit uncertain, go through this chapter once more – perhaps a bit more slowly this time.

QUESTIONS AND PROBLEMS

1. What is a hypothesis?
2. Give an example of a null hypothesis and its corresponding alternative hypothesis.
3. What are the five steps in hypothesis-testing?
4. Under what circumstances would you formulate directional hypotheses?
5. What are the two types of errors associated with hypothesis-testing and how do they relate to one another?
6. Explain what the terms 'significance level' and 'region of rejection' mean.
7. What are the main considerations in selecting an appropriate significance test?
8. When would you use a one-tailed test versus a two-tailed test? Use an example to illustrate your answer.
9. Why should statistical significance be distinguished from substantive significance?
10. What is statistical power, and why should you bother with it?
11. How can you use your knowledge of hypothesis-testing to impress your least-liked relative?

FURTHER READING

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (2016). *What If There Were No Significance Tests?* New York: Routledge Classic Editions. A great selection of readings on the pros and cons of significance testing together with suggestions for alternative approaches.

Henkel, R. E. (1976). *Tests of Significance.* London: Sage. While (very) old, this is a very good introduction to the principles of hypothesis-testing, highlighting key concepts and techniques; the issues relating to the interpretation of significance tests are particularly worth noting.

Kanji, G. K. (2006). *100 Statistical Tests*, 3rd edition. London: Sage. The title speaks for itself. This is a great reference book that, in some 200 pages, manages to cram in just about every test you are ever going to need (plus a few more!); excellent for choosing the correct type of test.

10
Simple things first: *One* variable, *one* sample

SINGLE-SAMPLE HYPOTHESES

In this chapter, we start applying the basic ideas of hypothesis-testing that you learned in Chapter 9 to specific problems (otherwise, you would bitterly complain that you get too much theory and too little practice!). As you may be a bit stressed from Chapters 8 and 9, we are going to take things easy here and leave the trickier stuff for later. Specifically, we are going to look at formulating and testing hypotheses relating to one variable at a time as applied to a single sample; such hypotheses focus upon the characteristics of a single population. After you have mastered these, we will show you how to make comparisons between different groups/measures (Chapter 11), how to investigate relationships between variables (Chapter 12), and how to perform such analyses by taking into consideration multiple variables at the same time (Chapter 13). We know this is riveting stuff, and the mere thought will excite you. Perhaps you want to do a bit of yoga to calm down at this stage.

So, what kinds of hypotheses can we develop for a single population? Well, quite a few, actually.

For starters, we can set up a hypothesis concerning the *distribution* of the variable; for example, we may hypothesize that our variable follows a normal distribution in the population and then use our sample data to test whether this is likely to be the case. Statistical tests aimed at determining the extent to which the distribution of a variable follows some pre-specified functional form in the population are known as **goodness-of-fit tests**; they compare the extent to which the observed (i.e., empirical) frequencies 'fit' the expected (i.e., theoretical) frequencies.

Another hypothesis we can set up relates to the *central tendency* of the variable; for example, we may hypothesize that the mean of our variable has a certain value in the population and then use our sample information to test whether this is likely to hold true. Statistical tests aimed at determining whether a population mean or median takes on a particular value are known as **tests for location** (by the way, realtors always say location is key: what counts is 'location, location, and location'); the example used throughout Chapter 9 to illustrate the nature of hypothesis-testing utilized such a test (namely the z-test for a population mean).

Third, we can set up hypotheses concerning the degree of *dispersion* in a variable; for example, we may hypothesize that the variance of our variable has a certain value in the pop-

ulation and then test for this using our sample data. Statistical tests aimed at testing whether a population variance is of a certain magnitude are known as **tests for variability**.

Fourth, with dichotomous variables (if you have forgotten what these look like, have a quick peek at Chapter 3), we can set up hypotheses relating to *proportions* (or percentages); for example, we may hypothesize that only π% of the population has a certain characteristic (while $100-\pi$ does not) and then test this hypothesis using our sample data. Tests of this type are commonly known as **tests for proportions**.

Lastly, it would only increase your hatred for statistical testing if we told you that there are also **tests for randomness**, which, among other things, can be used to check that a given sample is indeed a random one. So, we won't tell you about them! – Only kidding; we definitely will talk about this later in the chapter.

Figure 10.1 summarizes the above discussion by providing an exceptionally outstanding overview (yes, we are truly proud of this figure!) of the various population characteristics for which we can formulate hypotheses and associated techniques for testing them. You should use Figure 10.1 as a 'navigation chart' each time you want to test a hypothesis concerning a single population. Similar charts will be provided free of charge in Chapters 11 and 12 (and can also be delivered for a very modest price in mahogany frames with gold rims for the connoisseur reader).

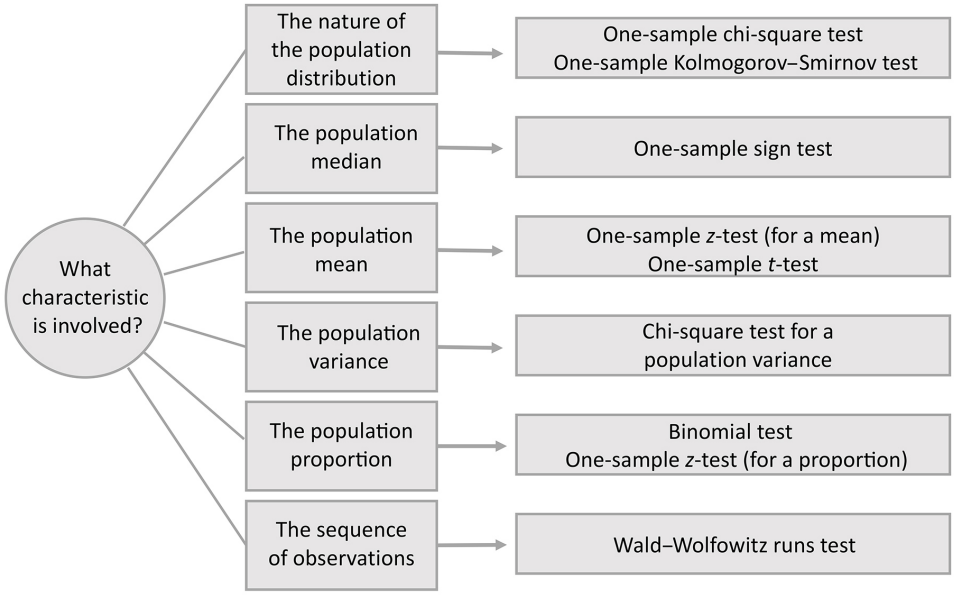


Figure 10.1 Statistical techniques for hypotheses involving population characteristics

ASSESSING FIT

At an intuitive level, the notion of 'fit' is perhaps grasped best by trying to imagine a situation where there is an 'ideal' scenario and an 'actual' turn of events. Say, for example, that you are

in your final year of studying Spanish medieval ecclesiastical history and are considering your lecture program. Now, you know that your total weekly lecture load is going to be 60 hours (only), but you do not yet know when these lectures will be scheduled. Let us assume that you would ideally like to spread your lecture program evenly through the week (i.e., attend a mere 12 hours of lectures, every day, Monday to Friday inclusive). When you actually get your timetable, you find out that you have 12 hours of lectures on Monday, 10 on Tuesday, 14 on Wednesday, 6 on Thursday, and 18 on Friday. How close is the actual timetable to your ideal one? In other words, what is the *fit* between what you had wished for and what you got? Would a timetable with 12 hours Monday to Wednesday, none on Thursday, and 24 hours on Friday (sorry, no sleep) represent a better or worse fit?

It is questions of this kind that we try to answer when we use goodness-of-fit tests. We will consider two of these tests here: the **one-sample chi-square test** and the **one-sample Kolmogorov–Smirnov (K–S) test** (rumors have it that Kolmogorov and Smirnov also invented a well-known vodka, but this is presumably fake news). The chi-square test is based on a consideration of *absolute* frequencies and, thus, it can be applied to any variable irrespective of the level of measurement (although it is typically used with categorical data). In contrast, the one-sample K–S test utilizes *cumulative* frequencies and, thus, implicitly assumes that the variable concerned is, at least, ordinal (see Chapter 6 if you cannot remember the distinction between absolute and cumulative frequencies).

The one-sample chi-square (χ^2) test

This is the test to go for whenever you want to compare a set of **observed frequencies** with a set of **theoretical frequencies**. By observed frequencies, we mean frequencies calculated from the empirical data (see Chapter 6), reflecting the *actual* distribution of the variable concerned in the sample (hence they are also known as 'actual frequencies'). For example, we could randomly select 100 students who bought this book, calculate how many of them are males and how many are females, and end up with observed frequencies of 35 males and 65 females. By theoretical frequencies, we mean frequencies generated on the basis of prior knowledge or theoretical considerations, reflecting our *expectations* regarding the distribution of the variable in the population. (Hence they are also known as 'expected frequencies'.) In the previous example, we would have no compelling reason to expect an uneven distribution between male and female buyers, so the theoretical frequencies would be 50/50. The question then arises whether the differences between observed and theoretical frequencies are significant.

The null hypothesis under the one-sample chi-square one-sample test is that no difference exists between observed and theoretical frequencies (or, amounting to the same thing, that the observed frequencies are equal to the theoretical frequencies). If the observed frequencies depart *significantly* from the theoretical frequencies (i.e., they cannot be reasonably explained by sampling fluctuations), we take this as evidence for the rejection of the theory/speculation that gave rise to the theoretical frequencies. If, on the other hand, we find that the differences between observed and expected frequencies are small and non-significant, then we have no grounds for rejecting the theory. Note that a 'reverse' hypothesis-testing logic characterizes all tests assessing fit: the interest is in *maintaining* (i.e., *not* rejecting) the null hypothesis; that is,

obtaining *non*-significant results rather than finding support for the alternative hypothesis via a significant finding (on this point, see the discussion on substantive significance in Chapter 9).

OK, we know that you are craving an example, so here is one. Say that you want to test TV viewers' preferences for late-night horror movies on the following four British TV channels: BBC1, BBC2, ITV, and Channel 4. Let us assume that you have no reason to believe that one channel would be preferred to another since all appear to provide an excellent selection of gory stuff. If this is indeed the case, you would expect equal preferences; that is, approximately the same number of people would prefer BBC1, BBC2, and so on; put in disgustingly complicated statistical jargon, your null hypothesis is that viewing preferences in the population follow a *uniform* distribution (see Chapter 7). You now collect some data on the viewing preferences of a random sample of 111 individuals and code them on the variable TVHORROR as follows: 1 = BBC1, 2 = BBC2, 3 = ITV, and 4 = Channel 4. Given that you hypothesize equal preferences, you would expect that individuals will be evenly distributed across the four categories, with approximately 28 individuals selecting each channel as being 'the best' (i.e., 111 individuals spread equally over four channels or 111/4 = 27.75). However, what you observe is that 11 respondents prefer BBC1, 64 BBC2, 16 ITV, and 20 Channel 4. To test whether these frequencies are significantly different from your expectations, you run a one-sample chi-square test using your trusted computer package; what you should get is something like Table 10.1 (this was produced by SPSS; other statistical programs produce similar output).

Table 10.1a An example of the one-sample chi-square test: TV horror

	Observed N	Expected N	Residual
1.00 [BBC1]	11	27.75	−16.75
2.00 [BBC2]	64	27.75	36.25
3.00 [ITV]	16	27.75	−11.75
4.00 [Channel4]	20	27.75	−7.75
Total	111		

Table 10.1b An example of the one-sample chi-square test: summary

Total N	111
Test Statistic	64.604[a]
Degree of Freedom	3
Asymptotic Sig. (2-sided test)	.000

a. There are 0 cells (0%) with expected values less than 5. The minimum expected value is 27.75.

According to the results, the hypothesis of equal viewing preferences is not supported, as the test statistic ($\chi^2 = 64.604$) is highly statistically significant ($p<0.001$). Many statistical programs, including SPSS, print only three places after the decimal point. That is why we cannot see *exactly* how (highly) significant this result is. It may be 0.0009 but it could also be 0.000000001. If you really need to find out – and usually this should not be necessary – you need to tease

more decimal places out of your program. In any event, going back to our example, the highly significant result is not surprising given the large discrepancies between the 'Observed' and 'Expected' values as indicated in the 'Residuals' column. In this context, more people preferred BBC2 than expected (positive residual) and fewer people than expected preferred BBC1, ITV, and Channel 4 (negative residuals); in absolute terms, the largest discrepancy is noted with BBC2 (36.25) and the smallest with Channel 4 (7.75). Overall, viewing preferences across TV channels are not equal.

In the above example, a uniform distribution of viewing preferences was assumed, specifying that expected frequencies across channels expected to be the same (each channel should be preferred by approximately 28 people or 25% of the whole sample). This is by no means a requirement of the chi-square test, and *any* theoretical (expected) frequencies can be specified and compared to observed (actual) frequencies. To illustrate this point, Table 10.2 shows the results of a one-sample chi-square test applied to the same data, but under the null hypothesis that 90% of individuals prefer BBC2 while the remaining 10% is equally distributed across the other channels. According to the results, this hypothesis must also be rejected as the theoretical frequencies do not 'fit' the observed data well. A summary of the steps involved in applying the one-sample chi-square test is generously provided in Table 10.3.

Table 10.2a Another example of the one-sample chi-square test: TV horror

	Observed N	Expected N	Residual
1.00 [BBC1]	11	7	4
2.00 [BBC2]	64	90	−26
3.00 [ITV]	16	7	9
4.00 [Channel4]	20	7	13
Total	111		

Table 10.2b Another example of the one-sample chi-square test: summary

Total N	111
Test Statistic	45.511[a]
Degree of Freedom	3
Asymptotic Sig.(2-sided test)	.000
a. There are 0 cells (0%) with expected values less than 5. The minimum expected value is 7.	

Table 10.3 Applying the one-sample chi-square test

1.	Specification of the categories of the variable of interest.
2.	Determination of the expected (theoretical) frequencies for each category (based on theory, prior research, wild guesses, etc.).
3.	Determination of the observed (actual) frequencies for each category.
4.	Comparison of observed versus expected frequencies and calculation of the chi-square statistic and its associated *p*-value.
5.	Examination of significance of the chi-square (χ^2) statistic based on the pre-specified αlevel and rejection (or non-rejection) of null hypothesis.

When the number of categories in your variable is greater than two (as in our example above), you should not use the chi-square test if (a) more than 20% of the expected frequencies are smaller than 5 or (b) any expected frequency is less than 1. (Don't ask us why; as with all tests, in order to get valid results, there are some silly conditions that have to be met even if this means that they make life complicated.) If you face any of the above problems (not an unusual situation, particularly with small samples and/or too many categories), then try to combine *adjacent* categories (assuming, of course, that such combinations are meaningful). This usually sorts things out. Something else to watch out for is that if you only have two categories (i.e., you are dealing with a dichotomous variable), then both expected frequencies must be 5 or larger; if they are not, combining categories is obviously not an option (as you would end up with a single category) and you have to use another test (such as the binomial test, discussed later in this chapter).

WARNING 10.1 Check that your expected frequencies meet the minimum levels required by the one-sample chi-square test before applying it to your data.

Before we part company with this illustrious test, you should note that the one-sample chi-square test is also often used to examine the hypothesis that a variable follows a normal distribution (or, more precisely, that the sample data have been drawn from a population that is normally distributed). While this is a frequent application, the K–S test of goodness of fit is more powerful than the chi-square test used for the same purpose. Bearing this in mind and given that we will be discussing the K–S test next, we see little benefit in demonstrating the use of chi-square for normality-testing purposes (check out the Further Reading section if you are truly desperate to find out).

The one-sample Kolmogorov–Smirnov (K–S) test

This tongue twister is *the* test to use in order to check that the distribution of a variable follows a particular form. By far its most common application is in testing whether observed values can reasonably be thought to have come from a normally distributed population (although it can be used to test for other distribution types, such as uniform or Poisson). In this context, the one-sample K–S test is an excellent check to apply *before* using statistical procedures that rest on the assumption of normality (see Chapters 5 and 9 and also the discussion on the one-sample *t*-test later in this chapter). By doing this, you will know whether the assumption of normality is tenable for the variable in question (thus pre-empting vicious criticism from statistical purists who desperately try to catch you out!).

HINT 10.1 Use the one-sample K–S test to check whether your variables are likely to be normally distributed in the population *before* applying statistical procedures that assume normality of distribution.

Note at this point that the assumption of normality permeates many statistical techniques and implies that the sampling distribution of the estimates (e.g., a mean) is normally distributed. With large enough samples, the assumption of normality can be considered tenable. This is due to the Central Limit Theorem (CLT) mentioned earlier in Chapter 8. According to the CLT, the sampling distribution of the estimates will converge toward a normal distribution as the sample size increases to infinity. This will hold true regardless of whether the source population is normal or skewed, so long as the sample size is sufficiently large (usually $n > 30$). We had to emphasize this because the K–S test will often indicate significant departures from normality (especially in large samples), and we do not want you to freak out for no reason!

The null hypothesis under the one-sample K–S test is that no difference exists between the observed distribution and a specified theoretical distribution (e.g., normal). The test computes the cumulative relative frequencies that would occur under the theoretical distribution and compares them with the observed cumulative relative frequencies. It then determines the point at which the two sets of frequencies diverge the most; that is, 'the maximum deviation' between the cumulative theoretical frequency distribution and the cumulative observed frequency distribution (see Chapter 6 for a memory refreshment on cumulative frequency distributions). If this maximum deviation is significant (i.e., it cannot be reasonably explained by sampling fluctuations), there is evidence that the observed distribution does not follow the specified form. If, on the other hand, the maximum deviation is not significant, there is no reason to believe that the observed distribution departs from the theoretical specification. As the one-sample K–S test is concerned with the degree of agreement between the distribution of a set of observed scores and some specified theoretical distribution, it is also a goodness-of-fit test (as is the one-sample chi-square test).

Let us continue using our example on TV-viewing habits to illustrate the application of the one-sample K–S test. Suppose that, in addition to TV channel-viewing preferences, you also obtained data from your random sample of 111 individuals on the number of horror films that they watched over the past year. Now you want to test whether this variable (let's call it FILMOHOR) is normally distributed in the population. You perform a one-sample K–S test and you get the results shown in Table 10.4 (any decent statistical package should be able to do a one-sample K–S test; if not, throw it away!).

The first thing to note from Table 10.4 is that, in this example, you are comparing the distribution of your variable against a normal distribution, which has the same mean (114.31) and the same standard deviation (83.94) as your sample. This is the **best-fitting normal distribution** for your data since no other hypothetical normal curve would provide a better fit (indeed, the best-fitting normal distribution for any observed frequency distribution is the one having the same mean and the same standard deviation as those computed from the observed data).

Table 10.4 An example of the one-sample Kolmogorov–Smirnov test

One-sample Kolmogorov–Smirnov test		
		FILMOHOR
N		111
Normal Parameters[a,b]	Mean	114.31
	Std. Deviation	83.94
Most Extreme Differences	Absolute	.131
	Positive	.131
	Negative	−.092
Kolmogorov–Smirnov Z		1.381
Asymp. Sig. (2-tailed)		.044
a. Test distribution is Normal		
b. User-Specified		

The second thing to note is that the test statistic (K–S $z = 1.381$) is significant ($p < 0.05$), indicating that the absolute 'Most Extreme Difference' (i.e., the maximum deviation) is too large to have come about if the distribution was normal. Thus, according to the results, the number of horror films watched over the past year does not follow a normal distribution.

Note that you do not *have* to test for normality of distribution by using the best-fitting normal distribution (as done above). You can specify the parameters of the theoretical distribution and then test your observed distribution against the theoretical distribution with the specified parameters distribution (notice the 'b' superscript in Table 10.4 indicating that the distribution parameters are user-specified). To appreciate this application, imagine that somebody else had undertaken a similar study in the past and concluded that the average number of horror films watched was 95 with a standard deviation of 66. You are dying to prove him wrong, so you apply the K–S one-sample test to *your* data specifying a normal distribution with *his* parameters; the results are shown in Table 10.5. Not surprisingly, the fit of this distribution is also poor and, as an inspection of the absolute 'Most Extreme Difference' against that in Table 10.4 shows, even worse than that obtained using the best-fitting normal distribution as the theoretical model.

Table 10.5 Another example of the one-sample Kolmogorov–Smirnov test

One-sample Kolmogorov–Smirnov test		
		FILMOHOR
N		111
Normal Parameters[a,b]	Mean	95
	Std. Deviation	66
Most Extreme Differences	Absolute	.157
	Positive	.084
	Negative	−.157

One-sample Kolmogorov–Smirnov test	
Kolmogorov–Smirnov Z	1.647
Asymp. Sig. (2-tailed)	.009
a. Test distribution is Normal	
b. User-Specified	

Finally, note that the one-sample K–S test is not only limited to testing whether a variable fits a normal distribution. Typically, statistical software allows you to also compare the observed distribution of a variable against a Poisson, uniform, or exponential distribution. If this is the case, the parameter that you need to specify for the Poisson and exponential distributions is the mean, while for the uniform distribution you need to give the minimum and maximum values. The interpretation of the results is similar to the examples above. The steps involved in applying the one-sample K–S test are summarized in Table 10.6.

Table 10.6 Applying the one-sample Kolmogorov–Smirnov test

1. Specification of the theoretical distribution of interest (uniform, normal, Poisson, or exponential).
2. Determination of the cumulative relative frequencies for the theoretical distribution.
3. Determination of the cumulative relative frequencies for the observed distribution.
4. Comparison of the two sets of frequencies and calculation of maximum absolute deviation.
5. Examination of significance of K–S *z*-statistic based on the pre-specified α level and rejection (or non-rejection) of the null hypothesis.

TESTING FOR LOCATION

You will recall from Chapter 7 that one aspect of summarizing data involves the calculation of 'average' or 'typical' values, such as means and medians, which describe the *central location* in a variable. In many situations, we are less interested in the average values we obtain from our sample data and more concerned with the corresponding population parameters. Thus, in the horror films example discussed earlier, what may be of particular importance to, say, a horror film producer is not the average number of films watched by your specific sample of 111 individuals, but the average number of films watched by the UK population as a whole. Now, as you know from Chapter 8, one thing you can do is *estimate* the latter from your sample data by setting up a confidence interval. But you can also *test* whether the population parameter (i.e., mean or median in this case) takes on a particular value; for example, you may want to test the hypothesis that, on average, UK residents watch 100 horror films a year. Or you may want to test the hypothesis that, on average, the number of horror films watched is greater (or fewer) than 300.

It is hypotheses like these (not necessarily constrained to horror films) that location tests are designed for. Table 10.7 provides a general overview of the various hypotheses that can be tested with location tests. We will consider two of the most useful tests here, the **one-sample**

sign test and the **one-sample *t*-test**; the former requires data that are at least ordinal in nature, whereas the latter assumes at least interval-level measurement.

Table 10.7 Hypotheses for one-sample location tests

Null hypothesis	Alternative hypothesis	Type of test
$M = M_0$	$M \neq M_0$	2-tailed
$M \leq M_0$	$M > M_0$	1-tailed
$M \geq M_0$	$M < M_0$	1-tailed

Note: M = population mean (μ) or median (m); M_0=hypothesized value.

The one-sample sign test

If you have a hypothesis about a *median*, this test should do the trick. Let us assume that you want to find out whether a median, m, takes a certain value in the population (i.e., your null hypothesis is that $m = m_0$). Remember that the median splits the distribution into two equal parts (see Chapter 7), so if the null hypothesis is correct, about one-half of the observations should be larger than the specified median value and about one-half should be smaller than the median. What the sign test does is (a) replace each observation less than m_0 with a minus (−) sign, (b) replace each observation greater than m_0 with a plus (+) sign, and (c) count the number of plus and minus signs (while ignoring zero differences). It then determines whether so many (or so few) plus signs could have come about given a median value of m_0. If yes, then we have no grounds for rejecting the null hypothesis (and we can go to the pub to celebrate); if not, then we have to conclude that the population median is not equal to the hypothesized value, m_0 (and we can go to the pub to drown our sorrows!).

For no other reason but to make life complicated, the test statistic of the one-sample sign test depends on the size of the sample (yes, we agree that whoever discovered this should be locked away, but that's the way it is). With small samples ($n < 30$), the test uses the binomial distribution and provides an exact probability value (see also Chapter 8 and the section on testing proportions later in this chapter). With larger samples ($n > 30$), the normal approximation to the binomial is used, which gives a z-value and an associated probability. As we are perfectionists of the highest possible standards, we shall illustrate both versions of the test using our celebrated example of horror films (to hone your skills further, you can later repeat the tests with Bollywood movies).

Let us assume that you want to test the hypothesis that the median number of horror films watched over the past year is 29; you picked this number while reading the *Transylvanian Times* arts page, which reported on a study of horror film viewership in Romania (and, having nothing better to do, you want to see whether UK residents have similar habits). Let us further assume that you have randomly split your total sample of 111 individuals into one small sub-sample of, say, 20 respondents and a larger sub-sample containing the rest. Tables 10.8 and 10.9 show the output obtained from testing the hypothesis $m = 29$ on the small and large sub-samples, respectively. Surprise, surprise, we are again using SPSS, which, however, does not include a ready-made procedure and requires a little tweak. The best way to do it is

to create an additional pseudo-variable in your data set, which will act as a reference point. The value of this new variable should be constant and equal to the hypothesized or test value, and it should be used in a pair with the variable of interest in a two-related-samples sign test (namely the **Wilcoxon signed-rank test**; see Chapter 11). Thus, in our example, we created a new variable with 111 cases, all of which correspond to the value of 29 (i.e., our hypothesized median value).

Table 10.8a An example of the one-sample sign test: small sample, frequencies

		N
MEDVAL - FILMOHOR	Negative Differences[a]	12
	Positive Differences[b]	7
	Ties[c]	1
	Total	20
a. MEDVAL < FILMOHOR		
b. MEDVAL > FILMOHOR		
c. MEDVAL = FILMOHOR		

Table 10.8b An example of the one-sample sign test: small sample, test statistics[a]

	MEDVAL - FILMOHOR
Exact Sig. (2-tailed)	.359[b]
a. Sign Test	
b. Binomial distribution used	

Other than the differences in the test statistics field, the output of the test is identical for the small and the large sub-samples, indicating the number of positive and negative differences (i.e., plus and minus signs) and the number of ties. MEDVAL is a constant, which has been set to 29, consistent with the hypothesis that the median of FILMOHOR in the population takes this value. Note that, as far as the small sub-sample is concerned (see Table 10.8), we cannot reject the null hypothesis that $m = 29$ since the test is non-significant ($p>0.10$). In contrast, for the larger sub-sample (see Table 10.9), the test turns out to be highly significant ($p<0.01$), implying that the median number of films watched by UK individuals is not 29; in fact, as there are a lot more minus signs, it seems that the population median is less than 29. Note that if we were to test this directional hypothesis (i.e., that $m < 29$) on the smaller sub-sample, we would again obtain a non-significant result (even at the 10% significance level) since the corresponding one-tailed probability would come to 0.359/2 = 0.180. In contrast, a directional hypothesis on the bigger sub-sample would correspond to an even even stronger result (significant at the 0.1% significance level). (See Chapter 9 for the calculation of one-tailed versus two-tailed significance levels.) The various steps involved in applying the one-sample sign test are summarized in Table 10.10.

Table 10.9a An example of the one-sample sign test: large sample, frequencies

		N
MEDVAL - FILMOHOR	Negative Differences[a]	59
	Positive Differences[b]	28
	Ties[c]	4
	Total	91
a. MEDVAL < FILMOHOR		
b. MEDVAL > FILMOHOR		
c. MEDVAL = FILMOHOR		

Table 10.9b An example of the one-sample sign test: large sample, test statistics[a]

	MEDVAL - FILMOHOR
Z	−3.216
Asymp. Sig. (2-tailed)	.001
a. Sign Test	

Table 10.10 Applying the one-sample sign test

1.	Specification of the hypothetical median value and formulation of the desired null hypothesis (see Table 10.7).
2.	Comparison of the hypothetical median value with the individual observations.
3.	Replacement of the differences with plus and minus signs and count of the number of positive and negative signs.
4.	Examination of binomial exact probability (small samples) or significance of *z*-statistic (large samples) based on the pre-specified α level and rejection (or non-rejection) of the null hypothesis.

As a final point, it is worth mentioning that the one-sample sign test can also be applied to test hypotheses concerning the population mean (assuming that the distribution of the latter is roughly symmetrical). The procedure is identical to the one discussed above, the only difference being that the plus and minus signs reflect departures from the hypothesized mean rather than the median. Although there are other more powerful tests for testing hypotheses involving means (such as the one-sample *t*-test discussed next), the one-sample sign test is worth a try when the sample is small *and* the distribution in the population is non-normal (i.e., when you are truly desperate!).

THE ONE-SAMPLE *T*-TEST

If you want to test a hypothesis concerning a mean, you cannot go wrong if you use the one-sample *t*-test. Say that you want to test whether the population mean exceeds a certain value (i.e., that $\mu > \mu_0$ which, in null hypothesis terms, implies that H_0: $\mu \leq \mu_0$ – see Table 10.7).

The one-sample *t*-test (a) computes the difference between the sample mean and the hypothesized value (i.e., $\bar{x}-\mu_0$), (b) takes into account the sample size, *n*, as well as the likely variability in the population (using the standard deviation of the sample, *s*, as a proxy of the population standard deviation, σ), and (c) determines whether the sample is likely to have come from a population whose mean value exceeds μ_0. If yes, this is reflected in a significant one-tailed test statistic (*t*-value), and we can reject the null hypothesis. If, on the other hand, the *t*-value is non-significant, we have no grounds to conclude that the population mean is indeed greater than the pre-specified value, μ_0.

Let us stick with our tried and tested example of horror films and use the one-sample *t*-test to test the hypothesis that, on average, UK consumers watch more than 30 horror films annually. Our variable is again FILMOHOR, and the output obtained is gracefully displayed in Table 10.11.

Table 10.11a An example of the one-sample *t*-test: one-sample statistics

	N	Mean	Std. Deviation	Std. Error Mean
FILMOHOR	111	32.02	11.906	1.130

Table 10.11b An example of the one-sample *t*-test: one-sample test

	Test Value = 30					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
FILMOHOR	1.786	110	.077	2.018	-.22	4.26

You get quite a bit of information from a typical *t*-test output. This includes the sample size, the sample mean, the sample standard deviation, and, as a bonus, the standard error of the mean (go back to Chapter 8 if you are not entirely sure what the standard error is). As you probably noticed, the 'test value' is a constant set to the hypothesized mean population value (here 30). Next, the value of the *t*-statistic is shown together with the degrees of freedom (df = $n-1$) and a two-tailed level of significance (which is the default in the SPSS output). Finally, the 'mean difference' is displayed (which is the difference between the sample mean and the pre-specified value, i.e., 32.02–30), along with its associated 95% confidence interval. The *t*-value is simply computed by dividing the 'Mean Difference' by the standard error (i.e., $t = 2.018/1.130$).

What can we make out of all this? First, as we are testing a directional hypothesis, we need to halve the *p*-value associated with our *t*-statistic (see Chapter 9). This comes to 0.077/2 = 0.0385, which indicates a (one-tailed) significant result (at $p<0.05$). However, before we can proudly announce that the mean number of horror films watched in the population exceeds 30, we must look at the *direction* of the mean difference (to ensure that we are at the 'correct' tail of the *t*-distribution; in this case, to the right of the hypothesized mean). Here the mean difference is indeed in the hypothesized direction, and, therefore, we find support for the hypothesis that the mean number of horror films watched annually exceeds 30. Note that although halving the *p*-value to adjust for one-tailed hypothesis works just fine, the confidence intervals

of the mean difference obtained become uninterpretable. Indeed, the one-tailed *p*-value is significant at a 5% level, but the 95% confidence intervals around the mean difference include zero. However, do not even *think* of halving the lower and upper limits of the confidence intervals! It does *not* work (trust us, we know). What you should do instead is to re-run the analysis specifying a 90% confidence level and simply pick up the values of the confidence interval produced. This should do the trick as the two-tailed critical value for the 90% confidence level equals the one-tailed critical value for the 95% confidence level (if you don't believe us, go and look at the *t*-distribution tables yourself, as if you have nothing better to do). A summary of the procedure associated with applying the one-sample *t*-test is given in Table 10.12.

Table 10.12 Applying the one-sample t-test

1.	Specification of the population mean value and formulation of the desired null hypothesis (see Table 10.7).
2.	Calculation of the sample mean value and associated standard error.
3.	Calculation of the difference between the sample mean value and the hypothesized population value and computation of *t*-statistic.
4.	Examination of significance of *t*-statistic based on the pre-specified α level and rejection (or non-rejection) of the null hypothesis.

At this stage, you are no doubt wondering why we have been using the one-sample *t*-test to test a hypothesis concerning a mean, while in Chapter 9 we used the one-sample *z*-test for exactly the same purpose. The reason for this is not a deliberate attempt to confuse you, but the fact that statisticians, in their infinite wisdom, recommend that different tests are appropriate, depending upon (a) whether the underlying population is normally distributed or not, (b) whether the sample size is large or small, and (c) whether the variance of the population is known or unknown. Thus, if we wanted to be 'statistically correct', we should consider no fewer than eight possibilities simply in order to decide how to test a hypothesis concerning a single mean.

WARNING 10.2 When testing directional hypotheses and using one-tailed significance tests, make sure that any significant result is in the correct direction. Otherwise, you may support exactly the opposite of what you hypothesized!

In the interests of your sanity (and ours), we decided to take a more pragmatic approach here. Observe the following.

First, it is a *very* rare occasion when the variance in the population is known; although in some exceptional circumstances one may have access to the population variance (e.g., from a previous study), in most instances we have to rely on the sample variance. Indeed, for all it is worth, the authors have never come across a situation where the population variance was known but the population mean was not. As the one-sample *z*-test is generally used when we know the population variance, its range of application is limited, particularly with small samples (where the assumption of a normal population must also hold).

With large samples, we can go ahead and use the *z*-test as an approximation, irrespective of the nature of the population distribution (we unashamedly did so in Chapter 9). Recall also

from Chapter 8 that, as the sample size increases, the *t*-distribution (on which the *t*-test is based) follows more and more closely the normal distribution (on which the *z*-test is based). Assuming that you have a reasonably sized sample ($n > 30$), the one-sample *t*-test and the one-sample *z*-test will give very similar results from a practical point of view.

Finally, in cases where the sample size is small *and* the population is decidedly non-normal, it is safest to go for a non-parametric test (such as the one-sample sign test discussed earlier) rather than risk being lynched by statistical hard-liners for using either the *t*-test or the *z*-test.

Bearing the above observations in mind, you should (a) use a one-sample *t*-test whenever you have a large sample, irrespective of the form of the distribution, (b) still use a *t*-test if you have a small sample but a normally distributed population, and (c) opt for a non-parametric test in all other instances. Phew!

A 'fringe benefit' of the *t*-test output (see Table 10.11) is that it contains all the information needed (i.e., sample mean, standard error, etc.) to calculate a confidence interval for the population mean (see Chapter 8). This is yet another opportunity to impress your client, boss, teacher, neighbor, spiritual advisor, and so on by casually demonstrating how to kill two birds with one stone (i.e., use the *t*-test output to do *both* hypothesis-testing *and* interval estimation).

TESTING FOR VARIABILITY

Complementary to location tests are tests that allow us to examine hypotheses relating to the degree of *variability* in the population. Here the emphasis is on the extent of fluctuation or dispersion in the values of the variable of interest rather than on 'typical' or 'average' values. In this context, you should recall from Chapter 7 that the notion of variability is more meaningful when metric (i.e., interval or ratio) data are involved and that the most useful measure of variability is the standard deviation. Accordingly, we shall limit ourselves here to an illustration of a hypothesis test concerning a population standard deviation.

Let us assume that, unless you test the hypothesis that 'the standard deviation of the number of horror films watched in the UK annually does not exceed 90', you will be unable to sleep at night. Thus, your null hypothesis is that $\sigma \leq 90$, and you want to use your survey data to test it against the alternative hypothesis that $\sigma > 90$. Now, here's the bad news: SPSS won't do it for you (although you might be able to pull it off through the syntax or with a more versatile statistical package like *R*). 'Why not?', you may ask with righteous indignation.

There are two points worth making for this most unsavory situation. The first is that one cannot test a standard deviation directly; one must test the variance instead (don't ask us why, it's the statisticians who are to blame – again!). The second point is that doing a **variance test** is dead simple to do yourself. Here's how.

First, translate your hypothesis from standard deviation terms into variance terms. This involves simply squaring your hypothesized value, as the standard deviation is defined as the square root of the variance (see Chapter 7). Thus, in our illustrious example, the null hypothesis becomes $\sigma^2 \leq 8{,}100$ and the alternative hypothesis $\sigma^2 > 8{,}100$. Next, calculate the sample variance s^2 (yes, you can get this from SPSS). Finally, divide the sample variance, s^2, by the

hypothesized population variance, σ^2, and multiply the ratio by the sample size n minus 1; in other words, calculate the following quantity:

$$(n-1)\frac{s^2}{\sigma^2}$$

The beauty of the above is that it follows a chi-square distribution with $(n-1)$ degrees of freedom and can thus be used as a test statistic; once you have calculated it, you can refer to tables of the chi-square distribution (you can find these online or in any statistics text, such as those suggested in the Further Reading section) and check out its significance. In our example, if we go back to Table 10.11, we see that $n = 110$ and $s^2 = (83.943)^2 = 7{,}046.427$. Plugging these values into the above formula gives us

$$(110-1)\times(7{,}046.427/8{,}100)=94.82$$

A chi-square value of 94.82 with 109 degrees of freedom is not significant (if you doubt this, check your statistical tables), which means that we cannot reject the hypothesis that the variance of horror films watched annually is no more than 8,100. Or, translating back to our original hypothesis, we can conclude that the standard deviation of horror films watched does not exceed 90.

Two additional points need to be mentioned in connection with the above procedure. First, the test makes the assumption that the underlying population is normal; violation of this assumption can lead to misleading results, as the test statistic will not be distributed as a chi-square (unfortunately, this test is not particularly robust). Second, when a two-sided hypothesis is involved, a complication arises in the calculation of the *p*-value because, unlike the normal and *t*-distributions, the chi-square distribution is not symmetrical (see also Chapter 7 on the notion of symmetry in distributions). This means that we cannot simply double the *p*-value if we want a two-tailed test. Instead, we have to follow the advice of statistical gurus who suggest that for two-sided tests based on asymmetrical distributions it is advisable to report the one-tailed *p*-value together with a statement about the direction of the observed departure from the null hypothesis.

TESTING FOR PROPORTIONS

Sometimes we want to test hypotheses concerning a variable that only takes two values (i.e., is dichotomous). An obvious example of such a variable is gender, with male and female being the two categories (although, as noted in Chapter 3, much more detailed gender classifications are also possible). Variables of this nature are naturally discussed in terms of *proportions* (or percentages), and if we know that the proportion of cases that fall into one category is *P*, then we also know that the proportion in the other category is $1-P$. For example, if the proportion of male students at the Frankenstein Institute of Technology in Düsseldorf is 0.78, then obviously the proportion of female students *must* be 0.22.

WARNING 10.3 Make sure that the normality assumption is justified before you apply variance tests (see also Hint 10.1). Also, remember that with asymmetrical distributions (such as chi-square), doubling of one-tailed *p*-values is not legitimate.

The kinds of hypotheses that can be tested when proportions are involved are directly analogous to those associated with location tests (see Table 10.7 earlier) and are summarized in Table 10.13. Probably the most widely tested hypothesis is when the population proportion, π, is set to 0.5, which is the same as saying that the proportions in both categories of the variable are equal in the population.

Table 10.13 Hypotheses for one-sample proportion tests

Null hypothesis	Alternative hypothesis	Type of test
$\pi = \pi_0$	$\pi = \pi_0$	2-tailed
$\pi \leq \pi_0$	$\pi > \pi_0$	1-tailed
$\pi \geq \pi_0$	$\pi < \pi_0$	1-tailed

Note: π = population proportion, π_0 = hypothesized value.

The statistical procedure you should use to test a hypothesis concerning a population proportion is known as the **binomial** test. The test has two versions, depending on the size of the sample, n, and the (hypothesized) population proportions μ and $1-\pi$. When $n\pi$ or $n(1-\pi)$ is smaller than 5, the test uses the binomial distribution and provides an exact probability value ('small-sample' case). When $n\mu$ and $n(1-\mu)$ are both greater than 5, the test uses the normal approximation to the binomial (see Chapter 8) to calculate a probability value ('large-sample' case). Incidentally, the latter version of the binomial test is often referred to in standard statistics texts as the ***z*-test for a proportion** (see Figure 10.1 and Further Reading section). Moreover, if only to confuse matters even more, some texts discuss the 'large-sample' case under parametric tests and the 'small-sample' case under non-parametric tests. While we could all do without such complications, thankfully most computer packages automatically determine from the input specifications which version of the test is appropriate.

To illustrate the application of the binomial test, let us test the hypothesis that the proportion of men who have a life subscription to *Horror Movie Weekly* is 0.70. Thus, in null hypothesis terms, we set $\pi = 0.70$, the alternative hypothesis being that $\pi \neq 0.70$. Moreover, let us assume that we have just conducted two separate surveys using random samples of subscribers to this quality magazine; the first sample consists of 24 subscribers based in Fries, Virginia, and the second of 86 subscribers in Whynot, North Carolina. Tables 10.14 and 10.15 show the results of testing this hypothesis.

Table 10.14 An example of the binomial test: small sample

Binomial test						
		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
GENDER	Male	1	20	.83	.70	.222
	Female	0	4	.17		
	Total		24	1.00		

Table 10.15 An example of the binomial test: large sample

Binomial test							
		Category	N	Observed Prop.	Test Prop.	Z Approximation	Exact Sig. (2-tailed)
GENDER	Male	1	71	.83	.70	2.424	.015
	Female	0	15	.17			
	Total		86	1.00			

The output of the test is similar for the small and the large samples, the only difference being an extra column and the distribution used to calculate the *p*-values. Gender is coded as a dichotomous variable, with 1 = male and 0 = female (of course, we could have coded this variable using *any* two different numbers; if you are not sure why, go back to Chapter 3). The output shows both the *hypothesized* proportion in the population (referred to as 'Test Prop.') and the proportion *observed* in the sample (referred to as 'Observed Prop.'). As yet another caveat of SPSS, there is no direct procedure to perform a *z*-test for a single proportion. Users first need to install an extension (i.e., confidence interval proportion) and then run the test through the 'Utilities' function.

