ROUTLEDGE

Business Statistics Using Excel

A Complete Course in Data Analytics

R. Panneerselvam

Business Statistics Using Excel

This book gives readers a hands-on understanding of Excel-assisted statistical techniques to take effective business decisions. It showcases applications of the tools and techniques of statistics for analysing business data from the domain of business statistics.

The volume provides an exhaustive introduction to the application of statistics in solving business problems and implementing data analytics for effective decision making in all kinds of business situations around the world. With an emphasis on simplicity in presentation of concepts of statistical methods and associated Excel functions, the volume explores the implementation of Excel functions through well-defined sequences of steps. It covers an array of key topics which include

- Discussions on real-world problems, decision support systems, scope of business statistics, types, and steps of research;
- Introduction to Excel and its mathematical and preliminary statistical functions; usage of different types of average functions; mean, median, and mode functions; measures of variation; measures of skewness of Excel;
- In-depth discussions on probability distributions, sampling distributions, testing of hypothesis, chi-square test, non-parametric tests of Excel;
- Extensive coverage on correlation and covariance, forecasting, analysis of variance, charts in Excel; and
- Analysis of the concept of linear programming, problem formulations, and techniques of linear programming, followed by the application in Excel.

Comprehensive in scope and simple in approach, this book will be key for students and researchers of business studies, business administration, economics, finance, commerce, data analytics/science, and computer science. This will also serve as useful guidebook for business executives and working professionals across the globe.

R. Panneerselvam has more than 40 years of teaching and research experience and served as faculty at Anna University, India, for seven years and then served at Pondicherry University, for 34 years. He has been Head of Department of Management Studies three times and Dean of the School of Management, Pondicherry University. He specialises in operations, quantitative techniques, and computer systems. He holds B.E. (Mech.), M.E., and Ph.D. degrees in industrial engineering from Anna University and has guided 13 doctoral students in management. Dr Panneerselvam has published more than 100 research articles in leading national and international journals, along with 21 textbooks. He has carried out an overseas consultancy for CEMS, Malaysia, and has received several accolades.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Business Statistics Using Excel

A Complete Course in Data Analytics

R. Panneerselvam

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Designed cover image: © Getty Images

First published 2024
by Routledge
4 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

Disclaimer: The views and opinions expressed in this book are solely those of the authors and do not necessarily reflect those of the publisher. The analyses, data and figures based on research material are intended here to serve general educational and informational purposes and not obligatory upon any party. The authors have made every effort to ensure that the information presented in the book was correct at the time of press, but the authors and the publisher do not assume and hereby disclaim any liability with respect to the accuracy, completeness, reliability, suitability, selection and inclusion of the contents of this book and any implied warranties or guarantees. The authors and publisher make no representations or warranties of any kind to any person, product or entity for any loss, including, but not limited to special, incidental or consequential damage, or disruption alleged to have been caused, directly or indirectly, by omissions or any other related cause.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-032-52191-6 (hbk)
ISBN: 978-1-032-67171-0 (pbk)
ISBN: 978-1-032-67170-3 (ebk)

DOI: 10.4324/9781032671703

Typeset in Sabon
by Apex CoVantage, LLC

Dedicated to

My great grandfather, the late Mr. Ramasamy,

Madras Railway (British period)

Contents

Figures

Tables

Glossary

Absolute addressing of a cell requires prefixing the $ symbol to the row and to the column defining that cell.

ANOVA model consisting of relevant components to test their effects on the response variable gives a clear picture of the design of experiment.

Arithmetic mean of grouped data uses mid-points of intervals and the corresponding frequencies to estimate it.

Artificial variable is introduced in a ≥ or = constraint just to serve as the basic variable, which should not be present in the final solution.

Autocorrelation deals with the correlation of a given set of data with another set of data, which is derived from the first set of data with a specified period of lag.

AVEDEV function finds the average of the absolute deviations of observations (X_i, i = 1, 2, 3, . . . , n, where n is the number of observations) from their arithmetic mean. Alternatively, this may be called mean absolute deviation (MAD) of a set of observations.

Average deviation is the mean of the absolute deviations of the observations from the mean of those observations.

AVERAGE function determines the arithmetic mean of a set of observations.

AVERAGEA function finds the average of a desired set of data, that is, numeric, Text, and False as 0 and TRUE as 1.

AVERAGEIF function finds the average of a set of numeric observations stored in a range of cells for a criterion.

AVERAGEIFS function finds the average of a set of observations stored in a range of cells for two criteria.

Bar (column) chart is in the form of vertical bars placed against different values/instances of a variable of interest on the X axis.

Biased estimator of standard deviation contains n as the denominator.

Binomial distribution comes under discrete probability distribution. It is based on the Bernoulli process.

Block in ANOVA brings homogeneity in its rows or columns.

Bowley's coefficient of skewness (CS) is computed for grouped data with open-ended class intervals.

Charts/graphs form an alternative way of representing the data when compared to tabular form.

Chi-square distribution is a distribution when the distribution of S^2 is taken from a normal population with variance σ^2 and it is has $(n - 1)$ degrees of freedom, where n is the sample size.

Chi-square test for categorised data deals with two categories of data, Category A and Category B, each with a specified number of levels, to check whether there is dependency among the observed frequencies under different combinations of the levels of those categories.

Coefficient of range based on the range is the ratio between the range and the sum of the highest value of a set of observations and the lowest value of that set of observations.

Coefficient of skewness is called a characterisation of the degree of asymmetry of a distribution around its mean. It is zero for a symmetrical distribution. When it is negative, then the distribution will have a thinner tailed portion on the left tail of the distribution. When it is positive, then the distribution will have a thinner tailed portion on the right tail of the distribution.

Coefficient of variation aims to check the consistency of the observations of a variable of interest.

Complete factorial experiment contains two factors with replications for each experimental combination of the treatments of the factors.

Confidence interval is the range of values of a statistic.

Correlation coefficient is defined as the degree of association between two variables. The range of the correlation coefficient is from –1 to + 1.

COUNT function counts the number of cells in a given range of cells of an Excel sheet that contains numbers.

COUNTA function determines the number of cells that are not empty in a given range of cells.

COUNTBLANK function determines the number of cells that are blank (empty) in a given range of cells.

COUNTIF function finds the number of cells within a given range of cells satisfying a given criterion.

COUNTIFS function is similar to the COUNTIF function, except it has more than one criterion.

Covariance is the average of the products of the deviations of the means from their respective observations.

Cyclical component (C) of the forecast is similar to seasonality, but its cycle time will be more than a year.

Data Analysis button can be invoked in Excel using a procedure.

Decision support system (DSS) aims to handle semi-structured decisions of middle-level managers of organisations.

Delphi method is a qualitative forecasting method, which aims to predict the future states of qualitative events, that is, culture of the society, level of computer technology, life style of people, value system, and so on.

Exact algorithm gives the optimal solution.

Excel is a handy tool for data analytics.

Excess kurtosis is equal to kurtosis of the distribution minus 3.

Exponential probability distribution is a continuous probability distribution which is used in queuing theory to represent the service time spent on customers in the queuing system. This is actually called negative exponential distribution.

Exponential smoothing method of forecasting aims to forecast the demand of an item for the next period based on the demand and forecast of the current period.

F distribution is a ratio of two chi-square variables.

Factor in an experiment is a parameter or entity which is suspected to have an effect on the response variable.

Filter function presents a subset of rows in a given range of cells, which has data based on one or more criteria applied to one or more columns, respectively.

Fixed factor means that the inferences of a selected set of levels of a factor, which is a subset of the total possible levels of that factor, are restricted to only to the selected set of levels of that factor.

Forecasting means the projection of an event, say, demand of an item based on its past data.

Goodness-of-fit test is the process of fitting a given set of data to an assumed fitting probability distribution.

Grouped data have frequencies.

Heuristic is a rule of thumb to find a near-optimal solution for a combinatorial problem.

Hypothesis is an assumption about a population.

Kolmogorov-Smirnov (K-S) test is an alternative test to the χ^2 test.

***K*-sample median test** is used if the number of samples is three or above, say, *K* samples.

Kurtosis of a distribution gives information about the heaviness in terms of peaked-ness or flatness of the distribution at tails.

Kruskal-Wallis (*H*) test is like the *K*-sample median test in which the objective is to test whether the *K* samples are drawn from K identical populations. This test is an alternative approach for ANOVA with a single factor.

Large sample in a run test means that the value of n_1 or n_2 or both is/are more than 20.

Large sample size in one-tailed/two-tailed sign test considers a random sample of *n* units with the condition that *np* as well as $n\ (1 - p)$ is greater than or equal to 5.

Large samples in two-sample sign test considers two random samples, each with size *n*, with the condition that *np* as well as $n\ (1 - p)$ is greater than or equal to 5.

Latin square design has a single factor with two blocks without replications.

Leptokurtic is a degree of heaviness of a distribution, when the excess kurtosis is positive.

Line chart is in the form of piecewise linear graph constructed on an *X-Y* plane.

Linear programming is a mathematical programming technique, which optimises a measure of performance of a system of interest under a given set of constraints imposed by the management.

Mann-Whitney U test is an alternative to the two-sample *t* test, and it is powerful. It is also known as the rank-sum test.

MAX function finds the maximum among a given set of observations.

MAXA function finds the maximum among a given set of observations, that is, numeric value, logical, and text. The TRUE of the logical is assumed to be 1, and the FALSE of the logical is assumed to be 0. Any other text is assumed to be 0.

Median function determines the middlemost observation of a given set of data.

Median test checks whether the two samples which are independent have been drawn from two populations with the same median.

Mesokurtic is the degree of heaviness of a distribution, when the excess kurtosis is zero.

MIN function finds the minimum among a given set of observations.

MINA function finds the minimum among a given set of observations, that is, numeric value, logical and text. The TRUE of the logical is assumed to be 1, and the FALSE of the logical is assumed to be 0. Any other text is assumed to be 0.

Mode is a kind of measure of central tendency. The mode of a given set of data (observations) is the item of that data which has the maximum frequency.

Moving average method aims to estimate the demand of an item in short run, say based on the demand of the past three or four weeks.

Multi-bar (column) chart is in the form of vertical bars for *multiple instances*, placed against different values/instances of a variable of interest on the *X* axis

Multiple-line chart is in the form of several piecewise linear lines constructed on the *X*-*Y* plane. Normally, an independent variable will be taken on the *X* axis, and a set of dependent variables will be taken on the *Y* axis.

Multiple linear regression equation contains more than one independent variable on its right-hand side.

Negatively skewed distribution will have a thinner tailed portion on the left tail of the distribution.

Non-parametric test is applied to data which do not have the estimate(s) of parameter(s).

Normal distribution has zero excess kurtosis. It is a symmetrical distribution, and the value of the random variable (*X*) varies from – ∞ to + ∞.

Normal probability density function with a mean of 0 and variance of 1 is designed using the central limit theorem by substituting the normal random variable *X* with a standard normal variable *Z*.

Null and alternate hypotheses for the goodness of fit test are: H_0: The given data follow the assumed probability distribution. H_1: The given data do not follow the assumed probability distribution.

One-sample sign test assumes the number of plus signs as the value of the random variable *X* of the binomial distribution, with $p = 1/2$ and the number of trials *n* to compute the probability that *X* is more than the number of plus signs to check different hypotheses. When the sample size is small, it assumes a small random sample from a non-normal population, and then it is tested against a median value (μ) such that the observations in the sample are more or less than that median (μ) at a significance level of α using binomial distribution.

Open-ended interval means that the lower limit of the first interval and the upper limit of the last interval are absent.

***p* value for the computed chi-square value,** if less than the significance level (α), means to reject the null hypothesis; otherwise, accept the null hypothesis.

Paired *t* test is used for a situation in which the value of a random variable at a particular setting may be different from another setting.

Parametric test is applied to data which have the estimate(s) of parameter(s).

Percentile is a value in the given range of values such that a given percentage of observations fall below that value.

Pie chart is in the form of a circle, in which 360° will be divided proportionately according to the frequencies of the variable of concern.

Platykurtic is a measure of a distribution to study the heaviness of a distribution in terms of the presence of frequency.