According to the findings for the Fries sample (Table 10.14), we cannot reject the null hypothesis that $\pi = 0.70$ as the test is non-significant ($p>0.10$). However, the test returns a significant result (at $p<0.05$) when applied to the Whynot sample (Table 10.15), indicating that the proportion of male subscribers to *Horror Movie Weekly* is not 0.70. In fact, judging from the observed (i.e., sample) proportion, the relevant population proportion is likely to be greater than 0.70 (whether the conflicting results for the Fries and Whynot samples reflect differences in location-specific factors, differences in statistical power as a result of grossly unequal sample sizes, or both is something that need not concern us here). Table 10.16 summarizes the steps involved in using the binomial test.

Table 10.16 Applying the binomial test

1.	Specification of the proportion value and formulation of the desired null hypothesis (see Table 10.13).
2.	Comparison of the hypothesized proportion with the observed proportion in the sample.
3.	Examination of binomial exact probability (small samples) or normal approximation (large samples) based on the pre-specified α level and rejection (or non-rejection) of the null hypothesis.

At this stage, you may be getting the strange feeling that the binomial test is rather similar to the one-sample chi-square test for testing goodness of fit (albeit for dichotomous variables). This is very much so and, provided that the assumptions underlying the one-sample chi-square test are met (see earlier discussion), one could translate a hypothesis regarding proportions into expected frequencies and perform a one-sample chi-square test. However, given the somewhat 'special' nature of dichotomous variables in terms of the level of measurement (see Chapter 3), as well as their natural interpretation as proportions, the binomial test is more appealing. Its appeal is further enhanced by noting that both one- and two-tailed *p*-values can readily be obtained with the binomial test, whereas the asymmetric nature of the chi-square distribution makes matters a bit more complicated in this respect (see also Warning 10.3 above).

TESTING FOR RANDOMNESS

If you have made it this far, you will be pleased to know that there is only one more test to consider before you can put this terrifying experience of single-variable, single-sample hypothesis tests behind you (of course, this is nothing compared to the horrors awaiting you in Chapters 11 and 12!). The test in question is the **Wald–Wolfowitz runs test** for randomness (and we treat with contempt any suggestion that its name is a joke or indicative of the eating and sanitary habits of its developers).

HINT 10.2	For testing hypotheses on dichotomous variables, the binomial test is preferable to the one-sample chi-square test.

Now, why would anybody in their right mind be interested in formulating and testing hypotheses concerning randomness? Goodness of fit, OK. Location and variability, sure, it makes sense. Proportions, well, alright. But *randomness*?

Recall from Chapters 2 and 6 that an important requirement in statistical inference (whether in the form of estimation or hypothesis-testing) is *random* sampling. Thus, it is quite handy to have a method for actually testing the hypothesis that our sample is indeed a random sample; this is precisely what the runs test allows us to do. Given a set of observations, the runs test examines whether the value of each observation influences the values taken by later observations; if the observations are independent, the sequence is considered random. More specifically, the test looks at the number of **runs** present in the sample, a run being defined as a sequence of like observations; too few or too many runs suggest dependence between observations (i.e., lack of randomness).

Consider the following three sequences reflecting the gender of 10 respondents queuing at a cinema ticket counter (for the premiere of *Nightmare on Downing Street III*):

MMMMM FFFFF
M F M F M F M F M F
FFF MM F MM F M

In the first two sequences (containing 2 and 10 runs, respectively), we suspect a lack of randomness, as there appears to be a pattern running through the observations (i.e., 'bunching' and 'alternating' of the genders, respectively). In contrast, the third sequence (containing six runs) appears to be well mixed and gives little cause for suspecting a lack of randomness.

Now that you have the general idea, let us apply this test to check out whether your sample of 24 Fries subscribers to *Horror Movie Weekly* (see Table 10.14 earlier) is indeed a random one with regard to respondent gender. Table 10.17 shows the results. (Yawn, yawn, it's SPSS output again.)

Table 10.17 An example of the runs test

Runs test	
	GENDER
Test Value[a]	1
Cases < Test Value	4
Cases ≥ Test Value	20
Total Cases	24
Number of Runs	9
Z	.650
Asymp. Sig. (2-tailed)	.515
a. Median	

The 'Test Value' in the output indicates the way in which the runs are defined; in this case, this has been set equal to 1, as we have coded GENDER as 1 = male, 0 = female (and we are interested in seeing whether the sequence of male and female respondents in our sample of 24 individuals is indeed random). On this basis, there was a total of nine runs and, according to the test statistic, this number was neither too low nor too high to suggest non-randomness (the *z*-value is non-significant). Thus, we cannot reject the hypothesis that, as far as respondent gender is concerned, the sample is random (which is just as well, otherwise we would have trouble justifying the use of this sample in our previous analyses).

When the variable involved is dichotomous (as in the above example), the runs are straightforward to define. What happens if the variable is continuous (or discrete but with many values)? No problem – the runs test can cope with such data just as well. All one has to do is specify a 'cutting point' for dichotomizing the variable being tested (usually the mean, median, or mode). This cutting point then becomes the 'Test Value'; that is, values less than the cutting point form one category and values greater than or equal to the cutting point form the other category. Then the runs are calculated exactly as before. Table 10.18 shows an application of the runs test to a continuous variable with the mean being used as the 'Test Value'; the variable concerned is the age of the 24 Fries subscribers to *Horror Movie Weekly*.

Table 10.18 Another example of the runs test

Runs test	
	AGE
Test Value[a]	41.05
Cases < Test Value	13
Cases ≥ Test Value	11
Total Cases	24
Number of Runs	14
Z	.520
Asymp. Sig. (2-tailed)	.603
a. Median	

You should note with great delight that the usefulness of the runs test is not restricted to testing the null hypothesis that a given sample is random (although this is its most common application). The test can be applied to any sequence, no matter how the sequence was generated. For example, you could apply it to test for randomness of the presence or absence of snow over a given period, the wins and losses of the local rugby team, the correct and false answers to a multiple-choice test, and so on. In fact, whenever you see a sequence, look at it as a not-to-be-missed opportunity for applying the runs test.

SUMMARY

In this chapter, we took the first concrete steps into hypothesis-testing by examining various hypotheses relating to a single variable and a single sample. We started by looking at hypotheses concerned with goodness of fit and had a jolly good time fooling around with the one-sample chi-square test and the Kolmogorov–Smirnov test. We then moved on to testing hypotheses regarding location and had a ball with the one-sample sign test and the one-sample *t*-test. Hypotheses concerning variability were next in line, and these were swiftly dealt with (in old-school fashion, may we add) via the variance test. Dichotomous variables were tackled subsequently, and the binomial test was used to test hypotheses concerning proportions. The icing on the cake was none other than the (somewhat unfortunately named) runs test, which is *the* test to use where randomness is the issue. We are now ready to enter more complex territory and have some fun making comparisons between groups and between measurements.

QUESTIONS AND PROBLEMS

1. Explain in non-mathematical terms the underlying hypotheses for the following tests:
   a. goodness-of-fit tests
   b. tests for location
   c. tests for variability
   d. tests for proportions
   e. tests for randomness.
2. What type of measurement does a variable require to apply the one-sample chi-square test?
3. What is the null hypothesis under the one-sample chi-square test?
4. What is meant by the terms 'theoretical frequencies' and 'expected frequencies'?
5. Does the one-sample chi-square test require that all the expected frequencies are identical?
6. Which test is commonly applied to find out whether observed values have come from a normally distributed population?
7. Explain when you would use (a) a one-sample sign test and (b) a one-sample *t*-test.
8. You have taken a random sample of 70 Buenos Aires taxi drivers and found that 90% admit to exceeding the speed limit by 60 miles per hour on a regular basis. Which test would you use to check whether this result is likely to apply to all taxi drivers in Buenos Aires?
9. In which situations (other than to impress friends at cocktail parties) would you use the runs test?
10. Have you watched any good horror movies lately?

FURTHER READING

Field, A. (2017). *Discovering Statistics Using SPSS IBM Statistics*. London: Sage. If you are going to use SPSS for your analysis, this is the text to get. It covers most statistical techniques discussed in this book in an SPSS environment.

Ho, R. (2014). *Handbook of Univariate and Multivariate Data Analysis with IBM SPSS*. Boca Raton, FL: CRC Press. A bit 'drier' than Field (2017) but very comprehensive.

Siegel, S. & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition. New York: McGraw-Hill. This is the 'bible' on non-parametric statistics, first published in 1956 (ah, those were the days!). Essential reading.

11
Getting experienced: Making comparisons

THE PLEASURE OF COMPARING

Now that you have mastered the fundamental ideas about data analysis and are able to squeeze various statistical tests out of your computer, it is time to apply your skills to something more pleasurable, namely comparisons. In fact, it is often only comparisons that provide meaning. Concepts like small and large, young and old, or ugly and beautiful only make sense if used in a comparative sense. (For example, a pot-bellied pig is uglier than a rhinoceros – we purposely use animal examples so as not to be accused of body shaming.) Similarly, figures like the sales revenue of a waterbed manufacturer in Tombstone, Arizona, the proportion of Italian men using wallpaper strippers, or the number of US dog owners who buy them special Christmas dinners become more illuminating if compared to, for example, waterbed manufacturers in Fishguard, Wales, men using wallpaper strippers in Angola, and the number of Canadian dog owners who buy them special Christmas dinners. In all these examples, *different groups* are compared on a particular characteristic.

Comparisons can also involve contrasting *multiple measures*; for example, comparing the average importance given to different criteria when buying a car on a five-point rating scale (e.g., price, fuel consumption, number of windscreen wiper speed options, and so on). Here, a single group (i.e., sample of respondents) provides two or more measurements, which are then contrasted. Often, such measurements are obtained at different points in time and are aimed at detecting change. A typical example of such a comparison involves data of the 'before and after' variety; for example, comparing the percentage of Scotsmen suffering from diarrhea before and after eating 11 portions of haggis, or attitudes toward smoking before and after being exposed to a 'shock' television advert (depicting an elderly smoker spontaneously combusting as a result of lighting up!).

From the above, it should be clear that comparisons can take two basic forms: either two (or more) different *groups* are compared on a given characteristic, or two (or more) *measures* of the same group are compared to one another. This distinction is important because it has direct implications for the choice of statistical technique. As extensively discussed in Chapter 5, with **independent measures** (or 'independent samples', as they are confusingly referred to in the statistical literature), different units of analysis are compared on the same variable. In contrast, with **related measures** (or 'related samples'), the same units are compared on differ-

ent variables. As Figure 11.1 shows, different techniques are appropriate for making comparisons between independent measures as opposed to related measures; note that we have only listed the most important and widely used techniques, and we refer you to the Further Reading section if what's on offer in Figure 11.1 is not enough for you.

In what follows, we shall first explain the delights of comparing independent measures (in our experience, the most common situation) and then, more briefly, introduce you to techniques for undertaking comparisons between related measures. Since you are now 'old hands' at hypothesis-testing and given the number of techniques that we must try to cover (see Figure 11.1), we shall go through the various tests in a bit less detail than we did in Chapter 10; this should make this chapter more palatable and also avoid unnecessary repetition of points already covered.

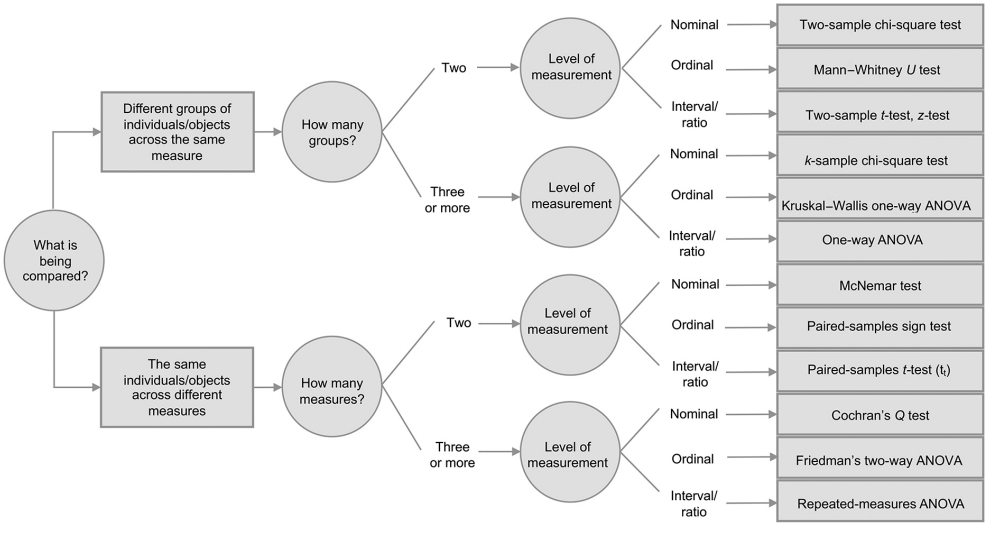


Figure 11.1 Statistical techniques for making comparisons

INDEPENDENT MEASURES: COMPARING GROUPS

This is a very typical situation in data analysis, which occurs whenever (a) one has data from more than one sample (e.g., from two separate surveys of Internet surfing habits of Algerian and Czech dentists) or (b) one decides to split a single sample into two (or more) sub-samples on the basis of some characteristic (e.g., creating sub-samples of male and female Algerian dentists and subsequently comparing their Internet surfing habits). In both cases, there are two or more *groups* involved, and the interest lies in identifying similarities and differences between them. Clearly, any observed differences are only of importance if they are likely to apply to the respective *population*. If it is only sampling error that is responsible for the patterns observed, then any conclusion that the groups compared differ with respect to the characteristic of interest is unwarranted (see Chapter 9). Thus, in comparing groups, it is nec-

essary not only to identify differences based on the sample data but also to test the statistical significance of such differences.

WARNING 11.1 Before making comparisons, you should establish whether you are dealing with independent or related measures. Failure to do this is likely to result in an incorrect choice of statistical technique.

In hypothesis-testing terms, comparisons between groups can be conveniently expressed as follows:

- There is no difference between the two (or more) groups in terms of the characteristic of interest (null hypothesis).
- The groups differ with respect to the characteristic of interest (alternative hypothesis – exploratory).
- One group has more/less of the characteristic of interest than the other group(s) (alternative hypothesis – directional).

Note that you should always compare the *same* variable (i.e., characteristic) between the groups; that is, do not attempt to compare the average age of Austrians taking sleeping pills before going to bed with the average length of marriage among Germans who take sleeping pills. Instead, conduct two analyses; one comparing the average age and one comparing the average length of marriage between the groups.

Depending upon the nature of the characteristic, which is the focus of comparison, hypotheses can be formulated relating to frequencies (nominal data), rank orders (ordinal data), and mean levels (interval/ratio data) across the groups. Unfortunately, as is clear from Figure 11.1, the level of measurement of the characteristic of interest is not the only determinant of the appropriate statistical technique for carrying out the comparison. The *number* of groups to be compared also needs to be taken into consideration, since different techniques are used for two-group comparisons (e.g., comparing a sample of Indian households owning electric curling tongs with an equivalent sample from Cameroon) from those for k-group comparisons (e.g., comparing Swedish, Norwegian, and Finnish owners of pet tarantulas). While you may be inclined to have a nervous breakdown at this stage as even more tests have to be learned, there is a bright side – really! All the tests for making k-sample comparisons (where $k > 2$) are straightforward extensions of the equivalent tests for making two-sample comparisons. Even better, the two-sample tests are themselves often based on the single-sample tests that you so enjoyed learning about in Chapter 10. So, please do not give up – make yourself a herbal tea or go get a massage and then read on.

Now that we have lulled you into a false sense of security, we can start looking into different statistical tests for making comparisons. In order to do this as painlessly as possible, we will first discuss the two-group case for each level of measurement and immediately follow it with the k-group case. To ease your suffering even further, we will set $k = 3$ for our examples; that is, deal with three groups only (this results in no loss of generality in that you can use exactly the same techniques to compare any number of groups you like – assuming, of course, that the

relevant test assumptions are satisfied). If during the following discussion you feel that you are getting lost or slowly drifting to sleep, a quick glance at our fantastic Figure 11.1 will remind you where you are. Once again, you may want to buy the figure for a very modest price in a mahogany frame with gold rims in order to display it for posterity in your office.

THE TWO-SAMPLE CHI-SQUARE (χ^2) TEST

If you want to compare two groups on a variable, which is measured on a nominal scale, this is the test you need. As its name implies, the **two-sample chi-square test** is the big brother (or big sister, if you prefer) of the one-sample chi-square test (see Chapter 10) and, like its baby sibling, it is based on a comparison of observed versus expected frequencies.

The null hypothesis tested by the two-sample chi-square test is that no difference exists between the two groups with respect to the *relative* frequency with which group members fall into the various categories of the variable of interest. The reason for focusing on relative rather than absolute frequencies is that the two groups may have unequal sample sizes and, therefore, the calculation of expected frequencies needs to take this into account. If the observed frequencies depart significantly from the expected frequencies (i.e., cannot be dismissed because of sampling error), we conclude that the two groups differ along the variable of interest. If, on the other hand, we find that the discrepancies between observed and expected frequencies are small and non-significant, then we have no evidence of differences between the groups.

The best way to visualize all this is by means of a **contingency table**; that is, the crosstabulation you obtain from tabulating your group variable against the characteristic of interest. For example, if you want to compare Greeks and Nigerians on a variable called 'natural headache remedies' (indicating their typical approach to curing headaches), you could visualize the data as illustrated in Table 11.1.

Table 11.1a An example of the two-sample chi-square test: HEADACHE × COUNTRY crosstabulation

			COUNTRY		Total
			Greece	Nigeria	
HEADACHE	Sleeping	Count	584	100	684
		Expected Count	581.5	102.5	684.0
		% of Total	39.9%	7%	46.8%
	Drinking ouzo	Count	599	105	704
		Expected Count	598.5	105.5	704.0
		% of Total	41.1%	7.1%	48.2%
	Juju music	Count	59	14	73
		Expected Count	62.1	10.9	73.0
		% of Total	4%	1%	5.0%

		COUNTRY		Total
		Greece	Nigeria	
Total	Count	1,242	219	1,461
	Expected Count	1,242.0	219.0	1,461.0
	% of Total	85.0%	15.0%	100.0%

Table 11.1b An example of the two-sample chi-square test: chi-square tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1.081[a]	2	.582
Likelihood Ratio	1.014	2	.602
Linear-by-Linear Association	0.488	1	.484
N of Valid Cases	1,461		

Note: [a] 0 cells (0.0%) have an expected count less than 5. The minimum expected count is 10.943.

Table 11.1c An example of the two-sample chi-square test: symmetric measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.027	.582
	Cramer's V	.027	.582
N of Valid Cases		1,461	

From Table 11.1a, you can see that the Greek sample consists of 1,242 people, whereas the Nigerian sample has only 219. The column percentages indicate the relative composition of the sample (i.e., 85% Greeks and 15% Nigerians, respectively), whereas the row percentages indicate the overall popularity of each of the three headache remedies across both samples (with 'drinking ouzo' being the most popular and 'juju music' least popular). Within the cells of the table, two pieces of information are given, namely (a) the observed number (i.e., 'count') of Greeks and Nigerians who are curing their headaches with one of the three remedies and (b) the number of individuals in each sample that would be expected to prefer each remedy if there was no difference in preferences between the two groups (i.e., the 'Expected Count'). Note that we can calculate the expected frequencies by simply multiplying the total observed count in the row and column that correspond to a given cell and then dividing this product by the total sample (i.e., expected frequency$_{ij}$ = (N_{row_i}*N_{column_j})/N_{total}). For example, the expected frequency of Greeks opting for 'juju music' as a headache cure (i.e., cell$_{3,1}$) is (73×1,242)/1,461, which rounds up to 62.1 (see Table 11.1a).

So, where does the chi-square test come in? It makes its appearance in the comparison of the actual versus the expected frequencies. You can basically think of the chi-square (χ^2) as a test statistic that quantifies the overall deviation between observed and expected frequencies under the null hypothesis. If the null hypothesis is true (see earlier), there should be no real difference between the actual and expected counts in each cell of the table, and the resulting χ^2 statistic should be non-significant. In our example, the (Pearson) chi-square statistic comes to

1.081, with an observed probability of 0.582. Since the latter is not smaller than 10% ($p<0.1$), 5% ($p<0.05$), 1% ($p<0.01$), or 0.1% ($p<0.00l$), which are the typical cut-off points for rejecting the null hypothesis (see Chapter 9), we would conclude that there are no statistically significant differences between the Greek and Nigerian populations in terms of their favorite headache remedies. Along with the χ^2, SPSS, by default, offers a number of alternative statistics for you to indulge in. These practically answer the same research question and rarely do not converge. Briefly, the 'Likelihood Ratio' relies on maximum-likelihood theory and, while it tends to be identical to the Pearson χ^2 for large samples, it is often recommended for analyses in small samples. The 'Linear-by-Linear Association' is best suited for investigating linear trends in contingency tables and, unlike the previous two metrics, assumes that the variables are of an ordinal nature (even if they are binary). Finally, note the 'symmetric measures' in Table 11.1c. These are effect size measures and quantify the strength of association between the variables in a standardized way, ranging from 0 to 1 (with 0 indicating no association and values close to 1 showing a strong association). For 2 × 2 tables, Phi and Cramer's *V* will be the same. Yet, for reasons we won't bother you with, in tables where either variable has more than two dimensions (as in our example), the latter is preferred. (We will revisit Phi and Cramer's *V* in Chapter 12, so it's not goodbye forever yet.) Needless to say, had we obtained a larger chi-square value and a lower probability (e.g., a chi-square of 7.378 with a probability of 0.025), we would have been able to reject the null hypothesis. Specifically, we would have argued that the differences between observed and expected frequencies are so large that they would occur only 2.5% of the time if, in the respective populations, the three headache remedies were equally popular. Altogether very easy and straightforward!

WARNING 11.2 In applying the two-sample chi-square test, ensure that no more than 20% of the cells have expected frequencies of less than 5 and no cell has an expected frequency of less than 1.

HINT 11.1 If more than 20% of the cells have expected frequencies of less than 5 and/or at least one cell has an expected frequency of 1 or less, try to combine categories so as to reduce the number of cells in the contingency table.

However, before you go out and celebrate at your local Senegalese restaurant that you understood when and how to use the two-sample chi-square test, there are a few caveats that you have to remember. The chi-square test may not be correct if the expected frequencies are less than 5. In fact, you should always ensure that no more than 20% of the cells in your table have expected values of less than 5 and none of the expected values is less than 1 (see also Chapter 10 on similar limitations of the one-sample chi-square test). Try to combine adjacent categories (i.e., reduce the size of the table) if you face such difficulties, making sure, however, that such combined categories are still meaningful. For example, two separate categories for whiskey and gin could be combined into a new category called 'spirits', but it would be difficult to extract meaning from a combination of ouzo and juju music. As you can see from Table 11.1b (superscript 'a'), the output indicates the 'minimum expected count', so you get an early

warning about potential problems with small expected frequencies (in our example, this comes to 10.943, so no problems).

With a 2 × 2 table (i.e., when two groups are compared on a binary variable), it is often recommended (for reasons far too esoteric to explain here) that **Yates's correction for continuity** is applied to obtain a modified chi-square statistic (the latter is somewhat lower than the unmodified chi-square and, thus, results in a more conservative test). You should be able to routinely apply this correction if needed through your software, so you do not have to worry about it (see later example in Table 11.2).

While on the thrilling subject of 2 × 2 tables, it takes no great imagination to realize that if you have a small sample (say, 20 cases or fewer), the requirements associated with minimum expected frequencies are *bound* to be violated. As it is not possible to combine categories (you would be left without a table), what can you do? Three avenues are open to you, namely (1) dismiss the problem as being a 'minor detail', (2) burst into tears and hope that your supervisor/boss/customer takes pity on you, or (3) use the so-called **Fisher's exact test**. This test is a godsend when you have a small sample as it evaluates exactly the same null hypothesis as the chi-square test. Fisher's exact test is based on the feared **hypergeometric distribution**, which is yet another discrete probability distribution dreamed up by statisticians (yes, they have done a good job of coming up with enough distributions to drive us all round the bend!). The output of the test is simply an exact probability value that can be compared to a pre-set significance level (e.g., 5%) to determine whether significant differences exist between the groups. Being incredibly clever, computer packages tend to perform this test automatically when required; SPSS does so whenever the sample size in the 2 × 2 table is 20 or fewer.

Table 11.2 shows an example of Fisher's exact test applied to a crosstabulation of gender by remarrying intentions (REMARRY); the data relate to 17 Hungarian divorcees who were asked to indicate whether they intended to marry again within the next six months.

Table 11.2a An example of Fisher's exact test: REMARRY × GENDER crosstabulation

			SEX		Total
			Male	Female	
REMARRY	No	Count	4	2	6
		Expected Count	2.5	3.5	6.0
		% of Total	23.5%	11.8%	35.3%
	Yes	Count	3	8	11
		Expected Count	4.5	6.5	11.0
		% of Total	17.6%	47.1%	64.7%
Total		Count	7	10	17
		Expected Count	7.0	10.0	17.0
		% of Total	41.2%	58.8%	100.0%

Table 11.2b An example of Fisher's exact test: chi-square tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.487[a]	1	.115		
Continuity Correction[b]	1.127	1	.288		
Likelihood Ratio	2.506	1	.113		
Fisher's Exact Test				.162	.145
Linear-by-Linear Association	2.341	1	.126		
N of Valid Cases	17				

Notes:

[a] Three cells (75.0%) have an expected count less than 5. The minimum expected count is 2.47.

[b] Computed only for a 2x2 table

Table 11.2c An example of Fisher's exact test: symmetric measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.383	.115
	Cramer's V	.383	.115
N of Valid Cases		17	

Although the value of the chi-square test is still provided in the output, the fact that three out of four cells have expected frequencies less than 5 indicates that it should not be relied upon (either in its raw form or after adjusting with Yates's continuity correction). Instead, we take a look at the probability associated with Fisher's exact test; as neither the one-tailed nor the two-tailed probabilities are remotely close to reaching significance at conventional levels, we can conclude that male and female divorcees do not differ with respect to their remarrying intentions.

A useful feature of chi-square analysis of 2 × 2 tables is that it enables the testing of hypotheses concerning differences in *proportions* between two populations (see Table 10.1 in Chapter 10 for a summary of possible hypotheses). Given that there are two groups involved and a dichotomous variable of interest, the calculation of expected frequencies reflects a null hypothesis of equal proportions between the two groups (see earlier points on the computation of expected frequencies). Although there is another technique that one can use to test for differences between two population proportions (namely, a ***z*-test for differences in proportions** based on the normal distribution – see Further Reading for details), when there are two populations, and the characteristic of interest has two categories, it is possible to use the chi-square test as an alternative way of testing the null hypothesis that two population proportions are actually equal. Indeed, the chi-square test has the advantage that it can be applied to small samples, whereas the *z*-test for differences in proportions should only be applied if *each* group has 30 cases or more. On the other hand, the *z*-test is more flexible as the symmetrical nature of the normal distribution on which it is based allows both one-tailed and two-tailed

tests to be carried out, whereas the chi-square test does not (see Warning 10.3 in Chapter 10). So, in the end, it's all swings and roundabouts.

THE *K*-SAMPLE CHI-SQUARE TEST

We unreservedly apologize for the above heading: '*k*-sample' sounds positively dreadful and is likely to leave you with a bitter aftertaste in your mouth! However, statistics and research methods books frequently use this expression when they mean comparisons between three or more groups and, therefore, we thought that you should know what they are on about.

The ***k*-sample chi-square test** is nothing but an extension of the two-sample chi-square test when more than two groups need to be compared on a nominal variable. Thus if, in a flash of inspiration, we extended our analysis of natural headache remedies by including a sample from Sri Lanka – in addition to our Greek and Nigerian samples – our null hypothesis would be that there are no differences between the three groups with respect to the relative frequency with which group members prefer the various remedies. The approach for setting up the contingency table and for calculating expected frequencies would be exactly the same as for the two-group case and so would the output. Specifically, a chi-square statistic would be provided together with a probability of occurrence under the assumption that the null hypothesis was true; provided that this probability was smaller than our significance level, we would reject the null hypothesis and conclude that the three populations differ in terms of preferred headache remedies. Note that in large and more complex contingency tables (e.g., 3×5), getting a (non-) significant χ^2 test is not very informative. Imagine that you want to find out whether being single, married, or divorced differs across five countries. You run a chi-square test on your 3×5 table and get significant results. Well, you have sufficient evidence to claim that there is indeed an association between individuals' marital status and nationality. However, there are so many combinations between these two variables that you cannot really understand how exactly the differences pan out. In order to get more insight (assuming your software plays along), you can run **column proportion tests**. For instance, SPSS takes all column proportions (for each row) and compares them with one another in pairs using a *z*-test. This follow-up is particularly useful in spotting the key differences in the data.

The two-sample and *k*-sample chi-square tests are often referred to as **tests of homogeneity** since they test whether several groups are homogeneous with regard to the characteristic of interest. Alternatively, they can be seen as **tests of independence** from the perspective of testing whether the two categorical variables forming the contingency table are related or not. While the two interpretations are not strictly synonymous (for reasons better not delved into), from a practical viewpoint they can be treated as such.

As with the two-sample chi-square test, no less than 20% of the cells should have an expected frequency of less than 5, and no cell should have an expected frequency smaller than 1 (see Warning 11.2 and Hint 11.1 above). With a fixed sample size, the bigger the contingency table (i.e., the more groups that are compared and/or the greater the number of categories of

the variable of interest), the greater the chances that you will run into problems. Bear this in mind when deciding on your sample size (see also Hint 2.1 in Chapter 2).

HINT 11.2 Look at the largest crosstabulation (in terms of rows × columns) that you want to analyze before you collect data and adjust your sample size accordingly (or lower your ambitions!).

Now that you feel at home with all possible versions of the chi-square test (single-sample, two-sample, *k*-sample), here's some food for thought: the value of the chi-square statistic is dependent on the sample size. In other words, without changing the pattern in the data, you can get a significant result simply by increasing your sample! To illustrate this, if we multiply all the entries in Table 11.1 by 20, the chi-square value would also become 20 times greater and highly significant (a chi-square value of 21.63 with two degrees of freedom is associated with a *p*-value less than 0.0001!). This happens because statistical power increases dramatically as a result of the increase in sample size (see Chapter 9), and the test goes mad and declares as significant even minuscule differences between the two groups. Indeed, with large sample sizes, even very small differences between two groups may be statistically significant. It is therefore advisable to check the actual percentages in the table as well as the effect size measures to determine whether the statistically significant results are of any practical relevance.

WARNING 11.3 When using the chi-square test in group comparisons, bear in mind its dependence on sample size and go beyond an investigation of statistical significance; look at the magnitude of any revealed 'differences'.

Finally, let us enlighten you with regard to the 'DF' bit that is always part and parcel of any chi-square test output. This stands for the dreaded 'degrees of freedom' and is always equal to $(r - 1)(c - 1)$, where r = number of rows and c = number of columns in the contingency table. Thus, in our example in Table 11.1, DF = 2, given that there are three rows and two columns (since $(3 - 1) \times (2 - 1) = 2$), whereas in Table 11.2, DF = 1 as there are only two rows and two columns. The case of a single-sample chi-square test (discussed in Chapter 10) can be considered as a contingency table having only one row (or one column); the associated degrees of freedom are $r - 1$ (or $c - 1$). Indeed, if you go back to Tables 10.1 and 10.2, you will see that there are three degrees of freedom; that is, one less than the number of categories in the variable of interest.

THE MANN–WHITNEY *U* TEST

The **Mann–Whitney *U* test** (also confusingly known as the 'Wilcoxon rank-sum test' or 'Mann Whitney Wilcoxon Test') is very useful when you have two groups to compare on a variable that is measured at ordinal level. The test focuses on differences in central location and makes the assumption that any differences in the distributions of the two populations are due only to differences in locations (rather than, say, variability).

The null hypothesis tested by the Mann–Whitney *U* test is that there is no difference between the two groups in terms of location, focusing on the median as a measure of central tendency. Note that, in some statistics texts, the null hypothesis is simply stated as involving identical distributions for the two populations; this amounts to the same thing because, as noted above, the test assumes that, if they differ at all, the distributions of the populations differ only with respect to location. You should also bear in mind that, in the case of symmetrical distributions and given interval data, the test can also be used to draw conclusions about means (since the mean and median coincide in symmetrical distributions – see Chapter 7). Thus, the Mann–Whitney *U* test is a useful alternative to a parametric location test (such as the two-sample *t*-test discussed below) when assumptions about normality are violated and/or the sample sizes are small.

Consider the data in Table 11.3, indicating the quality rankings assigned by a sample of 47 Catholic padres and a sample of 56 Jewish rabbis to the Happy Nunnery (a night club specializing in entertainment for clergy); respondents were asked to rank the Happy Nunnery against three other well-known night clubs where 1 = best and 4 = worst.

Table 11.3 Quality rankings for the Happy Nunnery

HAPPYRANK * RELGROUP Crosstabulation					
			PERGROUP		Total
			Padres	Rabbis	
HAPPYRANK	1	Count	27	2	31
					30.1%
	2	Count	10	5	15
					14.6%
	3	Count	5	16	21
					20.4%
	4	Count	5	31	36
					35.0%
Total		Count	47	56	103
		% of Total	45.6%	54.4%	100.0%

We can see under the column 'Total' that 31 respondents placed the Happy Nunnery in the number one position, 15 in the number two position, and so on. If we want to find out whether the assigned quality rankings differ between Catholic padres and Jewish rabbis, the Mann–Whitney *U* test provides the answer; the relevant output is shown in Table 11.4.

Table 11.4a An example of the Mann–Whitney U test: ranks

	RELGROUP	N	Mean Rank	Sum of Ranks
HAPPYRANK	Padres	47	32.65	1,534.55
	Rabbis	56	125.43	7,024.08
	Total	103		

Table 11.4b An example of the Mann–Whitney U test: test statistics[a]

Mann-Whitney U	406.550
Wilcoxon W	1,534.550
Z	−1.283
Asymp. Sig. (2-tailed)	.000

The first piece of information to note from the output is the 'mean rank' for each group (i.e., the sum of the ranks divided by the number of cases). If there are no differences in the populations, we would expect similar ranks in the two groups; if either group has more of its share of either large or small ranks, then it would be unlikely that the respective populations would be the same. The Wilcoxon *W* statistic is simply the sum of ranks for the group with the smaller number of cases (indeed, if you divide *W* by the sample size of the smaller group, you will come up with its mean rank; go ahead, try it yourself). The *U* statistic shows how often a ranking of the group with the *largest* number of cases is smaller than a value in the other group. Note that the significance levels of the *W* and *U* statistics are the same and, provided the *total* sample size is greater than 30, they are obtained through a transformation to a *z*-value (i.e., a standard normal deviate); the latter includes an adjustment for cases with tied (i.e., equal) ranks. If the total sample size is less than 30, an exact significance level for *W* and *U* is also displayed (as the *z*-transformation may not be accurate with small samples).

The results in Table 11.4 are highly significant ($p<0.001$), and we can therefore reject the null hypothesis that the quality rankings of the two groups are the same. Moreover, remembering that the mean ranks represent the sum of the ranks divided by the number of cases, we can see that our 47 Catholic padres are ranking the Happy Nunnery higher than the 56 Jewish rabbis (recall that 1 represents the highest rank). Since some of you might not believe us (and who could blame you for doubting the authenticity of our examples!), look again at the cross-tabulation in Table 11.3, and you will see with surprising clarity that our padres 'dig' the Happy Nunnery more than our rabbis.

THE KRUSKAL–WALLIS (K–W) ONE-WAY ANALYSIS OF VARIANCE (ANOVA)

If you wish to become even more adventurous and decide to compare an ordinal variable (e.g., preference rankings of films, advertisements, hamburger joints, or, indeed, night clubs) across three or more independent groups, the **Kruskal–Wallis one-way ANOVA** is the test for you. In this context, you should note that the mere mention of the name of this test will, in the right circles, immediately qualify you as an intellectual. In contrast, the authors experienced remarkably little success at parties with chat-up lines such as 'Can I possibly interest you in a Kruskal–Wallis one-way analysis of variance?'

In any event, the K–W one-way ANOVA tests the same null hypothesis as the Mann–Whitney *U* test but across *k* rather than two groups (i.e., all groups have the same distribution in the population and any differences reflect differences in location only). In a manner similar to the use of the Mann–Whitney *U* test as a non-parametric alternative to the *t*-test for differ-

ences between two means, K–W one-way ANOVA can be considered as a viable alternative to the parametric one-way ANOVA procedure (to be discussed later in this chapter), when the assumptions of the latter are violated.

To illustrate the use of the K–W one-way ANOVA, we have extended our survey of padres and rabbis by asking a sample of 42 Buddhist monks (doing missionary work in Las Vegas) what they think about the Happy Nunnery night club; the new data are shown in Table 11.5, while Table 11.6 displays the results of the associated K–W one-way ANOVA run.

Table 11.5 Quality rankings for the Happy Nunnery: three groups

HAPPYRANK * RELGROUP Crosstabulation						
			PERGROUP			Total
			Padres	Rabbis	Monks	
HAPPYRANK	1	Count	27	2	9	40
						27.6%
	2	Count	10	5	18	33
						22.8%
	3	Count	5	16	7	28
						19.3%
	4	Count	5	31	8	44
						30.3%
Total		Count	47	56	42	145
		% of Total	32.4%	38.6%	29.0%	100.0%

The first thing that attentive readers (yes, they *do* exist) will have noticed from Table 11.5 is that the 'mean ranks' given for padres and rabbis differ from those given in the Mann–Whitney *U* results in Table 11.4, although the quality rankings of these two groups have not changed. There is a technical reason for this, namely that the test computations rest on the combined ranks of *all* groups. Since we added the 42 observations of the Buddhist monks, this has affected the mean ranks of the padres and rabbis. Of course, their actual assessment of the Happy Nunnery night club has not changed, as can be verified by comparing the crosstabulations *with* (Table 11.5) and *without* (Table 11.3) the Buddhist monk sample added.

Table 11.6a An example of Kruskal–Wallis one-way ANOVA: ranks

	RELGROUP	N	Mean Rank
HAPPYRANK	Padres	47	46.35
	Rabbis	56	99.92
	Monks	42	66.93
	Total	145	

Table 11.6b An example of Kruskal–Wallis one-way ANOVA: independent-samples Kruskal–Wallis test summary

Total N	145
Test Statistic	45.917[a]
Degree Of Freedom	2
Asymptotic Sig. (2-sided test)	.000

Note: [a] The test statistic is adjusted for ties.

As can be seen from the test statistic, the results are again highly significant ($p<0.001$), and we can, therefore, reject the null hypothesis that there is no difference in the ranking of the Happy Nunnery between the three groups; indeed, looking at the mean ranks, padres think the most of the Happy Nunnery and rabbis the least; the rankings of monks are in between.

The test statistic associated with K–W one-way ANOVA is based on an approximation of the chi-square distribution with $k - 1$ degrees of freedom, where k is the number of groups compared; an adjustment, taking into account tied ranks, is also performed (see superscript 'a' in Table 11.6b). In some programs, especially if $k = 3$ or any of the groups has five cases or fewer (which is a disgracefully low sample size), a so-called H statistic is reported; the interpretation of the results remains, of course, the same. Note that a statistically significant K–W test indicates that *at least one* group comparison is statistically significant. However, it does not identify *where* the differences actually lie or *how many* comparisons are significant. Hence, it is always a good idea to also conduct **multiple pairwise comparisons** of the k groups in which a Bonferroni p-value adjustment is applied to protect against familywise (i.e., multiple testing) error inflation. No joke: Carlo Emilio Bonferroni thought about this when he was not feeling quite well after a particularly indulgent meal of spaghetti in Florence, Italy. We will show what such multiple pairwise comparisons look like later in this chapter when we discuss the (parametric) version of ANOVA.

THE TWO-SAMPLE *T*-TEST

This is a straightforward extension of the one-sample t-test for a mean (discussed in Chapter 10) and should be used when one wants to compare two groups on a variable measured at interval (or ratio) level. Thus, whenever you are overcome with an urgent desire to compare the means of two groups, don't think twice: go for the **two-sample** (also known as 'independent-samples') ***t*-test**.

The null hypothesis tested by the two-sample t-test is that the two population means are equal (i.e., H_0: $\mu_1 = \mu_2$), the alternative hypothesis being that the means are not equal (H_1: $\mu_1 \neq \mu_2$). Of course, directional hypotheses can also be tested if one mean is *a priori* expected to be higher or lower than the other; the relevant hypotheses are then H_0: $\mu_1 \leq \mu_2$ versus H_1: $\mu_1 > \mu_2$ and H_0: $\mu_1 \geq \mu_2$ versus H_1: $\mu_1 < \mu_2$, respectively. Note that these hypotheses are analogous to those relating to one-sample location tests, as summarized in Table 10.7 in Chapter 10.

Let us assume that you live in lovely Schweinebratenhausen (beautifully located in the Tyrolean Alps) and you want to compare the mean age of the people living in your town with the mean age of residents of Underwood, North Dakota (that's in the United States for those of you who haven't had the chance to visit yet). The total lack of previous research on this important topic leads you to postulate an exploratory hypothesis, namely that there is no difference in the average ages of the two groups. Having drawn random samples of residents in the two towns, you then confiscate their birth certificates and determine their ages. Subsequently, you diligently input your data into your computer and ask it to perform an independent-samples *t*-test; the output you get is shown in Table 11.7.

Table 11.7a An example of the two-sample *t*-test: group statistics

	TOWN	N	Mean	Std. Deviation	Std. Error Mean
Age	Schweinebratenhausen	102	38.50	13.586	1.345
	Underwood	108	35.40	12.995	1.250

Table 11.7b An example of the two-sample *t*-test: independent-samples test

		Levene's Test for Equality of Variances		*t*-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Age	Equal variances assumed	.433	.511	1.691	208	.092	3.102	1.834	−.514	6.718
	Equal variances not assumed			1.689	205.866	.093	3.102	1.837	−.519	6.723

Table 11.7a shows the means, standard deviations, and standard errors for the two groups; the group variable is denoted as TOWN and coded as 0 for Schweinebratenhausen and 1 for Underwood. Table 11.7b shows the results of the actual *t*-test based on (a) a 'pooled variance estimate' and (b) a 'separate variance estimate'. In its purest form, the two-sample *t*-test is based on the assumption that the variances of the two groups *in the population* are the same; if this assumption holds, then a pooled variance estimate is used, and we look at the *t*-value, number of degrees of freedom (which equal $n - 2$, where n = total sample size), and two-tailed probability in the first row titled 'Equal variances assumed'. If the equal-variances assumption does not hold, then a separate variance estimate is used, and we look at the information in the second row, which reads (drumroll …) 'Equal variances not assumed' (the latter includes an adjustment in the degrees of freedom to take account of the inequality of variances).

But how do we *know* whether the two groups are likely to have the same population variances or not? In anticipation of this, good old SPSS provides us with a test, namely the **Levene's test for equality of variances**; this involves forming the ratio of the two sample variances and is based on the *F*-distribution (yet another probability distribution with known properties – see Chapter 7). If the *F*-value is close to unity, the sample variances are similar; the larger the *F*-value, the more dissimilar the sample variances. The probability beside the *F*-value indicates the likelihood of seeing a ratio at least as large as the one observed in the sample data (here $F = 0.433$) if, in reality, the variances are equal in the population. If the *F*-value is significant, we must use the *t*-test results for unequal variances; otherwise, we use the results where equal variances are assumed. Note that the *F*-test for equality of variances assumes that the underlying populations are normal; if this is not the case, the *F*-value may be suspect as the test is not particularly robust.

In our example, the *F*-value is not significant, so the variances are assumed to be equal. The *t*-statistic comes to 1.691 and just fails to reach significance at the 5% level, given that we are applying a *two-tailed* test (our hypothesis regarding mean ages was exploratory, remember?). We should therefore conclude that there is no evidence that the mean age of residents in Schweinebratenhausen is *different* from the mean age of Underwood residents. If we were considering a directional hypothesis, say that Schweinebratenhausen residents are older than Underwood residents, then the *p*-value would be half that observed in Table 11.7 and significant at $p<0.05$ (see also discussion on one- versus two-tailed *p*-values under the one-sample *t*-test in Chapter 10).

Note that there is an alternative to the two-sample *t*-test that can be used to test for differences in means between two groups. This is the ***z*-test for differences in means** (based on the normal distribution) and is discussed in the texts suggested in the Further Reading section. The reasons we are not considering it here are identical to those offered in Chapter 10 for preferring the one-sample *t*-test over the one-sample *z*-test (i.e., we usually do not know the population variances; with large samples, the *t*- and *z*-tests give practically identical results; and, with small samples and/or distinctly non-normal distributions, it is best to opt for a non-parametric test anyway – the Mann–Whitney *U* test is a good choice in this respect).

Before moving on to greener pastures, we should briefly touch on a final issue if only to demonstrate our commitment, our responsibility, our devotion – in fact, our reverence, sense of duty, and unfailing dedication as educators (sorry, we always get carried away at this point). The first is a rather obvious suggestion concerning the use of Levene's test for equality of variances. While we have presented this test purely as a filter for deciding which version of the *t*-test to use, there is nothing stopping you from using it to test *substantive* hypotheses concerning differences in variability between two populations. Thus you can (and should) use Levene's test in a fashion analogous to that of the (single-sample) variance test (see Chapter 10). Having said that, bear in mind that homogeneity of variance tests are quite sensitive to departures from normality and perform well when the samples involved are big and of equal size. However, as the sample sizes increase, the tests become very powerful and will produce significant results even when variances are practically similar. In contrast, when variances are actually different but small samples are involved, the tests have a high false-negative rate and may produce non-significant results. Go figure!

ONE-WAY ANALYSIS OF VARIANCE (ANOVA)

Having mastered the comparison of two means with the help of the two-sample *t*-test, you might find the newly discovered analytical power at your fingertips so tempting that you want to investigate whether three (or even more) population means are equal. This you can do by running a one-way analysis of variance; of course, as with a *t*-test, you need interval or ratio-level data in order to (meaningfully) calculate means.

The null hypothesis tested by one-way ANOVA is that k groups have equal means in the population ($k \geq 3$); the alternative hypothesis is that *at least one* mean is different from the others. Note that the alternative hypothesis does not indicate *which* groups may differ, only that the groups are not all the same; additional analysis is necessary to identify where the identified differences exist (we will return to this issue shortly).

As with the independent-sample *t*-test above, a number of assumptions must be met before you can legitimately run a one-way ANOVA. First, observations need to be independent, meaning that the scores of one respondent should not be affected by another. Second, each of the groups must be a random sample from a normal population (note that 'normal' refers to the shape of the distribution, not to a state of mind). Third, the variances of the groups must be homogeneous; that is, equal. As a minimum precaution, you should plot the data for each group in a histogram (see Chapter 7) and visually inspect whether group distributions are approximately normal. It is also advisable to have a look at the skewness and kurtosis values (see Chapter 8). To satisfy the more sophisticated statistical connoisseur, you could even run a Kolmogorov–Smirnov test to check whether the normality assumption holds (see Chapter 10). Note that with big samples, if the normality assumption does not *formally* hold, the one-way ANOVA test gives trustworthy results. This is because (a) in large samples, normality tests will produce significant results even for very minor violations (see Chapter 10) and (b) according to the Central Limit Theorem, normality can be assumed anyway with large samples. Moreover, if the sample sizes of the groups are relatively similar, the test is considered robust against violations of the equal-variances assumption.

WARNING 11.4 Before conducting a one-way ANOVA, check that the normality and equal-variance assumptions are (at least approximately) satisfied.

As an example, assume that you have collected 120 questionnaires across three countries (China, Peru, and Liechtenstein) in order to compare consumer attitudes toward Black Forest gâteau (a variable called KUCHEN) on a 10-point semantic differential scale ranging from 10 = 'ambrosial' to 1 = 'rotten'. Your grouping variable (typically referred to as 'factor' in the ANOVA context) is labeled COUNTRY (coded 1 = Liechtenstein, 2 = China, and 3 = Peru) and your null hypothesis is that there is no difference in the mean attitude scores of the consumers in the three countries. Table 11.8 shows the SPSS output resulting from specifying a one-way ANOVA of KUCHEN by COUNTRY. Before we go through the results, note how you should organize all 120 responses of the data set into three distinct groups; you basically

classify the data according to a *single* criterion – that is, according to the factor COUNTRY, hence the name 'one-way' ANOVA.

While at first sight Table 11.8 appears to be as overwhelming as a manual describing the inner workings of a space shuttle, it is really straightforward. There are basically four distinct parts comprising the overall output, namely (a) descriptive statistics on each group, (b) a check on the equality of variances assumption, (c) the results of the one-way ANOVA test itself (along with 'robustness' tests that adjust for variance heterogeneity, if requested), and (d) the results of multiple-comparison tests. Let's deal with the easy parts first.

Table 11.8a An example of one-way analysis of variance: descriptives

KUCHEN								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Liechtenstein	42	7.31	2.214	.342	6.62	8.00	1	10
China	40	6.23	2.616	.414	5.39	7.06	1	10
Peru	38	8.63	1.261	.205	8.22	9.05	4	10
Total	120	7.37	2.319	.212	6.95	7.79	1	10

Table 11.8b An example of one-way analysis of variance: test of homogeneity of variances

		Levene Statistic	df1	df2	Sig.
KUCHEN	Based on Mean	9.205	2	117	.000
	Based on Median	9.808	2	117	.000
	Based on Median and with adjusted df	9.808	2	103.736	.000
	Based on trimmed mean	9.750	2	117	.000

Table 11.8c An example of one-way analysis of variance: ANOVA

KUCHEN					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	113.073	2	56.537	12.557	.000
Within Groups	526.793	117	4.503		
Total	639.867	119			

Table 11.8d An example of one-way analysis of variance: robust tests of equality of means

KUCHEN				
	Statistic[a]	df1	df2	Sig.
Welch	15.848	2	72.189	.000
Brown-Forsythe	12.796	2	95.981	.000

Note: [a] Asymptotically F distributed.

Table 11.8e An example of one-way analysis of variance: multiple comparisons

Dependent Variable: KUCHEN						
Games-Howell						
(I) COUNTRY	(J) COUNTRY	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Liechtenstein	China	1.085	.537	.114	−.20	2.37
	Peru	−1.322*	.398	.004	−2.28	−.37
China	Liechtenstein	−1.085	.537	.114	−2.37	.20
	Peru	−2.407*	.462	.000	−3.52	−1.30
Peru	Liechtenstein	1.322*	.398	.004	.37	2.28
	China	2.407*	.462	.000	1.30	3.52

Note: * The mean difference is significant at the 0.05 level.

Table 11.8a provides a descriptive picture of the groups in terms of group means, standard deviations, standard errors, and minimum/maximum values. These summary measures provide us with an overall 'feel' of the data in each group. Moreover, 95% confidence intervals for each group mean are indicated, which gives us an idea as to whether and to what extent there is an 'overlap' between the three groups.

Table 11.8b provides several alternatives of Levene's test for the equality of variances assumption. Typically, the version that is based on the differences between observed values and the mean is used. However, in the presence of highly skewed distributions or strong outliers, using the median as a reference point is a better idea. Notwithstanding the somewhat limited informative value of such tests (see the relevant discussion relating to the two-group *t*-test in the previous section), if the results are *not* significant, we can assume that variances across groups are not different and proceed with our main one-way ANOVA test. In our case, however, Levene's test indicates that the homogeneity of variances *cannot* be assumed ($p<0.05$). Indeed, examining the standard deviations (s) associated with the different countries reveals that in relation to Peru, the variances (s^2) in the samples of the other countries are more than three times larger. Thus, it is a good idea to also request an adjustment to correct for any heterogeneity of variances (see Table 11.8d). Although two alternative adjustments are

reported here, we tend to have a preference for the Welch's *F* correction (and this is not just because we can hardly pronounce the second one!).

Now, here's the important part. The key portion of the output (Table 11.8c) shows whether the three groups are different or not. This is done is by *partitioning* the overall variability (i.e., variance) in the data into two sources: variability *between* the group means (denoted as 'Between Groups' in Table 11.8c) and variability of the observations *within* each group (denoted as 'Within Groups' in Table 11.8c). In practical terms, **between-groups variability** shows how much variation in the data can be accounted for by the factor COUNTRY; that is, by the fact that consumers belong to different countries. **Within-group variability**, on the other hand, refers to variability in the data that cannot be explained by the factor (often called 'residual variance' or 'error variance') and reflects unsystematic variation due to individual differences; that is, simply because people are different regardless of the group they belong to. Obviously, adding the between-groups variance to the within-group variance will give us the **total variability** in the data.

Between-groups variability is captured by the **between-groups sum of squares**, while a measure of *average* variability across groups is given by the **between-groups mean square**, which is formed by dividing the total sum of squares by the number of degrees of freedom (in this case, equal to $k - 1$ where k = number of groups). The corresponding measures for within-group variability are the **within-group sum of squares** and the **within-group mean square** (computed by dividing the within-group sum of squares by the number of degrees of freedom; the latter is equal to $N - k$, where N = total sample size and k = number of groups).

The test statistic that shows whether there are significant differences between groups in the *F*-ratio, which is formed simply by dividing the between-groups mean square by the within-groups mean square. In our example, the *F*-ratio is 12.556, which is highly significant ($p<0.001$) and signals that it is very unlikely to find such a value (or larger) if the null hypothesis is true. This is further confirmed by the Welch test for robustness. We therefore reject our null hypothesis and conclude that people in Liechtenstein, Peru, and China are, on average, likely to have different attitudes toward Black Forest gâteau.

Having unearthed major insights into the culture-bound appreciation of Black Forest gâteau (what a dazzling research topic!), your appetite for pushing the frontiers of knowledge even further will lead you to question *where*, exactly, the differences lie. So far, we only know that the population means are unlikely to be equal. Looking at the descriptive information on the groups (Table 11.8a), we can see that Peru has, on average, the most favorable attitude, while China has the least. However, we do not know which groups are significantly different from one another. For example, is the difference between Liechtenstein and China statistically significant but that between Liechtenstein and Peru not?

To address these questions, we need to look at the final part of the output in Table 11.8e. Here we are given the results of what are known as **post-hoc tests**, which pinpoint exactly between which groups the differences exist. In our example, there are significant differences between Liechtenstein and Peru and between China and Peru, while there is no difference between China and Liechtenstein. Armed with this knowledge, we can finally conclude that Peru has a more positive attitude toward Black Forest gâteau than either China or Liechtenstein, but no difference in attitude can be identified between the latter two countries.

WARNING 11.5 Do not replace multiple-comparison tests with a series of pairwise *t*-test comparisons. If you do, you will artificially (and wrongly) increase the chance of finding a statistically significant difference in your sample even if there is no difference between the means in the population.

The particular multiple-comparison procedure we applied in our example is called the **Games-Howell test**, which accounts for unequal variances. There are several other multiple-comparison tests available with such intriguing names as the Hochberg's GT2 test (which sounds positively dangerous), the LSD test (which ought to be outlawed), the Tuckey test (which might remind you of the mating call of a roadrunner), and the Bonferroni test (yes, as stated earlier in this chapter, this was the outcome of eating too much spaghetti). These procedures (there are more!) differ mainly in how they calculate the significance level, taking into account pairwise comparisons. Stick to the Games-Howell test and you can't go wrong. Note that the post-hoc comparisons reported in Table 11.8e are non-directional and, thus, the *p*-values are two-tailed. If you have *a priori* expectations about the direction of comparisons, you should divide the obtained *p*-value by 2 to derive the one-tailed equivalent. Finally, do *not* attempt to run independent-sample *t*-tests for all possible pairs of means instead of using an ANOVA with a multiple-comparison procedure. The more means you have to compare, the more likely it becomes that you will find a statistically significant difference even if, in reality, there is no difference between the means in the population. This is the very reason why multiple-comparison tests are used – to take into account that *multiple* (i.e., many) comparisons are made!

RELATED MEASURES: COMPARING VARIABLES

Sometimes in a data analysis project you may want to undertake comparisons between two or more variables rather than between two or more groups. A typical situation is when you have a research design in which you take some measurements of the same group (i.e., a sample of respondents) before and after an event. For example, you may question a set of respondents about their attitude toward a particular product (say, an electric toenail-clipper). Subsequently, you show these individuals an advert featuring the product being used by a celebrity (say, the Pope). Afterwards, you measure their attitude again to find out whether the advert had a positive effect on attitudes. Clearly, in this case, you are dealing with related (or dependent) measures since the *same* individuals are being compared, albeit on different variables (i.e., pre- and post-advertisement attitudes, respectively).

However, it is not only longitudinal designs (involving taking measurements at different points in time) that result in related measures. You often get the latter in cross-sectional designs as well, for example when you ask the same group of people to rate different attributes of a certain product and you want to compare the average importance attached to each attribute. Thus, you could have asked a sample of Norwegian consumers to rate three key attributes of an electric toenail-clipper (say, energy consumption, adjustable blade, and price)

on a five-point importance scale. Subsequently, by comparing the average ratings, you can find out which of the three attributes is considered more important than the others (so that you can emphasize it in your next multi-million-dollar advertising campaign). Again, you have a situation where the same group is measured on different variables; that is, you are again facing related measures.

As was the case when comparing groups, when comparing measures the interest lies in identifying differences that are likely to apply in the *population*. Thus, again, it is important that observed differences based on sample data are subjected to significance tests to ensure that they are not simply the product of sampling error.

Hypotheses involving comparisons between related measures take the following general form:

- There is no difference between the levels of two (or more) measures (null hypothesis).
- There is a difference in the levels of the two (or more) measures (alternative hypothesis – exploratory).
- The level of one measure is higher/lower than that of the other measure(s) (alternative hypothesis – directional).

The overall approach to undertaking comparisons between related measures is very similar to that involving comparisons between groups. Initially, you must establish a (null) hypothesis of no difference. Next, you must select a statistical test that is appropriate for (a) the number of measures to be compared (i.e., two versus more) and (b) the level of measurement involved (i.e., whether you want to compare nominal, ordinal, or interval/ratio measures).

Our amazing Figure 11.1 (see beginning of the chapter) lists the most important statistical tests for undertaking related measures comparisons. Conceptually, these are very similar to the tests we have just considered for investigating differences between groups; for example, the paired-sample sign test is the equivalent to the Mann–Whitney *U* test, while the paired-sample *t*-test is directly analogous to the two-sample *t*-test. This simplifies things considerably and allows us to cover ground more quickly. As we know that your tolerance for learning yet more tests is diminishing at an alarming rate, what we will do is (a) give examples of null hypotheses for each of the tests, (b) use a data set to test these hypotheses, and (c) present the relevant SPSS output with only the briefest of comments; for further details, you should consult our exquisite Further Reading section. Again, for each level of measurement, we will first discuss comparisons between two measures and, while the iron is hot, strike again and extend the discussion to *k*-measures (we will stick to three measures for illustration purposes); frequent reference to Figure 11.1 should ensure that you do not get lost (or, if you do, that you can find your way back!).

The data set we will be using in all that follows is based on a survey of 84 randomly selected shoppers in the island of Lesbos, Greece, who have been asked a number of questions concerning their lifestyles, product preferences, and shopping habits. The answers to these questions have been recorded in a variety of ways, yielding nominal, ordinal, interval, and ratio data. The overall interest lies in identifying similarities and differences in the patterns of responses by these 84 shoppers and testing to see whether these are likely to hold in the population. Note that the sample size may fluctuate somewhat in the various analyses owing to missing data.

THE MCNEMAR TEST

The **McNemar test** is very similar to the chi-square test as applied to a 2 × 2 contingency table and is used when two dichotomous variables are involved. The null hypothesis is that the proportions of subjects with the characteristic of interest are the same under two conditions/treatments.

In our shopper survey, suppose we asked the respondents whether they intended to go to the local cinema during the week (to watch *Freddy Got Married*) and recorded their answers as 1 = yes, 0 = no. We then gave them a coupon that would entitle them to a free tub of popcorn at the cinema, and asked them again if they now intended to go (coding their answers again as 1 = yes, 0 = no). The McNemar test allows us to test whether there was any change in their intentions to watch the movie *before* and *after* being given the coupon; such changes are captured in the upper-right and bottom-left cells of the 2 × 2 table in Table 11.9a. These cells tell us how many people changed their original intention (BEFORE) in response to getting the free popcorn incentive (AFTER).

Table 11.9a An example of the McNemar test

BEFORE	AFTER	
	No=0	Yes=1
No=0	10	37
Yes=1	6	31

Table 11.9b An example of the McNemar test: test statistics[a]

	BEFORE & AFTER
N	84
Chi-Square[b]	20.930
Asymp. Sig.	.000

Notes:

[a] McNemar test

[b] Continuity corrected

The McNemar test normally uses the familiar chi-square statistic (with one degree of freedom) to test for significance. However, if fewer than 10 cases have different values for the two dichotomous variables, the binomial distribution is used instead. Interpretation of the results is very straightforward: the significant test statistic tells us that there has been a change in the intentions as a result of the free coupon. Looking at the crosstabulation, we can see that of the 47 shoppers that did not originally intend to go and watch *Freddy Got Married* (total of the first row), 37 (78.7%) have changed their minds, whereas, from those that were originally intending to go (total of the second row), only six (16.2%) have now decided not to go (probably because they hate coupons, popcorn, or both!).

THE COCHRAN *Q* TEST

The **Cochran *Q* test** is nothing but an extension of the McNemar test to a situation when three or more dichotomous variables are involved. The null hypothesis is that the proportions of subjects are the same across the set of (three or more) measures.

For example, if we had asked our shoppers to indicate with a yes/no answer (where 1 = yes, 0 = no) whether they like (a) oysters, (b) prawns, and (c) crabs, we could use the Cochran *Q* test to test for equality of preferences across the three types of seafood. The output we would get is shown in Table 11.10.

Table 11.10a An example of the Cochran *Q* test: frequencies

	LIKE	
	No=0	Yes=1
Oysters	22	61
Prawns	16	67
Crabs	47	36

Table 11.10b An example of the Cochran *Q* test: test statistics

N	83
Cochran's *Q*	28.456[a]
df	2
Asymp. Sig.	.000

Note: [a] 1 is treated as a success.

The *Q*-statistic follows approximately a chi-square distribution with $k - 1$ degrees of freedom, where k = number of variables compared (here $k = 3$, so DF = 2). The significant result indicates that we have to reject the hypothesis of equal preferences; indeed, looking at the displayed frequencies, we can see that the preferences for crabs are considerably different from those relating to oysters and prawns. For some reason – which only need concern owners of seafood restaurants – only about 40% (36 out of 83) of the sample like crabs as compared with 73% for oysters and 80% for prawns.

THE PAIRED-SAMPLES SIGN TEST

If you want to compare two ordinal measures, this is the test to use. Casting your mind back to Chapter 10, you may recall that one of the location tests we considered was the one-sample sign test (we used that to test our sample's median against a fixed value). In fact, we do so by 'manipulating' the SPSS menu for the paired-samples sign test. So, essentially the **paired-sample sign test** (also called the 'Wilcoxon signed-rank test') tests the null hypothesis that the median *difference* of two measures is zero.

Table 11.11 shows the results of a paired-sample sign test applied to our shopper survey; specifically, respondents were asked to indicate on a five-point scale how often they go to the opera and the theater, respectively (the scaling format used was as follows: 5 = weekly, 4 = fortnightly, 3 = monthly, 2 = annually, and 1 = every seven years). The question of interest is whether the median attendance levels are the same for opera and theater or not.

Table 11.11a An example of the paired-sample sign test: ranks

		N	Mean Rank	Sum of Ranks
OPERA - THEATRE	Negative Ranks	14[a]	21.50	301.00
	Positive Ranks	28[b]	21.50	602.00
	Ties	39[c]		
	Total	81		

Notes:
[a] OPERA < THEATRE
[b] OPERA > THEATRE
[c] OPERA = THEATRE

Table 11.11b An example of the paired-sample sign test: test statistics[a]

	OPERA - THEATRE
Z	-2.160[b]
Asymp. Sig. (2-tailed)	.031

Notes:
[a] Wilcoxon signed ranks test
[b] Based on negative ranks

The test statistic is based on the normal distribution (i.e., a *z*-value is produced) as long as there are a reasonable number of cases; if the sample size is fewer than 30, the test uses the binomial distribution (this also applies to the one-sample sign test – see Tables 10.8 and 10.9 in Chapter 10). The significant result indicates that the median difference is unlikely to be zero; that is, opera and theater attendance are not equally frequent. In fact, looking at the differences between the pairwise ratings, the greater number of instances favoring opera indicates that attendance for the latter is higher than for theater (for more details on the rationale of the test and the interpretation of the signed differences, we refer you back to its single-sample equivalent in Chapter 10).

FRIEDMAN'S ANALYSIS OF VARIANCE (ANOVA)

You can think of **Friedman's ANOVA** as the equivalent of the Kruskal–Wallis one-way ANOVA for related measures. This is the appropriate test to use when several ordinal-level measures need to be compared to one another; conceptually, it is very similar to the Cochran *Q* test discussed earlier, the only difference being that instead of dichotomous, we have ordinal

(rank-order) data. The null hypothesis is, of course, that there are no differences among the set of measures.

Let us assume that we have asked our sample of shoppers to rank three local supermarket chains – Superbuy, Extrasave, and Megadeal – in order of preference (1 = most preferred, 3 = least preferred), and we are interested in finding out whether the three chains are equally preferred. Table 11.12 shows the results of applying a Friedman's ANOVA to the responses obtained.

Again, the test statistic follows an approximation of the chi-square distribution with $k - 1$ degrees of freedom (where k = number of variables); the significant result obtained indicates that preferences among shoppers for the three supermarkets are not equal. If the sample size n and the number of variables k are very small, the test statistic is not exactly distributed as a chi-square, but there are published tables for different combinations of k and n (such tables can be found on the Internet or – if you are lucky – in the texts listed in the Further Reading section).

Table 11.12a An example of Friedman's two-way analysis of variance: ranks

	Mean Rank
SUPERBUY	2.27
EXTRASAVE	1.95
MEGADEAL	1.78

Table 11.12b An example of Friedman's two-way analysis of variance: test statistics[a]

N	79
Chi-Square	9.930
df	2
Asymp. Sig.	.007

Note: [a] Friedman test

This test is the related-measures equivalent to the independent-samples *t*-test for differences in group means. It lends itself nicely to comparisons of two interval or ratio-level measures, the null hypothesis being that the mean difference in the population is zero.

Table 11.13 shows the results of a **paired-sample *t*-test** aiming at identifying whether there is a difference in the quality of service provided by two different restaurants (Nick's Garden Dungeon and Tony's Jolly Hellhole), as perceived by our sample of shoppers. A five-point semantic differential interval scale ranging from 1 = appalling to 5 = outstanding was used to register opinions. Those of you with an incredible memory will have noticed that the format of the output is very similar to that of the one-sample *t*-test (see Table 10.11 in Chapter 10).

Table 11.13a An example of the paired-sample *t*-test: paired-samples statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Hellhole	3.5412	85	.87359	.09475
	Dungeon	4.0235	85	.91906	.09969

Table 11.13b An example of the paired-sample *t*-test: paired-samples correlations

		N	Correlation	Sig.
Pair 1	Hellhole & Dungeon	85	.629	.000

Table 11.13c An example of the paired-sample *t*-test: paired-samples test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Hellhole & Dungeon	−.48235	.77324	.08387	−.64914	−.31557	−5.751	84	.000

You get all sorts of goodies from the output of the paired-sample *t*-test: the means and standard deviations of the two variables, their standard errors (so you can, yet again, calculate confidence intervals). Here SPSS also gives you a test of linear association; that is, the *correlation* between the two measures (we will talk about correlations to an excruciating degree in Chapter 12, so don't worry). Importantly, the output provides you with the difference between the means, its standard error, and the 95% confidence interval. The *t*-value and associated two-tailed probability are also shown.

The significant test statistic indicates that the service quality evaluations of the two restaurants are not the same. Indeed, judging from the mean values, Nick's Garden Dungeon is more highly rated than Tony's Jolly Hellhole (don't forget to visit the next time you're in Lesbos).

REPEATED-MEASURES ANOVA

The **repeated-measures ANOVA** is the extension of the paired-sample *t*-test to a situation where more than two variables are involved. For example, if our sample were asked to rate George's Wacky Tavern (coded as 1) alongside Nick's Garden Dungeon and Tony's Jolly Hellhole (coded as 2 and 3, respectively), we would make *three* paired comparisons of service quality ratings. In other words, instead of applying the paired-sample *t*-test three times (taking two restaurants at a time), we now integrate all these in a single procedure. It's as simple as that.

The logic of the repeated-measured ANOVA is similar to the ANOVA for independent groups; that is, it works by partitioning the total variance in the data into distinct components. The assumptions we need to conform to are identical to those in the ANOVA for independent groups, with the exception that in the one-way repeated-measures ANOVA, we do not deal with homogeneity of variance. Instead, we are concerned with the assumption of *sphericity*. Very simply put, sphericity takes into consideration the variance of the different dependent measures as well as the covariance (i.e., association) between them. To the extent that the variances across measures and the covariances between pairs of measures are relatively equal, sphericity can be assumed.

Table 11.14a An example of the repeated-measures ANOVA: descriptive statistics

	Mean	Std. Deviation	N
Wacky	1.91	1.723	85
Hellhole	5.84	1.503	85
Dungeon	4.45	1.524	85

Table 11.14b An example of the repeated-measures ANOVA: Mauchly's test of sphericity[a]

Measure: Quality_rating							
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
RESTAURANT	.945	4.694	2	.096	.948	.969	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

Notes:

[a] Design: Intercept. Within Subjects Design: RESTAURANT

[b] May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table (Table 11.14c).

Table 11.14c An example of the repeated-measures ANOVA: tests of within-subjects effects

Measure: Quality_rating						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
RESTAURANT	Sphericity Assumed	675.043	2	337.522	115.969	.000
	Greenhouse-Geisser	675.043	1.896	356.081	115.969	.000
	Huynh-Feldt	675.043	1.938	348.273	115.969	.000
	Lower-bound	675.043	1.000	675.043	115.969	.000
Error(RESTAURANT)	Sphericity Assumed	488.957	168	2.910		
	Greenhouse-Geisser	488.957	159.244	3.070		
	Huynh-Feldt	488.957	162.814	3.003		
	Lower-bound	488.957	84.000	5.821		

Table 11.14d An example of the repeated-measures ANOVA: pairwise comparisons effects

Measure: Quality_rating						
(I) RESTAURANT	(J) RESTAURANT	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	−3.929*	.291	.000	−4.640	−3.219
	3	−2.541*	.245	.000	−3.140	−1.942
2	1	3.929*	.291	.000	3.219	4.640
	3	1.388*	.247	.000	.786	1.991
3	1	2.541*	.245	.000	1.942	3.140
	2	−1.388*	.247	.000	−1.991	−.786

Notes:

Based on estimated marginal means

* The mean difference is significant at the .05 level.

[a] Adjustment for multiple comparisons: Bonferroni.

Table 11.14a provides the descriptive statistics (mean and standard deviation) for the quality rating measures of each restaurant. Assuming that you have checked for normality and independence (i.e., ratings across respondents do not influence one another), then the important thing is to examine the assumption of sphericity. As can be seen, Mauchly's W is not significant; thus, sphericity is met. Our null hypothesis that quality ratings for George's Wacky Tavern, Nick's Garden Dungeon, and Tony's Jolly Hellhole are equal is tested in Table 11.14c. Since Mauchly's test was not significant, we look at the *F*-value in the first row titled 'sphericity assumed' (no surprises here). Had we violated the assumption of sphericity, the typical option would be to go for the 'Greenhouse-Geisser' correction reported in the second row. Overall, the ANOVA results are statistically significant ($p<0.001$), indicating that the three quality ratings are not equal. But how do they differ?

Once again, ANOVA's *F* is considered to be an overall test that does not provide information about where the differences may lie. We thus have to examine the pairwise comparisons in Table 11.14d, which show that all three restaurants differ significantly from one another in terms of their quality-of-service ratings (we will leave it up to you to identify the restaurants scoring highest/lowest). Note that all comments relating to multiple comparisons in the context of the one-way ANOVA (e.g., that *p*-values are two-tailed) also apply to the pairwise comparisons associated with repeated-measures ANOVA.

As a final remark, always keep in mind that, in comparing related measures, the comparison undertaken must make *conceptual* sense. This may sound like a trivial comment, but you will be surprised at how often people perform an analysis correctly but on shaky conceptual grounds. For example, comparing the importance of price as a choice criterion when buying a car with the importance of reliability when buying a washing machine is obviously not a good

idea. Even if the same scale is used in both instances to measure importance, the comparison can hardly be meaningful – so *think* before you compare.

SUMMARY

This chapter introduced you to the undoubtedly pleasurable activity of making comparisons. We first reminded you of the difference between independent and related measures, as different statistical techniques are involved in each case. We then looked, in some detail, at different tests for undertaking comparisons between groups and followed this with a discussion of the most common tests for comparing sets of measures. You should now be in a position to select the right technique to carry out comparisons, given the number of groups/measures involved and the level of measurement of the variables (have a look back at Figure 11.1 for a quick confidence booster). If you are still feeling a bit uncertain and your head is spinning with 'Mauchly's test of sphericity' or 'Greenhouse-Geisser' correction, think about the immense pleasures you will experience in the future simply by asking unsuspecting colleagues whether they have applied these tests. You will spread shock and horror with the mere mention of these terms!

QUESTIONS AND PROBLEMS

1. List the considerations you have to go through to decide on the appropriate statistical technique for making comparisons.
2. You are comparing two samples of male and female train passengers on their drinking habits. The passengers drink either tea, coffee, vodka, or beer. Which test would you use to compare males with females in this setting?
3. Explain under which circumstances you would use Fisher's exact test.
4. Is it possible to apply test statistics used for higher levels of measurements to situations involving lower levels of measurement?
5. You have devised a service quality measure for waiters at your university's cafeteria. It is a three-point scale: 'none', 'lousy', 'rudimentary'. If you intended to compare a sample of 35 waiters over 40 years of age with a sample of 27 waiters who are all below 40, which statistical test would you use? Why?
6. You have obtained samples of 100 railway workers each in Nepal and Ireland, and you intend to compare them on the number of visits to beauty salons during the week prior to sampling. Which statistical test are you likely to use?
7. Construct an example built around samples of three different groups of workers employed by the Belgian Ministry of Agriculture that would enable you to run a Kruskal–Wallis one-way ANOVA.
8. Explain when you would use (a) a paired-samples *t*-test, (b) Friedman's ANOVA, and (c) Cochran's *Q* test.
9. How would you go about making comparisons across multiple related measures?
10. How does this chapter compare to peeling potatoes in terms of entertainment value?

FURTHER READING

Babbie, E., Wagner, W. E., & Zaino, J. S. (2018). *Adventures in Social Research: Data Analysis Using IBM SPSS Statistics*, 10th edition. Thousand Oaks, CA: Pine Forge Press. This is a hands-on guide for doing data analysis with SPSS; a major data set is used to illustrate the application of the various techniques and several computer screens are also included. Useful.

Malhotra, N. K., Nunan, D., & Birks D. F. (2017). *Marketing Research: An Applied Approach*, 5th edition. London: Pearson. Chapters 20 and 21 cover several tests for making comparisons in a succinct way.

Rowntree, D. (2018). *Statistics without Tears*. London: Pelican. Chapters 6 and 7 should do the trick – but feel free to cry if you must!

12
Getting adventurous: Searching for *relationships*

THE MYSTIQUE OF RELATIONSHIPS

Relationships with different people form the fundamentals of our lives. There are relationships, for example, between mother and child, between colleagues at work, between lovers, and even with the local public toilet attendant who always acknowledges 'regular' customers with a friendly nod, but whose name one does not know. Of course, the nature of these relationships varies considerably, as does their strength: some are quite casual, while others are very intense.

Relationships (or 'associations') also play an important role in data analysis. Very often we want to find out whether two variables are related and, if so, the nature and strength of this relationship. To do this, we employ what are known as **measures of association**. In this chapter, we (reluctantly!) restrain ourselves to **bivariate relationships** only (i.e., those involving two variables). However, as restraint is not exactly among our numerous virtues, we will give you a good taste of **multivariate relationships** (i.e., those involving three or more variables) in the next chapter.

In studying relationships, we frequently ask questions, such as: is advertising related to sales? Is number of children related to happiness? Is the number of royal family scandals related to newspaper circulation? Questions of this kind are primarily directed at discovering *whether* a relationship exists between variables. Note that an association between two variables implies that the variables tend to change together. For instance, as one variable increases or decreases, the other variable also increases or decreases; if a variable stays put while another variable changes (in any direction), then there is no association among them (no relationship).

In many cases, we are not only interested in the mere existence of a relationship but also in the way in which the two variables are related to one another; that is, in the *direction* of the relationship. For example, we may ask: is a decrease in price associated with an increase in sales? Is hair loss positively related to beer drinking? Is there a negative link between church attendance and attendance at rave parties? Here, we are interested in whether the variables change in the *same* direction (positive relationship) or in the *opposite* direction (negative relationship).

WARNING 12.1 Always accompany a measure of association with a statement regarding its significance.

Another aspect of a relationship we may be interested in is its *magnitude*. Thus, although we may find, say, a positive relationship between advertising and sales, this relationship may not be as strong as, say, between price reductions and sales. The magnitude (or 'strength') of a relationship tells us how closely two variables are related to one another. In this context, measures of association are usually calibrated to range between 0 and ±l, with 0 indicating no relationship between the variables (i.e., complete independence) and 1 a perfect relationship (the '+' or '–' sign indicating the direction of the relationship); intermediate values reflect different magnitudes of the relationship. Thus, measures of association also often function as effect size measures (see Chapter 9). Unfortunately, what constitutes a 'weak' or a 'strong' association is a rather complex matter, and while there are several rules of thumb providing T-shirt size-type guidelines (i.e., small, medium, large), the strength of an association should be considered in the context of the specific research design (e.g., small vs. large sample size) and the specific research field (e.g., advertising vs. pharmaceutical research). Note that, with real-life data, it is very unlikely that a measure of association will reach its extreme values (i.e., 0 or 1). For example, even if in the population the two variables are totally unrelated, a measure of association will generally produce a non-zero value owing to sampling error. Thus, when dealing with sample data, it is important to *test* whether a value produced by a measure of association does, in fact, reflect the existence of a 'true' relationship in the population. This is easily done as most measures of association are accompanied by a significance test, which tests whether the association between the variables in the population is indeed significantly different from zero.

Bearing all the above in mind, hypotheses involving relationships between two variables X and Y can be stated as follows:

- There is no relationship (i.e., zero association) between the two variables: X and Y are unrelated (null hypothesis).
- There is a relationship (i.e., non-zero association) between the two variables: X and Y are related (alternative hypothesis – exploratory).
- There is a positive/negative relationship between the two variables: X is positively/negatively related to Y (alternative hypothesis – directional).

WARNING 12.2 Do not attempt to interpret the direction of a relationship between nominal variables.

MEASURES OF ASSOCIATION

There are a large number of (bivariate) measures of association that can be used to examine relationships between variables (see Further Reading section for examples). After careful deliberation, many sleepless nights, close consultation with 43 Nobel laureates in statistics,

and three visits to a fortune-teller, we selected a few of them, which are definitely among the most widely used: **Cramer's *V*** for looking at the association between nominal variables; **Spearman's rank-order correlation** for capturing the relationship between ordinal variables; and **Pearson's product moment correlation** for examining linkages between interval- and/or ratio-scaled variables. We will consider each of them in turn and mention in passing some other association measures (i.e., the **Phi coefficient**, **Kendall's Tau**, and the dreaded **point-biserial correlation**). Finally, to spoil you even more, we will discuss some techniques that not only identify the underlying relationship between variables but also allow us to make specific **predictions** about the expected value of one variable based on the other. These are the bivariate **linear regression**, with which we can predict the expected value of a continuous (interval or ratio) variable, and the bivariate **logistic regression**, which is used to predict the expected value of a categorical (nominal or ordinal) variable.

Cramer's *V*

You may recall from Chapter 9 that statements concerning relationships between nominal variables, such as hair color, religious denomination, or nationality, are inherently limited by the nature of the measures. Specifically, it does not make sense to try to interpret the *direction* of an association between nominally scaled variables, as demonstrated by meaningless statements like 'as nationality increases, religious denomination decreases' or 'hair color is negatively related to religious denomination'. Since categories of nominal variables do not possess a meaningful order (e.g., blond is not more or less of a hair color than black), they cannot be related in a particular direction. Indeed, given the arbitrary scoring of nominal scales (see Chapter 3), whether a relationship turns out to be positive or negative depends purely on how categories are scored rather than on any intrinsic pattern in the data.

WARNING 12.3 The chi-square statistic can only establish whether two nominal variables are independent or not. It does not show the strength of the association between the variables.

However, what you can interpret is the *strength* of the dependence or relationship between two nominal variables, and this is where Cramer's *V* comes in handy. To understand Cramer's *V*, it is necessary to recall our titillating discussion of the chi-square statistic in Chapter 11. You may vaguely remember that we used chi-square to test the null hypothesis that there is no difference between the actual and expected frequencies in each cell of a contingency table; that is, that the two nominal variables forming the table are independent. Unfortunately, even if we reject the hypothesis of independence, we cannot conclude anything about the strength of the dependency between the two variables. This is because the value of the chi-square test statistic depends on the sample size and the size of the contingency table; as a result, the value of chi-square cannot distinguish between different relationships in terms of their strength.

But why spend time demonstrating which technique does *not* work instead of telling you one that *does*? And why does toast always land on the buttered side when it falls off the table? While we have still not solved the last problem, the answer to the first question is that

chi-square provides the basis for many measures of association developed by statisticians. While, on its own, chi-square can only test independence, it can be modified so that (a) it is not influenced by sample size and (b) its values fall in a range from 0 to 1 (where 0 indicates no association and 1 indicates perfect association).

Cramer's *V* represents such a chi-square-based adjustment. Its values always fall between 0 and 1 and, thus, can be interpreted as reflecting relationships of different magnitudes. Cramer's *V* is often provided as an additional option in statistical analysis packages; if your package does not include it, you can easily calculate it by hand if you know the chi-square value. The relevant formula is as follows:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

All you have to do is divide the chi-square value by *n* (the sample size) times (*k*–1), where *k* is the *smaller* of the number of rows or columns in the contingency table, and then take the square root. If either the number of rows or columns equals 2, it follows that you only need to divide the chi-square value by the sample size. Statistical connoisseurs (should one or two have accidentally picked up this book) will notice that, in this special case, Cramer's *V* is identical to the so-called **Phi (φ) coefficient** (which is a measure of association for 2 × 2 tables, i.e., when two dichotomous variables are related to one another – see also Chapter 11). While this is completely inconsequential for most of us, we thought we ought to mention it in a desperate attempt to raise the academic standard of the book!

HINT 12.1	Report both the chi-square and Cramer's *V* statistics to make interpretation clearer.