Poisson distribution is a discrete distribution which is used to capture, for example, the arrival rate of customers at a service station.

Positively skewed distribution contains a thin tailed portion on the right tail of the distribution.

Primary data collection is a direct observation method of data collection.

Quartile is the value of a random variable with respect to a specified percentile out of five different percentages, that is, minimum value, 25th percentile, 50th percentile, 75th percentile, and maximum value of the total frequency.

Quartile deviation (QD) is half of the difference between the third quartile and the first quartile.

Questionnaire consists of sections depending on the dimensions of the research problem, in which the first section contains demographic data of the respondent and the following sections represent macro dimensions of the research problem, that is, organisational climate, employees' welfare measure, labour productivity, and so on.

Random component (*R*) of the forecast will account for all explained factors in reality which will have impact on the demand of a product.

Random factor means that the inferences of a set of levels selected from the total number of levels for the purpose of conducting an experiment are extended to all the levels of that factor.

Randomised block design is ANOVA with a single factor and one block.

Range is a simple measure of variation, which is the difference between the highest value of a set of observations and the lowest value of that set of observations.

RANK.AVG function obtains the rank of a number among a given set of numbers either in descending or ascending order. If more than one value has the same rank, then the average of ranks of those numbers is obtained and treated as the rank of those numbers.

RANK.EQ function obtains the rank of a number among a given set of numbers either in descending or ascending order.

Regression is defined as the dependence of a variable of interest (dependent variable) on one or more other variables (independent variables).

Replications are repeated observations under the same experimental conditions.

Response in an experiment is a measurement of a dependent variable of interest, which may be influenced by the effects of one or more factors and their interactions.

ROUND function rounds the given decimal number to the nearest number with a desired number of decimal digits.

ROUNDDOWN function in Excel reduces the number of decimal places to a desired number of decimal digits without rounding.

ROUNDUP function in Excel reduces the number of decimal places of a decimal number with a desired number of decimal digits with rounding.

Run means the stream of data that is collected in a system with certain patterns.

Sampling deals with the selection of the respondents from a population.

Sampling distribution of a mean with an infinite population has the variance of σ^2/n. But the variance of the sampling distribution with a finite population will have a different variance, which can be obtained from the variance of the sampling distribution with an infinite population by multiplying by a finite population multiplier.

Seasonal component (*S*) in the forecast delas with the regular and predictable changes that happen in a year.

Secondary data deals with the collection of data from the sources, that is, published records of the company, journals, newspaper, and so on.

Semi-structured decision is the decision taken at the middle management level.

Simplex method was developed by George Dantzig in 1947 to solve the linear programming problem, which is an iterative procedure that begins with an initial feasible solution and continues until an optimality condition is reached.

Slack variable is a variable in a less than or equal to constraint, which equates the left-hand side of that constraint to the right-hand side of that constraint.

Solver function is used to solve linear programming models, which come under the decision-making environment of operations research.

SORT function helps to rearrange the content of the cells in a given range in a desired order, that is, ascending or descending order of a particular column.

Spearman's correlation coefficient (r_s) is the degree of association between two different streams of ranks. An example of the ranks may be the ranks given by two different judges for *n* units of a product of interest.

Spread distinguishes samples with the same mean.

Stacked bar chart is in the form of vertical bars such that each vertical bar is subdivided into smaller rectangles according to the instances of the respective value of the variable on the *X* axis.

Standard deviation is the square root of the variance. It represents the spread of the data around the mean of that data.

Statistics is a field of science which deals with the analysis of data of business firms, government, sports, and many other domains of real-world problems.

STDEVA function finds the standard deviation of the observations in a sample of a population including logical values and text.

STDEVP or STDEV.P function finds the standard deviation of a set of observations of a population, excluding text and logical values in it.

STDEVPA function finds the standard deviation of the observations in a population including logical values and text.

STDEVS or STDEV.S function finds the standard deviation of a set of observations of a sample excluding text and logical values in it.

Surplus variable is a variable in a greater than or equal to constraint, which equates the left-hand side of that constraint to the right-hand side of that constraint.

Symmetrical distribution has symmetricity of the left half of the distribution and the right half of the distribution.

T-distribution is for the normal population when its variance is unknown.

Time series data consists of trend (*T*), seasonal (*S*), cyclical (*C*), and random (*R*) components.

Treatment refers to different settings of a factor.

Trend *(T)* component in the forecast deals with the increase or decrease of a dependent variable with an increase in the independent variable.

Two-tailed one-sample sign test when sample size is small, assumes a small random sample taken from a non-normal population, and then it is tested against a median value (μ) such that the observations in the sample are not equal to that median (μ) at a significance level α using binomial distribution.

Ungrouped data do not have frequencies.

Uniform distribution is a continuous distribution which has an equal probability of occurrence for each of the values of the random variable in an interval with lower and upper limits of *a* and *b*, respectively.

Unstructured decision is a decision taken at the top management level.

Variance is a measure of variation, which is the average of the squares of deviations of the mean of a set of observations from individual observations of that set of observations.

Weighted average is computed for the data in reality with certain weights.

Yates' algorithm is a generalised algorithm which gives the sum of squares of different components of the model of the 2^n factorial experiment, where *n* is the total number of factors and each factor has two levels.

1 Introduction

Learning Objectives

After going through this chapter, you will be able to

- Understand real-world decision problems.
- Recognise the decision support system as a tool for decision making.
- Know business statistics and its scope.
- Understand different types of research, that is, exploratory research, conclusive research, and modelling research.
- Analyse the steps of research.
- Have an idea of statistical techniques.

1.1 Real-World Problems

Demand from households is the starting point of the entire business cycle. Businesses construct a variety of enterprises in a variety of industries to satisfy the needs of families. In addition to these businesses and sectors, there are several infrastructural needs for the general public to move around and live with a respectable standard of living. Therefore, public-sector organisations were established by both federal and state governments to satisfy public needs.

When the term “business” is brought up, there are many elements that come to mind, including resource mobilisation; designing an operating system made up of people, machines, and materials; and creating operating procedures to carry out the duties of the various functional subsystems of the business, including operations, finance, marketing, and human resource management.

Each of these functional subsystems has numerous decision-making challenges, as can be seen if one carefully examines the micro aspect of how they operate.

Examples of decision problems in the operation subsystem of any business, which belong to and are integral to the core of the business, include the decision to purchase materials, the decision to schedule jobs and machines, the decision to match operators with machines, design of quality assurance systems to meet customer expectations, and so on.

The finance subsystem of a business faces decision problems related to providing long-term and short-term capital for seamless integration of the business through all of its functions against various uncertainties, analysing the financial performance of the business, allocating funds to different business requirements in an optimal manner while taking into account various constraints, and so on.

DOI: 10.4324/9781032671703-1

Marketing is a business's strongest component. As a result, choosing the appropriate products and services to produce and provide by the business based on customer expectations, communicating demand to production facilities so they have enough lead time to supply the goods and services on schedule without exceeding the lead times specified by the customers, and conducting market research to understand the expectations of customers are all necessary.

Although the idea of an unmanned factory is taking shape thanks to widespread factory automation, it would be good to employ the nation's or the world's existing labour force to install organisational subsystems in order to save money. Judicious hiring of personnel, developing and arranging employee training programmes, scheduling employees, designing employee welfare programmes, and so on constitute human resource decision-making issues.

Numerous business interface areas, such as the company's digital platform, employee logistical support, and many more, as well as corporate social responsibility, will require the company's attention when making decisions.

The decision problems stated in the functional areas of business are classified into the following three categories.

- Completely structured decisions
- Semi-structured decisions
- Unstructured decisions

The majority of decisions made at the operational level are entirely structured decisions, including scheduling jobs, materials, and so on. Semi-structured decisions are those made at the medium management level. This category includes the company's choice to increase overtime. Unstructured decisions are those made at the highest levels of management. Top management decisions include things like choosing where to build a factory and whether to add a new product to the existing product list.

With the help of a decision support system (DSS), decisions that fall within semi-structured environments can be handled. Excel is a useful tool for suggesting decisions to middle-level managers.

1.2 Decision Support System

The decision support system's goal, as mentioned in the previous section, is to manage semi-structured decisions made by middle-level managers in organisations. The following tools can be used to make these decisions.

- Mathematical models, specifically operations research models
- Exact algorithms for polynomial problems
- Heuristics for combinatorial problems

Mathematical Models

An empirical model created from past data from a system, or an objective function subject to a set of constraints, makes up a mathematical model. The product mix problem in a company to decide the production volumes of a given set of products subject to resource constraints such that the total profit is maximised sets an example of linear programming model in operations research. A queueing model aims to find a set of system performances as listed in the following using the empirical formulas available for that model if the reality follows the assumptions of that model.

- Percentage utilisation of server
- Average number of customers waiting in the queue
- Average number of customers waiting in the system
- Average waiting time of customers in the queue
- Average waiting time of customers in the system

The optimal number of toll booths at a toll plaza can be designed using such a model subject to some system constraints:

- No vehicle should wait for more than three minutes to cross the toll plaza.
- The maximum number of vehicles in the system is less than or equal to five

Exact Algorithm

There are some problems that, despite their size being moderate or huge, can be resolved in a reasonable amount of time. This category includes determining the shortest path in a distance network between a source node and a destination node. Such problems are called polynomial problems. The optimal solution is provided by the exact algorithm for a problem. For this kind of problem, there are various algorithms, including Floyd's algorithm and Dijkstra's algorithm.

Heuristics

Combinatorial problems make up the majority of problems in the real world. Solving such problems will need an enormous amount of computer time for moderate and large problems. The travelling salesperson problem serves as an illustration of this group. Finding a route for the salesperson who departs from his base city, travels to each city exactly once or at least once, then returns to his base city while minimising the overall distance travelled is the goal of this problem. This problem falls under the combinatorial category. Such a problem can be solved using a heuristic, which gives a near-optimal solution in a reasonably short execution time. Managers can apply their intuition to the results that are provided by a DSS tool such as Excel after mapping real-life problems to the format of the mathematical model/exact algorithm/heuristic in it.

1.3 Business Statistics

The science of statistics deals with the study of data from businesses, governments, sports, and many other areas of real-world problems.

Business statistics is a field of study that uses statistical principles and methods for business organisations in order to make managerial decisions. The following is a list of some examples of business firm decisions.

1. Future demands of products manufactured by companies
2. Study of price fluctuation of raw materials used in firms

3. Attrition pattern of employees in industries
4. Projection of capital needs of organisations
5. Study of organisational climate
6. Study of labour absenteeism in organisations
7. Adherence to production schedules
8. Analysis of employee satisfaction
9. Analysis of life of equipment in companies
10. Study of labour productivity
11. Study of in-process inventory on shop floor
12. Study of quality issues on shop floor
13. Vendor rating analysis
14. Study of inter-week production data
15. Study of reach of advertisements
16. Media selection decisions
17. Projecting demand of a product after its launch
18. Predicting future rate of return of a company
19. Study of demographic data of customers

1.4 Types of Research

Business research focuses on various problems that the management of a business faces in order to find answers over the short-term, medium-term, and long-term planning horizons. The types of research that are carried out are classified into the following.

1. Exploratory research
2. Conclusive research
3. Modelling research

1.4.1 Exploratory Research

Exploratory research aims to collect data and analyse the data to explore as many relationships as possible among different variables of a system of interest without having knowledge on their end applications for the time being [1]. This research is in the category of general research. The relationships that are derived in this research would form the foundations for future research depending on the future requirements of real-world problems. The domain of data mining is the best example of exploratory research. In companies, data storage over a period of time grows exponentially. Data mining aims to analyse the pile of data after screening them and draw as many meaningful relationships as possible among several variables.

Exploratory research is classified into literature reviews, experience surveys, and studies of insight-stimulating examples.

Literature Surveys

A literature survey tries to gather material in a researcher's particular area of interest and analyse any forward progress. In order to aid themselves or other researchers in the future conduct additional research, the researcher will critically analyse the literature and compare and contrast it with other works in the field. Some examples of literature survey include the following.

1. Study of gross domestic product (GDP) data of a country
2. Study of evolution of culture and its effects on society
3. Demographic analysis
4. Population study of countries
5. Study of wholesale price index

Experience Surveys

Executives in an organisation develop experiences over a period of time. Their experiences are available to the organisation and helpful as long as they continue to work there. There will be a void in the organisation after they leave or go on to other organisations if no plans are made to fill the void with qualified executives. This behaviour is widespread throughout all organisations. This means that if a study is designed to record the experiences of senior executives from various organisations, either while they are still employed or after they retire, such an experience database will serve as a knowledge base for middle-level managers to make effective and efficient decisions, which is called an experience survey.

The knowledge base that is created through such a study will form a component of expert system of the field of interest, which is a high-level domain of management information management.

A sample set of an experience survey is as given in the following.

1. Material ordering systems and their implications in stores operation of a company
2. Maintenance management system in companies
3. Judicial mix of manpower planning
4. Budget exercise of companies as well as governments
5. Expert knowledge base in critical domain of medical field
6. Impact analysis of government schemes
7. Investment decisions
8. Portfolio management
9. Pricing of products
10. Technology forecasting

Studies of Insight-Stimulating Examples

The purpose of this type of study is to get knowledge about a specific area of research topics. The research employs a case study methodology. Think about disruptive innovation as an example. Many established businesses, or incumbents, will exist in a certain industry, like the steel industry. Over time, new businesses – referred to as entrants – will join the industry. Even though the market grows over time, the extra capacity created by new competitors will outpace market expansion in that sector.

In order to stop the rise of the incumbents, the newcomers will aim to take their market share. This is a conviction. Christensen created a notion that the incumbents in the steel sector lose market share to newcomers; however, this is not always the case in India [2]. However, India proves Christensen's idea of disruptive innovation accurate in the aviation sector [3, 4]. These investigations are based on a case study methodology using historical data.

1.4.2 Conclusive Research

Conclusive research tests the hypotheses of the problem of research formulated through exploratory research to draw inferences [1]. Later, a decision-making framework will be designed based on these inferences. Conclusive research is further classified into descriptive and experimental research.

Descriptive Research

Descriptive research focuses on particular goals; therefore, its findings will have conclusive implications. The characteristics of the respondents will be examined in a study with regard to a product or any other attributes of the entities of real-world systems, such as societal culture, regional governance patterns, the effects of government programmes on the general public, and, in particular, the expansion of the rural economy. Consider the public welfare programmes that a government runs. An investigator must establish a questionnaire that includes questions about the respondents' demographics, the schemes' stated goals, the impacts on the public in terms of their enhanced quality of life, the region's economic development, and the effectiveness of their implementation.

The questions in the questionnaire should be such that they provide data to test a set of hypotheses, which are oriented to testing the impacts of the government schemes created on the public.

A sample set of null hypotheses is presented in the following. The alternate hypotheses are just opposite to the following hypotheses.

1. There is no improvement in the quality of life of the public in the region through government schemes.
2. Most of the schemes are not necessary for the public.
3. Anything that is given free to the public does not bring the value of money of that scheme.
4. Provisions in the schemes help the public to misuse them.

Experimental Research

The response variable(s) of a system of interest is/are influenced by a variety of factors. The researcher will be keen to study the impact of those factor(s) on the response variable using analysis of variance (ANOVA) or multivariate analysis of variance(s) (MANOVA(s)) [5]. The researcher must design an experiment in a controlled environment with the right number of levels for each of the experiment's elements and conduct it with a minimum of two replications for each combination of levels of the factors. Then a complete ANOVA or MANOVA is to be performed in order to draw conclusions regarding the experiment's hypotheses.

A country is made up of many states. The goal is to put to the test various theories about how the general people in the country save. One can suppose that State is a factor.

Choose a state at random from each of the four directions: east, west, north, and south. The public is segmented into three levels within each state: the high-income group (HIG), middle-income group (MIG), and low-income group (LIG). Such a division of the public forms the factor Income. Data on annual deposits, which are gathered for two consecutive years for each combination of the levels of the factors State and Income, which constitute replications, serve as the experiment's performance indicator (response variable).

A sample design of the experiment for this problem is shown in Table 1.1. Let the factor State be Factor A and the factor Income be Factor B.

The model of the ANOVA of this experiment is shown here [5].

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + e_{ijk}$$

Where

Y_{ijk} is the k^{th} replication of the response variable for the i^{th} level of Factor A and the j^{th} level of Factor B

μ is the overall mean of the response variable

A_i is the effect of the i^{th} level of Factor A on the response variable

B_j is the effect of the j^{th} level of Factor B on the response variable

AB_{ij} is the effect of the i^{th} level of Factor A and j^{th} level of Factor B on the response variable

e_{ijk} is the random error associated with the k^{th} replication of the response variable under the i^{th} level of Factor A and j^{th} level of Factor B

The null and alternate hypotheses of this experiment are listed in the following.

Factor A

H_0 : There are no significant differences among the levels of Factor A in terms of the response variable Saving habit.

H_1 : There are significant differences among the levels of Factor A in terms of the response variable Saving habit.

Table 1.1 Design of Experiment to Analyse Saving Habits of Public

State (Factor A)	*Income (Factor B)*		
	HIG	*MIG*	*LIG*
Eastern state			
Western state			
Northern state			
Southern state			

Factor B

H_0 : There are no significant differences among the levels of Factor B in terms of the response variable Saving habit.

H_1 : There are significant differences among the levels of Factor B in terms of the response variable Saving habit.

Interaction AB

H_0 : There are no significant differences among the interaction terms of the i^{th} level of Factor A and j^{th} level of Factor B in terms of the response variable Saving habit.

H_1 : There are significant differences among the interaction terms of the i^{th} level of Factor A and j^{th} level of Factor B in terms of the response variable Saving habit.

1.4.3 Modelling Research

A model is an abstraction of a system. Modelling research aims to develop a model for a business scenario. It is further divided into mathematical modelling research and simulation research [1].

Mathematical Modelling Research

A mathematical model is the representation of a research problem of interest based on the parameters and variables of that research problem. A sample list of modelling research is as given in the following.

1. Forecasting model for demand estimation of a product in a company.
2. Linear programming model to maximise profit through manufacture and sales of a set of products subject to a set of constraints. This type of model will have an objective function and a set of constraints for the resources used to manufacture the products in the product mix. The objective is to determine the production volumes of the products such that the profit is maximised subject to the set of resource constraints.
3. Econometric model relating to the GDP of a country.
4. Model for selecting projects from competing projects such that the return for the organisation is maximised.
5. Model for manpower scheduling.

Simulation Modelling Research

A simulation is an experiment carried out on a system of interest. The last resort for analysing a situation is a simulation modelling study if the behaviour is probabilistic in nature and does not fit into any of the usual probability distributions that reflect the mean and variance of each variable in the system [1]. Any form of assumption can be used in a simulation. In general, simulation can be used to provide average estimates of the measures of interest to the researcher of the research topic if reality does not have an empirical model to provide its solution. Examples of simulation modelling research applications include queueing simulation, maintenance simulation, stock market simulation, and production scheduling simulation.

1.5 Steps of Research

The research process has the following steps.

1. Problem definition
2. Objectives of the research
3. Research design
4. Data collection
5. Data analysis
6. Interpretation of results
7. Validation of results

1.5.1 Problem Definition

Operations, finance, human resources, and marketing are the functional areas of management where there may be competing research concerns in an organisation. The organisation's R&D division should compile a list of every research issue and rank them in accordance with the current priorities, as these shift over time. Next, choose the issues one at a time based on the R&D wing's staffing capacity. Once a problem has been decided upon, it must be precisely defined. Any poorly defined challenge will fail because it is difficult to turn a vague research notion into a practical research problem. Lack of definition will result in a garbage-in, garbage-out situation. Therefore, the issue should be precisely described by taking into consideration all relevant practical limitations.

1.5.2 Objectives of the Research

There will be a list of requirements for the system for which the study is conducted. When developing the study's objectives, the investigator should take one or more system needs into consideration. The objectives of the research should be identified before finalising the research questions, hypotheses, and study boundaries.

The research questions should cover the following topics: the goal of the study, the setting in which it will be conducted, the current state of the research question, the method used to accomplish the goal of the study, and the rationale behind the choice of that specific approach.

A hypothesis is an assumption about a population. The research will warrant formulation of several hypotheses to test them. The hypothesis about an assumption will have two forms, the null hypothesis $\left(H_o\right)$ and alternate hypothesis (H_1). Generally, the null hypothesis will be in supportive nature of the research issue. Consider a situation where the HR manager of a company forms the opinion that the graduates of different colleges are not different from one another. A hypothesis to support the HR manager's claim is stated as follows:

"There are no significant differences among colleges in terms of quality of students, who appear for interview".

The study's boundaries are determined by the research's objectives, population size, and other parameters. The study's scope will depend on whether it is applied to all states in the country or just a few. This will depend on if the research is investigating the impact of household income level increases on the sales of a company's goods.

1.5.3 Research Design

After identification of the research problem and its objectives, the next step is research design, which gives complete guidelines for data collection. The elements of the research design are listed in the following.

1. Selection of the research methodology approach
2. Design of sampling plan
3. Design of experiment
4. Design of questionnaire

Selection of Research Methodology/Approach

As discussed in Section 1.4, the different research methodologies are exploratory research, conclusive research, and mathematical and simulation modelling research. Depending on the objectives of the study, a suitable research method is to be selected by the investigator.

Design of Sampling Plan

Respondents play a significant role in the research when it is survey based. The study's questionnaire cannot possibly be given to every person in the population. Census studies are time and money consuming since all members are treated as responders in the research. So, the respondents are to be sampled from their population. In such a situation, the following questions will arise.

1. Sample size
2. Sampling method

The determination of the sample size will be done after carrying out a pilot survey, which will provide variance of the responses given by the respondents. Based on these data, the sample size will be determined. After having decided the sample size, the investigator has to select a suitable sampling method. The sampling methods are classified into probability and non-probability sampling methods.

The probability sampling methods are as follows.

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling
5. Multi-state sampling

The non-probability sampling methods are as listed.

1. Convenience sampling
2. Judgement sampling
3. Quota sampling
4. Snowball sampling

Design of Experiment

Many response variables are present in a research investigation. One or more factors may have an impact on each response variable. Every factor will have a range. It is necessary to choose the number of replications for each combination of the levels of the factors impacting a response variable, which is a need for all such combinations. An accurate picture of the experiment's design is provided by an ANOVA model that includes pertinent components to assess their impact on the response variable. Before moving on to the next step of constructing the questionnaire, null and alternate hypotheses must be developed for each of the ANOVA model's components.

Design of Questionnaire

There are two types of data that must be gathered for a research study: primary data and secondary data. Primary data are those that are gathered through personal interviews, observational methods, telephone interviews, or postal surveys. For the first time, they are gathered by direct observation. Secondary data are those that are gathered from publicly available firm records, news articles, journals, and so on.

A questionnaire consists of sections depending on the dimensions of the research problem. The first section contains demographic data of the respondent. There will be a section for each macro dimension of the research problem, such as organisational climate, employees' welfare measure, and labour productivity. Each later section will contain a set of questions to facilitate data collection to test different hypotheses which have already been proposed pertaining to the questions of that section. The questions in the questionnaire should be simple and straightforward to facilitate the understanding of the questions by all respondents.

1.5.4 Data Collection

Data collection is an important step of the research process. If it is secondary data collection, the sources are published records of the company, journals, newspaper, and so on. The accuracy of data collection for this type of data depends on the commitment of the investigator. In the case of primary data collection, respondents play a vital role, and their cooperation and consistent answers to the questions in the questionnaire will provide reliable data. Primary data collection is a direct observation method of data collection. In most of the cases, in the midst of regular activities of the respondent, if one approaches him/her, there is a possibility of non-involvement of the respondent in providing the answer to the questions of the questionnaire. So, the availability of the respondent should be ensured before collecting data from him/her.

In some cases, the same respondent may be approached twice if certain data require some time to access before the respondent fills out the questionnaire.