To show you how easy it is to calculate Cramer's *V*, we will use the data in Table 11.1 in Chapter 11 (and we categorically dismiss any suggestion that the 'recycling' of Table 11.1 reflects laziness on our part to come up with a fresh example!). Just in case you cannot recall the relevant information from memory, the chi-square value was 1.08, the table had three rows and two columns, and the sample size was 1,461. Plugging this information into the above formula, we get the following:

$$V = \sqrt{\frac{1.08}{1461(2-1)}} = 0.027$$

This is as close to zero as you can get, so we can be fairly sure that there is no relationship whatsoever between the country of origin and preferred headache remedies. In any case, the fact that the chi-square statistic was not even remotely significant in the first place (its *p*-value was 0.582) would have warned us not to raise our hopes when looking at the strength of the relationship (because there is none!). In this context, you should look at and report both the chi-square statistic *and* Cramer's *V*, as the interpretation of the latter rests on your ability to

reject the null hypothesis of independence between the two nominal variables (as tested by the chi-square).

It is worth noting that an advantage of Cramer's V is that it enables us to compare contingency tables of different sizes (unlike Phi, which becomes unreliable for bigger than 2×2 tables) and, equally importantly, tables based on different sample sizes; the chi-square statistic allows us to do neither.

Spearman's rank-order correlation

If you are dealing with a situation in which both variables concerned are (at least) ordinal, you can investigate not only the strength of the association but also its direction (and, thus, distinguish between positive and negative relationships). Spearman's rank-order correlation coefficient (also sometimes referred to as 'Spearman's (ρ) rho') is an appropriate measure of association in this case. It ranges from −1 to +1, with values close to zero indicating little or no association between the variables concerned. Moreover, its sampling distribution under the null hypothesis is known (it follows approximately a t-distribution), so we can use this to test for significance.

Consider the following example for further enlightenment. We are trying to find out whether there is a relationship between liking the royal family and liking those strange little dogs (known as corgis) Her Majesty Queen Elizabeth II was often seen with. To do this, we asked a random sample of 145 Japanese golfers to indicate their liking for the (British!) royal family on the following scale: 1 = 'I don't like them', 2 = 'They are OK – just', and 3 = 'I would like to move in with them'. Liking for the corgis was captured by responses to the following scale: 1 = 'Would not waste a photo on such a dog', 2 = 'Would only take a photo from a safe distance', 3 = 'Would like to take lots of photos', and 4 = 'Would like to be photographed with such a dog on my lap'. In Table 12.1 we use the (by now hopefully familiar) crosstabulation to summarize the data and present the output of the Spearman's rho test.

Table 12.1a Liking the royal family and liking corgis: RoyalPREF × DogPREF Crosstabulation

Count						
		DogPREF				Total
		No photo	Distant photo	Lots of photos	Joint photo	
RoyalPREF	I don't like them	27	10	5	5	47
	They are OK – just	4	5	16	31	56
	I would like to move in with them	9	18	7	8	42
Total		73	40	33	28	44

Table 12.1b Liking the royal family and liking corgis: correlations

			RoyalPREF	DogPREF
Spearman's rho	RoyalPREF	Correlation Coefficient	1.000	.225*
		Sig. (2-tailed)	.	.005
		N	145	145
	DogPREF	Correlation Coefficient	.225*	1.000
		Sig. (2-tailed)	.005	.
		N	145	145

Note: *. Correlation is significant at the 0.01 level (2-tailed).

Requesting the Spearman rank-order correlation coefficient, we obtain a value of 0.225, which is highly significant (for the record, the corresponding *t*-value comes to approximately 2.77). This means that the observed correlation is unlikely to have come about *if* there was no association between the two variables in the population (i.e., if the population correlation coefficient was equal to zero). Thus, we can reject the null hypothesis that there is no relationship between liking the royal family and liking corgis. According to our findings, the two variables are positively related, with the strength of their relationship being relatively weak (as indicated by the magnitude of the coefficient).

Note that we would have reached exactly the same conclusion had we employed **Kendall's rank-order correlation** (also known as 'Kendall's (τ) tau'), which is another widely used (non-parametric) measure of association for ordinal variables (it would have produced a value of 0.187, which is also significant at $p<0.01$). While there are some technical differences between Spearman's and Kendall's coefficients, they tend to produce quite similar results in practical applications: use whichever makes you happy (or whichever you can get from your analysis package). When applied to a given data set, both measures will almost always reach the same conclusion and reject the null hypothesis at the same level of significance (as was the case in our example). However, Spearman's and Kendall's coefficients use different algorithms in their calculation and, obviously, their actual values cannot be directly compared.

WARNING 12.4 Don't 'mix and match' Spearman's and Kendall's rank-order correlation coefficients; numerically, they are not directly comparable to each other.

Pearson's product moment correlation

This is the most widely used measure of association for examining relationships between interval and/or ratio variables. Also known as 'Pearson's *r*', the product moment correlation coefficient focuses specifically on **linear relationships** and ranges from −1 (a perfect negative linear relationship) through 0 (no linear relationship) to +1 (a perfect positive linear relationship). The emphasis on 'linear' is important because if two variables are linked to one another by means of a **non-linear relationship**, the Pearson's correlation coefficient cannot really

detect it. For example, calculating Pearson's correlation coefficient based on the data displayed in Figure 12.1 is likely to produce a small and non-significant value; this would suggest no relationship between the variables despite the fact that a substantial (albeit curvilinear) relationship exists between the two variables.

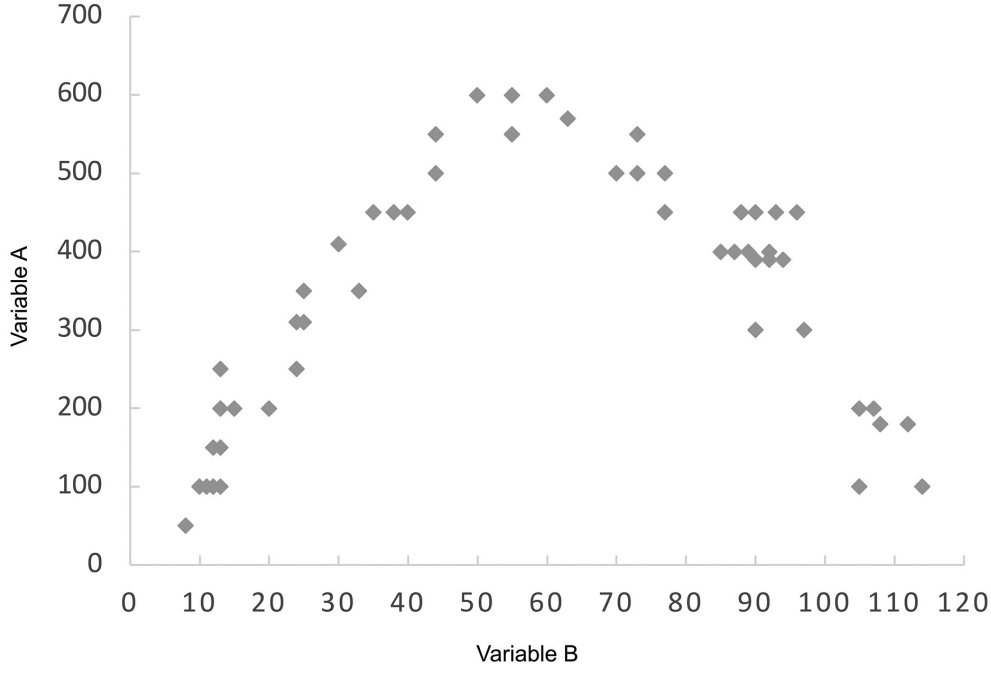


Figure 12.1 An example of a non-linear relationship

To avoid such pitfalls, it is always a good idea to plot the relationship between the variables in a **scatterplot** before applying Pearson's correlation coefficient. Figure 12.2 shows a scatterplot pattern for what is clearly a positive linear relationship between the sales volume (SALESVOL) for left-handed teacups in a sample of export markets served by Global Teapot Provisions Inc. (a manufacturer of teapots, cups, and assorted products) and the number of sales representatives (SALESREP) appointed in these markets. A visual inspection of the plot suggests that an increase in the number of sales representatives tends to be associated with an increase in sales and vice versa.

HINT 12.2	Plot your data whenever possible. A visual inspection helps to gain a better understanding of the relationship between two variables.

HINT 12.3 Use a one-tailed significance test of the correlation coefficient if you know in *advance* the directionality of the relationship between the variables. Otherwise, use a two-tailed test.

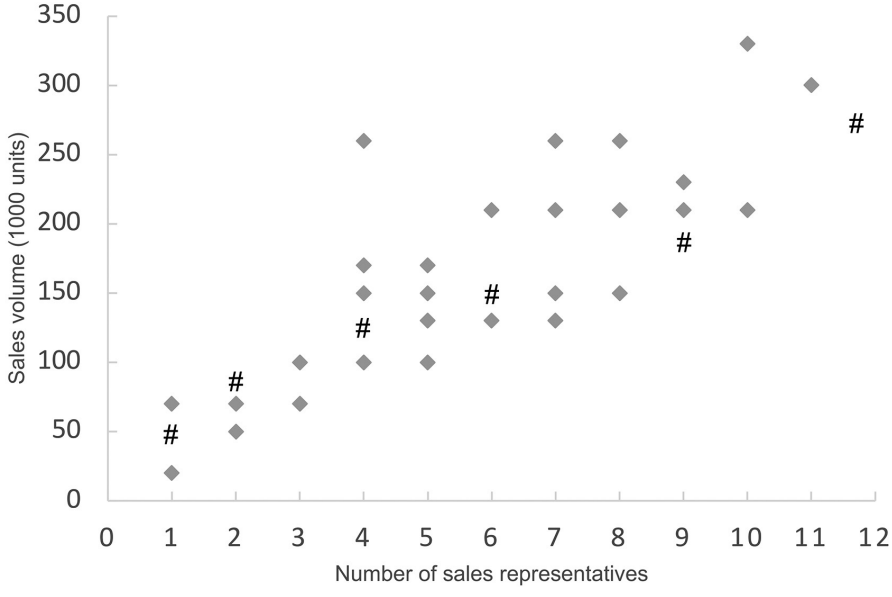


Figure 12.2 An example of a linear relationship. (# indicates two or more cases share the same values.)

If we now kindly ask SPSS for the Pearson product moment correlation coefficient, we should get something along the lines of Table 12.2. The coefficient is positive, of substantial magnitude (r = 0.75), and highly significant (p<0.001). Bear in mind (as mentioned previously) that the exact probability may be, say, p=0.00003 but SPSS limits the output to three decimal points. Staying with the p-value, you should further notice that one can usually choose whether the output reports one-tailed or two-tailed significance levels. Report the more powerful one-tailed test probability whenever you are confident about the direction of the relationship (whether positive or negative). In cases where you do not know the directionality of the expected relationship, you should report the two-tailed p-value (see also earlier discussion on choosing one- vs. two-tailed tests in Chapter 10).

HINT 12.4 If you are primarily interested in the direction of a (linear) relationship, use the Pearson correlation coefficient (r). If your main concern is the strength of the relationship, focus on r^2.

Table 12.2 An example of Pearson's product moment correlation

Correlations			
		SALESVOL	SALEREP
SALESVOL	Pearson Correlation	1	.751*
	Sig. (2-tailed)		.000
	N	45	45
SALESREP	Pearson Correlation	.751*	1
	Sig. (2-tailed)	.000	
	N	45	45

Note: *. Correlation is significant at the 0.01 level (1-tailed).

A Pearson correlation coefficient of 0.75 tends to suggest a 'moderate to strong' association between the two variables. However, we can go beyond this somewhat imprecise description and actually calculate the **proportion of variance** in sales volume that is shared with the number of sales representatives. This is easily achieved by simply squaring the correlation coefficient; that is, by calculating r^2. In our example, $r^2 = (0.75)^2 = 0.56$, which indicates 56% of common variation between sales volume and number of sales representatives. Both r and r^2 measure the strength of a linear association between two interval- or ratio-scaled variables. The r^2 plays an important role in the interpretation of relationships between variables as it shows the proportion of variance in one variable explained by the other; however, it tells us nothing about the direction of the relationship. In Figure 12.3, we present you (at no extra cost) with three different cases of association between two variables, x and y. Each panel uses a scatterplot and a **Venn diagram** to visualize the correlation (r) and shared variance (r^2) of the variables. Note that if these correlation coefficients were negative, their corresponding shared variance would be exactly the same.

At this point, if only to show off, we should mention the **point-biserial correlation**. This is mathematically equivalent to the Pearson correlation coefficient, where one variable is continuous (interval or ratio) and the other a nominal binary variable. For example, if we wanted to correlate gender (male vs. female) or shoe color (red vs. black) with the willingness to buy high-heeled shoes (measured in a five-point interval scale anchored by 1 = 'no way' and 5 = 'absolutely'), this is the correct association measure to use. To obtain the point-biserial correlation coefficient, we simply run a standard Pearson correlation analysis by including the binary variable in the menu. The result is interpreted in exactly the same way as described in relation to Table 12.2.

While we pledged to look at only two variables at a time, we would neglect our duties (and, no doubt, upset book reviewers) if we failed to mention one of the most widely used applications of the Pearson correlation coefficient, namely its use in creating **correlation matrices**. This will correlate a *set* of variables with each other (taking two at a time) and indicate which of the resulting relationships are statistically significant. Table 12.3 shows an example of such a correlation matrix for five variables.

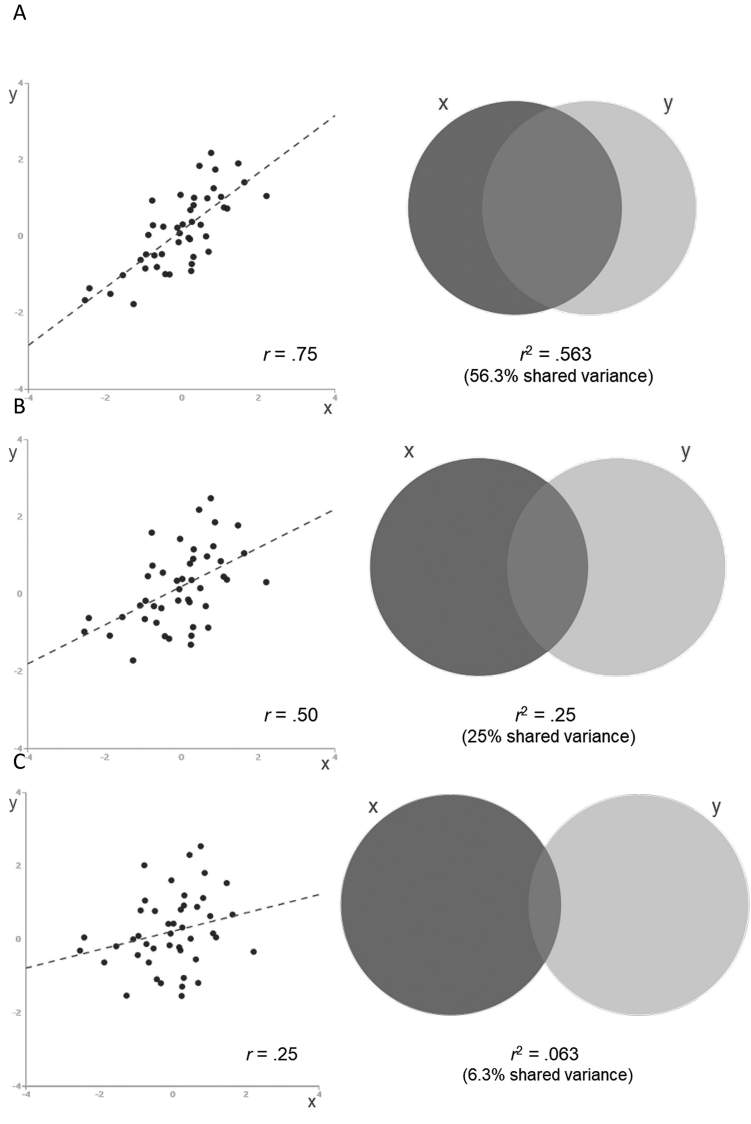


Figure 12.3 Visual representation of correlation and shared variance between two variables

Table 12.3 An example of a correlation matrix

Correlations						
		VAR1	VAR2	VAR3	VAR4	VAR5
VAR1	Pearson Correlation	1	.633*	.532*	.464*	.271
	Sig. (2-tailed)		.000	.000	.001	.071
	N	45	45	45	45	45
VAR2	Pearson Correlation	.633*	1	.703*	.606*	.210
	Sig. (2-tailed)	.000		.000	.000	.166
	N	45	45	45	45	45
VAR3	Pearson Correlation	.532*	.703*	1	.825*	.474*
	Sig. (2-tailed)	.000	.000		.000	.001
	N	45	45	45	45	45
VAR4	Pearson Correlation	.464*	.606*	.825*	1	.552*
	Sig. (2-tailed)	.001	.000	.000		.000
	N	45	45	45	45	45
VAR5	Pearson Correlation	.271	.210	.474*	.552*	1
	Sig. (2-tailed)	.071	.166	.001	.000	
	N	45	45	45	45	45

Note: *. Correlation is significant at the 0.01 level (2-tailed).

Note that all correlations on the diagonal are equal to 1; this should not surprise you, since the correlation of a variable with itself *has* to be perfect. As far as the off-diagonal elements are concerned, note that the correlations forming the two triangles above and below the diagonal are identical; thus, it is only necessary (and customary) to report the correlation results in one of the two triangles and leave the other blank.

WARNING 12.5 Do not blindly correlate all variables with each other – correlation matrices are always easy to produce but often difficult to justify and interpret!

Although the ease of producing a correlation matrix tempts many researchers – especially the inexperienced and the lazy – to correlate *all* variables in their study with each other in order to uncover all possible bivariate relationships (or, expressed less charitably, to desperately fish for results), we advise you to use this procedure with caution! First, such an indiscriminate approach is a giveaway that the study lacks conceptualization and shows that the researchers do not really know what they are looking for (see also our discussion on '*p*-hacking' in Chapter 10). Second, if you compute a large enough number of correlation coefficients, some

relationships will turn out to be significant by pure chance. For example, if you compute 100 correlation coefficients, you could expect about 10 of them to show significance levels at $p<0.10$ or better, even when there is no relationship among these variables in the population. Third, the amount of computer printout that you will generate will be so overwhelming that you will most probably have to spend all your weekends, public holidays, and entire summer vacation poring over the output! To put it in perspective, correlating four variables with each other results in six distinct correlations, while intercorrelating eight variables results in no fewer than 28 correlations (in general, given k variables, the number of distinct correlations is equal to $k(k-1)/2$ – not a pretty thought!).

Two final points concerning Pearson's correlation coefficient. The first is that it assumes that the *joint* distribution of the variables in the population is normal; that is, that we are sampling from a **bivariate normal distribution**. Note that this assumption affects the significance test of r, not the computation of r itself; in this context, bear in mind that it is not possible to make population inferences if the joint distribution in the population is not normal. However, there is nothing to prevent us from computing descriptive correlation measures for the sample. While there are ways of testing for bivariate normality, the procedures involved are way outside the scope of this book. For practical purposes, as long as the *individual* distributions are not markedly non-normal and the sample has a reasonable size, one can safely use Pearson's correlation coefficient (if either of these conditions does not apply, one might just as well rely on non-parametric measures of association such as Spearman's and Kendall's rank-order correlations).

WARNING 12.6 Always think about the substantive significance of any statistically significant correlation coefficient.

The second point has to do with interpretation. Having obtained a statistically significant result, you should pause and think about the *substantive* significance and the limitations of the findings (this is also a good excuse to take some time out and think about the meaning of life in general). Specifically, you should always keep in the back of your mind that, for very large sample sizes, even small correlation coefficients (e.g., 0.1 or 0.2) may be significant (see also discussion of power in Chapter 10). While such a result would, indeed, indicate a very small linear relationship between the variables, the relationship may not be particularly important from a substantive point of view; indeed, taking an r^2 perspective, only 1% and 4% of the variance is explained by the above relationships. Thus, you should consider the 'practical' significance of your correlation coefficient alongside its statistical significance (see also relevant discussion in Chapter 9).

SIMPLE LINEAR REGRESSION

Given your sharp intellect and ability to reflect on the material we have so very carefully presented in this chapter, you will have realized that when two variables are correlated, to a certain extent, we can use the values of one variable to make an educated guess or *predict* the values of the other. This is correct, and it is precisely these kinds of cunning conclusions that will

catapult you into the higher-income brackets in your future career. However, the correlation coefficient is a standardized measure and, as such, it cannot tell us how much a variable will change in response to a certain change of the other variable, nor does it allow us to predict the **expected value** of a variable for a given value of the other variable. This is where simple (bivariate) linear regression analysis comes into play. Regression analysis assesses the relationship between two variables while also keeping the original units of measurement, thus allowing us to make specific predictions. Unlike with correlation analysis in regression analysis, we distinguish between the **independent variable** (also called the 'predictor' or 'explanatory') and the **dependent variable** (also known as the 'criterion' or 'outcome'). We conventionally refer to the predictor and the outcome variable as *X* and *Y*, respectively. Regression analysis models the relationship between the two variables using the equation of the straight line. In practical terms, imagine that we plot the distribution of data for all pairs of the *X* and *Y* variables (i.e., we project the bivariate relationship in a scatterplot) and we try to identify a (straight) line that best 'fits' (i.e., describes) the data. This straight line is basically our regression model and is used to estimate the predicted values of the outcome ($\hat{Y}$) at given values of the predictor (*X*). Figure 12.4 visually represents this idea. By changing the slope (Panel A) and/or the intercepts (Panel B), we can come up with the optimal model.

$$\hat{Y} = \alpha + \beta X (+\varepsilon)$$

In short, regression equation assumes that the expected values of *Y* are a linear function of *X*, with α representing the **intercept** (i.e., the point at which the line crosses the y-axis), β representing the regression coefficient (that is, the slope of the line), and ε corresponding to the prediction error or **residual**.

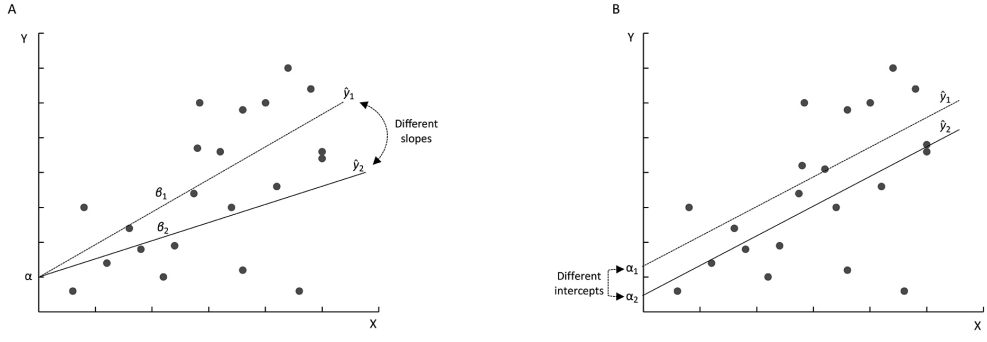


Figure 12.4 Regression lines with different (same) slopes and same (different) intercepts

The line best describing the data (i.e., the best-fitting model) is the one that passes the closest through all possible data points. This also means that the total **residuals** (the 'leftovers') associated with the regression model should be at a minimum (note in this context that it is unkind

to refer to your unmarried relatives as 'residuals'). To obtain the optimal line, one would take all deviations (i.e., residuals) between the observed data (the dots) and different regression lines ($\hat{y}$), square them (so that negative and positive residuals do not offset each other), add everything together to form the sum of squared residuals (SS_R), and finally select the line that corresponds to the least SS_R. This is the standard estimation method used in regression and it goes by the name **Ordinary Least Squares (OLS)**. Luckily, you don't have to carry out *any* of these steps yourself as your computer software will do it for you in no time.

Regression analysis can be thought of as a two-stage process. Typically, the first stage involves an assessment of the overall model fit, while the second stage focuses on the necessary parameters estimates, which you can plug into the equation of the line and make specific predictions.

To illustrate the application of regression analysis, suppose that we want to investigate whether the amount of time spent reading this book (BOOKTIME) predicts readers' intention to take up meditation classes (MEDITATE) and that we have surveyed a random sample of 40 people who bought this book (which might be close to the overall population of interest) about the amount of time they have spent on the book (BOOKTIME; measured in number of hours), as well as their intention to start meditation (MEDITATE; measured with a seven-point interval scale, anchored by 1 = 'Nope, my chakras are just fine' and 7 = 'Get me my mantra now!'). In order to answer this question, we can regress MEDITATE on BOOKTIME. Put differently, we can run a simple linear regression with BOOKTIME as the predictor variable and MEDITATE as the outcome variable. The main parts of the relevant SPSS output are presented in Table 12.4.

Table 12.4a An example of simple linear regression: model summary[a]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.367[b]	.134	.112	1.358

Notes:

[a] Dependent variable: MEDITATE

[b] Predictors: (Constant), BOOKTIME

Table 12.4b An example of simple linear regression: ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.878	1	10.878	5.897	.020[b]
	Residual	70.097	38	1.845		
	Total	80.975	39			

Notes:

[a] Dependent variable: MEDITATE

[b] Predictors: (Constant), BOOKTIME

Table 12.4c An example of simple linear regression: coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3.853	.604		6.375	.000	2.629	5.077
	BOOKTIME	.038	.016	.367	2.428	.020	.006	.070

Note: [a] Dependent variable: MEDITATE

Table 12.4a provides the first fit diagnostics of our model. In simple regression, the *r*-value is just the Pearson correlation coefficient between BOOKTIME and MEDITATE and shows a positive relationship ($R = 0.357$) between the variables. The R^2 is the **coefficient of determination** and indicates that 13.4% of the total variation in the dependent variable (MEDITATE) can be explained by the independent variable (BOOKTIME). The adjusted R^2 is particularly relevant in *multiple* regression analysis and will be discussed in the following chapter. Finally, the *standard error of the estimate* is the square root of the mean squares of the residuals (MS_R) and, being measured in the same units as the outcome variable, can be used as an indication of the accuracy of predictions made by the regression line (e.g., our regression line, on average, *mis*predicts MEDITATE scores by 1.358).

Table 12.4b tests the model's fit using the *F*-ratio statistic (which should be familiar to you from Chapter 11). As you might have guessed, model fit is tested by partitioning the total variance (as reflected in the total sum of squares into that explained by the model and that not explained by the model, i.e., the residual variance). In our example, the *F*-value is rather high (5.897) and significant ($p<0.05$), suggesting that the amount of time reading the book indeed predicts people's intention to take meditation classes.

Table 12.4c gives specific information about the model's parameters and it is necessary to construct the prediction equation. Initially, the results provide the estimated value of the intercept (labeled 'constant'), which is the starting point of the regression line. Technically, it shows the expected value of MEDITATE when BOOKTIME equals zero. In our case, the intercept is 3.853 and is associated with a significant *t*-test indicating that the intercept is different from zero. In the second row, we see the independent effect of our predictor in full detail. This is reflected in the **unstandardized regression coefficient** B (i.e., the slope) and its corresponding test of significance. B is expressed in the original units of measurement and shows that average change in MEDITATE is associated with one-unit increase in BOOKTIME. Thus, we can say that spending an additional hour reading this book increases people's intention to start meditation classes by an average of 0.038 units. By dividing the B (0.038) by its standard error (0.016), we get a *t*-value (2.428), showing that our regression slope is significantly different from zero ($p<0.05$). The *p*-values reported under the 'sig.' label refer to a two-tailed test. The output also kindly includes 95% confidence intervals for B (if you have forgotten what a confidence interval is, have another look at Chapter 8). The **standardized regression coefficient** (Beta) expresses the regression relationship in a standardized way. It can be read as: one standard deviation change in BOOKTIME results in 0.367 standard deviations change in MEDITATE.

Note that in simple (bivariate) linear regression, the β is essentially the correlation coefficient (*r* or *R*) between the predictor and the outcome. This does not hold true if multiple predictors are included in the regression model (where standardized regression coefficients are important to assess the relative strength of individual predictors).

Overall, our regression equation can be expressed as: MEDITATE = 3.853 + 0.038 × BOOKTIME. Knowing that our regression model is significant, we can now enter specific values of BOOKTIME in this equation to make predictions about the expected value of MEDITATE. For instance, the intention to start meditation classes of someone who has spent 40 hours reading the book is 5.373 (= 3.853 + 0.038 × 40). An exceptionally beautiful visual representation of our model can be found in Figure 12.5, along with a few annotations that will, hopefully, help things fall into place.

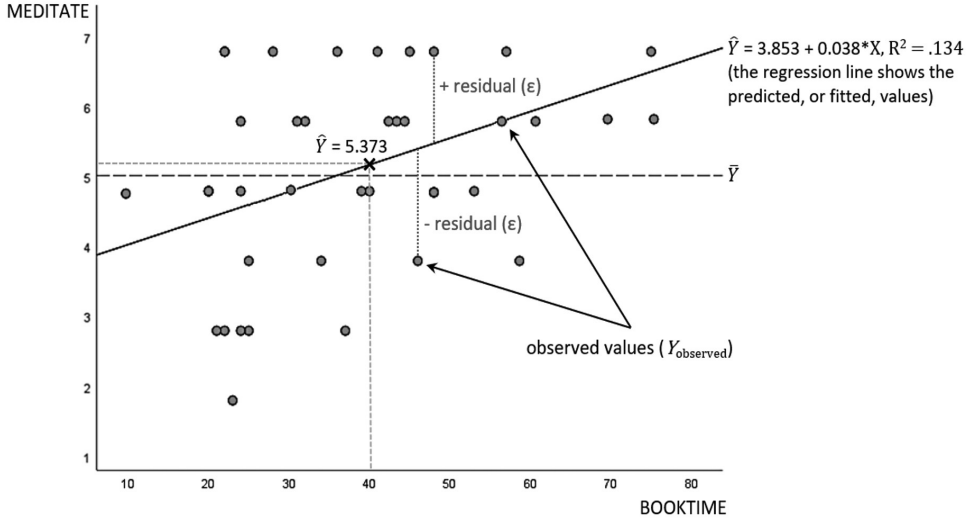


Figure 12.5 Intention of taking up meditation as a function of the amount of time spent reading this (wonderful) book

It should be noted that a regression equation that describes the observed values (i.e., it fits the data) does not necessarily mean that it can be generalized outside the sample. The validity and generalizability of a regression model depend on the extent to which specific underlying assumptions have been met. First, the dependent variable that we are trying to predict should be either interval or ratio. Second, the predictor and the dependent variable should be related in a linear way (this can be easily detected through a scatterplot). Third, there should be **independence of errors**; that is, the residuals across cases should not be intercorrelated (there is a test for this, as we will see in the next chapter). Fourth, the residuals should also be normally distributed (we can save the residuals in SPSS and examine their distribution using techniques described in Chapter 10). Fifth, we need to conform to the assumption of **homoscedasticity**

(try to perfect your pronunciation of 'homoscedasticity' next time you are having a long warm bath). This is similar to the homogeneity of variance assumption discussed in ANOVA and implies that the variance of the residuals should be relatively similar along with the different values of the predictor (i.e., that the regression model is equally predictive across all levels of the predictor). Again, SPSS gives you the option to save the residuals and plot them against the predicted values to investigate if this is indeed the case. Finally, you should be aware that significant **outlier values** (e.g., very extreme cases) may shift the regression line, practically invalidating what is considered to be the best-fitting line. Thus, cases associated with very high standardized residuals (typically close to or bigger than 2.58) warrant further consideration and even possible exclusion from the sample in order to get meaningful regression results.

We cannot possibly stress enough how wonderful regression analysis is. Believe us, it is truly amazing; like being free of unwanted chest hair after a visit to a waxing studio! Some PhD students even spend their summer holidays in research camps to learn more about the wonders of regression analysis (and even prefer this to visiting waxing studios). Jokes aside, regression is one of the most versatile and elegant analytical techniques. It has numerous capabilities and the basic model can be extended to include a large number of predictors (either simultaneously, individually, or in blocks), it can accommodate binary and multi-categorical predictions via different variable codings, it can incorporate interactions between two or more predictors, it can be specified to include non-linear effects of predictors (e.g., quadratic), and it can even deal with highly skewed and non-normal data via data transformations (e.g., logarithmic transformations). There are alternative forms of regression analysis that allow us to predict non-continuous, categorical variables. One such technique is discussed in the following section.

LOGISTIC REGRESSION

In many cases, what you want to predict might not be an interval or ratio variable but, instead, a categorical variable. In such situations, the simple linear regression model will not be valid, as the relationship between the predictor and the outcome can hardly be described in a linear way. In addition, a number of necessary assumptions of linear regression (e.g., homoscedasticity and normality, or residuals) are inevitably violated. The most common is **logistic regression**. Essentially, with logistic regression what you are trying to predict is the *likelihood* that the categorical variable will assume a given value as opposed to another one. For instance, if you want to test whether starting a diet can be predicted by people's weight, or if you want to use IQ scores to predict the likelihood that someone would go swimming wearing a bikini, a tuxedo, or a ninja costume, this is the method you should perform. Logistic regression can be used to tackle categorical dependent variables with two or more levels. Given that the main logic and interpretation are practically the same, we will go easy on you and describe its basic form, which involves a *binary* outcome variable with just two levels. This is known as **binary logistic regression**.

Suppose that we want to find out if the perceived popularity of a lecturer (ATTRACT, measured on a 10-point interval scale anchored by 1 = 'very low' and 10 = 'very high') predicts

whether students sign up for a course on Emotional Intelligence in Mechanical Engineering (COURSE, coded as 0 = 'no' and 1 = 'yes'). Given that the outcome variable is of a categorical, dichotomous nature, it would be incorrect to rely on the linear regression equation to answer this question. Logistic regression circumvents this problem by applying a series of transformations to the outcome variable so as to represent the non-linear relationship in a linear way (clever, right?). Specifically, it uses the **logistic function** to express the outcome in terms of probabilities (e.g., *P*(yes), probability that a student takes the course), then calculates the *odds* of the two possible outcomes (e.g., Odds = *P*(yes)/1−*P*(yes), the ratio of the probability of taking by not taking the course), and finally takes the natural logarithm of the odds ratio – that is, the **logit** – to arrive at the following equation:

$$\ln\left(\frac{P(yes)}{1-P(\text{yes})}\right)=\alpha+\beta X$$

The logit is a continuous variable that is now a linear function of the predictor X and satisfies the necessary criteria of linear regression. Importantly, the logit can easily be converted back to odds and from odds to probability, thus allowing us to make predictions about the likelihood of a certain outcome. Going back to our example, to test if perceived lecturer popularity predicts whether students take the course, we conducted a survey on a random sample of 120 college students in Fargo, North Dakota. (Note that Fargo is good for sampling because filling in questionnaires provides a welcome diversion from everyday life.) The logistic regression results obtained from SPSS are summarized in Table 12.5.

Table 12.5a An example of (binary) logistic regression, block 0: classification table[a,b]

	Observed		Predicted		
			COURSE		Percentage Correct
			No	Yes	
Step 0	COURSE	No	76	0	100.0
		Yes	44	0	.0
	Overall Percentage				63.3

Notes:

[a] Constant is included in the model

[b] The cut value is .500

Table 12.5b An example of (binary) logistic regression, block 0: variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	−.547	.189	8.324	1	.004	.579

Table 12.5c An example of (binary) logistic regression, block 0: variables not in the equation

			Score	df	Sig.
Step 0	Variables	ATTRACT	41.567	1	.000
	Overall Statistics		41.567	1	.000

Table 12.5d An example of (binary) logistic regression, block 1: omnibus tests of model coefficients

		Chi-square	df	Sig.
Step 1	Step	57.672	1	.000
	Block	57.672	1	.000
	Model	57.672	1	.000

Table 12.5e An example of (binary) logistic regression, block 1: model summary

Step	-2 Log Likelihood	Cox & Snell R Square	Nagelkerke R Square
1	100.046[a]	.382	.522

Note: [a] Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Table 12.5f An example of (binary) logistic regression, block 1: classification table[a]

Observed			Predicted		
			COURSE		Percentage Correct
			No	Yes	
Step 1	COURSE	No	68	8	89.5
		Yes	22	22	50.0
	Overall Percentage				75.0

Note: [a] The cut value is .500

Table 12.5g An example of (binary) logistic regression, block 1: variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1[a]	ATTRACT	1.396	.264	27.989	1	.000	4.040	2.408	6.776
	Constant	−11.276	2.101	28.810	1	.000	.000		

Note: [a] Variable(s) entered on step 1: ATTRACT

Tables 12.5a–c present the results of the baseline (or 'null') model, which is what happens if no predictor variable is included in the analysis. This information is necessary to subsequently calculate the fit of the logistic model (which is based on how much our prediction is improved by including the predictor). Table 12.5a shows that, according to the observed responses, only a minority take the course (44 out of 120), while most of the students do not take the course (76 out of 120). Thus, if we had to make a guess *without* any other information at our disposal, we would say that students are generally not likely to take the course and we would actually be correct 63.3% (i.e., 76/120) of the time. This is not a bad hit ratio in terms of classification accuracy. In fact, Table 12.5b tells us that this percentage is significantly better than simply guessing by chance (i.e., 50:50 chance). This is formally shown by the Wald statistic, which tests the null hypothesis that the intercept ('constant') in the baseline model ($B = -0.547$) is significantly different from zero (p=0.004). By default, SPSS also provides the Exp (B) (or e^{β}), which is readily interpretable and shows the odds ratio between students who take the course and those who do not take the course. In our example, students are 0.579 (i.e., 44/76) times less likely to take as opposed to not take the course. Finally, Table 12.5c gives us a heads-up by telling us whether the inclusion of predictor is *expected* to make a significant contribution to the model. This is particularly interesting when multiple predictors are included in the model and we want to explore the potential contribution of individual predictors (more on that in Chapter 13).

Moving on to Tables 12.5d–g, this starts with testing the fit of the full logistic regression model (i.e., including the predictor). Although a different estimation method is used here (i.e., **maximum likelihood estimation**), the rationale is similar to that in ordinary linear regression. Model fit is assessed by comparing the unexplained information associated with the baseline and the full model. If the unexplained information of the full model is smaller than the unexplained information of the baseline model, then their difference should be high, indicating an improvement in the prediction. This difference follows a chi-square (χ^2) distribution, with degrees of freedom equal to the number of predictors in the full model minus one, and can be used to make inferences about statistical significance. In our example, model fit improvement is significant with a chi-square of 57.672 and a p-value below 0.1% (Table 12.5d). Thus, the perceived popularity of the lecturer can significantly (better) predict whether a student will take the course Emotional Intelligence in Mechanical Engineering.

Table 12.5e offers two alternative pseudo-R^2 measures (i.e., Cox & Snell R^2 and Nagelkerke R^2) that are analogous to the coefficient of determination earlier discussed in the context of linear regression. Having said that, the interpretation of these measures is not really the same as the R^2 obtained from regression and should be used as measures of effect size when comparing competing models for the same data (with higher values indicating better model performance).

Table 12.5f indicates how well the independent variable (ATTRACT) predicts the outcome (COURSE). Essentially, it shows how many cases are correctly predicted by the full logistic model. Here, we see that the model would correctly predict 68 (out of 76) students who do not take the course and 22 (out of 44) students who take the course. This corresponds to a classification accuracy of 89.5% for the former and 50% for the latter. Thus, it seems that the model performs better in predicting those who do not take the course as opposed to those who do.

Overall classification accuracy is 75%, which is indeed an improvement compared to the 63.3% of the null model (and, as mentioned above, this improvement is statistically significant).

Table 12.5g provides all the details about the effect of the predictor. The estimated coefficient (B) shows the directionality, strength, and statistical significance of this effect. The corresponding Wald's χ^2 statistic (27.989) tests the null hypothesis that B equals zero. In our case, this is not so, revealing a positive and statistically significant relationship ($p<0.001$) between ATTRACT and COURSE. The coefficient B shows that a one-unit increase in the perceived popularity of the lecturer results in a 1.396-unit increase in the log-odds (i.e., logit) of the outcome variable. This doesn't really make much sense, does it? Well, that is why we use the exponentiated beta (i.e., Exp (B)), which shows the relevant change in the odds associated with taking versus not taking the course. Here, we can conclude that a one-unit increase in the perceived lecturer's popularity makes it approximately four times more likely that students will take the course. The following general rule applies: if $B > 0$, then Exp(B) > 1 and the predictor *increases* the odds of an event occurring by (Exp(B)−1) × 100%. If $B < 0$, then Exp(B) < 1 and the predictor *decreases* the odds of an event occurring by (1−Exp(B)) × 100%. In order to make specific predictions about the likelihood of taking the course for a given level of the lecturer's popularity, we can use the equation of the logit presented above and exponentiate the outcome to derive the odds. For example, given a popularity score of 8, students are 10% less likely to take the course (i.e., $e^{-11.276+1.396*8} = 0.90$). To communicate this in terms of probabilities, we would need to plug the values into the equation of the logistic function (i.e., $f(x) = 1/(1+e^{-(a+bx)})$). This would show that if a lecturer's popularity score is 8, there is approximately a 47.3% chance that students will take the course.

Concluding, let us point out a few things about binary logistic regression. First, the outcome variable should be dichotomous with mutually exclusive categories coded with 0 and 1 values. Second, there has to be a linear relationship between the predictor and the logit transformation of the dependent variable. Third, as mentioned at the beginning of this section, logistic regression can be extended to dependent variables with multiple categories (**multinomial logistic regression**). While the process is exactly the same as in the binary case, the analysis uses one level of the outcome variable as the reference category and performs one-by-one (binary) comparisons with the remaining levels. Therefore, if a categorical outcome has three levels (coded 1, 2, and 3) and the reference category is set to be the first one, then the output will produce two tables of the parameters estimates; one comparing Level 1 versus 2 and another one comparing Level 1 versus 3. Of course, you are free to specify any level to be the reference category.

A final remark about any type of regression analysis. Note that the regression coefficient of a **binary predictor** represents the difference between the category coded as 0 and the category coded as 1. Remember that regression coefficients show the average change in the outcome variable as a result of one-unit change in the predictor. Given that the predictor now takes the values 0 and 1, a one-unit increase simply reflects the change associated with going from Category 0 to Category 1. Using this so-called dummy coding, we can also recode a multi-categorical variable with k categories into k−1 dummy variables (e.g., Level 1 = no/yes, Level 2 = no/yes, and so on), which we can then include in the regression model to address all possible category combinations. This is a good approach to follow when one has a nominal variable with several categories and wants to include it as a predictor in a regression model.