1.5.5 Data Analysis

The gathering of data yields information for data analysis. When analysing data to evaluate and infer from presented hypotheses, fitting tools are used. The choice of data analysis tools ought to have been made at the time of research design. The proposed tools are to be used to analyse the data at this stage. The data analysis aims to compute and compare the following.

1. Computation of statistics such as mean, median, mode, variance, standard deviation, coefficient of variation, and coefficient of skewness.
2. Development of a regression model to relate a dependent variable to one or more independent variables of the study.
3. Determination of correlation coefficients of different pairs of variables of interest.
4. Carrying out hypothesis testing of various research issues identified in the research.
5. Performing factor analysis to reduce the number of variables of the study.
6. Carrying out discriminant analysis.
7. Performing conjoint analysis to find different product profiles, which will serve as different products.

1.5.6 Interpretation of Results

It is very important to translate the inferences of the study into real-world inferences, which can be understood by the end users of the system of the study.

When doing survey-based research, the researcher makes appropriate assumptions regarding the sample size, data variability, curve fitting, and so on. Inferences and outcomes are obtained following the study of the data using appropriate statistical tools. Because of the coding of the data and different assumptions made throughout the research, it could be difficult to directly relate the inferences or results of the study to the expected real-world issue. Therefore, the researcher must adapt the conclusions and findings from the data analysis stage to the relevant practical research question so that each and every interpretation is carried out with the necessary understanding.

In the case of modelling research, codification of the variables is done at the initial stage to solve the problem. So, at the end, each and very coded variable should be replaced by the real-world variables along with the values for those variables, which are obtained after fitting or solving a model.

1.5.7 Validation of Results

The interpretation of the results gives the outcome of the hypothesis testing. While performing testing of a hypothesis, a significance level is always assumed. So, the reliability of the inference is ensured to the level of 1 – the significance level. In experimental and survey-based research, the validation of the results is a straightforward procedure. In the long run, if there are deviations in the inferences drawn now, on a continual basis, the exercise may be repeated to make corrections in the inferences of the respective research issues.

In modelling research, the results of the model are to be compared with the results of the real-world system on an experimental basis. If there is a mismatch between the results of the model and those of the real-world system, the parameters of the model should be fine-tuned so that there is a closer match between the results of the model and those of the real-world problem.

1.6 Statistical Techniques

The field of statistics includes many techniques. The investigator has to select suitable statistical techniques that are required to analyse the data collected through questionnaire.

The different statistical techniques are as listed.

1. Mean
2. Median
3. Mode
4. Range
5. Quartile deviation
6. Average deviation
7. Variance
8. Standard deviation
9. Coefficient of variation
10. Skewness
11. Quartile deviations
12. Average deviation
13. Measure of skewness
14. Analysis of variance
15. Probability distributions
16. Sampling distributions
17. Tests of hypotheses
18. Nonparametric tests
19. Correlation
20. Regression and forecasting
21. Sampling and Monte Carlo simulation
22. Charts

The term “mean” refers to the average of a group of observations and is a measure of central tendency. The initial step in the majority of statistical measures is to compute the mean. Arithmetic mean and weighted arithmetic mean are additional categories for the mean. The average of a specific set of observations or data points is known as the arithmetic mean. Each observation is given weight in the computation of the weighted arithmetic mean. There are still other ways to compute the mean, including the mean for grouped and mean for ungrouped data.

In relation to the cumulative frequency of 50% of the total frequency, the median represents the value of the relevant variable. This metric falls under the category of measurements of central tendency. Both grouped and ungrouped data are used to calculate the median.

The value of the variable of interest in relation to the maximum frequency is called the mode. It serves as a core tendency indicator. Calculating the mode for a given collection of data is the easiest metric of central tendency to use. The mode is calculated for both grouped and ungrouped data, just like it is for the other two measures, arithmetic mean and median.

Range is the difference between the maximum value and the minimum value of the given set of observations. This measure gives an idea about the limits of the data.

Quartile deviation is the half of the difference between the third quartile and the first quartile of a given set of observations.

The average deviation of a given set of observations is the absolute mean of the difference between the individual observations and mean or between the individual observations and median. This measure gives a non-zero value for the mean deviation of the values of the variable from the mean of the given set of observations if at least one pair of observations are different in their magnitudes. But, in the arithmetic mean, the mean may be zero even though there are deviations of the observations from their mean, because a plus value of, say, 10 will cancel out a minus value of, say, –10.

Variance is the mean of the squares of the difference between the individual observations and their mean. This is another important measure next to mean, because for populations/samples with equal means, they are distinguished using their variances.

Standard deviation is the square root of the mean of the squares of the difference between the individual observations and their mean.

The coefficient of variation (CV) of a set of observations is the ratio of the standard deviation and the mean of those observations. This checks the consistency of data. If there are two or more samples, the sample which has the lowest coefficient of variation is said to have consistent observations.

Measure of skewness relates to the shape of the distribution of a given set of data. The shape may be either symmetrical or asymmetrical. These can be checked using the coefficient of skewness, which is in the range from –1 to +1.

ANOVA stands for analysis of variance, which tries to investigate the influence of various variables on an experiment's key response variable. The fundamental designs are completely randomised design, randomised complete block design, Latin square design and complete factorial design. The completely randomised design only takes into account one variable. A factor and a block are both present in the randomised complete block design. A block creates homogeneity among the observations included within it. Two or more components with interaction effects will be included in the analysis of the complete factorial design.

The investor should fit the data of an entity of the study to a fitting shape, which is called probability distribution. Later, to analyse the entity, the probability distribution will be used. The probability distributions in the literature are binomial distribution, Poisson distribution, exponential distribution, uniform distribution, normal distribution, t distribution, chi-square distribution, F distribution, and many other advanced distributions.

The study of a population of interest is carried using a sample selected from that population. It is assumed that the characteristic of the population is represented by its sample. So, sampling distributions for mean, variance and proportion are extended topics to analyse the population using its sample data. The types of probability distribution considered for sampling distributions are normal distribution, *t* distribution, chi-square distribution, and proportion.

A hypothesis is an assumption about a population. The hypothesis is classified into null hypothesis and alternative hypotheses. The investigator should propose null and alternative hypotheses for each and every question included in the questionnaire, which gives a scope for testing it. The testing of hypotheses forms an important aspect of statistics. Hypothesis testing is applied to mean, variance, and proportion. The testing of the hypothesis is broadly classified into testing of mean and testing of variance. The testing of mean is further classified into testing a single mean and testing of difference between two means. Each of these categories has a one-tailed test and two-tailed test. The types of distribution considered in hypothesis testing are normal distribution, t-distribution, chi-square distribution, and F-distribution. There are extended topics, that is, the chi-square test for checking independence of categorised data and goodness of fit test.

The mean and variance are the parameters of a population as well as its sample. In hypothesis testing, the parameters of the population/sample will be computed, which is a time-consuming process. Non-parametric testing tests the hypothesis without estimates of the parameters. Under this type, the different tests are the sign test, chi-square test, Kolmogorov-Smirnov test, and run test for randomness for one sample. Under the two-sample test, the tests are the two-sample sign test, two-sample median test, and Mann-Whitney U test. Under K-sample tests, the different tests are the K-sample median test and Kruskal-Wallis test (H-test).

Correlation is a study which analyses the association between two variables. The range of the correlation coefficient is from –1 to +1.

Based on historical data, regression creates a relationship between a dependent variable and a group of independent factors. An investigator can anticipate the values of the dependent variable for a particular setting of the independent variables with the use of a relationship in the form of a prediction model. Regression can be used for a variety of purposes, including forecasting future dividend rates, growth indices for businesses, and the GDP of nations.

Real-world occurrences tend to be probabilistic in nature. As a result, Monte Carlo simulation can be used to estimate the expected value of the event variable via sampling. Monte Carlo simulation samples the values of the target random variable using evenly distributed random integers. The estimate of the important variable is then provided by the mean of the sampled values.

The investigator will have a better understanding of the pattern of data from which valid conclusions may be derived with the aid of the visual display of the data in the shape of a curve or bars. Charts, such as pie charts, bar charts, multi-bar charts, stacked bar charts, line charts, and multiple-line charts, can be used to analyse data that have been gathered for a study.

Summary

- Decisions stated in the functional areas of the business are classified into completely structured decisions, semi-structured decisions, and unstructured decisions.
- A mathematical model consists of an objective function subject to a set of constraints or an empirical model designed based on the past data of a system.
- Business statistics is a branch of study which applies principles and statistical techniques to business firms for managerial decisions.
- Exploratory research aims to collect data and analyse the data to explore as many relationships as possible among different variables of a system of interest without having knowledge on their end applications for the time being.
- A literature survey aims to collect literature in a specific field of interest of a researcher and analyse the progress made so far.
- An experience survey is used, for example, if a study is oriented to capture the experiences of senior executives of several organisations while they are in service or after their retirement, which forms an experience database known as a knowledge base for middle-level managers to take effective and efficient decisions.
- A study of insight-stimulating examples aims to have insight into a select field of research topics.
- Conclusive research tests the hypotheses of the problem of research formulated through exploratory research to draw inferences.

- Descriptive research focuses on specific objectives; hence the outcomes of this research will have definite conclusions.
- Experimental research deals with the design of an experiment in a controlled environment by taking a suitable number of levels for each of the factors of the experiment and conducting an experiment with a minimum of two replications under each combination of the levels of the factors.
- A simulation is an experiment conducted on a system of interest.
- A hypothesis is an assumption about a population.
- Research design gives complete guidelines for data collection.
- Sampling deals with the selection of the respondents from a population.
- An ANOVA model consisting of relevant components to test their effects on the response variable gives a clear picture of the design of experiment.
- A questionnaire consists of sections depending on the dimensions of the research problem. The first section contains demographic data of the respondent. There will be sections for each macro dimension of the research problem, such as organisational climate, employee welfare measures, and labour productivity.
- Primary data collection is a direct observation method of data collection.
- Secondary data deals with the collection of data from sources, that is, published records of the company, journals, newspapers, and so on.
- Mean is the average of a set of observations, which belongs to measures of central tendency.
- Median is the value of the variable of interest with respect to the cumulative frequency of 50% of the total frequency.
- Mode is the value of the variable of interest with respect to the maximum frequency.
- Range is the different between the maximum value and the minimum value of the given set of observations.
- Quartile deviation is half of the difference between the third quartile and the first quartile of a given set of observations.
- Average deviation of a given set of observations is the absolute mean of the difference between the individual observations and mean or between the individual observations and median.
- Variance is the mean of the squares of the difference between the individual observations and their mean.
- The coefficient of variation of a set of observations is the ratio of the standard deviation and the mean of those observations.
- Measure of skewness relates to the shape of the distribution of a given set of data.
- ANOVA means analysis of variance, which aims to test the effects of factors on a response variable of interest in an experiment.
- Non-parametric testing tests the hypothesis without estimates of the parameters.
- Regression establishes a relationship between a dependent variable and a set of independent variables based on past data.
- Monte Carlo simulation uses uniformly distributed random numbers to sample the values of the random variable of interest, based on which the mean value of the variable will be estimated.
- Charts give a visual display of the data in the form of a curve or bars, which will help the investigator to have a better perception of the pattern of data through which meaningful inferences can be drawn.

Keywords

Unstructured decision is a decision taken at the top management level.

Semi-structured decision is a decision taken at middle management level.

Decision support system aims to handle semi-structured decisions of middle-level managers of organisations.

Exact algorithm gives the optimal solution.

Heuristic is a rule of thumb to find a near-optimal solution for a combinatorial problem.

DSS deals with the decisions which come under a semi-structured environment to assist middle-level managers in their decisions.

Statistics is a field of science which deals with the analysis of data of business firms, governments, sports, and many other domains of real-world problems.

Hypothesis is an assumption about a population.

Sampling deals with the selection of the respondents from a population.

ANOVA model consisting of relevant components to test their effects on the response variable gives a clear picture of the design of experiment.

A questionnaire consists of sections depending on the dimensions of the research problem, in which the first section contains demographic data of the respondent and the following sections represent macro dimensions of the research problem, that is, organisational climate, employee welfare measures, and labour productivity.