WARNING 12.7 Correlation results on their own can never prove causality. Interpreting correlation results as causal relationships is erroneous and misleading.

CORRELATION AND CAUSALITY

Whenever you are examining a relationship between two variables, there is always a temptation to draw **causal inferences** on the basis of correlational results. This temptation must be firmly resisted (as should all temptations – well, most of them!). For example, think about the link between sales and advertising. Most people would naturally assume that sales are caused by advertising. Sounds perfectly reasonable until you find out that many companies determine their advertising budget based on the previous year's sales figures – oops!

The fact that two variables are related, the fact that this relationship can be captured by an association measure, and the fact that this association measure may generate a statistically significant result is no evidence whatsoever that one variable actually causes the other. Especially in the context of regression analysis, no matter how 'intuitively appealing' a cause-and-effect explanation may be and no matter how 'obvious' the designation of the independent (predictor) and dependent (outcome) variable as cause and effect, the fact remains: *correlation does not prove causality*. All that is expressed by an association measure is the degree of covariation between two variables. Any notion of causality must come from practical knowledge or theoretical insights into the subject area, preferably supported by longitudinal data obtained under experimental conditions. (The need for longitudinal data becomes evident when you think that a cause *must* temporally precede an effect, while experimental conditions reflect a need to control for other variables so as to 'isolate' the effect from the influence of unwanted causes.) While a full-blown discourse on the conditions necessary for drawing causal inferences would take us into the realm of the philosophy of science, you should think twice before you present correlational measures among variables as reflecting causal processes.

Lastly, we should point out that there are several other specialized measures of association for gauging the strength of relationships between variables measured at different levels of measurement (such as 'polychoric' and 'polyserial' correlations); however, for most data analysis projects, the techniques described in this chapter and Chapter 13 are quite sufficient for dealing with any combination of variables. If you are still skeptical, here's a challenge for you: select any two variables of your choice, have them measured any way you like, and we *bet* you that somewhere in Chapter 11 or 12 (this one!) you will find a method for analyzing the linkage between them.

SUMMARY

In this chapter, we focused on uncovering and assessing relationships between two variables. We first distinguished between the existence, direction, and strength of a relationship and indicated the importance of accompanying measures of association with a significance test. We then considered several measures of association, with the level

of measurement being, again, a key factor determining the correct choice of technique. Moving on, we discussed key techniques that allow us to make specific predictions about a variable based on its association with another variable. Finally, we warned against interpreting correlation results in causal terms. We hope that you are now sufficiently physically and mentally prepared to face the challenges of multivariate analysis – if not, get into shape!

QUESTIONS AND PROBLEMS

1. What is the difference between the strength and the direction of a relationship?
2. Can the chi-square statistic be used to measure the strength of a relationship between two nominal variables?
3. Which values can be assumed by Cramer's V and how should these numbers be interpreted?
4. What is the null hypothesis under the Pearson correlation coefficient? What would a coefficient of 0.9 indicate?
5. Does a Pearson correlation coefficient close to zero *necessarily* imply no association between the two variables under consideration?
6. Under what circumstances would you use a one-tailed test to test the significance of a correlation coefficient?
7. Why can we not interpret significant correlation results as indicating causal relationships?
8. What do a standardized and an unstandardized regression coefficient indicate?
9. Could you explain simply what a logit is?
10. Why is competitive Brazilian waxing still not an Olympic discipline?

FURTHER READING

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates. This is probably *the* standard text on regression and correlation.

Field, A. (2017). *Discovering Statistics using IBM SPSS Statistics*, 5th edition. London: Sage. Covers everything we discussed at a lower level of silliness.

Pampel, F. C. (2021). *Logistic Regression: A Primer*, 2nd edition. London: Sage. More than most of you ever wanted to know about logistic regression.

13
Getting hooked: A look into multivariate analysis

THE NATURE OF MULTIVARIATE ANALYSIS

This and the next chapter are devoted to the connoisseurs of data analysis: those of you who have truly lost your fear of the subject, those of you who dream of becoming virtuosos of number-crunching, those of you who want to elevate the art of data analysis to new and dizzy heights, and, finally, those of you who would not even shy away from marrying a statistician! In short, it is written for the 2% of our readership who are interested in learning something about the techniques available for analyzing several variables *simultaneously*; that is, about **multivariate analysis** (yes, we do, of course, realize that the sudden mention of the term 'multivariate analysis' can result in nausea, stomach colic, and uncontrollable eye-twitching among uninitiated novices).

So, what is multivariate analysis? Well, let us start by saying that none of the statistical techniques we have discussed so far would qualify for the 'multivariate' title. Up until now, the techniques we have covered were concerned with either the analysis of one variable at a time (for example, a one-sample *t*-test) or the link between two variables at a time (for example, correlation analysis). In other words, we have been discussing techniques suitable for undertaking univariate and bivariate analysis but, alas, not for multivariate analysis. Multivariate analysis deals with multiple variables concurrently; for example, it enables the comparison of several groups in terms of several variables and the investigation of interrelationships among sets of variables. Multivariate analysis techniques are, in many instances, clear extensions of univariate and bivariate techniques. For example, **multiple regression** extends simple regression (see Chapter 12) by considering several predictors at the same time. Similarly, **multivariate analysis of variance (MANOVA)** enables the comparison of several groups in terms of multiple dependent variables; thus, it can be conceived as an extension of the analysis of variance (ANOVA) procedure discussed in Chapter 11 (in which three or more groups were compared on a single interval-scale variable). Indeed, the null hypotheses of the two techniques are quite similar: whereas in one-way ANOVA the null hypothesis is that there is no difference in the group means for the variable of interest, in MANOVA the null hypothesis is that there is no difference in the *sets* of means across the groups (since several variables are simultaneously compared). The latter is known as a **multivariate hypothesis** as its rejection or non-rejection refers to the set of variables *as a whole* rather than to any of the individual variables. To test

multivariate hypotheses, **multivariate significance tests** are employed, several of which, you will be thrilled to know, we will discuss in this and the next chapter.

HINT 13.1	Bear in mind that a multivariate hypothesis always refers to the rejection or non-rejection of a set of variables as a whole rather than to any of the individual variables.

But what is the benefit of undertaking multivariate analysis (other than showing off during dinner parties)? Given that multivariate analysis techniques are substantially more complex than their univariate and bivariate counterparts (as well as more computationally demanding), why not just repeat, say, a univariate analysis a number of times until all variables of interest have been considered? Why should we look at all the variables simultaneously? The answer to these questions is best given with an example.

Assume you have a *real* job, which is clearly an uncomfortable thought (we were lucky enough to avoid having real jobs most of our lives). To make things worse, further assume that you are a business analyst and you want to find out the key factors that influence the sales revenue of your firm (a manufacturer of hand-knitted horse blankets); the motivation behind your analysis is to generate some information that you can use to decide how best to allocate your marketing budget to advertising, in-store displays, personal selling, and so on. Say also that, on the basis of historical data in company records, you have found that the (Pearson) correlation between sales revenue and the amount spent on advertising comes to 0.48, and the correlation between revenue and number of sales personnel comes to 0.56. What can you conclude from this analysis, other than that both marketing variables appear to have a positive influence on revenue? Can you identify the *relative* importance of each marketing variable on sales revenue? Can you say something about the *joint* (i.e., total or combined) impact of advertising and personal selling on revenue?

Unfortunately, the answer to both questions is 'no'. Simple (i.e., bivariate) correlations do not permit us to deal with questions of this sort. While you may vehemently protest at this stage, since it is obvious that the number of sales personnel has a stronger correlation with revenue than does advertising (0.56 vs. 0.48), the inference that the influence of the sales personnel is more important is only correct *if* there is no correlation between it and the amount spent on advertising (i.e., that the two marketing variables are totally independent). In reality, however, this is highly unlikely to be the case (e.g., the decision to use more/fewer salespeople is often accompanied by a reduction/increase in the advertising budget). Unless we somehow take into account this interrelationship, we cannot isolate the 'true' impact of either marketing variable on sales revenue. For the same reason, we cannot compute the *combined* influence of advertising and personal selling on sales revenue from our bivariate results. It is not possible to simply add the r^2 values associated with advertising and salespeople together because if these variables are inter-correlated, each r^2 does not reflect only the relationship between the variable under consideration and sales revenue; part of each r^2 could be attributed to the other variable. Consequently, the proportion of variance in sales revenue that is jointly explained by advertising and personal selling is less than what would be obtained if the two r^2 values were simply added up (have a look at Figure 13.1, once again a beautiful piece of art, and you will

see immediately what the problem is). In short, unless you use a procedure that investigates the relationship between sales revenue and the two marketing variables while considering the interrelationship among the latter (i.e., unless we analyze all three variables simultaneously), you are stuck. Your job as a business analyst may come to an abrupt halt (and so will your income)! Similar difficulties would occur if you wanted to explore differences in, say, sales revenue and profit across three company branches. The interrelationship between the two dependent variables would somehow need to be accounted for to produce valid results.

WARNING 13.1 Do not attempt to compute the combined influence of two or more bivariate Pearson correlations (*r*) by simply adding them, as the underlying variables are usually correlated with each other.

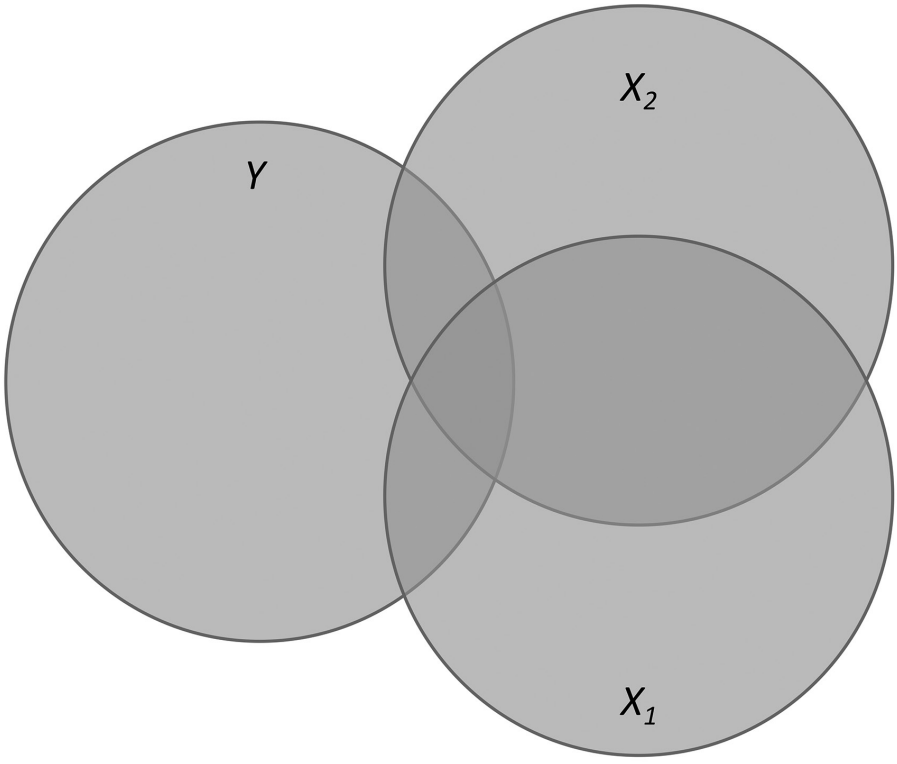


Figure 13.1 Relationships among three variables. Y = sales revenue; X_1 = advertising; X_2 = number of salespeople.

It is for such reasons – and to secure your livelihood as a business analyst – that you need to opt for multivariate techniques. Whenever you are facing a situation where multiple variables are involved, and the potential interrelationships among these variables need to be taken into

account in order to be able to answer our research questions (i.e., variables are interrelated in such a way that their effects cannot be meaningfully interpreted separately), multivariate analysis procedures are your only option. Moreover, from a practical perspective, many problems are multivariate in nature (e.g., sales performance may be a function of several potentially interrelated variables; product choice may be determined by sets of interdependent attributes; market segments may differ in terms of several interlinked characteristics; etc.). Consequently, researchers who have an appreciation of multivariate techniques will obtain a more realistic understanding of complex problems. They will also be able to show off more at dinner parties.

HINT 13.2 Most complex (read: real-life) problems are multivariate in nature. An understanding of multivariate analysis, although technically demanding, is likely to pay off in terms of a better understanding of complex problems.

TYPES OF MULTIVARIATE TECHNIQUES

Multivariate analysis has largely become popular through the widespread use of statistical computer packages (it is now probably as popular as the Xbox or PlayStation). Mystical calculations that would literally have taken months to perform by hand (a task only willingly undertaken by the sort of people who leave parties early in order to have fun with a random numbers generator) can now be managed with a few simple computer commands producing near-instantaneous results. The techniques outlined below are available on practically all major commercial statistical packages; specialized programs that focus only on one multivariate technique also exist, but you must be really in love with the technique (or a compulsive collector of statistical software) to justify buying them.

Multivariate techniques can be broadly divided into two main groups, namely **dependence methods** and **interdependence methods**. All dependence methods are characterized by a distinction between **dependent variables** and **independent variables**. Dependent variables are those that are predicted or explained by (yes, you've guessed it) the independent variables. In the first example mentioned earlier, sales revenue was the dependent variable, and the number of salespeople and advertising were the independent variables. In the second example, sales revenue and profit were the dependent variables, and the three company branches were different levels of a (categorical) independent variable. Interdependence methods do not make a distinction between dependent and independent variables; such techniques are frequently employed when it is still unclear which kind of relationships and structures exist in a given data set (we are still debating whether the techniques listed in the *Kama Sutra* should be treated as dependence or interdependence techniques – any ideas?).

To identify which multivariate technique is appropriate in a given situation, it is important that you ask the right questions. First, you should ask whether your data permit you to distinguish between dependent and independent variables. If the answer is yes, you will need a dependence method (see exquisite Figure 13.2); if the answer is no, you will need an interdependence method (see equally exquisite Figure 13.3). Once again, a reminder that, for a minor fee, all our figures can be delivered to the statistical connoisseurs on brass wall plaques

embedded in mahogany frames. These plaques can be tastefully exhibited in your office or may serve as gifts for birthdays, wedding or divorce anniversaries, or important religious holidays.

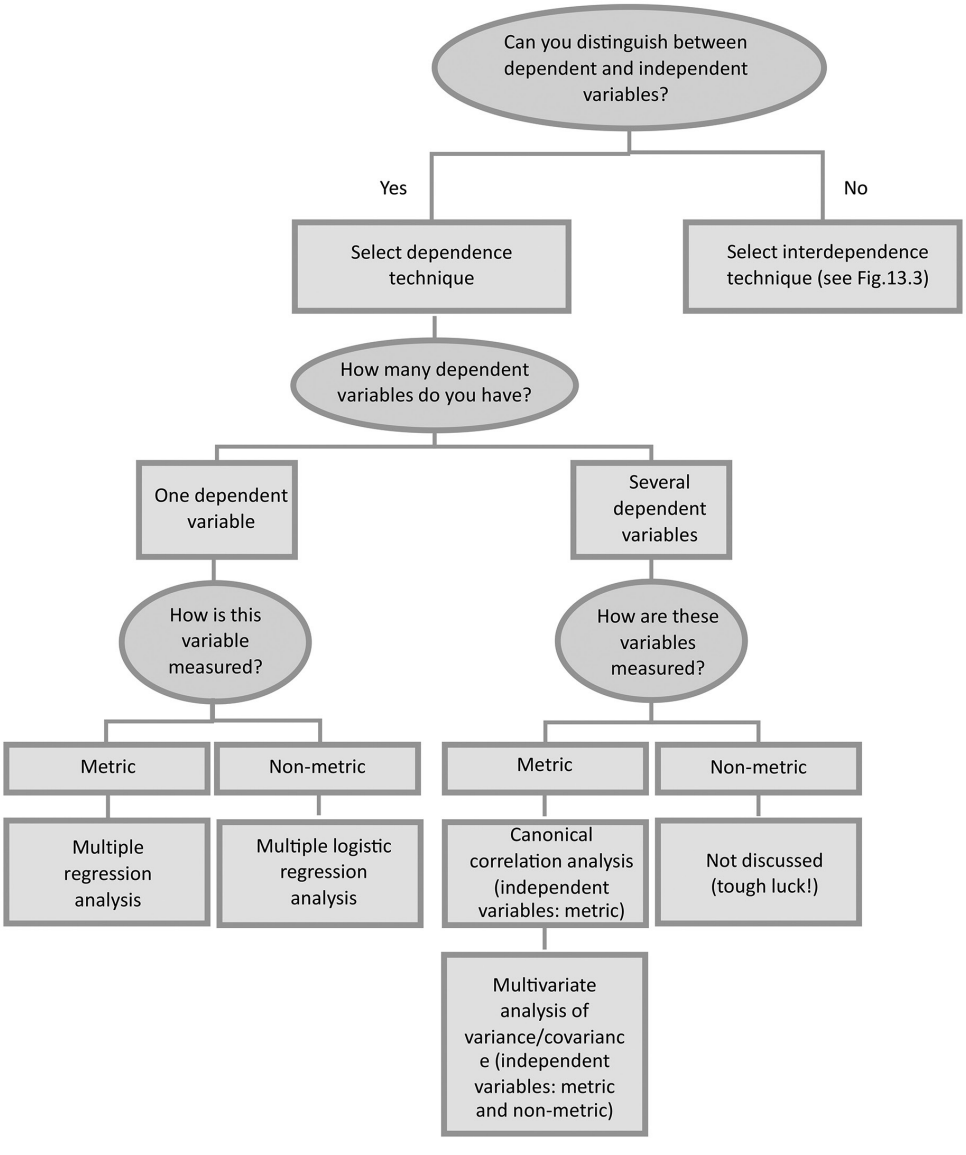


Figure 13.2 Selecting multivariate techniques: dependence methods

HINT 13.3	When undertaking multivariate analysis, do not confuse dependence methods and interdependence methods.

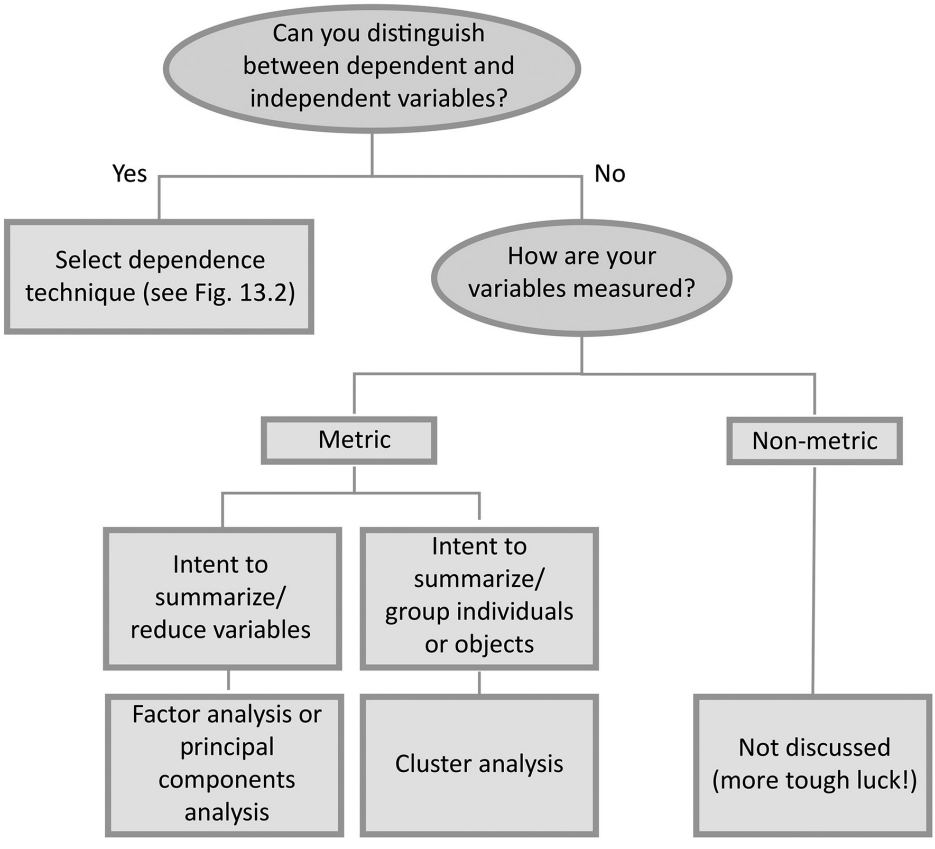


Figure 13.3 Selecting multivariate techniques: interdependence methods

Focusing on dependence methods, the first issue you need to clarify is the *number* and *measurement level* of the dependent variables. If, for instance, you are dealing with just one dependent variable and this one variable is measured on a metric scale, you most likely require a **multiple regression analysis**. If, in contrast, your one dependent variable is non-metric, you should have a look at **multiple logistic regression**. For both kinds of analysis, your multiple independent variables should be metric (inclusive of dichotomous variables with dummy variable scoring – see Chapter 3).

HINT 13.4	Use factor analysis/principal component analysis if you want to summarize/ reduce variables into a smaller number of dimensions. Use cluster analysis if you want to summarize/organize units (i.e., individuals or objects) into groups.

Turning to situations in which several metric dependent variables are involved, you will need either a **multivariate analysis of variance** or **canonical correlation analysis**. For the former, your independent variables can be non-metric (categorical), while for the latter, your independent variables must be metric.

Compared to selecting an appropriate dependence technique (and compared to assembling some of the horrible flat-packed furniture you can buy these days), choosing a suitable interdependence technique is relatively easy. Essentially, you only need to answer one question: do you want to summarize/reduce your *variables* into a smaller number of dimensions, or do you want to summarize/organize your *units* (i.e., individuals or objects) into groups? Depending upon your answer to this question, you have a choice between **factor analysis/principal component analysis** and **cluster analysis**.

In the next section, we provide a brief description of the dependence methods shown in Figure 13.2; interdependence methods (see Figure 13.3) will be the subject of Chapter 14. Note that, given the technical complexity of multivariate techniques, you should not expect to become an expert on them just by reading a couple of chapters in this book. While you will gain a good understanding of the circumstances in which these techniques might be useful, there are no shortcuts here: you *must* also consult specialized textbooks (such as the ones recommended in the Further Reading section) if you want to become a master of multivariate analysis. A romantic relationship with a statistician might also help.

DEPENDENCE METHODS I: MAKING (MORE COMPLEX) COMPARISONS

Analysis of covariance (ANCOVA)

In many cases, we are interested in testing for differences across group means, knowing that one or more additional variables influence the outcome. For example, you may want to investigate if consumers' purchase intentions vary across promotional strategies (e.g., no promotion (the control condition), buy one get one free, and get 50% off). However, you have reasons to believe that individual differences in price sensitivity (which, in your infinite wisdom, you have also measured in your study) will affect any potential variations. To account for such influences, you need to run an **analysis of covariance (ANCOVA)**. This method will enable you to compare the means across the three groups (i.e., the three promotional strategies) while adjusting for any impact of price sensitivity (known as the **covariate**).

WARNING 13.2 Do not underestimate the complexity of multivariate techniques. While they can be very powerful analytical tools in the right hands, they can also ruin an analysis in the wrong hands. Treat with respect.

ANCOVA is very similar to the one-way ANOVA for independent groups discussed in Chapter 11. The key difference is that it calculates **adjusted group means** based on the potential association between the independent variable and covariate as well as the association

between the dependent variable and the covariate. Thus, in our example, purchase intentions would be adjusted according to the relationship between price sensitivity and purchase intentions as well as the differences in price sensitivity across the three promotional strategies (i.e., the association between price sensitivity and promotional strategy).

Table 13.1 shows what the ANCOVA output for our example looks like. In this example, 125 consumers were randomly assigned into one out of three groups, each corresponding to a different promotional strategy. Purchase intentions and price sensitivity were both measured with seven-point interval scales anchored by 1 = low and 7 = high.

HINT 13.5 ANCOVA calculates adjusted group means based on the potential association between the independent variable and covariate as well as the association between the dependent variable and the covariate.

Table 13.1a An example of analysis of covariance (ANCOVA): descriptive statistics

Dependent Variable: PURCHASE			
Promotional strategy	Mean	Std. Deviation	N
Control	3.22	1.636	41
1+1	4.88	1.435	42
50% off	5.00	1.361	42
Total	4.38	1.678	125

Table 13.1b An example of analysis of covariance (ANCOVA): Levene's test of equality of error variances[a]

Dependent Variable: PURCHASE			
F	df1	df2	Sig.
26.164	2	122	.000
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.			

Note: [a] Design: Intercept + PR_SENS + PROMO

Table 13.1c An example of analysis of covariance (ANCOVA): tests of between-subjects effects

Dependent Variable: PURCHASE					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	205.621[a]	3	68.540	57.710	.000
Intercept	206.784	1	206.784	174.110	.000
PR_SENS	123.722	1	123.722	104.173	.000
PROMO	128.877	2	64.438	54.257	.000
Error	143.707	121	1.188		

Dependent Variable: PURCHASE					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Total	2,743.000	125			
Corrected Total	349.328	124			

Note: [a] R Squared = .589 (Adjusted R Squared = .578)

Table 13.1d An example of analysis of covariance (ANCOVA): estimated marginal means, estimates

Dependent Variable: PURCHASE				
Promotional strategy	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Control	2.916[a]	.173	2.574	3.258
1+1	4.810[a]	.168	4.477	5.143
50% off	5.368[a]	.172	5.027	5.708

Note: [a] Covariates appearing in the model are evaluated at the following values: Price sensitivity = 2.95.

Table 13.1e An example of analysis of covariance (ANCOVA): estimated marginal means, pairwise comparisons

Dependent Variable: PURCHASE						
(I) Promotional strategy	(J) Promotional strategy	Mean Difference (I–J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Control	1+1	−1.894*	.240	.000	−2.478	−1.311
	50% off	−2.452*	.248	.000	−3.054	−1.850
1+1	Control	1.894*	.240	.000	1.311	2.478
	50% off	−.558	.242	.068	−1.144	.029
50% off	Control	2.452*	.248	.000	1.850	3.054
	1+1	.558	.242	.068	−.029	1.144

Notes: Based on estimated marginal means

* The mean difference is significant at the .05 level.

[a] Adjustment for multiple comparisons: Bonferroni.

The essential difference from the output that we would have obtained from a simple ANOVA is that we now also get (a) information regarding the influence of the covariate and (b) the adjusted group means accounting for this influence. Table 13.1a shows the means and standard deviations corresponding to the three promotional strategies (without taking the covariate into consideration). Next, in Table 13.1b, Levene's test for the homogeneity of variances is given. In this case, results are significant, indicating that the variances across groups are not

equal. Be that as it may, remember from Chapter 11 that Levene's test does not really matter when group sizes are equal and a relatively large sample is involved.

Table 13.1c provides the overall effect of the ANCOVA model. In particular, we observe that the covariate price sensitivity (PR_SENSE) has a significant influence on consumers' purchase intentions ($F(1,121) = 104.17$, $p<0.001$) and so does the type of promotional strategy ($F(2, 121) = 54.257$, $p<0.001$). The latter indicates that after controlling for the effect of price sensitivity, consumers' purchase intentions vary significantly across the three promotional strategies. But how do they differ? Well, have a cup of coffee and something to eat before you continue because we do not want to over-exhaust you!

Having replenished your energy, you can explore how the three promotional strategies differ by looking at Tables 13.1d and 13.1e. These present the post-hoc pairwise comparisons (see Chapter 11), which are (unlike in ANOVA) now based on the *adjusted* group means. You can see that, compared to the unadjusted values (see Table 13.1a), the adjusted means for the first two groups ('control' and '1+1') are slightly lower, while the mean of the third group ('50% off') is noticeably higher. The footnote at the bottom of Table 13.1d indicates that the group means are adjusted based on the mean level of price sensitivity, which is 2.95. Also, it shows that the familywise error rate is accounted for by applying a Bonferroni correction to the significance level (see Chapter 11). In Table 13.1e you can see that making no promotional offer ('control') is associated with significantly lower purchase intentions compared to both promotional strategies. Although the 50% off promotion has the highest mean in terms of purchase intentions, we also see that this is not significantly different from the 1+1 (buy one get one free) promotion ($p=0.068$). Note that in this study, we did not specify the expected directionality among group means (i.e., which groups are expected to be higher/lower than the others). Thus, we rely on a two-tailed significance test. Had we expected that the 50% off promotion would be the most effective strategy, we would look into the one-tailed p-value (which is half that for the two-tailed value) and conclude that purchase intentions in response to a 50% off promotion are significantly higher ($p=0.034$).

A number of things are worth mentioning at this point. First, all assumptions mentioned in Chapter 11 for the simple ANOVA need to be met. In addition, ANCOVA makes the assumption of **homogeneity of regression slopes**. In simple words, it assumes that this relationship is pretty much similar for respondents, regardless of the group they belong to. Thus, if we regress purchase intentions (dependent variable) on price sensitivity (covariate) *within* each of the three groups, the slopes of the regression lines should be relatively similar. This assumption can be explored by creating a scatterplot between price sensitivity and purchase intention and fitting a regression line for each group individually (see Figure 13.4). In this case, we see that the slope for the 50%-off group is noticeably different from the other groups. There are also advanced analytical approaches to formally test for differences in slopes. While a discussion of these approaches is beyond the purpose of this book, in the highly unlikely event that your life is devoid of any other pleasures, we kindly direct you to the Further Reading at the end of this chapter if you really want to know more.

WARNING 13.3 Beware of the assumption of homogeneity of regression slopes in ANCOVA. To check it, create a scatterplot and fit a regression line for each group individually.

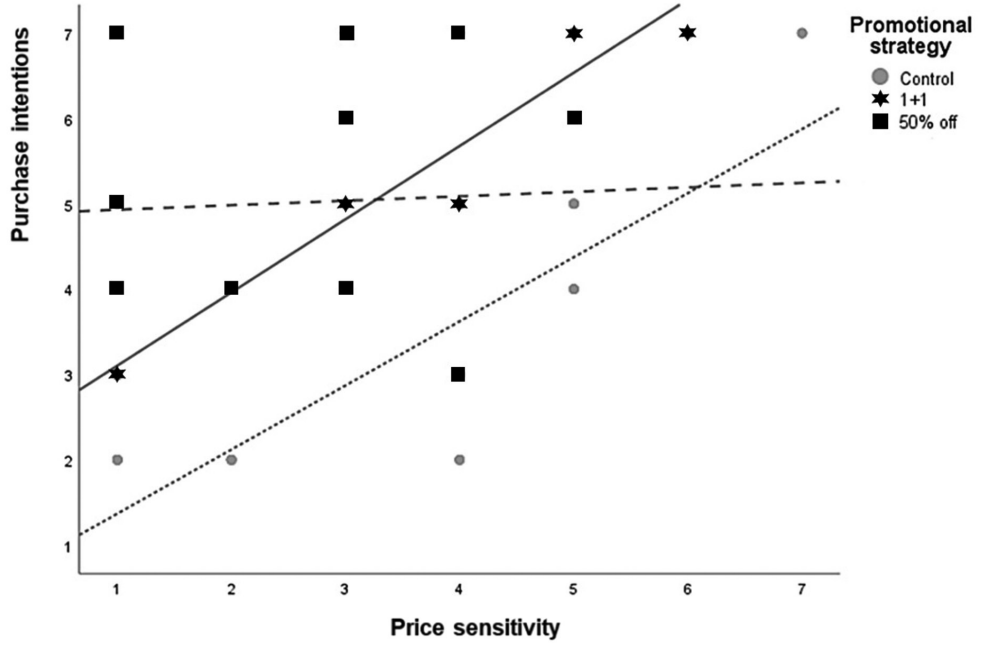


Figure 13.4 Visually inspecting the homogeneity of regression slopes

ANCOVA can simultaneously handle several variables as covariates. There is literally no limit. You can specify as many covariates as you want (as long as your sample is big enough to support the estimation procedure). Covariates are also often referred to as **control variables** because they do not represent the main focus of the analysis. They are incorporated to 'calibrate' the results relating to some other key independent and dependent variables. This does not imply that covariates are of secondary importance and, in many cases, covariates are essential to produce accurate and valid results. Indeed, it is not unusual that results with and without the covariates are considerably different, turning from non-significant to significant and vice versa.

Factorial ANOVA

Another extension of the simple one-way ANOVA that incorporates multiple variables in the analysis is **factorial ANOVA**. Factorial ANOVA simultaneously accounts for the influence of two or more categorical variables (i.e., factors) on a metric (interval or ratio) dependent variable. In doing so, it looks into the effect of each individual factor but also allows us to see

how the various factors *interact* with each other in influencing the outcome. When two or more factors interact with each other, it means that the effect of one factor on the dependent variables depends on the levels of another factor.

HINT 13.6 Bear in mind that when two or more factors interact with each other, the effect of one factor on the dependent variable(s) depends on the levels of another factor.

In order to see how factorial ANOVA works, we will consider a case in which just two factors are simultaneously investigated. It's quite alright to get (highly) excited at this point! Not all of your textbooks provide such exquisite entertainment value! The analysis is typically referred to as **two-way ANOVA**. Suppose that you are interested in identifying whether people's intention to go to the cinema (1 = not at all, 7 = absolutely) differs based on the type of movie ('Hollywood blockbuster', 'French romance', and 'independent Polish filmography') and their gender ('female' vs. 'male').

The first thing you need to do in order to understand the mechanics of factorial ANOVA is to distinguish between the main and the interaction effects. Given that your analysis includes two categorical factors (gender with two levels and movie type with three levels), you should expect two **main effects** (one for each factor) plus an **interaction effect**. A main effect for movie type implies that people's intention to go to the cinema (averaged between females and males) differs across the three movie types. Similarly, a main effect for gender implies that intention to go to the cinema (averaged across movie types) differs between females and males. In contrast, an interaction effect between movie type and gender implies that the magnitude and/or the directionality of mean differences across movie types depends on gender (or the other way around). To see how all this looks in our example, we ran a 2 (gender) × 3 (movie type) factorial ANOVA in SPSS. As the output is alarmingly long, we have split it into two parts, namely Table 13.2 and Table 13.3.

Table 13.2a An example of a two-way factorial ANOVA: main and interaction effects analysis, descriptive statistics

Dependent Variable: Intention to go to the cinema				
Gender	Movie	Mean	Std. Deviation	N
Female	Blockbuster	4.6875	.99791	16
	Romance	4.5625	1.27639	16
	Independent	2.4063	1.83683	16
	Total	3.8854	1.74197	48
Male	Blockbuster	4.0000	1.37840	16
	Romance	4.2500	1.09545	16
	Independent	3.8095	1.11216	21
	Total	4.0000	1.18484	53

Dependent Variable: Intention to go to the cinema				
Gender	Movie	Mean	Std. Deviation	N
Total	Blockbuster	4.3437	1.23417	32
	Romance	4.4063	1.18074	32
	Independent	3.2027	1.60926	37
	Total	3.9455	1.46953	101

Table 13.2b An example of a two-way factorial ANOVA: main and interaction effects analysis, Levene's test of equality of error variances[a,b]

		Levene Statistic	df1	df2	Sig.
Intention to go to the cinema	Based on mean	1.921	5	95	.098
	Based on median	.859	5	95	.511
	Based on median and with adjusted df	.859	5	60.905	.514
	Based on trimmed mean	1.716	5	95	.138
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.					

Notes:

[a] Dependent variable: Intention to go to the cinema

[b] Design: Intercept + GENDER + MOVIE + GENDER * MOVIE

Table 13.2c An example of a two-way factorial ANOVA: main and interaction effects analysis, tests of between-subjects effects

Dependent Variable: Intention to go to the cinema					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	54.728[a]	5	10.946	6.450	.000
Intercept	1,561.811	1	1,561.811	920.294	.000
GENDER	.452	1	.452	.266	.607
MOVIE	37.268	2	18.634	10.980	.000
GENDER * MOVIE	21.548	2	10.774	6.348	.003
Error	161.222	95	1.697		
Total	1,788.250	101			
Corrected Total	215.950	100			

Note:

[a] R Squared = .253 (Adjusted R Squared = .214)

Table 13.2d An example of a two-way factorial ANOVA: main and interaction effects analysis, pairwise comparisons

Dependent Variable: Intention to go to the cinema						
(I) Movie	(J) Movie	Mean Difference (I–J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Blockbuster	Romance	−.062	.326	1.000	−.856	.731
	Independent	1.236*	.316	.001	.466	2.006
Romance	Blockbuster	.062	.326	1.000	−.731	.856
	Independent	1.298*	.316	.000	.529	2.068
Independent	Blockbuster	−1.236*	.316	.001	−2.006	−.466
	Romance	−1.298*	.316	.000	−2.068	−.529
Based on estimated marginal means						

Notes:

* The mean difference is significant at the .05 level.

[a] Adjustment for multiple comparisons: Bonferroni.

In Table 13.2a, descriptive statistics for all possible combinations of movie type and gender are presented. This information is quite helpful, as it can give you an impression of whether the patterns of group means within each factor are similar or not (and thus suggest a potential interaction or not). Next, in Table 13.2b, we have (once again) Levene's test for the assumption of homogeneity of variances (note that SPSS now uses a slightly modified wording, 'Levene's test of equality of error variances', in a deliberate attempt to confuse you!). Notwithstanding the limitations associated with Levene's test (see Chapter 11), we observe that the relevant results are not significant (p=0.098), indicating that the variances across the six groups do not differ.

Table 13.2c provides the main and interaction effects of our ANOVA model. Once again, these tests are based on the F-statistic, which is calculated by partitioning the total variance in the data according to the factors of the model (see Figure 13.5 if you must).

The results reveal that there is no main effect for Gender ($F(1, 95) = 0.451$, p=0.607), which should not come as a surprise as the overall means for females and males, irrespective of movie type, are quite similar (see Table 13.2a). Thus, intentions to go to the cinema do not vary between female and male respondents. In contrast, movie type is associated with a high F-ratio (18.634) and a very low p-value (p<0.001). Therefore, we reject the null hypothesis and conclude that considering both female and male respondents, the intentions to go to the cinema are not equal across the three movie types. To see exactly how the group means vary, you can consult the corresponding pairwise comparisons (Table 13.2d). These clearly show that differences across movie types are mainly driven by 'independent Polish filmography', which appears to be significantly less appealing to respondents in general than the other two movie types. Importantly, the results also reveal a significant interaction between movie type and gender ($F(2, 95) = 10.980$, p=0.003). This suggests that the effect of movie type is not the

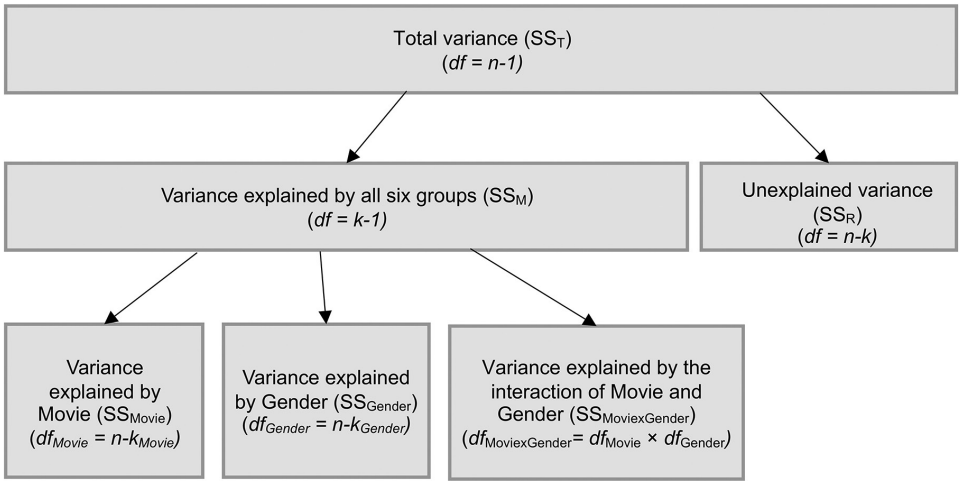


Figure 13.5 Partitioning total variance into the components of the factorial ANOVA (k = number of groups, df = degrees of freedom, n = sample size, SS = sum of squares)

same between the two gender levels (i.e., the pattern of the means across movie types is not the same for female and male respondents).

Now that we have established the presence of a **significant interaction** between the factors, we need to break down this interaction by investigating each factor at each level of the other (i.e., differences across movie types within each gender *and* differences in gender within each movie type). This is done through what is known as 'simple effect analysis', which is shown in Table 13.3.

Table 13.3a An example of a two-way factorial ANOVA: simple effect analysis, univariate tests (movie type in gender)

Dependent Variable: Intention to go to the cinema						
Gender		Sum of Squares	df	Mean Square	F	Sig.
Female	Contrast	52.635	2	26.318	15.508	.000
	Error	161.222	95	1.697		
Male	Contrast	1.762	2	.881	.519	.597
	Error	161.222	95	1.697		

Each F tests the simple effects of Movie within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

Table 13.3b An example of a two-way factorial ANOVA: simple effect analysis, pairwise comparisons (movie type in gender)

Dependent Variable: Intention to go to the cinema							
Gender	(I) Movie	(J) Movie	Mean Difference (I–J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
Female	Blockbuster	Romance	.125	.461	1.000	−.997	1.247
		Independent	2.281*	.461	.000	1.159	3.404
	Romance	Blockbuster	−.125	.461	1.000	−1.247	.997
		Independent	2.156*	.461	.000	1.034	3.279
	Independent	Blockbuster	−2.281*	.461	.000	−3.404	−1.159
		Romance	−2.156*	.461	.000	−3.279	−1.034
Male	Blockbuster	Romance	−.250	.461	1.000	−1.372	.872
		Independent	.190	.432	1.000	−.863	1.244
	Romance	Blockbuster	.250	.461	1.000	−.872	1.372
		Independent	.440	.432	.932	−.613	1.494
	Independent	Blockbuster	−.190	.432	1.000	−1.244	.863
		Romance	−.440	.432	.932	−1.494	.613
Based on estimated marginal means							

Notes:

* The mean difference is significant at the .05 level.