Primary data collection is a direct observation method of data collection.

Secondary data deals with the collection of data from sources, that is, published records of the company, journals, newspapers, and so on.

Review Questions

1. List and explain the types of decisions.
2. Define a mathematical model and give an example with its measures to be optimised.
3. What is an exact algorithm? Give an example.
4. Define heuristic and give an example of a heuristic.
5. List and explain the scope of business statistics.
6. What is exploratory research? Discuss research types under this category.
7. What is conclusive research? Explain different research methods under this category.
8. Explain experimental research using an example.
9. Explain the steps of research.
10. Give a brief account of statistical techniques.

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. Madhusudan, C. and Panneerselvam, R., 2017, Disruptive Innovation – A Review with Particular Reference to India, *Asian Journal of Management*, Vol.8, No.4, pp.1370–1378.
3. Madhusudan, C. and Panneerselvam, R., 2020a, Managing Disruptive Innovation: Contextual Evidence for Christensen’s Theory from India, *Studies in Indian Place Names*, Vol.40, No.10, pp.615–645.
4. Madhusudan, C. and Panneerselvam, R., 2020b, A Four-Stage Framework for Disruptive Innovation, *International Journal of Business Excellence*, Vol.22, No.4, pp.474–491.
5. Panneerselvam, R., *Design and Analysis of Experiments*, PHI Learning Private Limited, New Delhi, 2012.

2 Introduction to Excel

Learning Objectives

After going through this chapter, you will be able to

- Know the layout of an Excel sheet.
- Master the edit features of Excel.
- Understand the Excel operations using formulas.
- Format Excel sheets according to requirements.
- Know the formula for adding using the @sum function.
- Understand the procedure to copy a formula.
- Master commands, that is, SUMIF, SUMIFS.
- Learn the COMBIN and FACT commands.
- Understand MAX and MIN functions and their variations.
- Analyse data using the MAXIFS function and MINIFS function.
- Analyse data using the ROUND, ROUNDUP, ROUNDDOWN, and INT functions.
- Apply the SORT function for a given set of data.
- Understand the RANK.AVG and RANK.EQ functions.
- Implement Hide and Unhide commands for hiding/unhiding columns or rows in Excel sheet.
- Understand the implementation of the filter function.
- Use the PROB function in Excel 2019.

2.1 Introduction

The history of computing started with paper and pencil. Clark's table proved useful for elementary calculations in colleges up until the 1970s. Computing became simpler with the development of computer technology and programming languages. Students in engineering programmes used slide rules for computation in the classroom and during exams at the same time. The slide rule was employed by practitioners and consultants in numerous engineering applications.

In India, the microprocessor-based computing system with the CPM operating system was introduced in 1983. Later, in 1985, the DOS operating system gradually took over the role of CPM. The Windows operating system, which supports a wide variety of programming languages, database languages, and spreadsheets, currently rules the computing world. High-level languages for computational purposes, like Fortran, Basic,

DOI: 10.4324/9781032671703-2

C, Pascal, and many other languages, can be utilised; nevertheless, each one necessitates complete programming language proficiency.

To create Excel sheets for numerous managerial applications, such as accounting applications, shop floor applications, human resource applications, operations applications, and many other applications in interface areas of business, one need only spend a short amount of time learning the fundamental Excel commands, such as formulas and functions. For the majority of applications, Excel supports interactive computing. Additionally, it contains functions and formula features for numerous applications, including statistical and linear programming ones. The fundamental formulas and Excel functions are covered in this chapter.

Excel is a spreadsheet which has a set of columns and a set of rows. The columns are represented by letters starting from A, and the rows are represented by numbers from 1 onwards. A cell in Excel is represented by a combination of column label and row number. The maximum numbers of rows and columns in Excel are 1,048,576 and 16,384 (A to XFD), respectively.

2.2 Excel

Excel is a handy tool for data analytics. It has the features of mapping even a semi-structured decision environment to support business managers for their decision-making activities.

2.2.1 Excel Templates

A screenshot of Excel sheet is shown in Figure 2.1. It has rows and columns. The rows are numbered as 1, 2, 3, 4, . . . , 1048576 and the columns are labelled A to XFD, which accounts for 16,384 columns. A given combination of a column and a row

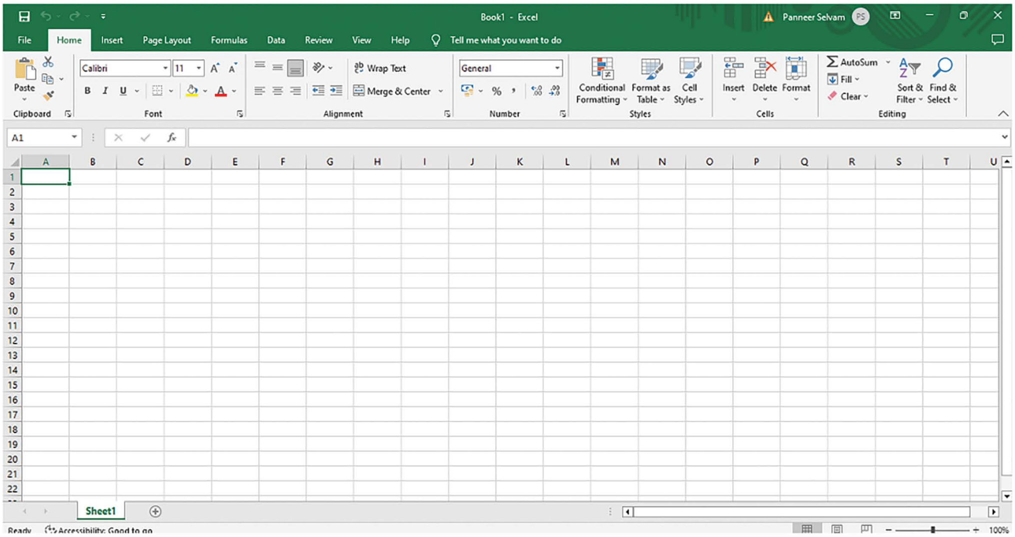


Figure 2.1 Screenshot of Excel template

defines a cell in the Excel sheet. The combination of column A and row 1 is defined as cell A1, and the combination of column E and row 7 is defined as cell E7. One can define the required number of cells in this way per a required pattern to solve a problem of interest in Excel. A pattern of cells means a square format of cells, rectangular format of cells, and so on.

2.2.2 Arithmetic Operators in Excel

The arithmetic operators used in Excel are listed in the following. A formula may begin with either a + sign or = symbol. Mostly the "=" is used to begin a formula.

+ is for adding the value of one cell to that of another cell (Example: = A1 + A2 or + A1 + A2).
– is for subtracting the value of one cell from that of another cell (Example: = A1 – A2)
/ is for dividing the value of one cell by that of another cell (Example: =A1/A2)
^ is for obtaining the power of the value of one cell for a given exponent. The exponent may be a constant or may be available in some cell (Examples: =A1^3, =A1^A2, etc.)

2.2.3 Changing Width of Cells

Excel's standard column width is 8.43 units. This may be viewed by putting the mouse in any cell and selecting the HOME icon, which displays a screenshot similar to Figure 2.2, before selecting the Format icon in the ribbon. When you select Column Width from the dropdown menu, a screenshot similar to Figure 2.3 appears, showing 8.43 as the default column width. By selecting Column Width from the dropdown menu of Figure 2.2, the column width can be modified to a specific value as needed by entering that value in the dropdown menu that displays in Figure 2.3.

2.2.4 Operations in Excel Using Simple Formulas

Consider the following tasks to demonstrate the basic capabilities of Excel.

1. Adding two numbers
2. Subtracting one number from another number
3. Dividing one number by another number

The format of a formula in Excel begins with either the "+" or "=" symbol, and it is followed by one or more actual operations such as addition, subtraction, and division.

Let the numbers to be added be 10 and 15. These numbers are entered in cells A1 and A2, respectively, as shown in Figure 2.4.

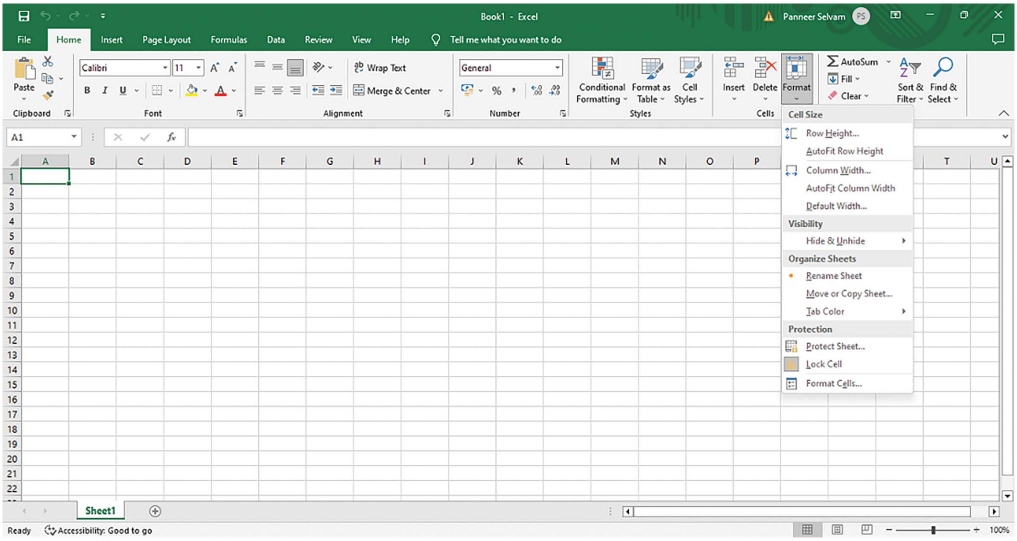


Figure 2.2 Screenshot after clicking Format button in the ribbon of Figure 2.1

The previous Case 1 is achieved by the following command entered in cell A3, which will give the result in the same cell, as shown in Figure 2.4.

$$= A1 + A2$$

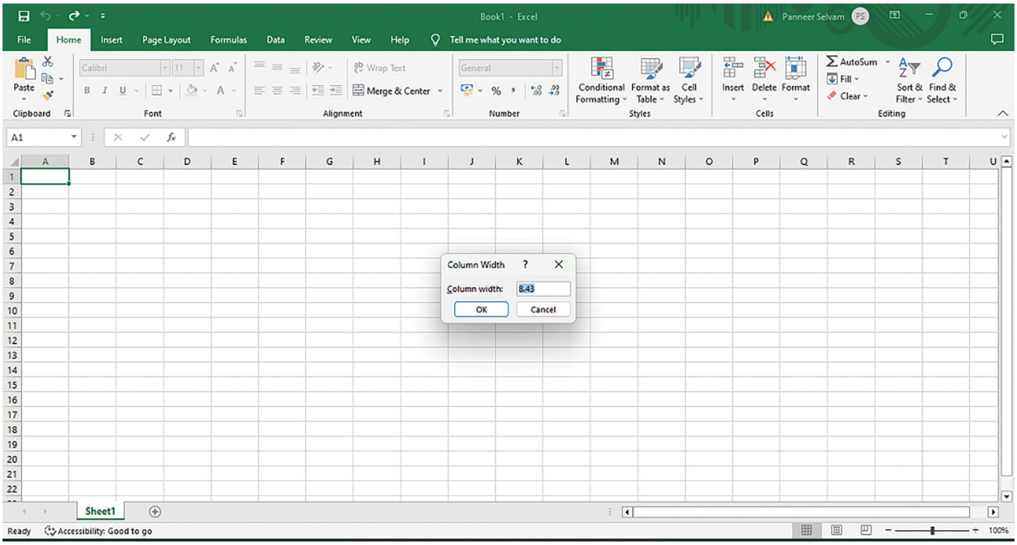
In Figure 2.5, a screenshot of the equation to add the two values is displayed. By putting the pointer in cell A3, we can see the formula in the ribbon's formula region (above columns D, E, F). The formula will be processed as a string and appear in cell A3, as illustrated in Figure 2.5, if the character "=" is included before the formula in the formula region.