[a] Adjustment for multiple comparisons: Bonferroni.

Table 13.3c An example of a two-way factorial ANOVA: simple effect analysis, univariate tests (gender in movie type)

Dependent Variable: Intention to go to the cinema						
Movie		Sum of Squares	df	Mean Square	F	Sig.
Blockbuster	Contrast	3.781	1	3.781	2.228	.139
	Error	161.222	95	1.697		
Romance	Contrast	.781	1	.781	.460	.499
	Error	161.222	95	1.697		
Independent	Contrast	17.882	1	17.882	10.537	.002
	Error	161.222	95	1.697		
Each F tests the simple effects of Gender within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.						

Table 13.3d An example of a two-way factorial ANOVA: simple effect analysis, pairwise comparisons (gender in movie type)

Dependent Variable: Intention to go to the cinema							
Movie	(I) Gender	(J) Gender	Mean Difference (I–J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
Blockbuster	Female	Male	.688	.461	.139	−.227	1.602
	Male	Female	−.688	.461	.139	−1.602	.227
Romance	Female	Male	.313	.461	.499	−.602	1.227
	Male	Female	−.313	.461	.499	−1.227	.602
Independent	Female	Male	−1.403*	.432	.002	−2.261	−.545
	Male	Female	1.403*	.432	.002	.545	2.261
Based on estimated marginal means							

Notes:

* The mean difference is significant at the .05 level.

[a] Adjustment for multiple comparisons: Bonferroni.

Table 13.3 starts by testing the differences across movie types individually for female and male respondents. Here (Table 13.3a) you can see that intentions to go to the cinema vary significantly across movie types for females but not for males. The pairwise comparisons in Table 13.3b show that female respondents have significantly lower intentions to attend independent Polish filmography compared to both Hollywood blockbusters and French romance movies, while male respondents are completely indifferent and appreciate all three genres equally. A possible explanation is that the likelihood of males going into the cinema is more strongly influenced by the quality of their companions than the genre of the films. However, we cannot explore this (intuitively quite appealing) proposition with the data at hand.

Tables 13.3c and 13.3d deal with differences between female and male respondents for each movie type. None of these comparisons are significant, except for the last one; male respondents' intentions to watch independent Polish filmography are significantly higher than those of female respondents. If you look closely at the results, you will realize that the *p*-values for the pairwise comparisons are identical to those in the univariate tests.

In factorial analyses, it is always a good idea to plot the results, as this can give you an immediate picture of how the means vary across factors. Figure 13.6 graphs the effects of movie type and gender on individuals' intentions to go to the cinema.

It should be highlighted that factorial ANOVA calculates mean comparisons on the basis of the **estimated marginal means**. The marginal means for one factor are the means for that factor averaged across all levels of the other factor. For example, the marginal mean for 'independent Polish filmography' is calculated by adding the corresponding means for females and males and dividing the result by 2 (which comes to 3.1079). Therefore, do not be surprised if there are some (usually minor) discrepancies between the means reported in the descriptive

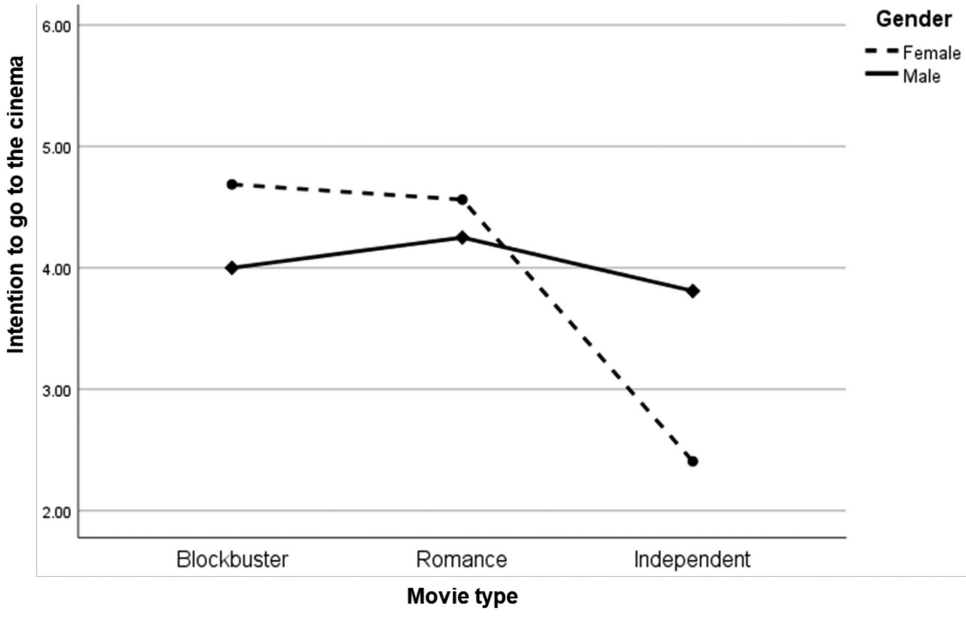


Figure 13.6 Intentions to go to the cinema as a function of movie type and gender

statistics of Table 13.2a (which are based on the overall sample size) and those based on estimated marginal means.

Factorial ANOVA can simultaneously handle more than two factors. However, it rarely makes sense to include more than three factors at the same time, as the corresponding models become very complex and the interpretation of **higher-order interactions** (e.g., 2 × 3 × 3) quite cumbersome. It is also possible to introduce (one or more) covariates in the analysis forming a **factorial ANCOVA**. The process is similar to the one described above, but the group means will also be adjusted based on the variance explained by the covariate(s). Finally, factorial ANOVA is not only limited to independent groups (or 'between-subjects') designs. The repeated-measures ANOVA described in Chapter 11 can also be extended to incorporate multiple factors in a **factorial repeated-measures AN(C)OVA**. What is more, we can combine measurements *between* independent groups of respondents with repeated measurements *within* respondents to form a **mixed factorial AN(C)OVA**. For instance, in our cinema example, we could have measured respondents' intentions to go to the cinema three times (one for each movie type), creating a repeated-measures factor, and then split respondents based on their gender, operationalizing the latter as a between-subjects factor (possibly also including relevant covariates). Those of you who wish to learn more about these more complex techniques can have a look at the Further Reading section (at your own risk, of course!).

Multivariate analysis of variance (MANOVA)

Those of you who have no fear and are really adventurous may want to simultaneously compare the means of *a set* of dependent variables that are interrelated. **Multivariate analysis of variance (MANOVA)** is the multivariate technique appropriate to do this and is an extension of the one-way ANOVA procedure we discussed in Chapter 11.

Assume, for example, that you have developed a new breath freshener for dogs (and it is high time that somebody develops more effective products than the dog breath fresheners that are currently on the market). Now, you would like to test the influence of three different product options (categorical variable) on both perceived product quality and satisfaction level (metric variables). Not surprisingly, perceptions of quality and evaluations of satisfaction are expected to be related to a certain extent. The three product options correspond to a spray, a liquid, and a tablet formulation. Perceived quality and satisfaction are assessed with seven-point interval scales with higher numbers indicating more of the property being measured. In case you were wondering, dogs (no matter how well trained they are) can hardly provide responses on our two self-report measures (perceived quality and satisfaction). Thus, we conducted the study with a random sample of 197 Sicilian dog owners of all sizes (dogs *and* owners are of all sizes), who presumably can decipher what their four-legged pals prefer.

Obviously, we could run two one-way ANOVAs and be done with the analysis in no time. Well, nobody prevents you from doing so, but remember from our magnificent discussion on the familywise error rate in Chapter 11 that when multiple separate tests are included in the same analysis, the Type I error is inflated. In addition, with two separate one-way ANOVAs, the relationship between perceived quality and satisfaction is completely ignored. It is in such cases that MANOVA comes to the rescue as it has the power to detect mean differences across groups along a *combination* of interrelated dependent variables.

Multivariate ANOVA tests the null hypothesis that the *vectors* of means on multiple dependent variables are equal across groups. This highly impressive hypothesis is tested by applying multivariate tests, which are then followed by individual univariate analysis for each dependent variable. Typically, one does not proceed to the univariate analyses if the initial multivariate tests are not significant. The results of the MANOVA in our example are shown in Table 13.4.

Table 13.4a An example of a one-way MANOVA: descriptive statistics

	Product formulation	Mean	Std. Deviation	N
Perceived quality	Spray	5.37	1.629	63
	Liquid	6.04	1.112	68
	Tablet	5.06	1.201	66
	Total	5.50	1.384	197
Satisfaction	Spray	4.67	1.751	63
	Liquid	4.62	1.779	68
	Tablet	3.58	1.692	66
	Total	4.28	1.804	197

Table 13.4b An example of a one-way MANOVA: Box's test of equality of covariance matrices[a]

Box's M	41.192
F	6.763
df1	6
df2	913,907.150
Sig.	.000

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

Note: [a] Design: Intercept + FORM

Table 13.4c An example of a one-way MANOVA: multivariate tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.953	1,967.218[b]	2.000	193.000	.000
	Wilks' Lambda	.047	1,967.218[b]	2.000	193.000	.000
	Hotelling's Trace	20.386	1,967.218[b]	2.000	193.000	.000
	Roy's Largest Root	20.386	1,967.218[b]	2.000	193.000	.000
FORM	Pillai's Trace	.149	7.798	4.000	388.000	.000
	Wilks' Lambda	.855	7.875[b]	4.000	386.000	.000
	Hotelling's Trace	.166	7.951	4.000	384.000	.000
	Roy's Largest Root	.134	13.006[c]	2.000	194.000	.000

Notes:

[a] Design: Intercept + FORM

[b] Exact statistic

[c] The statistic is an upper bound on F that yields a lower bound on the significance level

Table 13.4d An example of a one-way MANOVA: Levene's test of equality of error variances[a]

		Levene Statistic	df1	df2	Sig.
Perceived quality	Based on mean	5.802	2	194	.004
	Based on median	2.326	2	194	.100
	Based on median and with adjusted df	2.326	2	136.580	.102
	Based on trimmed mean	4.819	2	194	.009
Satisfaction	Based on mean	.097	2	194	.908
	Based on median	.111	2	194	.895
	Based on median and with adjusted df	.111	2	193.655	.895
	Based on trimmed mean	.110	2	194	.896

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

Note: [a] Design: Intercept + FORM

Table 13.4e An example of a one-way MANOVA: tests of between-subjects effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Perceived quality	34.020[a]	2	17.010	9.671	.000
	Satisfaction	49.901[b]	2	24.951	8.229	.000
Intercept	Perceived quality	5,931.598	1	5,931.598	3,372.316	.000
	Satisfaction	3,616.443	1	3,616.443	1,192.815	.000
FORM	Perceived quality	34.020	2	17.010	9.671	.000
	Satisfaction	49.901	2	24.951	8.229	.000
Error	Perceived quality	341.228	194	1.759		
	Satisfaction	588.180	194	3.032		
Total	Perceived quality	6,329.000	197			
	Satisfaction	4,254.000	197			
Corrected Total	Perceived quality	375.249	196			
	Satisfaction	638.081	196			

Notes:

[a] R Squared = .091 (Adjusted R Squared = .081)

[b] R Squared = .078 (Adjusted R Squared = .069)

Table 13.4f An example of a one-way MANOVA: pairwise comparisons

Dependent Variable	(I) Product formulation	(J) Product formulation	Mean Difference (I–J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
Perceived quality	Spray	Liquid	−.679*	.232	.011	−1.239	−.119
		Tablet	.304	.234	.582	−.260	.869
	Liquid	Spray	.679*	.232	.011	.119	1.239
		Tablet	.984*	.229	.000	.430	1.537
	Tablet	Spray	−.304	.234	.582	−.869	.260
		Liquid	−.984*	.229	.000	−1.537	−.430
Satisfaction	Spray	Liquid	.049	.304	1.000	−.686	.784
		Tablet	1.091*	.307	.001	.350	1.832
	Liquid	Spray	−.049	.304	1.000	−.784	.686
		Tablet	1.042*	.301	.002	.315	1.768
	Tablet	Spray	−1.091*	.307	.001	−1.832	−.350
		Liquid	−1.042*	.301	.002	−1.768	−.315
Based on estimated marginal means							

Notes:

* The mean difference is significant at the .05 level.

[a] Adjustment for multiple comparisons: Bonferroni.

Table 13.4a shows descriptive statistics for the dependent variables broken down by the factor (i.e., independent variable) of interest. Of particular importance here are the group sample sizes. When sample sizes are equal, MANOVA becomes quite robust against the violation of certain statistical assumptions that otherwise would require rather sophisticated approaches to address them. In our case, sample sizes across groups (for both outcome variables) are very similar, so we are fine. MANOVA also assumes that the variances across groups are homogeneous for each outcome variable and that the covariances between the outcome variables are similar in all groups. This latter assumption is tested with Box's test of equality of covariance matrices (Table 13.4b). The results are statistically significant, suggesting that the assumption is violated. However, given that our sample sizes are fairly equal, we should not worry too much about this. We can thus proceed to the multivariate tests in Table 13.4c, all of which are resilient to violations of multivariate normality. Although Wilks' λ (lambda) is the simplest and perhaps more commonly reported statistic, we will opt for Pillai's trace (Pillai-Bartlett trace *V*) because when group sizes are equal, this test is rather bulletproof to violations of the homogeneity of variance-covariance matrices assumption. We observe that there is a significant effect of product formulation on perceived quality and satisfaction ($V = 0.149$, $F(4, 388) = 7.798$, $p<0.001$), indicating that consumer judgments of quality and satisfaction *taken together* are different across the spray, liquid, and tablet options.

HINT 13.7 Make your life easier by keeping the group sample sizes as equal as possible when conducting a MANOVA analysis.

A significant multivariate test does not say anything about *how* product options differ between them, nor does it reveal whether the multivariate effect is mainly driven by perceptions of quality, satisfaction levels, or both. We thus need to examine the individual univariate tests (tests of between-subjects effects, Table 13.4e), as well as the post-hoc comparisons that follow. First, we see that both univariate *F*-statistics (SPSS runs simple one-way ANOVAs) are significant (and, yes, we also get the dreaded Levene's test of equality of error variances at no extra cost in Table 13.4d). Therefore, there are significant differences across groups for both perceived quality ($F(2, 194) = 9.671$, $p<0.001$) and satisfaction ($F(2, 194) = 8.229$, $p<0.001$). Next, the post-hoc comparisons (pairwise comparisons, Table 13.4f) show that, in terms of perceived quality, the liquid product option is superior to the other two options, which are not significantly different from one another. In terms of satisfaction, respondents are indifferent between the spray and the liquid form but significantly less satisfied by the tablets (given to their dogs – hopefully). Once again, plotting the results is very useful for quickly communicating the relevant findings (see Figure 13.7).

As with all tests involving comparisons between independent groups, MANOVA has to conform to all standard statistical assumptions (i.e., independence, normality, and homogeneity of variance). However, being a more sophisticated technique, it also requires some additional quality checks. In addition to the previously mentioned assumption of homogeneity of variance-covariance matrices, we also need to make sure that there is no **multicollinearity** (i.e., correlation that is too high) between the dependent variables. MANOVA performs best under moderate correlations among the dependent variables. If the variables are excessively

Figure 13.7 Perceived quality and satisfaction means across product formulations

correlated (say, $r \geq 0.80$), then some of the dependent variables will be redundant; if variables are almost completely unrelated, then running two separate one-way ANOVAs is a more powerful approach. In our example, the correlation between perceived quality and satisfaction is significant ($p<0.01$) with a correlation coefficient of $r = 0.231$, which is fine.

As with regression analysis (Chapter 12), outlier values may distort the results. **Multivariate outliers** can be identified by looking at a weird metric called 'Mahalanobis Distance'. If you are unable to get to sleep unless you know everything about this metric, the Further Reading section should do the trick.

If you have more than one independent variable, you can run a **factorial MANOVA**. Going back to the previous example, imagine that you wanted to test the influence of a number of alternative product characteristics on perceived quality and satisfaction level. For instance, along with the three product options, you want to consider the choice between a blue package (for male dogs) and a pink package (for female dogs), as well as two different brand names: Freshhound and Dogsmile. Thus, there are $3 \times 2 \times 2 = 12$ possible combinations of product characteristics (formed by 'crossing' the three factors), and you want to identify the optimal combinations on perceptions of quality and level of satisfaction. MANOVA would allow you to do this by indicating the impact of each product characteristic as well as the potential interactions between pairs of characteristics (known as 'two-way interactions') and between all three characteristics at once (known as 'three-way interactions'). Yes, it is (considerably) more complicated than running a 'simple' MANOVA, but it is doable.

Another extension of MANOVA is a **multivariate analysis of covariance (MANCOVA)**, which should be used when the confounding (i.e., unwanted) influence on the dependent variables of one (or more) variables (measured on a metric scale) needs to be accounted for in

the study. If you think that the age of the dog, for example, could possibly influence the perceived quality and/or satisfaction level with the dog breath freshener, then 'age of dog' could be treated as a covariate and its effect removed before the effects of the (categorical) independent variables (e.g., color or brand name) are analyzed. Alternatively, you may have an independent variable that consists of repeated measures (e.g., respondents are measured at different time points). In this case, you can use a **repeated-measures MANOVA**. You can even combine the previous approaches in a **factorial (mixed-design) MAN(C)OVA**. Having said that, you must *really* know your stats to set up more complex MAN(C)OVA models properly, as there are several conditions that have to be met in order to get valid and reliable results. This is not to discourage you from using them but to encourage you to use them *correctly* (which is easier said than done!).

Congratulations for having made it so far into the chapter! Before you go on, take another break to prepare yourself for learning how to investigate even more complex relationships! If you have a dog, use the break to smell the breath of Mr. or Ms. Canis (yes, we are good at Latin too!) and consider whether a dog freshener should be urgently added to your shopping list.

WARNING 13.4 Make sure you know what you are doing when playing around with complex MAN(C)OVA models. If you don't, make sure you know someone who does!

DEPENDENCE METHODS II: INVESTIGATING (MORE COMPLEX) RELATIONSHIPS

Partial and semi-partial correlation analysis

Partial correlation analysis involves the examination of multiple metric variables (that are linearly related), where one or more 'third' variables are assumed to be affecting a focal relationship between two variables. In other words, partial correlation is the proper technique to use when you want to assess the relationship between two variables, 'controlling' for the effect that one or more variables may have on it (we usually refer to these 'third' variables as 'control variables' or 'covariates').

As we discussed in Chapter 12, when two variables are correlated, it means that they have some common variance. This shared variance is captured by the squared correlation coefficient (r^2) and, in Figure 13.8, is visually represented by the intersection of X and Y (Panel A). When a third variable, Z, that is somehow correlated with X and Y, is introduced to the analysis, what is *uniquely* shared between the X and Y variables changes. Partial correlation focuses on the unique relationship between X and Y by 'partialing out' any variance accounted for Z (shaded areas c, b, and d in Panel B). Note that the initially observed shared variance between X and Y (i.e., $a + b$) is now reduced to the shaded area, a (also implying a lower correlation coefficient between the two variables). Thus, the squared partial correlation shows the amount of variance shared by X and Y *after* removing the influence of Z from both variables.

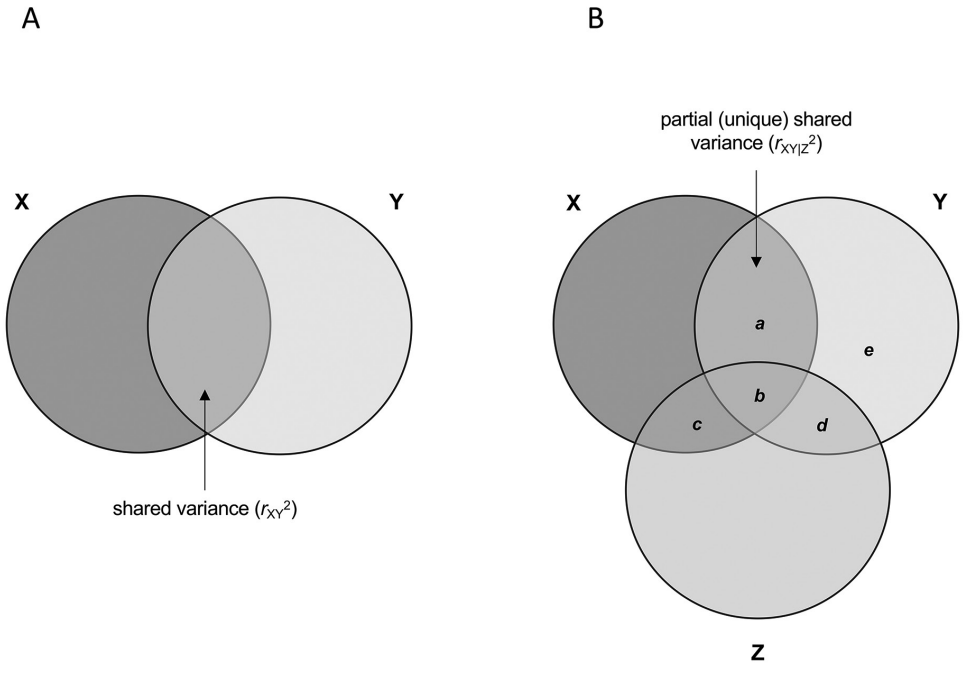


Figure 13.8 Shared variance in simple and (semi-)partial correlations

Suppose that we want to investigate the relationship between the frequency of practicing meditation and weight loss (both measured on interval scales anchored by 0 = 'not at all' and 10 = 'a lot'). We obtained data from a random sample of 100 Mongolian investment bankers, ran a (Pearson) correlation analysis, and found that there is a significant positive relationship between meditation and weight loss ($r = 0.389$, $p < 0.001$). However, we were kind of skeptical about this too-good-to-be-true finding, and we thought that we should somehow 'purify' this relationship by controlling for the simultaneous influence of 'exercise intensity' (measured in a similar way as the other variables). Arguably, if the original relationship persists after partialing out all of the variance accounted for by exercise intensity, our trust in the association between meditation and weight loss would increase (and we would meditate at least twice daily in a desperate attempt to improve our figures *even* further). Table 13.5 shows the results of the partial correlation analysis.

Table 13.5 An example of partial correlation

Correlations				
Control Variables			WEIGHT	MEDITATE
EXERCISE	WEIGHT	Correlation	1.000	.086
		Significance (2-tailed)	.	.399
		df	0	97
	MEDITATE	Correlation	.086	1.000
		Significance (2-tailed)	.399	.
		df	97	0

The SPSS output provides an adjusted intercorrelation matrix between meditation and weight loss, which is relatively self-explanatory. The results show that after controlling for respondents' intensity of exercising, there is no relationship ($r = 0.086$, $p = 0.399$) between meditation and weight loss. Partial correlation analysis does not have a limit in terms of the number of control variables that can be included. However, if you have several control variables, make sure that your sample size is big enough to provide trustworthy results and that all relevant statistical assumptions are met (see Chapter 12).

There are cases where we may want to adjust for the influence of control variables with respect to only *one* of the focal variables. For instance, going back to Figure 13.8, we may be interested in excluding the influence (shared variance) of *Z* only on *X*, while keeping *Y* intact. This can be done by employing the **semi-partial correlation** (SPSS labels it as 'part correlation'), which indicates the proportion in the total variance of *Y* that is uniquely accounted for by *X*. Unfortunately, there is no stand-alone option to get a semi-partial ('part') correlation from SPSS. However, as you will shortly see, you can indeed obtain semi-partial correlations through the regression menu of SPSS, so there is no need to panic (yet).

Multiple linear regression analysis

Multiple linear regression analysis is used to analyze the relationship between a single dependent variable and a number of independent variables. Both the dependent and the independent variables need to be metric (i.e., measured at interval or ratio level). Yet, as we saw in Chapter 12, categorical predictors can also be included in the analysis in the form of dummy variables. To illustrate the technique, assume that you want to predict the amount of money people will donate to Data Addicts Anonymous (a newly created charity that aims to pave the way back to normality for statisticians) during the next year. Assume further that you expect three variables to influence the amount (in thousand euros) donated, namely the number of times people are asked to give money per week (TIMES), the monthly income (INCOME) of the potential donors (also in thousand euros), and their gender (male = 0, female = 1). Multiple regression analysis would be the ideal technique for this task. It would not only enable the prediction of the dependent variable but also provide an assessment of the *relative* impact of each of the three independent variables. Moreover, it would indicate the *combined* ability of

the independent variables in explaining the variation in the dependent variable. Impressed? You should be!

The logic of multiple regression is similar to that of simple linear regression, as previously discussed in Chapter 12. However, now we are not fitting a regression 'line' to the data; instead, we fit a **regression 'plane'** like the one shown in Figure 13.9 (clearly one of the more impressive figures in our text, which undoubtedly demonstrates the fluid boundaries between statistics and abstract painting). Imagine that the data points are now dispersed in space, whose dimensions are defined by the number of independent variables we have. For example, Figure 13.9 shows how a multiple regression with two predictors (X_1 and X_2) would look. In multiple regression, we have one intercept (the starting point of the plane) and multiple regression coefficients (slopes). By changing the intercept and 'rotating' the plane along each dimension (predictor) individually (i.e., changing the individual slopes), we can find the **optimal multiple regression model**. The best model would be reflected by the plane that passes as close as possible to all observed values.

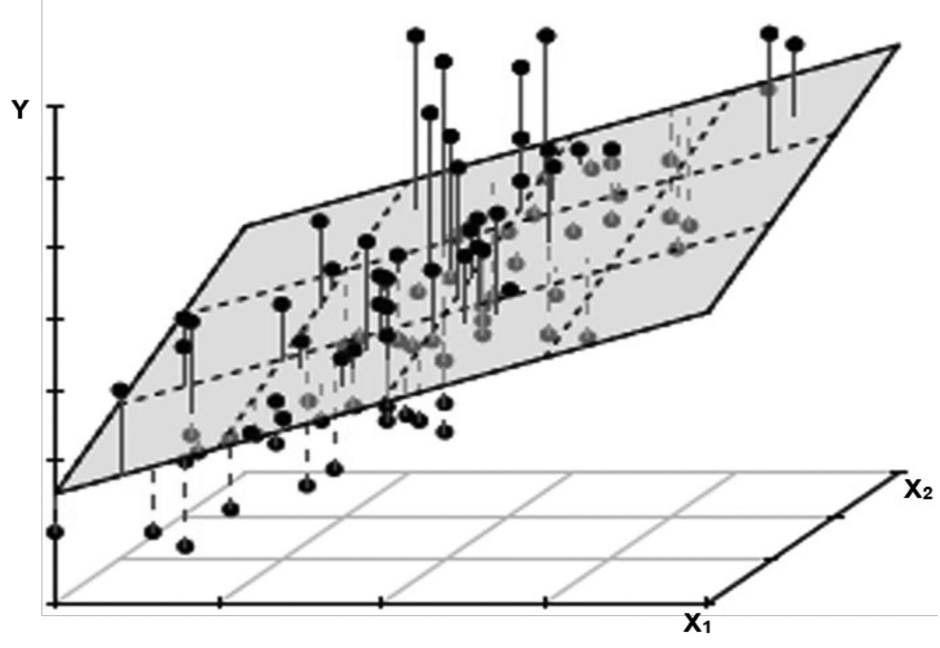


Figure 13.9 Visual representation of a multiple regression with two predictors

Since we now have multiple predictors to consider, a pertinent decision we need to make is how to introduce all these independent variables in the regression model. SPSS offers a number of options that can be classified into three different methods: enter, hierarchical/blockwise, and stepwise. The first method, **'enter'**, forces all predictors simultaneously in the model. The second method, **'hierarchical'**, allows users to add predictors in sequential blocks. For example, you may be interested in testing how a number of characteristics that relate

to consumers, products, and retailers can predict individuals' willingness to pay a premium price. In this case, you can group the characteristics into three blocks, which can then be entered sequentially in the model, thus allowing you to see the relative contribution of each block to the overall model fit. Finally, **'stepwise'** methods involve different variations of steps, whereby variables are included in and/or excluded from the model based on some mathematical criterion (typically the semi-partial correlation), capturing the unique contribution of each predictor to the predictive validity of the overall model. In theory-based research and hypothesis-testing, the first two methods are most appropriate as they are intrinsically linked to researchers' expectations and hypothesized model structure. In contrast, in exploratory studies as well as in cases where the aim is solely to identify a set of predictors that maximize the amount of explained variance in an outcome, stepwise methods may be handier.

HINT 13.8 In theory-based research and hypothesis-testing, incorporate independent variables in your multiple regression model via the 'enter' or 'hierarchical' methods.

In our donation to Data Addicts Anonymous example, we will use a hierarchical method, putting gender in the first block and the remaining two predictors in the second block. The results are shown in Table 13.6.

Table 13.6a An example of multiple (hierarchical) linear regression: model summary[a]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.134[b]	.018	.014	1.272	.018	4.563	1	248	.034	
2	.463[c]	.214	.205	1.143	.196	30.720	2	246	.000	2.095

Notes:

[a] Dependent variable: Amount of money people will donate (DONATE)

[b] Predictors: (Constant), GENDER

[c] Predictors: (Constant), GENDER, TIMES, INCOME

Table 13.6b An example of multiple (hierarchical) linear regression: ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.388	1	7.388	4.563	.034[b]
	Residual	401.528	248	1.619		
	Total	408.916	249			
2	Regression	87.632	3	29.211	22.366	.000[c]
	Residual	321.284	246	1.306		
	Total	408.916	249			

Notes:

[a] Dependent variable: Amount of money people will donate (DONATE)

[b] Predictors: (Constant), GENDER

[c] Predictors: (Constant), GENDER, TIMES, INCOME

Table 13.6c An example of multiple (hierarchical) linear regression: coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	5.720	.099		57.564	.000	5.524	5.915					
	GENDER	.362	.169	.134	2.136	.034	.028	.696	.134	.134	.134	1.000	1.000
2	(Constant)	4.699	.262		17.954	.000	4.183	5.214					
	GENDER	.266	.153	.099	1.744	.082	−.034	.567	.134	.110	.099	.993	1.007
	INCOME	.325	.044	.424	7.458	.000	.239	.410	.423	.429	.421	.990	1.010
	TIMES	−.090	.032	−.160	−2.832	.005	−.152	−.027	−.137	−.178	−.160	.997	1.003

Note:

[a] Dependent variable: Amount of money people will donate (DONATE)

The first thing to notice is that all parts of Table 13.6 consist of two models (1 and 2), which correspond to the two blocks of our hierarchical method. Model 1 is a simple regression model using only GENDER to predict the amount donated (DONATE), while Model 2 is the full model with INCOME and TIMES added to the previous step. As in the simple regression (see Chapter 12), the output starts with a 'model summary' (Table 13.6a). Given that Model 1 includes only one predictor, the interpretation of the first row is identical to that discussed in simple linear regression (Chapter 12). However, the interpretation of the second row (Model 2) is more sophisticated. The first column (R) now reports the **multiple correlation coefficient**, R, which is a measure of association between the outcome variable and the best linear combination of the three predictors. Right next to it is the **squared multiple correlation** (R^2; also known as 'coefficient of determination'), which shows the proportion of variance in the dependent variable that is explained by the full model (Model 2); the latter can explain 21.4% of the variance in the amount of money people intend to donate.

The R^2 is a bit naughty as it tends to overestimate model fit in multiple regression. As one continues to add predictors in the model, the variance explained will always become larger. However, this is more likely to be attributable to chance variation in a particular sample and does not necessarily represent what holds true in the population. The **adjusted R^2** ($_{adj.}R^2$) attempts to provide a more accurate estimate of the variance explained in the population. If the number of observations is small and the number of predictors is large, there will be a greater difference between R^2 and $_{adj.}R^2$, whereas when the number of observations is fairly large in relation to the number of predictors, the two measures will be very similar. The difference between R^2 and $_{adj.}R^2$ observed in Table 13.5 is less than 1%, which suggests that the model is fairly stable. In general, as the complexity of a model increases, the chances that one may be capitalizing on the idiosyncratic characteristics of the sample also increases. This is called **overfitting** and results in particularly high R^2 values that do not really represent the situation in the population and, thus, are not likely to replicate. Although the $_{adj.}R^2$ is not purposely

designed to identify overfitting problems, the difference between R^2 and $_{adj.}R^2$ can, to a certain extent, alert you to potential overfitting.

WARNING 13.5 Beware of overfitting, which results in particularly high R^2 values. This may reflect idiosyncratic characteristics of the sample and may not replicate.

The output also shows the *additional* variance explained in the dependent variable as a result of adding the two additional independent variables in Model 2. This is labeled ***R²* change** and is accompanied by a formal test indicating whether the increase in the variance explained is significant. The R^2 change is tested with an *F*-test (*F*-change). A significant *F*-change value means that the variables added in that step significantly improve prediction. In our example, the variance explained jumps from 1.8% to 21.4%, and this, of course, is a highly significant improvement in predictive ability ($F(2, 246) = 30.720$, $p<0.001$).

Table 13.6b tests the overall model fit. This is done by comparing the variance explained by the model to the unexplained variance. For both models, model fit is significant, as shown by the relevant *F*-values. Table 13.6c provides the intercepts (i.e., the point where the regression plane crosses the *Y*-axis or, put differently, what the expected value of *Y* would be if the predictors were zero) and informs us about the unique influence of each predictor. Each regression coefficient (standardized and unstandardized) tells us the effect of a given predictor on the outcome if the effects of all other predictors are held constant (i.e., fixed to zero). This why regression coefficients in multiple regression analysis are often referred to as **partial slopes**. Looking at the results, we see that as additional predictors are included in the model, the individual influence of gender reduces and becomes non-significant (two-tailed *t*-test at the 5% significance level). Controlling for all other predictors, we see that income has a positive significant effect on the amount donated ($B = 0.352$, $p<0.001$). This means that a one-unit increase in the monthly income results, on average, in a 0.325-unit increase in the amount respondents are willing to donate. In contrast, the more you bother people by repeatedly asking them to donate money, the less money they can be expected to donate; this is indicated by a negative *B* (unstandardized coefficient) of −0.090 ($p<0.01$). Controlling for gender and income, one additional request to donate per week is therefore expected to reduce the amount of donation by −0.090. Note that the standardized coefficients (β) work similarly but show by how many standard deviations the dependent variable will change in response to one standard deviation change in a given predictor (adjusting for the influence of all other predictors).

WARNING 13.6 If you *really* think that a certain predictor is redundant, then you should run a new regression model excluding that predictor and use this new model to make the forecast.

As previously noted, respondents' gender does not have a significant, unique influence on the outcome in Model 2 but, being a binary variable, it is worth seeing how it is interpreted. Recall from Chapter 12 that the expected difference associated with a one-unit increase in a binary variable practically shows the difference between the two binary levels (i.e., male = 0, female =

1). Thus, the β coefficient for gender shows that females, compared to males, are expected to donate 0.266 (thousand) dollars more.

The equation associated with the multiple regression model is a variation of the equation of the straight line associated with simple linear regression (see Chapter 12), as follows:

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$$

We now only need to replace the *X*s with specific values of our predictors and the βs with the unstandardized coefficients from Table 13.6 to make a specific forecast/prediction. For example, on average, the amount of money that a female respondent who earns $1,000 per month and is being asked to donate five times per week is expected to donate is $329.52 (= 4.699 + 0.266 × 1 + 0.325 × 1,000 − 0.090 × 5). Note that some people exclude non-significant predictors from a multiple regression equation when they are making predictions. Although this may sound reasonable at a superficial level, it is erroneous and should be avoided. The reason is simple. The model you test consists of *all* parameters and, thus, any forecasting calculation should reflect this in order to be complete. The best-fitted model, the coefficients, and the intercept are estimated based on a certain specification, so any modification (by excluding predictors) would basically result in a different model. Even if a predictor does not significantly influence the outcome, it may be related to other predictors and, consequently, influence the results.

As mentioned earlier, the multiple regression coefficients (partial slopes) show the unique effect of each predictor on the dependent variable, controlling for the influence of all other predictors in the model. Given that the standardized coefficients (β) are expressed in the same unit of measurement for all predictors, they can be used to convey the relative strength of each predictor. In this context, the raw (unstandardized) regression coefficients can be misleading in conveying the magnitude of an effect. For example, in Table 13.6, the raw coefficient of GENDER is noticeably larger than that of TIMES. However, bringing them in comparable units of measurement (by standardizing both coefficients), this relationship is completely reversed. Thus, you should always consult the standardized regression coefficients to explore the magnitude of the various individual effects. Although this is true, note that it only allows us to make statements about the relative predictive strength on a descriptive basis. If you want to make a more conclusive statement about which predictor is significantly stronger, then you need to apply a formal test for the difference between regression slopes. (This is *not* provided by default in most statistical packages and requires the application of more custom-made techniques – have a look at the Further Reading in the highly unlikely event that you are interested.)

Another extremely useful measure, which sadly researchers rarely utilize, is based on the semi-partial correlations between the predictors and the outcome. As described in the previous section, semi-partial correlations remove all the variance in a predictor that is shared with all other predictors, thus producing the **unique shared variance** of this predictor and the outcome. Thus, by squaring the semi-partial correlation of each predictor (labeled as 'part' in the SPSS output), we can quantify the percentage of variance in the dependent variable

accounted for by the independent variable. In our example, we can say that INCOME alone accounts for 17.7% (i.e., 0.421^2) of the variance in the amount of money people are willing to donate (see Table 13.6c).

And how about including one or more **control variables** in our regression model? Well, we can 'control' for the influence of other variables by simply including them in the regression model as additional predictors. Remember that regression coefficients capture the unique effect of each predictor, adjusting for the influence of the others. Going back to our example, suppose that we wanted to control for the frequency with which people generally donate to similar benevolent causes (e.g., to such well-known charities as Support the Aliens or the American National Cattle Women's Retirement Fund). As you can see in Table 13.7, our proposed covariate (FREQUENCY) is simply added to the 'Coefficients' table along with all other predictors. Overall, covariates or control variables are statistically treated as any other predictor, and therefore whether an independent variable is designated as a theoretically relevant predictor or a covariate/control variable depends on the particular focus of the study.

Table 13.7 Including covariates in multiple linear regression models

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
	(Constant)	4.526	.260		17.435	.000	4.015	5.037					
	INCOME	.303	.045	.392	6.798	.000	.215	.390	.435	.399	.375	.915	1.092
	TIMES	−.103	.031	−.183	−3.298	.001	−.164	−.041	−.145	−.207	−.182	.992	1.008
	GENDER	.155	.151	.057	1.024	.307	−.143	.453	.130	.065	.056	.965	1.036
	FREQUENCY	.108	.034	.182	3.122	.002	.040	.176	.292	.196	.172	.896	1.116

Note: [a] Dependent variable: Amount of money people will donate (DONATE)

There are some circumstances where examining what happens to the dependent variable by looking at how individual predictors change one at a time (while the others remain fixed) is not sufficient. For example, we may be interested in seeing what happens to donations as *both* INCOME *and* TIMES increase at the same time, or have reasons to believe that INCOME will not have a similar effect between males and females. To account for such joint influences, we must specify **interaction effects** in our model. Interaction analysis can reveal whether the effect of one predictor (e.g., INCOME) on an outcome variable (here, DONATE) changes at different levels of another predictor (e.g., TIMES or GENDER) (remember that we introduced the notion of interaction in the factorial ANOVA section earlier in this chapter).

An important first step to understand how interaction analysis works is to select a focal predictor as a reference point and then see whether its effect is influenced by another predictor; we typically refer to the latter as the **'moderator'** variable and use the term **moderation analysis** to describe the analysis of regression models with interaction terms. If we cannot explicitly identify a moderator, we can treat both independent variables in an interaction similarly and

absolutely no difference for the analysis, it is advisable to set your eyes on one of the predictors as a primary focus to facilitate interpretation. To demonstrate, let us assume that we are investigating the interaction between two independent variables on some dependent variable (Y). Also, let us consider the first independent variable as being the focal predictor (X) and the second one as the moderator (M). Figure 13.10 shows two different cases of interaction effects.

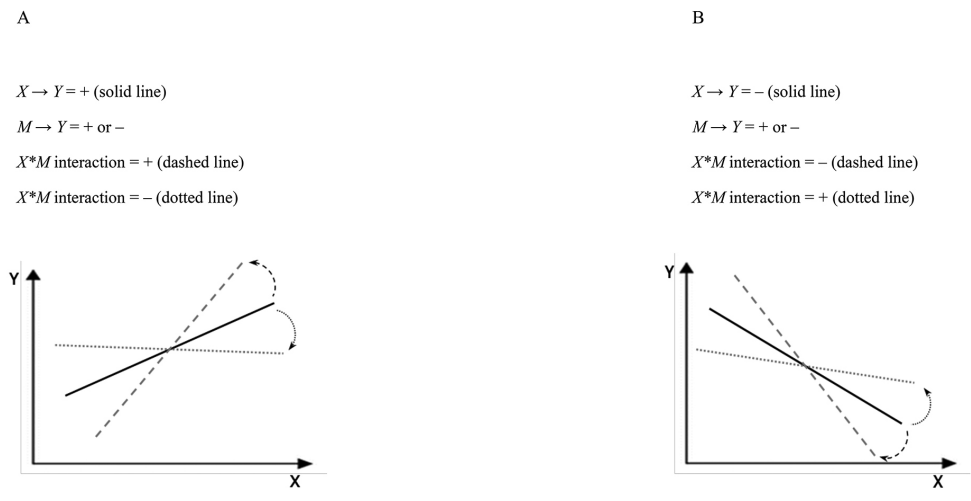


Figure 13.10 Visual representation of interaction effects between two variables

In Panel A, we see that the focal predictor (X) has an individual positive effect on the dependent variable (Y). The individual effect of the moderator (M on Y) does not concern us at this point, and it may be either positive, negative, or even non-significant. Regardless of its effect on Y, M may interact positively or negatively with X. In the former case, we say that M positively moderates the effect of X on Y. This means that the already positive $X \rightarrow Y$ effect (solid line) will be even stronger at higher levels of M (dashed line). In the latter case, the opposite holds true. If M negatively moderates the effect of X, then the positive $X \rightarrow Y$ effect (solid line) will become less positive as M also increases (dotted line). Note that, depending on its magnitude, this negative interaction may weaken, eliminate, or even reverse the initial positive effect of X on Y.