Case 2 of subtracting 15 from 10 can be achieved by using the following formula in cell A3, which gives the results in the same cell, as shown in Figure 2.6.

$$= A1 - A2$$

Case 3 of dividing 10 by 15 can be achieved by using the following formula in cell A3, which gives the results in the same cell, as shown in Figure 2.7.

$$= A1 / A2$$

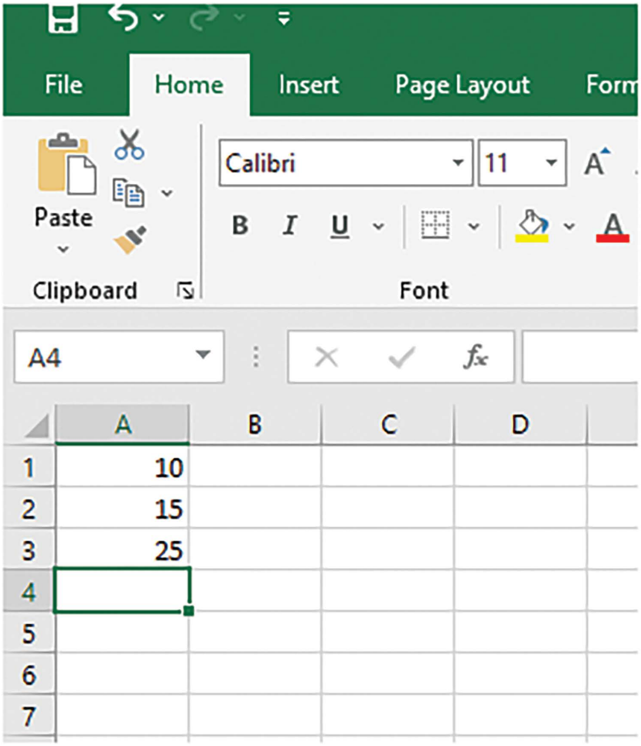


Column Width ? ×

Column width: 8.43

OK Cancel

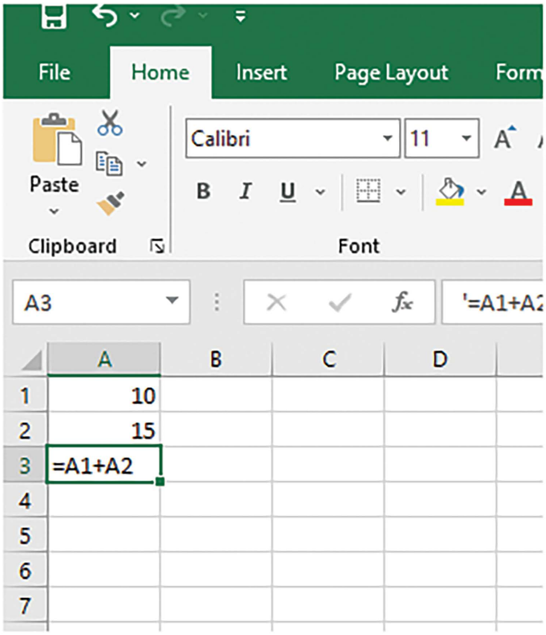
Figure 2.3 Screenshot after clicking Column Width in the dropdown menu of Figure 2.2



A4

	A	B	C	D
1	10			
2	15			
3	25			
4				
5				
6				
7				

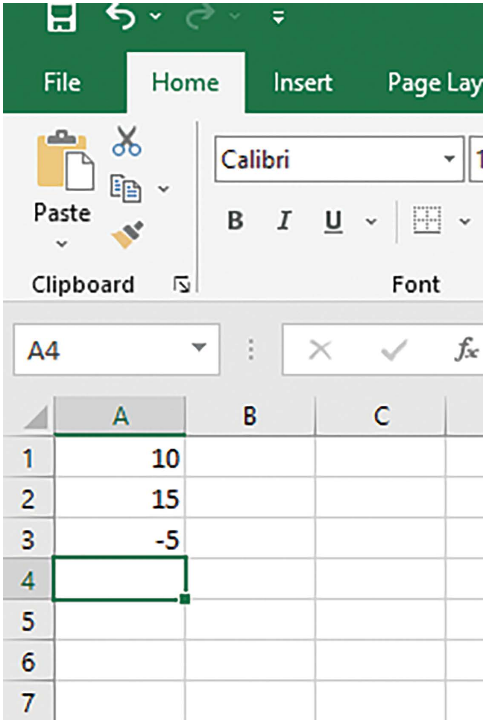
Figure 2.4 Screenshot of adding two numbers with the result in Cell A3 (Case 1)



A3 '=A1+A2

	A	B	C	D
1	10			
2	15			
3	=A1+A2			
4				
5				
6				
7				

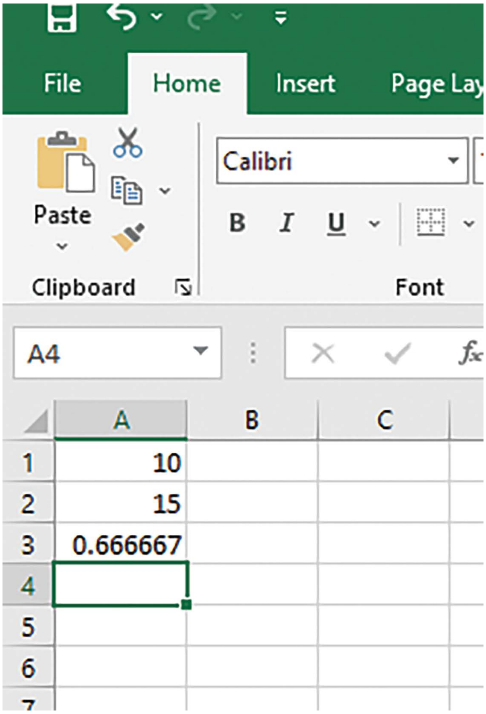
Figure 2.5 Screenshot showing formula to add two numbers (Case 1)



A4

	A	B	C
1	10		
2	15		
3	-5		
4			
5			
6			
7			

Figure 2.6 Screenshot of subtracting Cell A2 from Cell A1 and showing result in Cell A3 (Case 2)



A4

	A	B	C
1	10		
2	15		
3	0.666667		
4			
5			
6			

Figure 2.7 Screenshot of dividing 10 by 15 and showing result in Cell A3 (Case 3)

2.2.5 Formatting of Excel Templates

Strings can be used to format the Excel template in the desired manner. As was previously mentioned, the entire entry in a cell will be handled as a string if it starts with an apostrophe.

Think about the scenario where two numbers, *X* and *Y*, which are 10 and 15, respectively, are added. For a better understanding, the working are displayed by formatting the Excel sheet with strings, as seen in Figure 2.8.

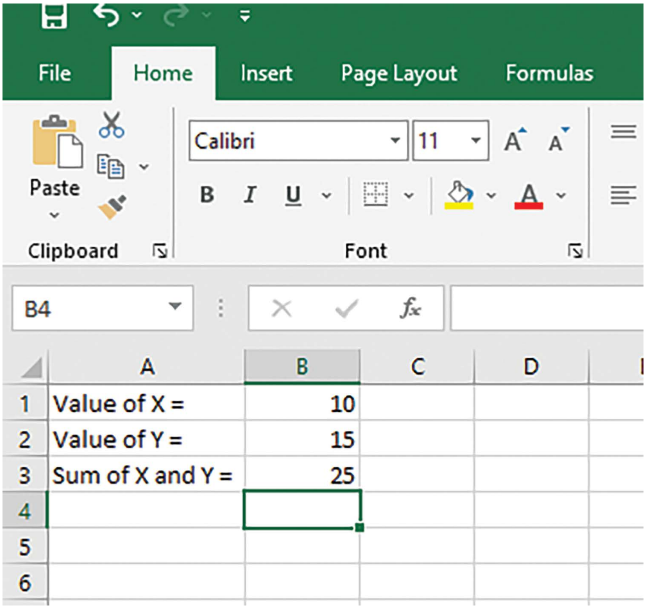
2.2.6 Adding Values in a Range of Cells by Including Cells in Formulas

Consider the case of adding the sales of a product during the past six months, as shown in Table 2.1.

These data can be copied and pasted starting from cell A2, as shown in Figure 2.9. In the screenshot, the extra formatting is done in cell A9 to show the six monthly sales. In Figure 2.9, entering the following formula in cell B9 gives the sum of the values from cell B3 to cell B8 in the same cell.

$$= B3 + B4 + B5 + B6 + B7 + B8$$

A screenshot of the formula in cell B9 is shown in Figure 2.10.



	A	B	C	D
1	Value of X =	10		
2	Value of Y =	15		
3	Sum of X and Y =	25		
4				
5				
6				

Figure 2.8 Screenshot of adding *X* and *Y* and showing the result in Cell B3 with formatting of the Excel sheet with strings in Cells A1, A2, and A3 for better understanding

Table 2.1 Sales Data for Product

Month (i)	*Sales in Units*
1	1000
2	1200
3	1150
4	1300
5	1400
6	1350

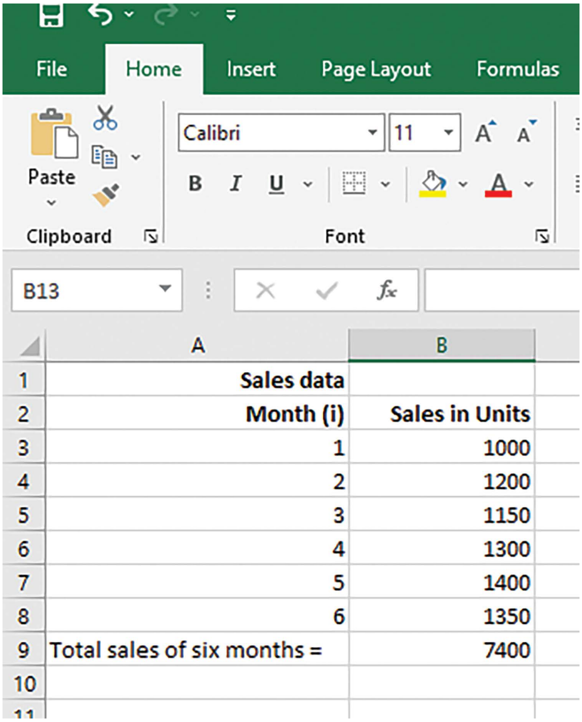
2.2.7 Adding Values in a Range of Cells Using the @SUM Function

Consider the case of adding the sales of a product during the past 12 months, as shown in Table 2.2.

In the snapshot shown in Figure 2.11, cell A2 is the beginning point for copying and pasting the data from Table 2.2. The sum of the values from cell B3 to cell B14 in the same cell is shown in Figure 2.11 by entering the following formula in cell B16. Figure 2.12 displays a snapshot of this formula in cell B16.

@SUM(B3:B14)

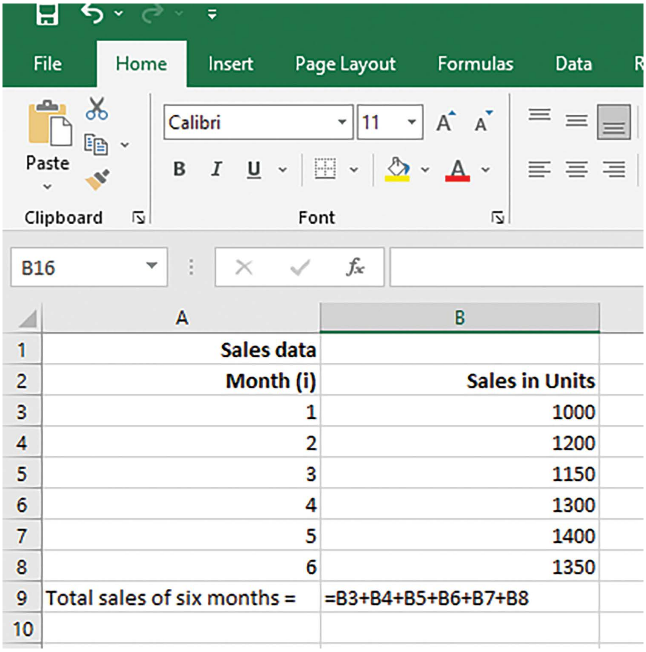
By keeping the cursor in the range's first cell and dragging it to the range's end cell, the range of cells can be inserted in the formula following the open parenthesis. After entering the formula in the final cell of that range, a close parenthesis must be typed.



B13

	A	B
1	Sales data	
2	Month (i)	Sales in Units
3	1	1000
4	2	1200
5	3	1150
6	4	1300
7	5	1400
8	6	1350
9	Total sales of six months =	7400
10		

Figure 2.9 Screenshot showing result of sum of values from Cells B3 to B8



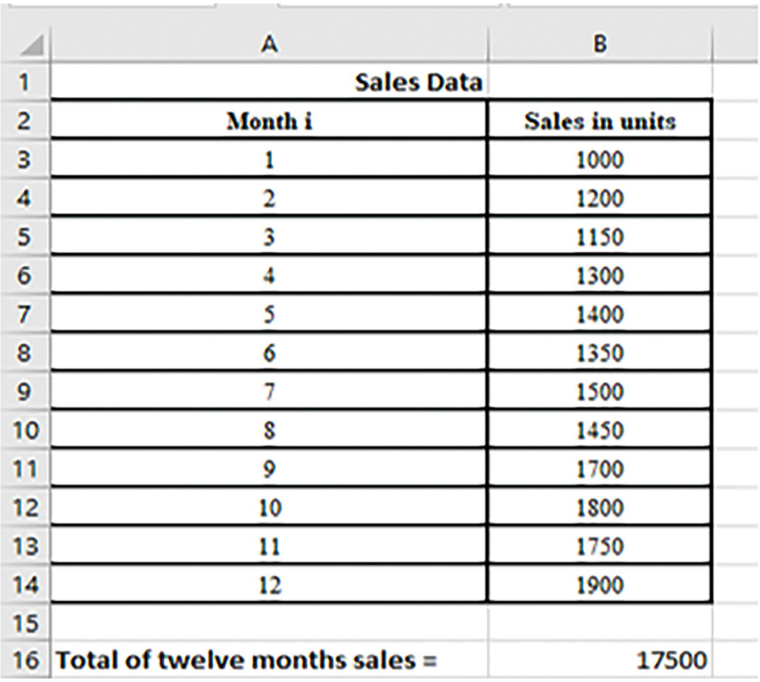
B16

	A	B
1	Sales data	
2	Month (i)	Sales in Units
3	1	1000
4	2	1200
5	3	1150
6	4	1300
7	5	1400
8	6	1350
9	Total sales of six months =	=B3+B4+B5+B6+B7+B8
10		

Figure 2.10 Screenshot of formula to obtain sum of values from Cells B3 to B8

Table 2.2 Sales Data for Product

Month (i)	*Sales in Units*
1	1,000
2	1,200
3	1,150
4	1,300
5	1,400
6	1,350
7	1,500
8	1,450
9	1,700
10	1,800
11	1,750
12	1,900



	A	B
1	Sales Data	
2	Month i	Sales in units
3	1	1000
4	2	1200
5	3	1150
6	4	1300
7	5	1400
8	6	1350
9	7	1500
10	8	1450
11	9	1700
12	10	1800
13	11	1750
14	12	1900
15		
16	Total of twelve months sales =	17500

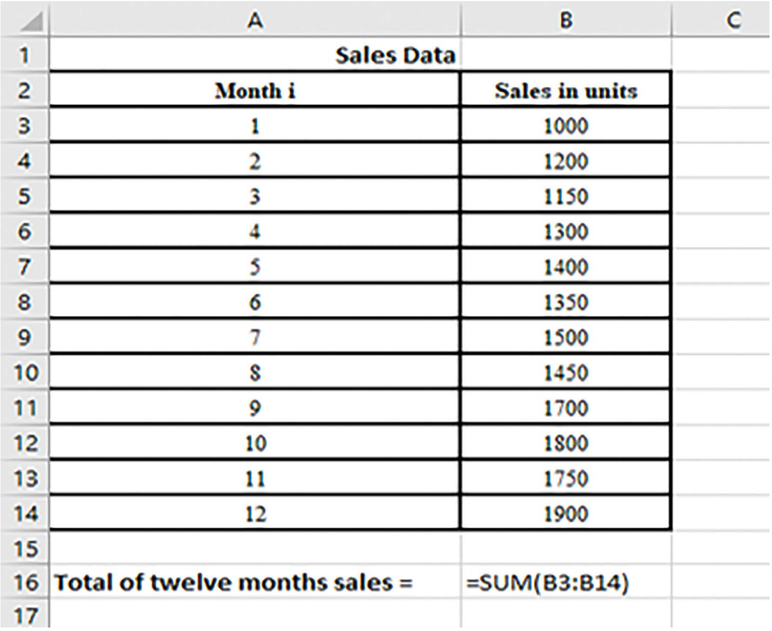
Figure 2.11 Screenshot of adding range of cells using @SUM function

2.2.8 Copying Formulas

Formulas in Excel can be copied if the same formula is to be used in a range of cells. Consider an example of the product code, number of units sold per month, and unit price of the products of a company, as shown in Table 2.3.

The formula to find the total sales of the product *i* is given by the following.

$$Total\,sales\,of\ product\ \ i\left(TSP_i\right) = D_i \times p_i, i = 1,2,3,\ldots,10$$



	A	B	C
1	Sales Data		
2	Month i	Sales in units	
3	1	1000	
4	2	1200	
5	3	1150	
6	4	1300	
7	5	1400	
8	6	1350	
9	7	1500	
10	8	1450	
11	9	1700	
12	10	1800	
13	11	1750	
14	12	1900	
15			
16	Total of twelve months sales =	=SUM(B3:B14)	
17			

Figure 2.12 Screenshot showing formula to add range of cells using @SUM function

Table 2.3 Data for Products of the Company

Product Code *(i)*	*Number of Units Sold* *(D_i)*	*Unit Price in ₹* *(p_i)*
C01	1,000	100
C02	1,200	90
C03	800	80
C04	500	40
C05	750	78
C06	900	150
C07	600	50
C08	550	25
C09	650	89
C10	200	200

Later, to find the sum of the total sales of the products, the following formula is to be used.

$$\text{Sum of total sales of the products} = \text{Sum}\left[TSP_i, i = 1, 2, 3, \ldots., 10\right]$$

As seen in the screenshot of Figure 2.13, the information from Table 2.3 has now been copied into an Excel spreadsheet. The total sales of product C01 are now calculated in cell D4 by multiplying the number of units sold in cell B4 by the unit price in cell C4 using the following method, as illustrated in the screenshot in Figure 2.13,

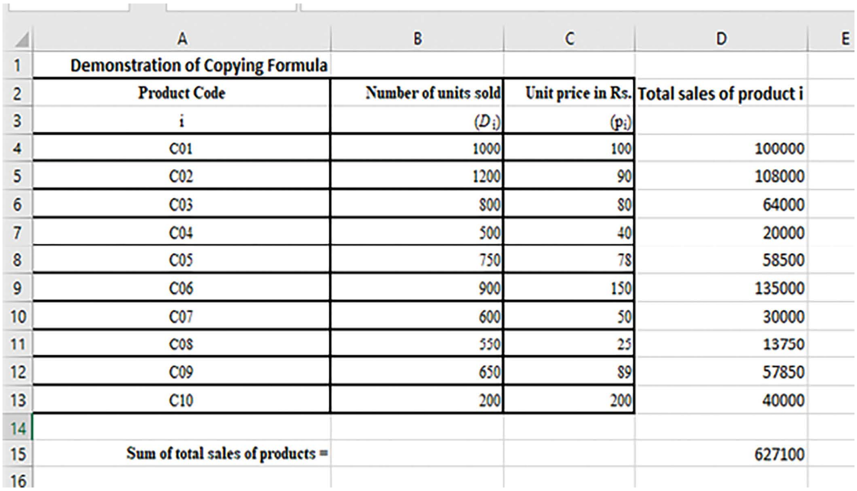
=B4+C4. The remaining products from C02 through C10's total sales can be calculated using the previous procedure. To do this, keep the pointer on cell D4's bottom right corner and move it up to cell D13. When the formula is copied, the results are the total sales for products C02 through C10 from cells D5 through D13, respectively.

Later, the grand total of these multiplied values is to be obtained using the following formula for the @SUM function. This is the sum of the total sales of all ten products.

@SUM (D4 : D13)

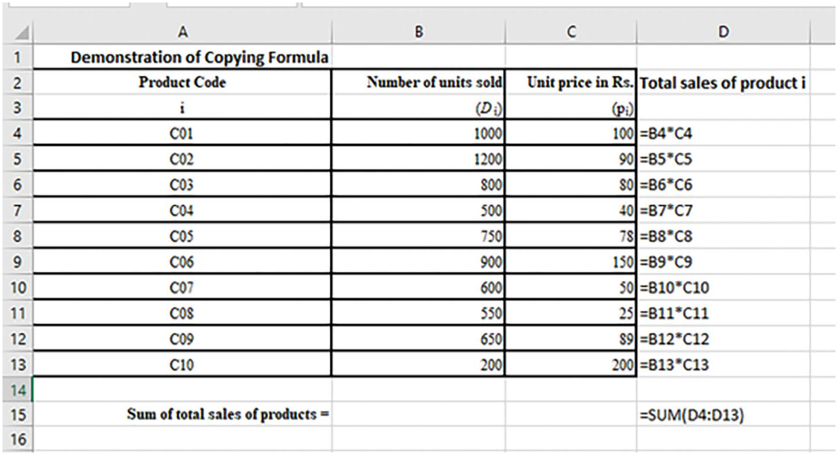
This function contains the range of cells from which the values are to be added. This formula is entered in cell D15.

The formulas for all the calculations in Figure 2.13 are shown in Figure 2.14.



	A	B	C	D	E
1	Demonstration of Copying Formula				
2	Product Code	Number of units sold	Unit price in Rs.	Total sales of product i	
3	i	(D_i)	(p_i)		
4	C01	1000	100	100000	
5	C02	1200	90	108000	
6	C03	800	80	64000	
7	C04	500	40	20000	
8	C05	750	78	58500	
9	C06	900	150	135000	
10	C07	600	50	30000	
11	C08	550	25	13750	
12	C09	650	89	57850	
13	C10	200	200	40000	
14					
15	Sum of total sales of products =			627100	
16					

Figure 2.13 Screenshot of working with copying of formulas for the example given



	A	B	C	D
1	Demonstration of Copying Formula			
2	Product Code	Number of units sold	Unit price in Rs.	Total sales of product i
3	i	(D_i)	(p_i)	
4	C01	1000	100	=B4*C4
5	C02	1200	90	=B5*C5
6	C03	800	80	=B6*C6
7	C04	500	40	=B7*C7
8	C05	750	78	=B8*C8
9	C06	900	150	=B9*C9
10	C07	600	50	=B10*C10
11	C08	550	25	=B11*C11
12	C09	650	89	=B12*C12
13	C10	200	200	=B13*C13
14				
15	Sum of total sales of products =			=SUM(D4:D13)
16				

Figure 2.14 Screenshot of guidelines for formulas of the working with copying of formulas for the example given

2.2.9 SUMIF Function

Think about a situation where various types of houses require varied amounts of electricity. Consider the two types of housing categories for the discussion: hut and concrete house. There are 15 residences in a street that fall under these categories. The electricity board is interested in learning how much electricity is used overall in each type of residence. Table 2.4 displays the information on power consumption in KWH.

The SUMIF function helps to find the total power consumption in each of the categories of houses. The button sequence ⟹ Formulas ⟹ Math & Trig ⟹ SUMIF gives the screenshot, as shown in Figure 2.15.

In the dropdown menu of Figure 2.15, the range should be entered as B5:B19, the criteria should be entered as =CH, and the sum range should be entered as C5:C19, which are as shown in the dropdown menu of Figure 2.16 to find the total power consumption in concrete houses. Clicking the OK button in the dropdown menu of Figure 2.16 will give the total power consumption by concrete houses, as shown in Figure 2.18.