Panel B shows the corresponding regression lines for a positive and a negative interaction when the individual effect of X on Y is negative. Overall, the following general rule (which you should memorize and be able to recite backward) applies: an interaction among predictors (X_1, X_2 … X_n) accentuates (i.e., strengthens) the effect of predictors with the same sign and attenuates (i.e., weakens) the effects of those with an opposite sign. We should note that Figure 13.10 is a crudely simplified visualization, just to help us get the message across. How the regression line changes depends, of course, on the relevant estimates (intercept and regression coefficients) of the specific model. Note that nothing prevents you from testing higher-order interactions (e.g., two-way interactions); however, the interpretation of the results gets more complicated (see discussion on factorial ANOVA earlier in the chapter).

Going back to our example on donations to Data Addicts Anonymous, assume that we now also want to test the interaction between the income of respondents (INCOME) and the number of times they are asked to donate (TIMES). To do this, we simply need to manually compute an **interaction term** (also known as a 'product term') between the two predictors by multiplying them and include their product as an additional predictor in the model. *Before* you do so, however, there is another step you may need to take, namely to **mean-center** the two predictors by subtracting the respective mean from every individual value. This rescales the values of the predictors around a central location of zero so that each observation now corresponds to a *distance* from the mean (retaining the original units of measurement). For example, a mean-centered value of 1.5 implies that the value of this particular case is 1.5 units above the mean.

In the past, mean-centering was widely thought to address problems of multicollinearity between predictors. However, recent methodological literature has cast doubts on this claim (an insight that will surely make you sleep better at night). A more legitimate reason for mean-centering is because it facilitates the interpretation of the results. Regression analysis estimates each coefficient by setting other predictors to zero. However, there are cases where zero does not represent a meaningful value. Think about people's weight, blood pressure, or (strictly speaking) even an interval scale ranging from one to five. In all of these cases, zero is outside the bounds of the measurement scale, thus relevant claims about the effect of one predictor while setting the others to zero make little sense (despite being statistically valid). Mean-centering the predictors prior to computation of their product yields estimates that are within range and readily interpretable, showing the expected value of an outcome at the mean value of the predictors. At the end of the day, whether one mean-centers the predictors or not is a decision taken according to the purpose of the analysis and not a requirement of the analysis *per se*.

HINT 13.9	Always consider whether mean-centering predictors makes sense when forming interaction terms.

In our example, we decided to mean-center INCOME and TIMES before calculating their interaction term and running the analysis. Table 13.8 shows the resulting coefficients. We can clearly see that the interaction effect (mcTIM_INC) is positive and significant at the 5% significance level. Following the rule mentioned earlier, we can conclude that as both predictors increase simultaneously, the positive effect of INCOME is enhanced and the negative impact of TIMES is attenuated.

Table 13.8 Including interactions between metric variables in multiple linear regression models

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.735	.089		64.541	.000
	GENDER	.269	.152	.100	1.776	.077
	mcINCOME	.313	.044	.409	7.177	.000
	mcTIMES	−.084	.032	−.150	−2.664	.008
	mcTIM_INC	.038	.019	.115	2.016	.045

Note: [a] Dependent variable: Amount of money people will donate (DONATE)

You will be overjoyed to learn that one can also calculate interactions between a metric and a binary variable. Thus, in our example, we might want to see if the influence of the number of times (TIMES) someone is asked to donate is intertwined with GENDER (which, remember, is coded as male = 0, female = 1). We followed the same procedure (i.e., computing the interaction term and entering it in the regression model). The results are shown in Table 13.9. Note that GENDER is not significant on its own (p=0.58). Thus, the amount of money people are willing to donate to Data Addicts Anonymous does not differ between male and female respondents. However, GENDER significantly moderates the influence of TIMES such that the latter has a significantly more negative effect on donations among female as opposed to male respondents (*why* this is the case we will leave it to you to ponder).

Table 13.9 Including interactions between a metric and binary variable in multiple linear regression models

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.749	.089		64.866	.000
	GENDER	.289	.152	.107	1.903	.058
	mcINCOME	.400	.055	.522	7.274	.000
	mcTIMES	−.090	.031	−.161	−2.865	.005
	mcGENDER_TIMES	−.196	.089	−.159	−2.211	.028

Note: [a] Dependent variable: Amount of money people will donate (DONATE)

You can break down the effect of an interaction even further by performing what is called **simple slopes analysis** or **spotlight analysis**. This is a pick-a-point approach, whereby the researcher picks a value of the moderator and examines the **conditional effect** of the predictor at that value. For instance, we can estimate the effect of TIMES on DONATE separately for

male and female respondents (upper graph in Figure 13.11) or estimate how TIMES influences DONATE at low, medium, and high values of the distribution of INCOME (lower graph in Figure 13.11). The bad news is that this type of analysis requires several troublesome computational steps to do in SPSS. However, there are specialized SPSS add-ons and more sophisticated statistical packages that offer ready-made solutions to perform such an analysis (yes, see Further Reading before you break down – break down the effects of an interaction, that is, not yourself after reading this unforgivably demanding section).

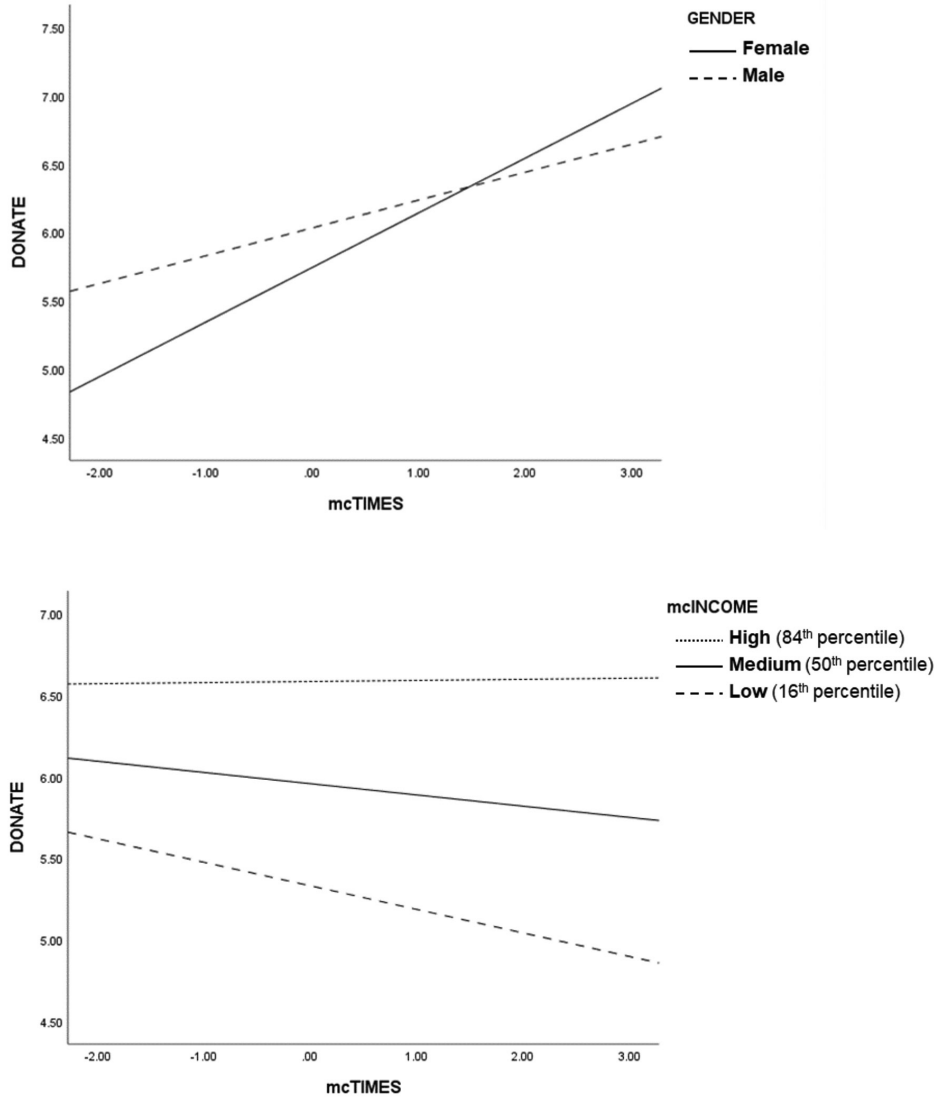


Figure 13.11 Visual representation of simple slopes effects

In Chapter 12, we pointed out that regression analysis needs to conform to a number of statistical assumptions to provide valid and generalizable results. This holds true regardless of the exact specification (e.g., including interaction terms or not). Specifically, the dependent variable should be linearly related to predictors, whose combined effect is best described by adding their individual effects together (i.e., linearity and additivity). Also, the dependent variable should be measured on a continuous scale (interval or ratio), the predictors must not have a zero variance, and the residuals should be independent, homoscedastic (i.e., have similar variance across the levels of the predictors), and normally distributed.

Independence of residuals (sometimes referred to as '**autocorrelation**') can be checked with the Durbin–Watson statistic (see Table 13.6a). Durbin–Watson tests for correlations in sequential pairs of residuals and values within 1.5 and 2.5 indicate no autocorrelation problems. To explore the residuals further, we can save their standardized values through SPSS and examine their distribution. First, we can check the normality of residuals using skewness and kurtosis scores or by visually inspecting their histogram. Moreover, we can create a scatterplot of the standardized predicted values and the standardized residuals to identify any unusual patterns. These plots are mainly used to identify violations of **homoscedasticity**, which tend to manifest through a funnel-like dispersion such as the one shown on the left-hand side of Figure 13.12. If your model is valid and good enough, then it should predict some reasonable amount of the dependent variable and leave out something that is unpredictable; that is, random error. Thus, if the residuals are not randomly scattered around zero along the range of the predicted values (note that zero error indicates perfect prediction), then something may be wrong! We know that you always suspected this although you never wanted to talk about it in public. In particular, clusters of negative residuals imply that your model **overestimates** the predicted values, while clusters of positive residuals indicate **underestimation**.

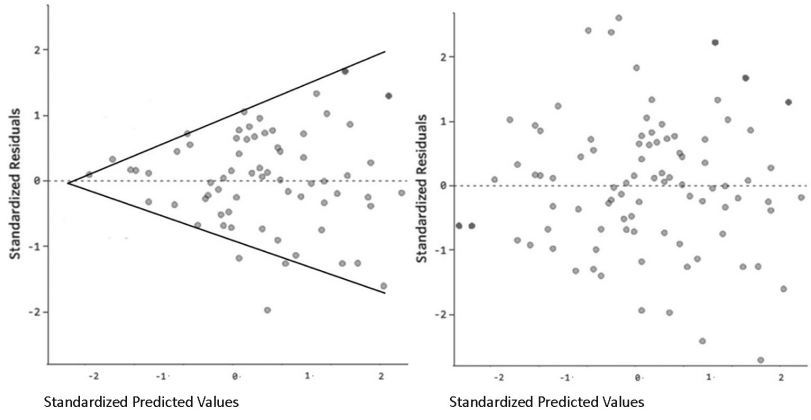


Figure 13.12 Scatterplot of standardized residuals against standardized predicted values

In models with multiple predictors, there should be no multicollinearity. That is, the predictors should not be excessively correlated with each other. There a number of options to

investigate this assumption. For example, multicollinearity can be examined with the **variance inflation factor** (VIF) and **tolerance values** (see Tables 13.6c and 13.7 under 'Coefficients'). The VIF shows how much the variance of a coefficient is inflated due to the fact that predictors are correlated and should preferably be below 5 (although more relaxed standards can also be found in the literature). Tolerance, on the other hand, can be roughly seen as the percentage of variance of an independent variable that is *not* explained by other predictors and should not be below 0.20. Thankfully, in our example, all of these conditions were satisfied.

Finally, the regression menu in SPSS offers you the option to investigate whether specific cases (e.g., outliers) exert a strong influence in the model. To identify such **influential values**, researchers use the Mahalanobis Distance (see the section on 'MANOVA' earlier) or the Cook's distance, where values above 1 may be suspicious. To our mind, however, the most straightforward measure is the DFFit (and Standardized DFFit), which represents the difference between the predicted value for a case when the model includes versus does not include this particular case. Cases that are not influential should be associated with a DFFit close to zero. Alternatively, the more influence an individual case exerts on the model, the higher its DFFit value; to overcome misinterpretations due to differences in the units of measurement, the standardized version of DFFit should be used.

While multiple linear regression analysis is extremely useful (never leave home without it!), we realize that this section has been heavy-going, to put it mildly. If you still haven't had enough of multiple regression, there are surely psychological helplines you can call! We have definitely had enough and are ready to move on to another riveting topic.

MULTIPLE LOGISTIC REGRESSION ANALYSIS

Multiple logistic regression is a technique that draws on multiple independent variables to predict a categorical dependent variable. It is a straightforward extension of the (binary) logistic regression analysis we already discussed in Chapter 12. To show the fundamental logic behind this method, we will focus on a dichotomous dependent variable; the principles can be easily extended to multi-categorical dependent variables.

Assume that we are interested in investigating whether a number of characteristics associated with coffee houses can predict whether a consumer will prefer one over the other. More specifically, we want to see if the laziness of the waiters (LAZINESS), the rudeness of the waiters (RUDENESS), the perceived quality of the coffee (QUALITY), and the perceived value for money (VALUE) can predict whether a consumer will choose to drink their coffee at Coffee Zombie (coded as 0) or Coffee Mania (coded as 1). All predictors are measured on seven-point interval scales with higher values indicating more of the property being measured. Table 13.10 shows the relevant output produced by SPSS.

Table 13.10a An example of multiple logistic regression analysis, block 0: classification table[a,b]

	Observed		Predicted		
			Coffee house choice		Percentage Correct
			Coffee Zombie	Coffee Mania	
Step 0	Coffee house choice	Coffee Zombie	40	0	100.0
		Coffee Mania	35	0	.0
	Overall Percentage				53.3

Note:

[a] Constant is included in the model

[b] The cut value is .500

Table 13.10b An example of multiple logistic regression analysis, block 0: variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	−.134	.231	.333	1	.564	.875

Table 13.10c An example of multiple logistic regression analysis, block 0: variables not in the equation

			Score	df	Sig.
Step 0	Variables	QUALITY	5.696	1	.017
		LAZINESS	10.339	1	.001
		VALUE	.480	1	.486
		RUDENESS	12.559	1	.000
	Overall Statistics		22.031	4	.000

Table 13.10d An example of multiple logistic regression analysis, block 1: method = enter (omnibus tests of model coefficients)

		Chi-square	df	Sig.
Step 1	Step	24.231	4	.000
	Block	24.231	4	.000
	Model	24.231	4	.000

Table 13.10e An example of multiple logistic regression analysis, block 1: method = enter (model summary)

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	79.407	.276	.369

Note: [a] Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Table 13.10f An example of multiple logistic regression analysis, block 1: method = enter (classification table[a])

	Observed		Predicted		
			Coffee house choice		Percentage Correct
			Coffee Zombie	Coffee Mania	
Step 1	Coffee house choice	Coffee Zombie	33	7	82.5
		Coffee Mania	10	25	71.4
	Overall Percentage				77.3

Note: [a] The cut value is .500

Table 13.10g An example of multiple logistic regression analysis, block 1: method = enter (variables in the equation)

								95% C.I. for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	QUALITY	−.497	.577	.742	1	.369	.609	.197	1.884
	LAZINESS	.190	031	6.428	1	.011	1.209	1.044	1.400
	VALUE	−.030	031	.892	1	.345	.971	.913	1.932
	RUDENESS	.380	.131	6.418	1	.004	1.462	1.131	1.889
	Constant	−1.736	1.159	2.243	1	.134	.176		

Note: [a] Variable(s) entered on step 1: LAZINESS, QUALITY, VALUE, RUDENESS

In block 0, we obtain the results associated with the **baseline (or null) model** (i.e., the predicted choice of coffee house in the absence of the predictors). As can be seen in Table 13.10a, the majority of people (40 respondents) choose Coffee Zombie, as opposed to Coffee Mania (35 respondents), for their coffee. Thus, the baseline model would make the prediction that consumers prefer Coffee Zombie, correctly classifying 53.3% of the cases (40 out of 75). This classification accuracy is not very different from a 50% guess. In fact, Table 13.10b shows that the baseline model does not predict the outcome significantly better than chance (B=−0.134, p=0.564). The probability of getting such, or more extreme, results assuming that respondents choose both coffee houses equally (i.e., the null hypothesis) is too high to be considered as evidence against the null hypothesis. On the other hand, Table 13.10c shows us the individual

contribution of each potential predictor to the baseline model. The table shows that all predictors, except for VALUE, would significantly improve the predictive strength of the baseline model.

The 'omnibus tests of model coefficients' (Table 13.10d) test the **full model**. The results show that model fit is significant with a χ^2 (4) = 24.231, p<0.001. Thus, including the four predictors helps us make a significantly better prediction. Table 13.10e and specifically the Nagelkerke pseudo-R^2 tells us that the model performs well; remember that the latter is (very loosely) interpreted as the amount of variability explained by the model. Nagelkerke pseudo-R^2 ranges from 0 to 1, with higher values indicating better fit.

Table 13.10f clearly shows that the overall classification accuracy is materially improved (77.3%) in relation to the baseline model (53.3%). In detail, the model predicts correctly 82.5% (33/40) of respondents who choose the Coffee Zombie coffee store and 71.4% (25/35) of those who would rather go for Coffee Mania. Thus, the model is slightly more accurate when it comes to classifying Coffee Zombie fans.

The final part of the output (Table 13.10f) shows the directionality, magnitude, and significance of the unique effects of each predictor (controlling for the others). Remember that the raw beta (B) is hardly interpretable because of the transformations involved in logistic regression; it basically reflects changes in the logit (log-odds), resulting from one unit change in the predictors. Therefore, we turn to the exponentiated beta (Exp(B) or e^B), which communicates the change in the odds resulting from a unit change in the predictors. The results show that, holding all other predictors constant, waiters' laziness is (scandalously!) associated with a 20.9% increase in the likelihood of choosing Coffee Zombie (e^B = 1.209) over Coffee Mania, while a one-unit increase in perceived waiters' rudeness makes it 1.46 times more likely that consumers will get their coffee from Coffee Zombie (e^B = 1.462) as opposed to from Coffee Mania. Value for money as well as perceptions of quality do not have an influence on respondents' choice (the relevant Wald χ^2 tests are not significant) and seem to be considerably less important in the presence of grumpy, slow-moving service (we don't blame them!).

Everything we discussed in the previous section on multiple linear regression with regard to including covariates and interaction terms between predictors is also applicable to multiple logistic regression. Furthermore, note that logistic regression does not assume linearity between the dependent and independent variables. Also, the residuals (i.e., error terms) need not be normally distributed or be homoscedastic. This, however, does not mean this technique is assumption-free. Logistic regression assumes linearity between the independent variables and the logit (log-odds). Independence of observations also needs to be ensured, while multicollinearity should not be a problem between predictors. With reference to the latter, note that VIF and tolerance values are not automatically produced through the logistic regression menu in SPSS. But there is nothing stopping you from running a 'mock' linear regression analysis, simply to get the relevant scores. In addition, as with ordinary (linear) regression, you can save the standardized residuals as new variables and examine any bias attributed to highly influential cases.

A logistic regression model typically requires a larger sample compared to an ordinary linear regression model. A commonly used rule of thumb found in the literature specifies a minimum of 20 cases per predictor with a minimum overall sample of 60. We need to

highlight that these (and other similar) rules of thumb should be seen as very crude starting points. Sample size determination involves several considerations (see Chapter 3), including the important question of statistical power (see Chapter 9).

CANONICAL CORRELATION ANALYSIS

Canonical correlation analysis, the last analysis we will discuss before you can tick off this chapter, can be used when we are interested in multiple dependent *and* multiple independent variables; both sets of variables have to be metric. This technique can also be viewed as an extension of multiple regression analysis, the key difference being the number of dependent variables. The technique derives separate **linear combinations** of the independent and dependent variables so as to maximize the correlation between the sets of independent and dependent variables. Put differently, the technique creates a new set of composites (called **canonical variates**) and optimally correlates the dependent canonical variates to the independent canonical variates.

Let us assume that we want to examine the joint influence of the number of new product introductions, the advertising budget, and the salary of the chief executive officer (CEO) on export profits and export market share among companies selling organic peanut butter to Guatemala. SPSS does not include a default option to perform canonical correlation analysis, and to run it we need to tweak the code (i.e., syntax) of the MANOVA option (which is closely related). The main results are presented in Table 13.11.

Table 13.11a An example of canonical correlation analysis: multivariate tests of significance

Test Name	Value	Approx. F	Hypothesis df	Error df	Sig.
Pillais	.39249	20.02143	6.00	492.00	.000
Wilks	.61142	25.58403	6.00	488.00	.000
Hotellings	.62912	22.77492	6.00	490.00	.000
Roys	.38224				

Note: F statistic for Wilk's lambda is exact

Table 13.11b An example of canonical correlation analysis: dimension reduction analysis

Roots	Wilks L.	F	Hypothesis df	Error df	Sig.
1 to 2	.61142	22.77492	6.00	490.00	.000
2 to 2	.98975	1.27416	2.00	246.00	.282

Table 13.11c An example of canonical correlation analysis: univariate F-tests

Variable	Sq. Multiple R	Adj. R-sq.	Hypoth. MS	Error MS	F	Sig.
Profits	.34830	.34036	43.28429	.98766	43.82531	.000
Market share	.17099	.16088	27.35817	1.61760	16.91282	.000

Table 13.11d An example of canonical correlation analysis: raw canonical coefficients for dependent variables

	Function No.	
Variable	1	2
Profits		−.58586
Market share	.23726	.74879

Table 13.11e An example of canonical correlation analysis: standardized canonical coefficients for dependent variables

	Function No.	
Variable	1	2
Profits	.82185	−.71688
Market share	.32941	1.03964

The output begins with the 'multivariate test of significance' (Table 13.11a) for the entire model using the same statistical tests we encountered earlier in the MANOVA section. These test the null hypothesis that the canonical correlations are zero; that is, there is no linear relationship between the two *groups* of variables. In our example, the relationship between the three independent variables and the two dependent variables is highly significant, no matter what test is involved. The following part ('dimension reduction analysis', Table 13.11b) determines how many dimensions (i.e., canonical variates) are actually required to describe the covariation between the two groups of variables. Note that the maximum number of canonical variates is equal to the number of variables in the smaller set (here, the export performance variables). Thus, in our example, we can derive two canonical dimensions. Do they both independently contribute to the overall model? In our case, Root 1 to 2 is significant, indicating that the canonical relationship is not equal to zero. This does not hold true for the second dimension, which, by itself, does not correspond to a significant relationship.

The part 'univariate *F*-tests' (Table 13.11c) give us the R^2 (squared multiple correlation coefficient) and adjusted R^2 for predicting (a) export profits and (b) export market share from the three company characteristics (i.e., number of new product introductions, size of the advertising budget, and salary of the CEO). Note that we would get the same fit statistics if we just ran two separate multiple regressions.

Finally, we obtain the **raw and standardized canonical coefficients** (Tables 13.11d and 13.11e) that define the linear relationship (direction and magnitude) between the canonical variates comprising export performance and the variates formed by the company character-

istics. The raw coefficients are interpreted similarly to regression coefficients. For example, we can say that a one-unit increase in the first variate comprising company characteristics corresponds to a 0.67-unit increase in profits.

Now the usual health warning: as with MANOVA, multiple regression, and related techniques, to derive valid results, canonical correlation analysis rests on the assumptions of multivariate normality, homogeneity of variance, and low multicollinearity. It also assumes a linearity between all variables as well as between the variables and the canonical variates. Last, given its complexity and the fact that it typically handles a large number of variables at the same time, canonical correlation analysis requires large sample sizes to provide robust results. The analysis should also never be conducted on an empty stomach.

SUMMARY

This chapter has provided you with a first insight into multivariate analysis. First, we looked at multivariate analysis as an extension of univariate/bivariate analysis and explained why multivariate techniques overcome the inherent limitations of simpler methods. Subsequently, we distinguished between two types of multivariate techniques, namely dependence and interdependence methods. We then introduced a range of dependence methods that can be used to make more complex comparisons between groups and investigate more complex relationships between variables. In doing so, we tried to give you a good understanding of the logic underlying some of the most useful and commonly used dependence techniques so that you are well equipped to tackle more advanced readings.

QUESTIONS AND PROBLEMS

1. Explain, in general terms, the difference between univariate, bivariate, and multivariate analysis.
2. What can you do with multivariate methods that you cannot do with univariate and bivariate methods?
3. Explain the difference between so-called dependence methods and interdependence methods.
4. Which multivariate technique would you use to determine how two groups differ in terms of several variables?
5. Give an example of a research situation for which multiple regression analysis would be appropriate.
6. What are the differences and similarities between a predictor and a covariate in regression analysis?
7. What do we really investigate in interaction (or moderation) analysis?
8. Which analysis is commonly used to correlate simultaneously multiple metric dependent variables with multiple metric independent variables?
9. When would you use multivariate analysis of variance (MANOVA)?
10. Should our young and overly enthusiastic co-author Georgios Halkias be physically reprimanded for the length of this chapter?

FURTHER READING

Affi, A., May, S., Donatello, R. A., & Clark, V. A. (2019). *Practical Multivariate Analysis*, 6th edition. Boca Raton, FL: CRC Press. A nice guide to various multivariate techniques also covering some more advanced topics. Best read after Black et al. (2018) below.

Black, W. C., Hair, J. F., Anderson, R. E., & Babin, B. J. (2018). *Multivariate Data Analysis*, 8th edition. London: Pearson Prentice Hall. An excellent introduction to multivariate analysis with good examples.

Hayes, A. F. (2018). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, 2nd edition. New York: Guilford. In this book, as well as the website (https://www.processmacro.org) developed by the author, you can find alternative options to easily conduct interaction analyses.

14
Getting obsessed: A further look into multivariate analysis

In this chapter, we continue our magical journey into the magnificent world of multivariate analysis by having a look at some popular **interdependence techniques**. As already noted in Chapter 13, such techniques do not distinguish between independent and dependent variables but use the interrelationships among a set of variables to either (a) group the original variables under a smaller number of dimensions or (b) allocate the units of analysis (e.g., individuals, brands, advertisements) to distinct groups.

INTERDEPENDENCE METHODS: IDENTIFYING STRUCTURES IN VARIABLES

Factor analysis

Factor analysis (also referred to as 'exploratory factor analysis' or 'common factor analysis' in the – often unintelligible – statistical literature) aims to describe a larger number of (metric) variables by means of a smaller set of variables (so-called factors) and to aid the substantive interpretation of the data. Factor analysis focuses on the *common* (i.e., shared) variance among the original variables and seeks to identify underlying dimensions (surprisingly known as **common factors**). To the extent that subsets among observed variables reflect a common 'core' (i.e., they are measuring the same underlying construct), the derived dimensions should be meaningful and interpretable. The original variables can then be described in terms of their common underlying dimensions. This technique is particularly useful in the context of measure development, as it enables an assessment of the dimensionality of multi-item scales. (For example, if a scale is truly unidimensional, all items comprising the scale should share a single common factor.) Needless to say, the term 'factor' as used in the current context has nothing to do with the use of the term in (M)AN(C)OVA, where it refers to a categorical independent variable; yes, this can indeed be confusing, but who wants a simple life anyway?

HINT 14.1	If you are interested in the common dimensions among a large number of (metric) variables, factor analysis is the correct technique for you.

Imagine that you are interested in measuring a concept such as advertising creativity (which is extremely important if you want to promote such highly exciting products as toilet paper,

paper clips, and frozen beetroot, to name but a few). You might look into the relevant literature, use theoretical knowledge, maybe do some qualitative research, and eventually come up with a set of items (say, 35 in all) that *supposedly* capture the notion of advertising creativity. However, the resulting number of items is quite large, and soon you realize that you are dealing with a rather complex data set. How easy would it be to link creativity to other variables? For instance, would it make sense to use all 35 variables in a multiple regression to see if creativity can predict, for example, advertising effectiveness? What about exploring differences in advertising creativity across different product categories? And, most importantly, do all these 35 items reflect the same concept (i.e., are they all indicators of advertising creativity)?

These questions make you wonder: 'Do I *really* need to use all these variables?' Maybe not all items are equally effective in measuring the construct of creativity. Maybe some items are not related to the other items, or maybe some of them are redundant and thus can be excluded altogether. Perhaps some items tap into different aspects of creativity and can therefore be seen as forming distinct groups. Faced with these seemingly insurmountable challenges (yes, we like dramatizing a situation), what you need to do is reduce the complexity and simplify the data set by placing the items into homogeneous sets and identifying any potentially problematic items. Factor analysis helps you to do this based on the idea that correlations between some variables exist because variables have something in common; they *reflect* (i.e., correspond to) the same **latent factor** (e.g., the same dimension of creativity). Note that these so-called latent factors are *unobserved* (i.e., we have not directly measured them) and only exist at a conceptual level.

To demonstrate the use of factor analysis, imagine that we have conducted a study in order to investigate consumers' preferences toward an organic tequila brand by asking 250 octogenarian consumers to signify their agreement/disagreement with the following highly profound statements (1 = strongly disagree, 7 = strongly agree):

1. Products of this brand are reliable
2. This brand is not expensive
3. Buying this brand is a good deal
4. This brand is a status symbol
5. This is a well-made brand
6. This is a high-quality brand
7. This brand makes me feel good
8. I enjoy using this brand
9. This brand is reasonably priced
10. This is an exciting brand
11. This is an exclusive brand
12. My friends admire this brand

We now want to see whether these 12 items can be grouped into distinct dimensions and then use the latter in subsequent analysis rather than the original items. But before we actually conduct the analysis, we need to investigate the 'factorability' of our data. To provide meaningful solutions, factor analysis requires that variables are sufficiently intercorrelated. We can check this by looking at the glorious Kaiser–Meyer–Olkin (KMO) **measure of sampling adequacy (MSA)**, which indicates the proportion of variance in the variables that might be

attributed to common variance (see Table 14.1). MSA ranges from 0 to 1, with higher values indicating that the observed data are suitable for factor analysis. Preferably, we look for values ≥ 0.80, and if values are less than 0.50, factor analysis probably does not make much sense. MSA values are provided for the overall model (see 'KMO and Bartlett's test', Table 14.1a) as well as for each individual variable; the latter are found on the diagonal of the 'Anti-Image Correlation' matrix (Table 14.1b). As with the overall MSA, values of individual MSAs should be above 0.50, while values off the diagonal should be small. The anti-image correlation matrix contains the negative partial correlation coefficients in the off-diagonal (i.e., the part of the variable that *cannot* be predicted by the other variables); thus, these values need to be small (implying high correlations between variables). Note that MSA values increase as the sample size and the number of variables included in the analysis increase.

Bartlett's **test of sphericity**, on the other hand, uses a χ^2 approximation to test the null hypothesis that the correlation matrix is an identity matrix (i.e., a matrix with ones on the main diagonal and zeros elsewhere). A non-significant result indicates that the variables are practically unrelated and therefore not suitable for factor analysis (of course, you only have yourself to blame for this – you should have generated a better pool of items in the first place!). In our example, all preliminary checks yield satisfactory results. The overall MSA is fairly good (KMO = 0.786) and Bartlett's test shows that the intercorrelations between variables are not zero ($\chi^2(66) = 2{,}354.406$, $p<0.001$). There are a couple of 'inflated' values in the off-diagonal of the anti-image matrix, but they are not really alarming.

Table 14.1a An example of common factor analysis: examining the factorability of the data, KMO and Bartlett's test

Kaiser–Meyer–Olkin Measure of Sampling Adequacy		.786
Bartlett's Test of Sphericity	Approx. Chi-Square	2,354.106
	df	66
	Sig.	.000

Table 14.1b An example of common factor analysis: examining the factorability of the data, anti-image matrices

Anti-image Correlation	Products of this brand are reliable	This brand is not expensive	Buying this brand is a good deal	This brand is a status symbol	This is a well-made brand	This is a high-quality brand	This brand makes me feel good	I enjoy using this brand	This brand is reasonably priced	This is an exciting brand	This is an exclusive brand	My friends admire this brand
Products of this brand are reliable	**.792**[a]	−.070	.140	.136	−.523	−.192	.104	−.052	−.028	−.018	−.159	−.093
This brand is not expensive	−.070	**.761**[a]	−.385	.086	.092	−.032	−.120	.077	−.487	.050	.025	−.180
Buying this brand is a good deal	.140	−.385	**.791**[a]	.055	.007	−.124	.039	−.078	−.405	.064	−.086	−.022
This brand is a status symbol	.136	.086	.055	**.801**[a]	−.148	−.137	.016	−.041	−.168	−.036	−.161	−.482
This is a well-made brand	−.523	.092	.007	−.148	**.749**[a]	−.496	−.112	.091	−.085	.009	.055	−.002
This is a high-quality brand	−.192	−.032	−.124	−.137	−.496	**.828**[a]	−.011	−.042	.111	.059	.030	−.037
This brand makes me feel good	.104	−.120	.039	.016	−.112	−.011	**.822**[a]	−.418	.131	−.384	.038	−.043
I enjoy using this brand	−.052	.077	−.078	−.041	.091	−.042	−.418	**.770**[a]	−.007	−.577	.024	.035
This brand is reasonably priced	−.028	−.487	−.405	−.168	−.085	.111	.131	−.007	**.752**[a]	−.049	.034	.138
This is an exciting brand	−.018	.050	.064	−.036	.009	.059	−.384	−.577	−.049	**.786**[a]	−.063	−.020
This is an exclusive brand	−.159	.025	−.086	−.161	.055	.030	.038	.024	.034	−.063	**.831**[a]	−.388
My friends admire this brand	−.093	−.180	−.022	−.482	−.002	−.037	−.043	.035	.138	−.020	−.388	**.775**[a]

Note: [a] Measures of Sampling Adequacy (MSA)

Having established that our data are suitable, we now turn to the main part of the factor analysis. This includes three somewhat painful-sounding steps: (a) **factor extraction**, which determines how many common factors are identifiable from the data, (b) **factor rotation**, which specifies the arrangement of the extracted factors and the relationships among them, and (c) **factor interpretation**, which allocates the individual variables to the different factors and identifies their common themes.

Factor extraction deals with how many factors best explain the covariation between the observed variables. In the interests of parsimony, we ideally aim at a small number of factors that explain a large amount of shared variation among the variables. There are several methods of factor extraction with highly impressive names, the most common ones being the **maximum-likelihood** and **principal axis factoring** approaches. These methods are based on different algorithms, but we really see no reason to get into the details of their differences (as usual, we direct strong-willed readers to the Further Reading). In our example, we opted for principal axis factoring (no, we are not going to tell you why!). The relevant output is shown in Table 14.2.

Table 14.2a An example of common factor analysis: factor extraction, communalities

	Initial	Extraction
Products of this brand are reliable	.645	.666
This brand is not expensive	.691	.774
Buying this brand is a good deal	.659	.739
This brand is a status symbol	.557	.586
This is a well-made brand	.735	.903
This is a high-quality brand	.650	.695
This brand makes me feel good	.797	.834
I enjoy using this brand	.838	.899
This brand is reasonably priced	.697	.782
This is an exciting brand	.836	.891
This is an exclusive brand	.435	.484
My friends admire this brand	.609	.833
Extraction Method: Principal Axis Factoring		

Table 14.2b An example of common factor analysis: factor extraction, total variance explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.725	31.043	31.043	3.447	28.728	28.728	2.619	21.828	21.828
2	3.455	28.795	59.838	3.283	27.357	56.085	2.304	19.201	41.030
3	1.791	14.922	74.760	1.590	13.247	69.331	2.265	18.872	59.902

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
4	1.052	8.770	83.530	.766	6.381	75.713	1.897	15.811	75.713
5	.496	4.132	87.662						
6	.328	2.730	90.392						
7	.314	2.616	93.008						
8	.227	1.893	94.900						
9	.193	1.609	96.509						
10	.186	1.547	98.057						
11	.131	1.092	99.148						
12	.102	.852	100.000						
Extraction Method: Principal Axis Factoring									

HINT 14.2 SPSS and other statistical packages also give you the option to 'force' the number of extracted factors, should you have *a priori* expectations about the number of dimensions underlying the data.

The imperial-sounding criterion used to determine the appropriate number of factors is called **Kaiser's criterion** and relies on the **eigenvalues**. The eigenvalues show the amount of common variance explained by the identified factors. Given that we conduct the factor analysis based on the correlation matrix, the variables are standardized, which means that they have a variance of 1. Thus, the total amount of variance in the data will equal the number of variables included in the analysis (here 12). That said, an eigenvalue of 1 means that the extracted factor explains as much variance as a single variable – this is no good, is it? Kaiser's criterion suggests that a factor should account for *at least* as much variance as a single variable (otherwise, it is no better than a single variable). Therefore, the number of extracted factors is determined by how many of them have eigenvalues greater than 1. In our case (see Table 14.2b), we clearly get a solution with four factors, which account for about 83.5% of the *total* variance in the data. Needless to say that a solution explaining less than 50% of the total variance is definitely not desirable. The columns in the second and third panel of Table 14.2b follow the same logic. The part labeled 'Extraction Sums of Squared Loadings' shows how much *shared* (as opposed to total) variance the extracted factors account for, while the 'Rotation Sums of Squared Loadings' simply redistributes the common variance explained following the rotation of the factors (to be discussed shortly). Note that the cumulative percentage of common variance accounted for is exactly the same before and after rotation and only the *relative* percentages corresponding to each factor change.

Going back to Table 14.2a, the initial **communalities** represent how much of the variance in one variable is explained by all remaining variables together. This is similar to the multiple R^2 we would have obtained had we regressed, for example, the first variable against all other variables. The communalities under the label 'Extraction' show the proportion of variance in

each variable that can be explained by the retained factors – that is, the variance explained had we regressed the extracted factors on a given variable. Factor analysis is based on the idea that variables consist of **common variance** and **unique variance**. With that in mind, variables should have more common than unique variance and, thus, communality values greater than 0.50 are desirable. Low values point toward potentially problematic items. In our example, all communalities meet this criterion, with only a single case being marginally low (0.484).

Table 14.3a An example of common factor analysis: rotation and interpretation, factor matrix[a]

	Factor			
	1	2	3	4
Products of this brand are reliable	.688	.212	–.304	.235
This brand is not expensive	.454	–.524	.533	.094
Buying this brand is a good deal	.413	–.524	.533	.103
This brand is a status symbol	.641	.301	.060	–.285
This is a well-made brand	.774	.253	–.333	.359
This is a high-quality brand	.726	.213	–.242	.252
This brand makes me feel good	–.115	.815	.376	.125
I enjoy using this brand	–.141	.825	.426	.132
This brand is reasonably priced	.426	–.563	.512	.146
This is an exciting brand	–.129	.838	.405	.092
This is an exclusive brand	.557	.214	.045	–.355
My friends admire this brand	.719	.283	.075	–.479
Extraction Method: Principal Axis Factoring				

Note: [a] 4 factors extracted. 16 iterations required.

Table 14.3b An example of common factor analysis: rotation and interpretation, rotated factor matrix[a]

	Factor			
	1	2	3	4
Products of this brand are reliable			.771	
This brand is not expensive		.859		
Buying this brand is a good deal		.845		
This brand is a status symbol				.681
This is a well-made brand			.921	
This is a high-quality brand			.783	
This brand makes me feel good	.894			
I enjoy using this brand	.935			
This brand is reasonably priced		.865		
This is an exciting brand	.923			

	Factor			
	1	2	3	4
This is an exclusive brand				.660
My friends admire this brand				.871
Extraction Method: Principal Axis Factoring Rotation Method: Varimax with Kaiser Normalization				

Note: [a] Rotation converged in 6 iterations.

Table 14.3 shows the **factor loadings** of the unrotated and rotated solutions. Factor loadings simply show the correlation between the variables and the factors. As such, they range from −1 to +1, with higher absolute values implying that certain variables load stronger on (are related to) a given factor. Most literature suggests that factor loadings above 0.50 or 0.40 are considered substantial and should therefore be retained in a factor. This, however, also depends on the sample size; factor loadings are correlations, and we know that even small correlations can be statistically significant in large samples. In general, factor analysis requires large samples (well above 100) to provide trustworthy results, so don't try to factor analyze 167 variables on a sample of 17 respondents! In our example, the sample size was 250, and because we wanted a *very* clear structure (we are perfectionists, remember?), we used 0.50 as a cut-off point.

Very often, interpreting the loadings right after the extraction of the factors is very difficult. Indeed, looking at the 'factor matrix' in Table 14.3a, it is hardly evident which variable is mostly relating to which factor. Factor rotation facilitates the interpretation of loadings by generating a simpler structure. It does so by rotating the factors in the geometric space in order to find the optimal arrangement that maximizes the loadings of variables on one factor while minimizing loadings on all other factors. There are two types of rotations, which are visually represented in Figure 14.1. **Orthogonal rotation** rotates the factors assuming that these are completely uncorrelated, while **oblique rotation** assumes that factors are allowed to correlate with each other. In our example, we followed the crowd and used the most popular orthogonal rotation method (the terrifying Varimax approach).

HINT 14.3 Apply an orthogonal rotation if you want the final factors to be uncorrelated and oblique rotation to allow correlations among the final factors.

Overall, clear factor structures are characterized by high and unique factor loadings. If a variable has a low factor loading or loads strongly on more than one factor, then it is not effective and should be dropped. Going back to Table 14.3, note that we have asked our software to exclude any loading that does not meet our specified criterion (i.e., a minimum loading > 0.50) from the results. This a very useful function to reduce visual clutter and quickly reveal the resulting structure (particularly when the original number of variables is large). From the 'rotated factor matrix' (Table 14.3b), it can be seen that the loadings are high enough, and there are no cross-loadings.

We can now examine the groups of variables and identify their common underlying themes. This is what factor interpretation is all about. The factor structure suggests four distinct product dimensions that seem to capture perceptions of brand affect (Factor 1), price perceptions (Factor 2), perceived quality (Factor 3), and brand prestige (Factor 4). The variables

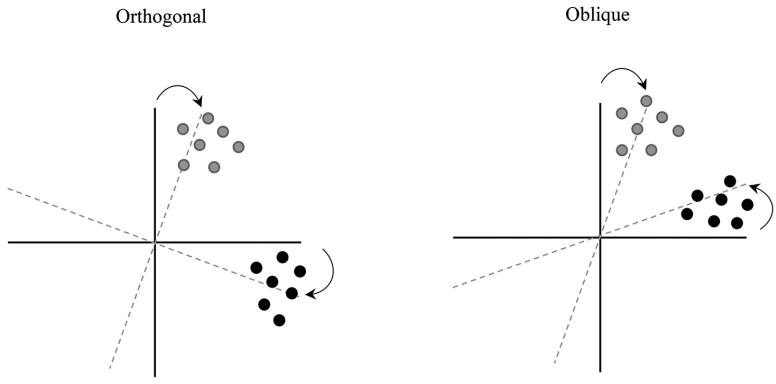


Figure 14.1 Visual representation of factor rotation

comprising each factor could now be aggregated or averaged to generate distinct composite scales, which can be used in further analysis (e.g., as predictors to see their relative influence on consumer preferences). The internal consistency of the resulting scales could also be checked using Cronbach's alpha (α) coefficient as described in Chapter 3.

Principal component analysis (PCA)

Principal component analysis is very similar to common factor analysis but focuses on the *total* variance and seeks to reduce the original set of variables into a smaller set of composite variables (called **principal components**), which are uncorrelated to one another. Each principal component is formed by linearly combining the observed variables, the key objective being to explain as much of the total variance in the data as possible by few principal components. Thus, the first principal component extracted accounts for as much of the variability in the data as possible, and then the analysis proceeds by constructing a subsequent component that explains as much of the *remaining* variability as possible, and so on.

HINT 14.4	If you only want to reduce a large number of (metric) variables into a smaller set of uncorrelated composite variables, principal component analysis is the correct technique for you.

As an example of principal component analysis, consider 20 variables cast into a five-point Likert scale describing attitudes toward pet iguanas. After running a PCA, you may find that the computer program has reduced the original 20 variables into, for example, three components, which together explain 70% of the total variance in the original variables. By looking at the individual variables that are strongly associated with each component, you may find that the first captures attitudes toward 'companionship', the second attitudes toward 'iguana hygiene', and the third attitudes toward 'iguana obedience'.

WARNING 14.1 Do not confuse principal component analysis with (common) factor analysis. They are different techniques and have different purposes.

The steps needed to run a PCA are (thankfully!) exactly the same as for common factor analysis (i.e., extraction, rotation, and interpretation). The output is also very similar, the only difference being that instead of common factors, we now obtain principal components. For instance, instead of a 'factor matrix' and a 'rotated factor matrix', SPSS produces a 'component matrix' and a 'rotated component matrix' (the interpretation of loadings is exactly the same). That said, the statistical model underlying factor analysis and principal components is quite different. This difference is illustrated beyond reasonable doubt in Figure 14.2. In common factor analysis, observed variables consist of a common latent factor and some unexplained unique variance (variable = loading × factor + error). In contrast, PCA assumes that all observed variables together compose a component (component = sum of loadings × variables). In practical terms, the key difference between PCA and common factor analysis is that, in the former case, the sole aim is to reduce the original set of variables into a smaller set of composite variables (components); it is simply a **data reduction technique** that makes no assumptions regarding the underlying structure of the data. In contrast, as previously discussed, common factor analysis focuses explicitly on the interrelationships among the original variables and seeks to describe them in terms of common underlying dimensions; thus, the focus is on explaining the patterns of relationships among the original variables by means of a factor structure. Both types of analyses are widely used; the choice between them is governed by the specific needs of the researcher (i.e., measure development vs. data reduction). Tossing a coin is also an alternative, though not recommended.

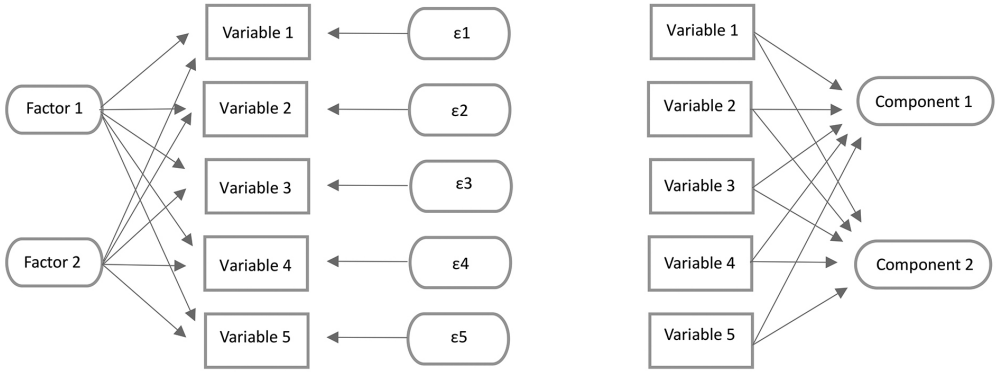


Figure 14.2 Statistical models underlying common factor analysis and principal component analysis

HINT 14.5 If you want to group objects (e.g., individuals, products, advertisements) based on their similarity, cluster analysis is the correct technique for you.

INTERDEPENDENCE METHODS: IDENTIFYING STRUCTURES IN OBJECTS

Cluster analysis

Cluster analysis seeks to group the *objects* (e.g., individuals, products, advertisements) for which measurements have been obtained. In essence, cluster analysis is a technique for grouping cases based on the similarity of responses on several variables of interest. By looking at the similarities and differences between the scores on the variables, each case is grouped with others having similar scores into what are known as **clusters**. The latter are mutually exclusive groupings formed in such a way that objects within the same cluster are homogeneous (similar) with each other and heterogeneous (dissimilar) with objects in other clusters.

Cluster analysis is often used in market segmentation studies, where segments are formed on the basis of customer characteristics (e.g., income, number of children, age, shoe size, and past purchase record of live chickens) or according to consumers' product perceptions, attribute preferences, and so on. By appropriately identifying 'similar' consumers on the basis of these variables, distinct groups (i.e., clusters) are formed, which can then be targeted with appropriate marketing strategies. Cluster analysis is particularly helpful in profiling and understanding people's behavior in that individuals with similar interests and attitudes will typically belong to the same cluster (e.g., based on their music preferences, grandmothers preferring Croatian heavy metal music are likely to form one group while grandmothers who are into Belgian hip-hop are likely to form another). Market researchers also use this technique to cluster products and brands (or even companies) in the market and identify niche markets and opportunities for new product offerings (e.g., such as identifying a market gap for single-wheel motorcycles).

WARNING 14.2 Scaling differences between input variables may affect the cluster solution. Consider rescaling as an option prior to running the cluster analysis.

Most clustering procedures can be classified into **hierarchical clustering** and **non-hierarchical clustering** approaches. Hierarchical procedures merge cases one at a time in sequential steps to derive homogeneous clusters. Non-hierarchical procedures start with a predetermined number of clusters and then assign each case to the nearest cluster. This implies that the researcher is fairly confident about the number of clusters in the final solution before running the analysis. If this is not the case, the analysis must be run again and again in order to derive and compare each possible solution. Hierarchical clustering offers the advantage that we can compare an increasing number of clusters in the same analysis. Hierarchical clustering is the most commonly used approach and can be used with binary, ordinal, and metric (interval and ratio) data. Having said that, it is generally *not* advisable to combine variables with different levels of measurements as scaling differences may affect the cluster solution (there is always a snag, isn't there?). If the variables used as input are very different, you had better rescale

them. SPSS offers a number of alternative standardization methods (e.g., *z*-scores) that can be automatically implemented through the hierarchical cluster analysis menu.

Hierarchical cluster analysis can be performed in two main ways. **Agglomerative clustering** starts with as many clusters as the number of cases (*n*) and ends with a single big/massive/enormous/humongous cluster. In the first step, it basically allocates all cases to their own individual clusters and, in sequential steps, merges similar cases (and then clusters) to superordinate clusters until every object falls under one cluster. In contrast, **divisive clustering** starts with one all-inclusive cluster (containing all cases) and ends with *n* clusters. Thus, all objects are initially members of one (very) large cluster, and gradually they are separated into groups of clusters until each case represents an individual cluster. The divisive method is more computationally demanding and does not really offer any distinct advantages over the agglomerative approach, so we will stick with the latter in the rest of this chapter.

In conducting an (agglomerative) cluster analysis, one first needs to select the variables upon which the clustering will be based, then select a **distance measure** or **similarity measure**, and, finally, use a criterion to determine the final number of clusters. Note that although both options exist, distance measures – as opposed to similarity measures – are usually preferred (by law, the reasons for such preference can only be revealed to people with at least two PhD degrees in statistics – sorry!). There are many different measures available for assessing the distance between cases but the default and most commonly used option is the **squared Euclidean distance**, which basically captures the sum of squared differences between pairs of variables for each case. Obviously, the smaller this distance, the more similar the cases, and vice versa.

To group cases together, one can choose among several **linkage measures**. Some measures are based on the distance between two clusters based on the smallest (e.g., nearest neighbor or single linkage) or largest distance (e.g., furthest neighbor or complete linkage) between pairs of cases belonging to different clusters. Others, such as the average (or between-group) linkage measure, average all distances between pairs of cases from different clusters. The logic behind all these approaches is similar in that they all begin with as many clusters as there are cases and, by successively forming clusters based on the criterion specified, end up with one superordinate cluster containing all cases.

To see how cluster analysis works in practice, imagine that we want to profile 30 consumers based on how important they consider convenience, price, quality, and product support to be for their purchase decisions of recyclable toothpicks. The idea is that if consumers are driven by the same motivations, the pattern of their responses will look similar across these measures. In this example, we applied a hierarchical, agglomerative cluster analysis, using the squared Euclidean distance measure and an average (between-groups) linkage clustering method. Note that all variables to be included in the cluster analysis have been measured on five-point interval scales, so we do not have to standardize them. The output is provided in Table 14.4. While selecting different distance and linkage methods may result in a different final cluster solution, the output is interpreted in a similar way.

Table 14.4 An example of cluster analysis: agglomeration schedule, average linkage (between groups)

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	15	30	.000	0	0	24
2	14	29	.000	0	0	17
3	13	28	.000	0	0	18
4	12	27	.000	0	0	18
5	11	26	.000	0	0	17
6	10	25	.000	0	0	22
7	9	24	.000	0	0	22
8	8	23	.000	0	0	23
9	21	22	.000	0	0	10
10	7	21	.000	0	9	16
11	5	20	.000	0	0	20
12	4	19	.000	0	0	21
13	3	18	.000	0	0	20
14	2	17	.000	0	0	19
15	1	16	.000	0	0	19
16	6	7	.000	0	10	23
17	11	14	.044	5	2	24
18	12	13	.044	4	3	25
19	1	2	.044	15	14	21
20	3	5	.150	13	11	26
21	1	4	.171	19	12	26
22	9	10	.193	7	6	27
23	6	8	.193	16	8	27
24	11	15	.391	17	1	25
25	11	12	1.216	24	18	29
26	1	3	1.394	21	20	28
27	6	9	1.445	23	22	28
28	1	6	5.728	26	27	29
29	1	11	14.010	28	25	0

The **agglomeration schedule** in Table 14.4 shows the sequence of how cases (here the 30 toothpick buyers) are clustered together. Each stage (i.e., row) combines pairs of cases in a cluster, using an algorithm determined by the squared Euclidean distance and the between-groups linkage criterion (see 'Cluster Combined'). The stages of the agglomeration schedule include all cluster formations until only one cluster remains after the last stage. Thus,

the number of stages is one fewer than the number of cases being clustered. For example, in Stage 1, Cluster 15 and Cluster 30 are combined to form a joint cluster, which, by default, is labeled after the first cluster, namely '15'. Note that in the beginning, each case is a cluster on its own, so Cluster 15 and Cluster 30 basically consist of Case 15 and Case 30. This is why the corresponding cells in the panel 'Stage Cluster First Appears' include zero values. In fact, the columns in 'Stage Cluster First Appears' show the stage in which the cluster was formed with values of zero, indicating that the corresponding cluster was a single case prior to this stage. Notice that zero values completely disappear at Stage 17, from which point onward clusters consist of more than a single case. Thus, Stage 17 tells us that Cluster 1 (labeled '11') was first formed in Stage 5, and Cluster 2 (labeled '14') was formed in Stage 2 (see 'Cluster Combined'). To confuse you even further, remember that cluster labeling is based on the *first* column in the 'Cluster Combined' panel.

The **cluster coefficients** at each stage represent the squared Euclidean distance between the two clusters being combined. The pairs with the *minimum* distance (coefficient) are first clustered together. Indeed, over the first several stages, the coefficients are very small, indicating that the combined clusters are pretty similar. Note that, by default, SPSS reports three decimal points (unless we specify otherwise); the actual values are not necessarily exactly zero. As the agglomeration schedule progresses, the coefficients increase, implying that the clusters being combined are increasingly more heterogeneous compared to previous stages. Through the agglomeration schedule, we can thus assess the point beyond which the clusters being combined are considered noticeably different enough to form a homogeneous group; this is reflected in a big 'jump' in the coefficient values between consecutive stages. A large difference implies that the clusters being merged from that point onward are rather dissimilar and, thus, it would be a good idea to stop the clustering process before things get out of hand. In Table 14.4, we can observe that the first big jump takes place after Stage 27. The coefficient from Stage 27 to Stage 28 almost triples in magnitude, corresponding to a difference of 4.283 points. To the trained (as well as untrained) eye, this suggests stopping the clustering procedure at Stage 27, which is associated with a three-cluster solution. In general, the number of clusters equals the sample size (n) minus the stopping stage (here 27).

In large samples, it can become very cumbersome to compare the relevant coefficient differences between stages. To overcome this problem, we can examine the coefficients and stages in a **scree plot**, as shown in Figure 14.3. This is simply a visual representation of the agglomeration schedule, which makes it easier to detect where the first big difference in coefficients is located. The 'elbow' of the scree plot (the point at which the slope of the line changes considerably) indicates the potential stopping point. Looking at the scree plot in our example, the line 'breaks' (the elbow) clearly at Stage 27.

A hierarchical cluster analysis is perhaps best illustrated using a **dendrogram**, which is shown in Figure 14.4. The vertical lines represent the stages of the agglomeration schedule. For example, at the first vertical level of the dendrogram, we see that the first cluster is between Cases 15 and 30, the next one is between 14 and 29, and so on (you can double-check with the agglomeration schedule in Table 14.4). The vertical lines also indicate the distance between two combined clusters ('Rescaled Distance Cluster Combine') based on the linkage method one has selected. The more heterogeneous the clusters being merged, the farther to the right

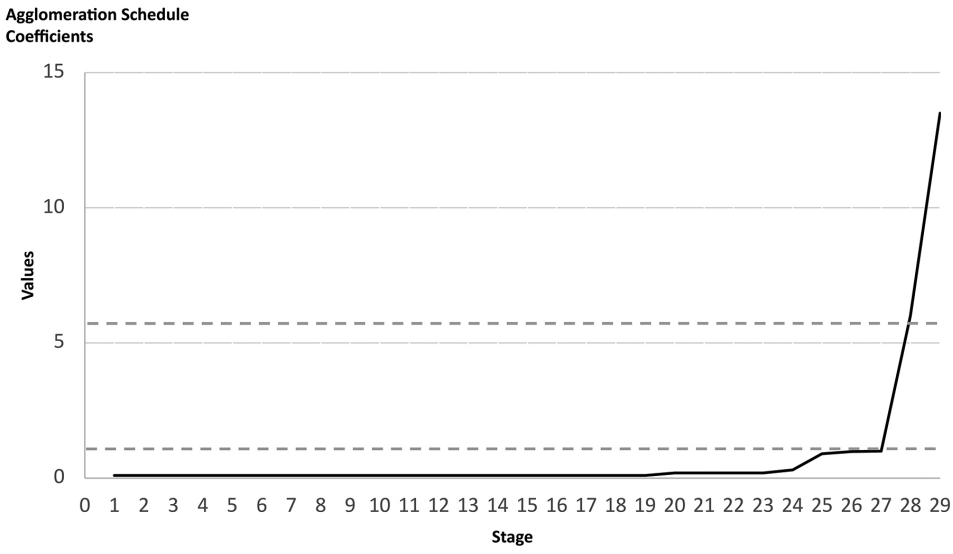


Figure 14.3 Scree plot of agglomeration coefficients

the vertical line linking them will be located. The horizontal lines, on the other hand, connect all cases that belong to a cluster. They also show the relative distance between clusters, with longer horizontal lines indicating more dissimilarity between the two clusters being merged. The length of the horizontal lines can also be used as a reference to identify the optimal stopping point of the clustering procedure. In general, when the vertical and horizontal lines are close to one another, there is high homogeneity in the formed clusters. Before the interconnecting lines start to spread noticeably, we should stop the clustering process and have a cup of tea (see the dashed vertical line in Figure 14.4).

As you have probably realized by now, there is no 'formal' stopping rule for hierarchical cluster analysis (and this can often become a problem with this technique). Thus, the best way to determine the final number of clusters is to integrate all information provided in the output. In our example, this is fairly easy as the agglomeration schedule, the scree plot, and the dendrogram converge with regards to the final number of clusters. In other applications, however, things may not be very clear and choosing the number of clusters may be more difficult (don't ask us how to do it, we don't know either!).

Once we have the final number of clusters, we can ask SPSS to save a 'single solution' setting the number of clusters to three (we can also specify a range of clusters, but this makes sense only if we are unsure and we want to compare alternative solutions). This will create a new variable in SPSS indicating the cluster to which each case is assigned. We can subsequently use this categorical variable to examine how consumers in the different clusters score with regard to the initial clustering variables and thus (with a bit of luck) make sense of the clustering solution. To illustrate, Table 14.5 shows the mean values of the three customer clusters according to the importance placed on convenience, price, quality, and product support. We first see that our toothpick buyers are evenly distributed across the three clusters (clearly, this is a coinci-

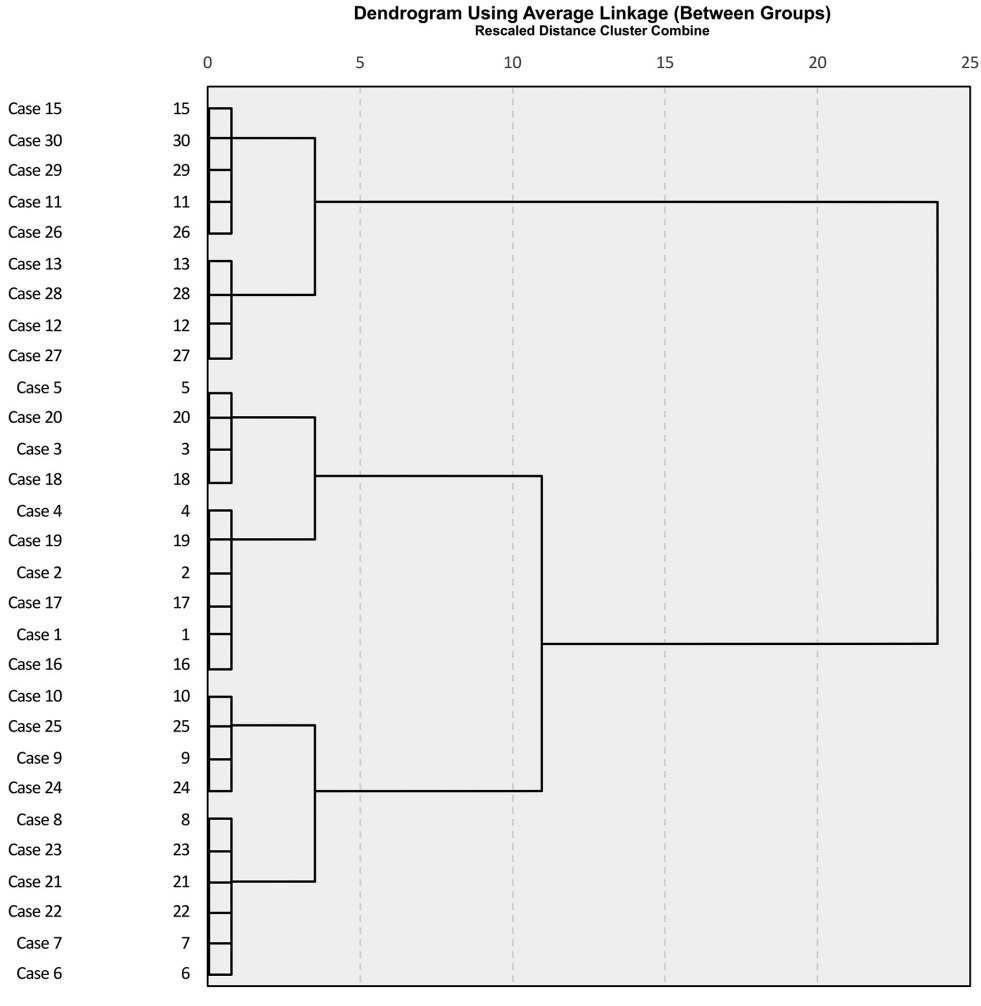


Figure 14.4 Dendrogram (dashed line indicates stopping location)

dence, and you should not always expect it to happen unless you are as lucky as we are). Also, looking at the pattern of the means, we can discern a group of buyers that are primarily driven by price (Cluster A), a group that is heavily influenced by convenience (Cluster B), and a final group that is mainly concerned with performance-related features (Cluster C). Having identified the common themes associated with our buyer clusters, we can devise different marketing mix strategies to target each one of them (e.g., offer quantity discounts to the price-conscious buyers of Cluster A or demonstrate the virtually unbreakable nature of the toothpick to the performance-oriented Cluster C).

Table 14.5 Mean scores and standard deviations (in parentheses) on the variables of interest across the three final clusters

Cluster	n	Convenience	Price	Quality	Product support
A	10	3.19 (0.81)	4.20 (0.89)	3.81 (0.92)	2.99 (0.89)
B	10	4.54 (0.93)	4.01 (0.81)	3.65 (0.81)	3.86 (0.90)
C	10	3.05 (0.93)	3.70 (0.82)	4.25 (0.65)	4.29 (0.64)

As we mentioned at the beginning of this section, non-hierarchical approaches (e.g., *k*-means clustering) make more sense if the researcher already has an informed idea about the potential number of clusters and perhaps wants to compare how solutions with a different number of assumed clusters perform. This is not very often the case, which is why we focused on the hierarchical approach, which does not require us to pre-specify the number of clusters. Having said that, researchers may often use non-hierarchical cluster analysis to further corroborate the solution of a hierarchical method. Thus, in our example – and assuming we have nothing better to do – we could run such an analysis by *a priori* fixing the number of clusters to three and seeing whether results converge with that provided by the hierarchical cluster solution in Table 14.4. SPSS also includes a **two-step cluster analysis** option, which combines the previous two approaches by first running pre-clustering and then running a hierarchical method. Feel free to try this in your spare time.

WARNING 14.3 Approach cluster analysis with (great) caution, and always remember that depending on how you run it, you may end up with (very) different results.

Before we conclude, we need to bring to your attention a number of things you should be aware of when you employ data-driven (as opposed to theory-driven) techniques, such as cluster analysis. First, as mentioned earlier, there are several variations of cluster analysis, and each one of them will likely end up with different results. With that in mind, it is important to consider the research context of the analysis and carefully select the method that is best suited for the purpose of the study. Moreover, given that clustering is based on (dis)similarity between cases and clusters, changes in the data, such as excluding outlier values or including additional cases, may produce quite different solutions. Remember that there are no formal criteria (e.g., significance tests) one can apply to arrive at the final solution. As such, there is *a lot* of subjectivity involved in these types of analysis, and subsequent interpretation should be performed very conscientiously. Finally, bear in mind cluster analysis will *always* create clusters, regardless of whether any structure exists in reality!

SUMMARY

In this chapter, we had a good look at interdependence methods of multivariate analysis. We kicked off by looking at factor analysis as a method for identifying common themes among a set of variables and exploring their underlying structure. Next, we discussed principal component analysis as a method for data reduction and highlighted its key dif-

ferences from factor analysis. Finally, we introduced cluster analysis as a method for grouping cases and illustrated its potential uses. Unbelievable as it may seem, cluster analysis is the very last method included in this book, so if you are reading this, you have almost completed your journey. Why almost? The answer is revealed in the next and final chapter.

QUESTIONS AND PROBLEMS

1. What is a communality?
2. How does one decide on the number of factors to retain following a factor analysis?
3. What is the meaning of explained 'cumulative common variance'?
4. Why are high-factor cross-loadings undesirable?
5. What is the purpose of rotation in factor analysis?
6. How would you decide between using orthogonal and oblique factor rotation?
7. What, if any, is the difference between common factor analysis and principal component analysis?
8. Which multivariate technique is widely used to classify people or objects into a smaller number of meaningful groups? When would such a technique be useful?
9. How would you go about choosing between a hierarchical versus non-hierarchical clustering procedure?
10. Which one of the three co-authors of this book should be the lead character in the new multi-million-dollar Netflix series on non-hierarchical clustering?

FURTHER READING

Everitt, B. S., Landau S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*, 5th edition. London: John Wiley & Sons Inc. The classic on the topic. Anyone who is doing cluster analysis and hasn't read it should be arrested.

Holmes Finch, W. (2019). *Exploratory Factor Analysis*. London: Sage. Short and sweet, it tells you all you want to know about factor analysis.

Verma, V. P. (2015). *Data Analysis in Management with SPSS Software*. New York: Springer. This covers not only the techniques discussed in this chapter but also just about every other technique we've covered. We only draw it to your attention now as it is too late to buy it instead of our book!

15
It's all over … or is it?

Congratulations! You finally completed your path-breaking analysis of secret Internet usage in Amish communities (or whatever research project you have been working on). You squeezed every possible bit of information out of your data set. You compared your findings with your initial objectives, bearing in mind all the words of infinite wisdom contained in Chapter 5. You produced enough computer printouts to feel personally responsible for the deforestation of Latin America. Yes, you have truly *finished* your analysis! Great achievement! Go and celebrate! But not for too long, since an important piece of work is still ahead of you – the effective *communication* of your results.

Most researchers are, quite understandably, proud of their work. However, they are usually so closely involved with every intricate detail of their analysis that they sometimes find it difficult to keep the 'big picture' in mind. Consequently, they are at risk of confusing **statistical significance** with **theoretical importance** and/or **managerial importance**, finding it hard to write punchy and relevant **research reports**, and having problems with making effective **oral presentations**. In what follows we give you some hints on how to cope with the above; after all, what's the point of doing a fantastic analysis if nobody appreciates it?

WARNING 15.1 In reporting your findings, focus on their theoretical importance and managerial relevance, not simply their significance.

THE WRITTEN RESEARCH REPORT

Before you write down the first line, you need to think about the requirements of your *audience*. If your research is part of your research thesis and you are presenting the findings to your supervisor, the sophistication and degree of detail required are likely to be relatively high. But while the technical sophistication of your audience determines the maximum depth of your report, you should not automatically assume that all readers who understand the technical details will also be *interested* in them or, indeed, have the time (or even the inclination) to read them. This is why you must make a distinction between the statistical characteristics of the results (e.g., whether they are significant or not) and their theoretical and/or managerial substance. Not every single result that is statistically significant is also important from either a theoretical or a practical point of view (see also Chapter 9, particularly Warning 9.7). Be very aware of this distinction and focus on those results that impact management decision-making

or, in an academic context, are important for theory development purposes. Also, do not forget that it is often more important that a particular estimate or relationship is *not* significant. Thus, do not fall into the trap of reporting only findings that are significant; also discuss non-significant results that provide managerial or theoretical insights.

HINT 15.1 Try to understand the specific requirements of your audience before you start writing your report. If you have multiple audiences, prepare separate reports or, at the very least, split your documents into a technical and a non-technical section.

Whenever possible, you should try to find out which style and format your target audience prefers *before* you start writing up your results. Most companies like very short and succinct reports that focus on the managerial implications of the research findings. A PhD thesis, on the other hand, has to discuss methodological details in depth. Consequently, when you are in a situation where you have to please multiple audiences (for example, your thesis supervisor and a sponsoring company), it is best to produce separate reports for the different audiences you plan to address. Although this means extra work (and may involve getting up at the crack of dawn, i.e., not your usual 11.00 a.m. leisurely rise), the effort usually results in documents that are far more useful to the different types of readers.

HINT 15.2 Always include a summary at the beginning of your report. This should be self-contained and highlight the key points in the report.

Most reports should be preceded by a short **summary** (frequently referred to as an 'executive summary'). In a corporate environment, this is often the only part that is read by all of your readers (people have *work* to do, you know!). Therefore, the summary may be the most important part of your research report and you need to ensure that it is 'polished'. A good summary has to be able to stand on its own; that is, it has to make sense without having to read the rest of the report (all 2 300 pages of it).

Following the summary, your report should open with an **introduction**. Here, the main aim is to familiarize the reader with the purpose and scope of the research and place the study into context (provide some historical background, indicate whether similar studies have been conducted previously, discuss past findings, etc.). In marketing terms, the introduction is a *positioning* exercise, in which the reader needs to be convinced of the need for and importance of your study.

In academic studies, the introduction is normally followed by a separate **literature review**, which extends the discussion of related research. If the pertinent literature is extensive, it is helpful to locate and refer to review articles. (In such circumstances, it may also be useful to construct summary tables of the findings of previous studies.)

The main body of the research report comprises details on methods, results, and limitations. Again, there is a difference between academic reports and management reports. Academic reports almost always contain a separate **methodology** chapter. This would include a precise definition of the population that was studied, as well as detailed information on the sample size, sampling procedure, response rates, research instrument (e.g., questionnaire or interview

schedule), and the way the variables were measured. In a management report, information on the methodology is usually much less detailed, often included in the discussion of the results, and frequently relegated to an appendix.

HINT 15.3	The introduction section should state the aims and the scope of the research and position it against similar/previous studies.

The **results** section (also often called 'analysis' or 'findings') is the place to stun your readers with all your nice tables, three-dimensional color figures, and compelling arguments (at your discretion, you may also include poems or dried flowers). Treat the results section as a *story* that needs to be unfolded in a *logical* sequence (i.e., do not first discuss the relationship between variable *X* and variable *Y* and *then* provide descriptive statistics on X and Y). Sometimes you may want to use sub-headings to provide the reader with signposts, indicating how your analysis has been structured. In some academic papers, you will encounter a results section followed by a discussion section. This usually means that the results section depicts key results in a succinct, factual manner, whereas the following discussion section, as the name suggests, provides a discussion of the meaning of the results in terms of theory development, managerial implications, or methodological advances. Whether you combine the reporting of results with the discussion of results in a *single* section or add a *separate* discussion section is a question of personal preference and depends, once again, on the audience or the requirements of the journal in which you aim to publish your study.

It is often suggested that the presentation of the results should be followed by a section on the study's **limitations**. This should draw the reader's attention to possible biases in the findings (for example, the exclusive sampling of organic vegetable farmers in a survey aimed at identifying the most influential opera production of the year) and clarify how far the findings can be generalized. It is usually more advantageous to be open and frank about the limitations inherent in a research study than to leave them to the reader to discover. In fact, an open admission of research limitations will demonstrate that you are honest and able to distance yourself from your work. Thus, readers may actually think more highly of the study.

An alternative approach is to discuss the limitations of the study either in the methodology section or in the conclusion section. In the latter case, specific limitations can be linked to a discussion of **future research needs**. Remember that any investigation worth its salt always raises more questions than it answers (you do not want to lose your job as an analyst or academic, do you?).

The **conclusions and recommendations** section rounds off the research report. A clear link should always be established between the original study objectives (as described in the introduction) and the respective conclusions. Sometimes you may even (briefly) restate the original objectives of the study and then present the conclusions relating to each individual objective. In a commercial setting, studies are initiated to help make specific decisions, and therefore your conclusions should demonstrate how the results of your analysis will help make those decisions. Similarly, in an academic environment, research is usually conducted to build theory. Here, it is crucial to show the implications of your findings for theory building.

HINT 15.4 Always demonstrate how the results of your analysis impact the decision(s) management needs to make or, in an academic setting, contribute to theory building. Make sure that what you 'promise' in the introduction section tallies with what you 'deliver' in the conclusion section!

Unfortunately, there is no one best way of structuring a research report; the nature of the audience, the research topic, space limitations, and so on will all have an impact on its final form. However, if only to illustrate different styles, Table 15.1 presents two 'generic' reports. The first is more suitable for an academic environment, while the second is more compatible with a company setting. These 'generic' reports are followed by a more 'action-oriented' report on a specific topic. Note that we have chosen relatively 'dry' headings for the first two examples. This is not to suggest that these types of reports cannot use more imaginative headings, but only to illustrate their structure more clearly.

THE ORAL PRESENTATION

In addition to one or more written reports, you may also have to give an oral report of your research. In fact, oral presentations are quite common at the **research proposal** stage and in the form of **progress reports** at various stages during lengthy projects. Depending on the specific corporate environment or academic setting, the formal oral report at the conclusion of your project may precede or follow the distribution of the written report(s). In any event, it is important to recognize that many senior managers (and even some busy academics) will judge the quality and usefulness of your work virtually exclusively on the basis of your oral presentation – perhaps supported by a quick glance at the summary and/or conclusions section of your written report over a cup of coffee. While this insight can be quite disheartening (particularly if you have been working full steam for five years to complete your study), it follows that oral reports are of the utmost importance and it is in your interest that they are done right.

So what should you do to make your oral presentation as effective as possible? The first step is no different from preparing your written report, namely that you get as much information about your likely audience as possible. Find out about your audience's interest, likely involvement, and technical (statistical) sophistication level. Find out about the possible setting. Will your presentation be held in a small conference room or in a large auditorium? Ask yourself which parts of your research are likely to be the most interesting, relevant, and important from *the audience's point of view*.

WARNING 15.2 Never underestimate the importance of an oral research report.

Next, you should develop a clear objective for your presentation. Usually, you want to persuade the audience that your findings do or do not support certain decisions/theories. Consequently, you need to carefully select the **key arguments** that best support your conclusions. The emphasis here is on *selection*; do not try to present everything you have done. A far

Table 15.1 Examples of research reports

Generic academic report
1. Title page
2. Table of contents
3. Summary
4. Introduction
5. Literature review
6. Methodology
7. Findings
8. Conclusions and recommendations
9. Limitations and future research
10. Bibliography
11. Appendices
Generic company report
1. Title page
2. Table of contents
3. Executive summary
4. Introduction
5. Body (results)
6. Conclusions and recommendations
7. Appendices (including methodology)
Action-oriented report
1. The Amish Internet secrets (title)
2. The shocking facts (summary)
3. Why the world needs to know (introduction – positioning)
4. How we revealed the truth (methodology)
5. Inside the secret Amish computer camp (body)
6. What the Amish download (body)
7. Amish software preferences (body)
8. An untapped market opportunity (conclusion)
9. How we can reach the New-Age Amish (recommendations)

more effective approach is to focus on no more than three key arguments. (People forget and get bored quickly.) For each key argument, you should subsequently consider which evidence (finding) is most suitable to support it. Try to illustrate the point with audiovisual material, draw comparisons, give examples, or use authoritative quotes; above all, *keep your audience interested*.

HINT 15.5 In preparing an oral presentation, put yourself in the shoes of your audience.

In contrast to written reports, there are no obvious headings and sub-headings in an oral presentation. It is therefore extremely important that you set clear **markers** to ensure that the audience notices when you have finished one point and are moving on to the next. Usually, these are short but important 'linking' sentences such as 'Having discussed the organization of the secret Amish computer camp, I shall now tell you which software they are using.' All too often, markers are overlooked by the presenter. This results in the audience getting lost and wondering what a particular point has to do with what you said before.

WARNING 15.3 Do not forget to use appropriate 'markers' throughout your presentation. Otherwise, you risk losing your audience.

With your objectives firmly established and the key points supporting your objectives developed, your presentation can now take shape. It is always helpful to remember the old adage: 'Tell them what you are going to tell them (introduction), tell them (core presentation), and tell them what you told them (summary).' Thus, you can think about the introduction as providing a preview and the summary as providing a review. The only other two points to focus on now are (a) finding an attention-grabbing opener for your presentation (for example, a quote or an anecdote) and (b) ensuring that the audience clearly understands what actions flow from your report. Table 15.2 shows a suggested structure for an oral presentation.

WARNING 15.4 Do not read your report when making an oral presentation – it is a sure way of boring/alienating your audience.

Having finalized your oral report, it is imperative that you practice it a few times. Ideally, you should learn as much of your oral report by heart as possible so that you can *talk to your audience*; at a minimum, you should know the beginning and the end. Under no circumstances should you *read* your report. (After all, your audience can probably read just as well as you can!) If you must, have a few notes on index cards for consultation but refrain from using a full-text copy of your presentation.

Before leaving the discussion of oral presentation, we should emphasize that **visual aids** are essential for effective communication. Use widely available presentation software like PowerPoint, Prezi, Keynote, and so on. However, beware their pitfalls and temptations. Most of these packages have an abundance of options, which allow you to animate your presentations, embed videos, or pan and zoom from one page to the next. This is all very well, but resist being carried away by the near-endless choices you have at your disposal. If every new word enters your presentation only after zigzagging across the screen, changing its size and color several times in the process, and then 'lands' to the tune of Bach's Toccata and Fugue in D Minor, the novelty wears off quickly. Thus, don't go wild with animations; it is your findings that should be the center of attention, not the animations! Also, do not go overboard with the number of slides you are presenting. If you click through your slides at a rate of one slide per minute, even if your individual slides are appealing, you will inevitably lose your audience. Less is more! Always remember the wary question: 'Do you have a PowerPoint presentation, or do you have something to say?' Finally, do not overload individual slides. While the information put on a single slide obviously depends on the circumstances (e.g., setting of the presentation,

Table 15.2 Suggested structure for an oral presentation

1.	Introduction
	(a) Attention-grabbing opener
	(b) State objectives of your presentation
	(c) Tell them what you are going to tell them
2.	Core presentation (tell them)
	(a) Supporting evidence no. 1
	(b) Marker to clarify the structure
	(c) Supporting evidence no. 2
	(d) Marker to clarify the structure
	(e) Supporting evidence no. 3
•	
•	
•	
3.	Summary
	(a) Review: tell them what you told them
	(b) Indicated actions: tell them what they need to do

size of the room, size of the audience, etc.), you should usually aim for short bullets rather than include running text.

HINT 15.6 Nicely prepared audiovisual aids help bring your presentation to life. However, do not go wild with too many animations.

A good oral presentation is as much of an art as it is a science. In the framework of this book, we can obviously only scratch the surface of the topic. If you think you need more help, we strongly advise you to buy one of the numerous books on presentation skills and/or watch relevant videos online (these also usually include guidelines on the preparation of visual aids, voice projection, eye contact, etc.). You can also join public speaking classes at your local adult education college or join Toastmasters, a group that meets specifically to develop the public speaking skills of its members. (Giving impromptu talks on your research at your local supermarket or sauna club is also an alternative, albeit less conventional.)

Well, this is it! Nothing more. No more hints. No more warnings. No more statistics. And no more jokes! We profusely apologize to all readers and book reviewers who found our jokes immature, politically incorrect, or downright silly. We also apologize for having a go at statisticians. (We just couldn't resist!) Finally, we apologize to our editor for all the hassle we caused.

You are now on your own. Hopefully we have convinced you that you do not need a PhD in mathematical statistics to understand and conduct sensible data analysis. Perhaps we have even managed to demonstrate that learning about data analysis can be fun. But, above all, we hope that we have been successful in *Taking the Fear Out of Data Analysis*. Take care.

SUMMARY

This final chapter demonstrated that successful data analysis should not end with terabits of computer files and piles of printouts that would make you an ecological outlaw. Only a professionally produced research report and a punchy and eloquent oral presentation will ensure that your data analysis efforts are fully appreciated. The key to successful reports, both written and oral, is an understanding of the audience's perspective. This requires a knowledge of its technical sophistication, interests, and priorities. We have given you some suggestions on how to structure written reports and some hints on how to make effective oral presentations. By now, you should be able to not only conduct a brilliant piece of data analysis but also convince other people that you have done a wonderful job!

QUESTIONS AND PROBLEMS

1. Provide an example illustrating the difference between statistical significance and theoretical or managerial substance.
2. Why does it sometimes make sense to report relationships that are *not* significant?
3. How does the nature of the audience affect the content and structure of a written research report?
4. Describe, in general terms, how you would structure (a) an academic report and (b) a managerial report.
5. In your opinion, which is the most important part of a written research report? Why?
6. What advice would you give to someone preparing for their first oral presentation?
7. Why shouldn't you just read your written report to your audience?
8. Why are 'markers' important in oral presentations?
9. Draw up a list of everything you think can go wrong with an oral presentation.
10. Are you brave enough to recommend this book on social media?

FURTHER READING

Becker, L. (2019). *Give Great Presentations*. London: Sage Publishing. A concise guide using several checklists, bullet points, and practical tips to power up your presentations. No student would regret reading this!

Evergreen, S. (2019). *Effective Data Visualization: The Right Chart for the Right Data*. Thousand Oaks, CA: Sage Publishing. Seeing is believing, and this book offers very useful step-by-step instructions to develop tables and figures that effectively convey your data.

Youknovsky, A. & Bowers, J. (2020). *Sell Your Research: Public Speaking for Scientists.* Cham: Springer Nature. Unless you analyze data for personal entertainment (we are not judging!), you will most likely have to present the results to an interested audience. This book is a great resource to take the (stage) fear out of your presentations!

INDEX