In the dropdown menu of Figure 2.15, the range should be entered as B5:B19, the criteria should be entered as =H, and the sum range should be entered as C5:C19, which are as shown in the dropdown menu of Figure 2.17 to find the total power consumption in huts. Clicking the OK button in the dropdown menu of Figure 2.17 will give the total power consumption by the huts, as shown in Figure 2.18.

These can be achieved by the following formulas.

The formula to find the power consumption by concrete houses is as follows.

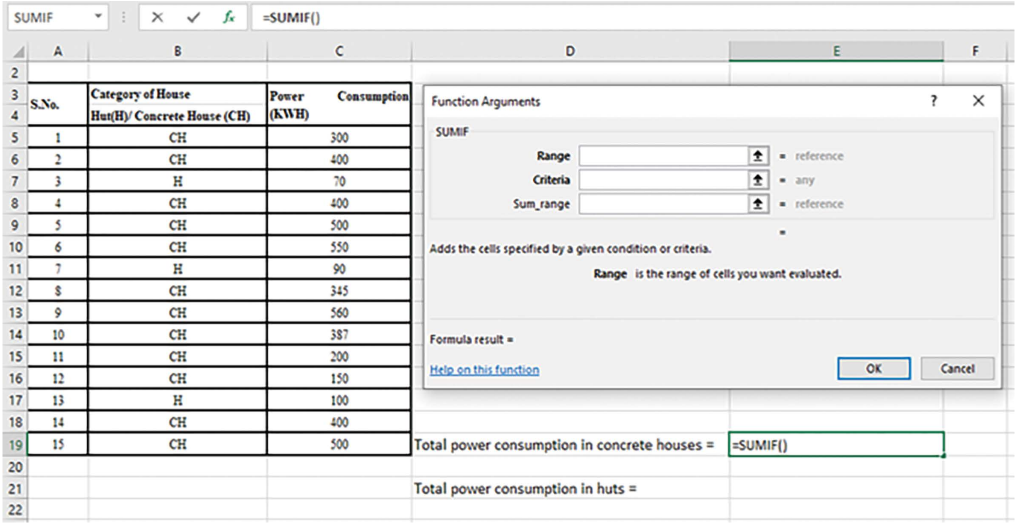
$$= SUMIF\left(B5:B19, " = CH'', C5:C19\right)$$

The formula to find the power consumption by huts is as follows.

$$= SUMIF\left(B5:B19, " = H', C5:C19\right)$$

Table 2.4 Data for Power Consumption

S. No.	*Category of House* *Hut (H)/Concrete House (CH)*	*Power Consumption (KWH)*
1	CH	300
2	CH	400
3	H	70
4	CH	400
5	CH	500
6	CH	550
7	H	90
8	CH	345
9	CH	560
10	CH	387
11	CH	200
12	CH	150
13	H	100
14	CH	400
15	CH	500



SUMIF =SUMIF()

S.No.	Category of House Hut(H)/ Concrete House (CH)	Power Consumption (KWH)
1	CH	300
2	CH	400
3	H	70
4	CH	400
5	CH	500
6	CH	550
7	H	90
8	CH	345
9	CH	560
10	CH	387
11	CH	200
12	CH	150
13	H	100
14	CH	400
15	CH	500

Total power consumption in concrete houses = =SUMIF()

Total power consumption in huts =

Function Arguments

SUMIF

Range = reference

Criteria = any

Sum_range = reference

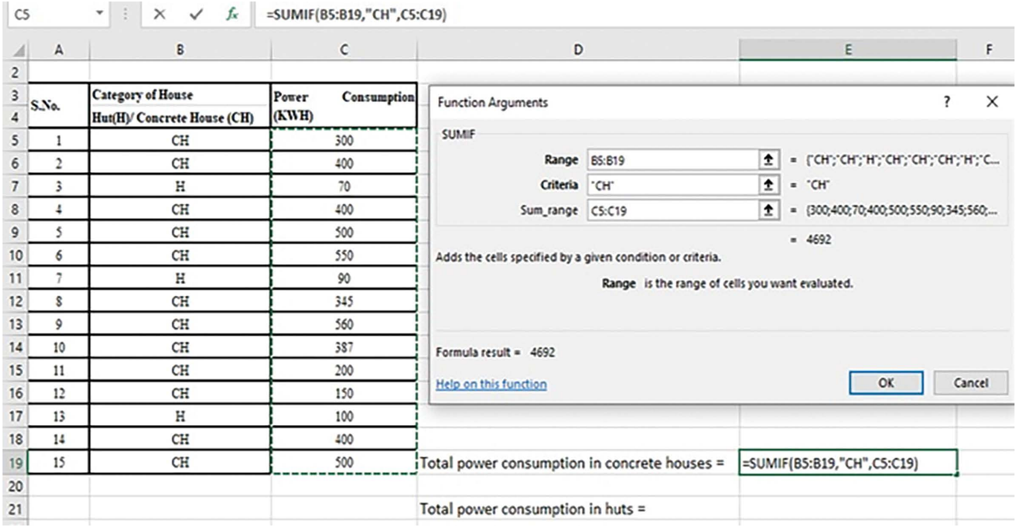
Adds the cells specified by a given condition or criteria.

Range is the range of cells you want evaluated.

Formula result =

Help on this function

Figure 2.15 Screenshot for the sequence of buttons Formulas $\Longrightarrow$ Maths & Trig $\Longrightarrow$ SUMIF, along with the data



C5 =SUMIF(B5:B19,"CH",C5:C19)

S.No.	Category of House Hut(H)/ Concrete House (CH)	Power Consumption (KWH)
1	CH	300
2	CH	400
3	H	70
4	CH	400
5	CH	500
6	CH	550
7	H	90
8	CH	345
9	CH	560
10	CH	387
11	CH	200
12	CH	150
13	H	100
14	CH	400
15	CH	500

Total power consumption in concrete houses = =SUMIF(B5:B19,"CH",C5:C19)

Total power consumption in huts =

Function Arguments

SUMIF

Range B5:B19 = {"CH";"CH";"H";"CH";"CH";"CH";"H";"C...

Criteria "CH" = "CH"

Sum_range C5:C19 = {300;400;70;400;500;550;90;345;560;...

= 4692

Adds the cells specified by a given condition or criteria.

Range is the range of cells you want evaluated.

Formula result = 4692

Help on this function

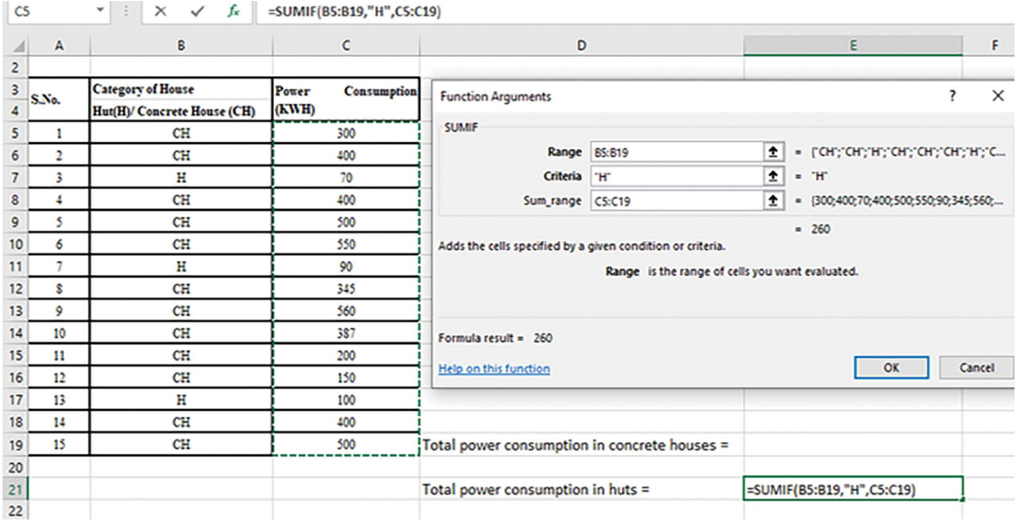
Figure 2.16 Screenshot after selecting range, Criteria "=CH" and Sum Range in the dropdown menu of Figure 2.15

2.2.10 SUMIFS Command

The SUMIFS command helps to find the sum of the values in a range of cells for the conditions of two criteria.

The syntax of SUMIFS is:

$$= SUMIFS\ (Sum\,range, Criteria\,1\,Range, Criteria\,1, Criteria\,2\,Range, Criteria\,2)$$



C5 =SUMIF(B5:B19,"H",C5:C19)

S.No.	Category of House Hut(H)/ Concrete House (CH)	Power Consumption (KWH)	
1	CH	300	
2	CH	400	
3	H	70	
4	CH	400	
5	CH	500	
6	CH	550	
7	H	90	
8	CH	345	
9	CH	560	
10	CH	387	
11	CH	200	
12	CH	150	
13	H	100	
14	CH	400	
15	CH	500	Total power consumption in concrete houses =
			Total power consumption in huts = =SUMIF(B5:B19,"H",C5:C19)

Function Arguments

SUMIF

Range B5:B19 = {"CH";"CH";"H";"CH";"CH";"CH";"H";"C...

Criteria "H" = "H"

Sum_range C5:C19 = {300;400;70;400;500;550;90;345;560;...

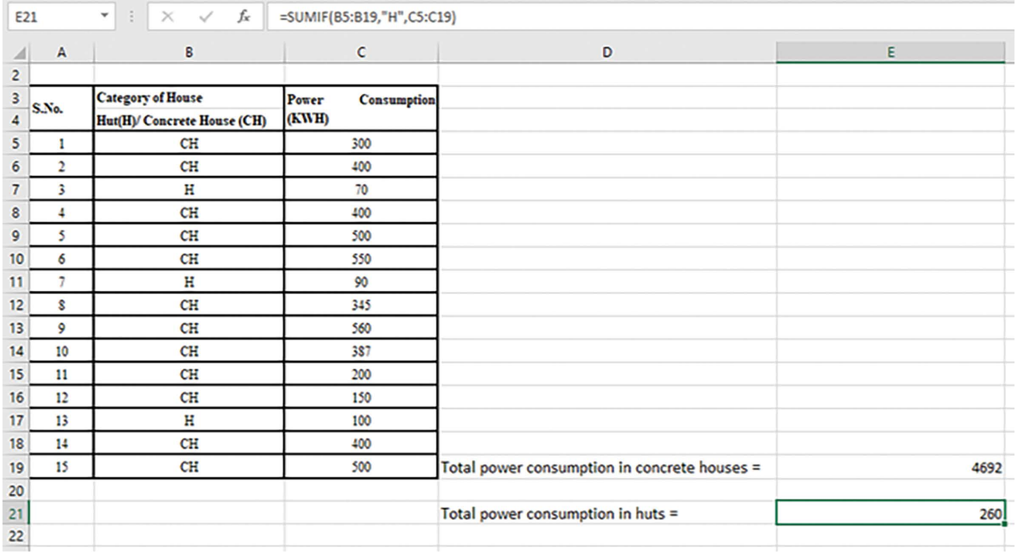
= 260

Adds the cells specified by a given condition or criteria.

Range is the range of cells you want evaluated.

Formula result = 260

Figure 2.17 Screenshot after selecting range, Criteria "=H" and Sum Range in the dropdown menu of Figure 2.15



E21 =SUMIF(B5:B19,"H",C5:C19)

S.No.	Category of House Hut(H)/ Concrete House (CH)	Power Consumption (KWH)		
1	CH	300		
2	CH	400		
3	H	70		
4	CH	400		
5	CH	500		
6	CH	550		
7	H	90		
8	CH	345		
9	CH	560		
10	CH	387		
11	CH	200		
12	CH	150		
13	H	100		
14	CH	400		
15	CH	500	Total power consumption in concrete houses =	4692
			Total power consumption in huts =	260

Figure 2.18 Screenshot after clicking the OK button in Figures 2.16 and 2.17

Note: "Criteria" should be "Criterion", but since it is used as "Criteria" in Excel, use "Criteria" wherever it appears.

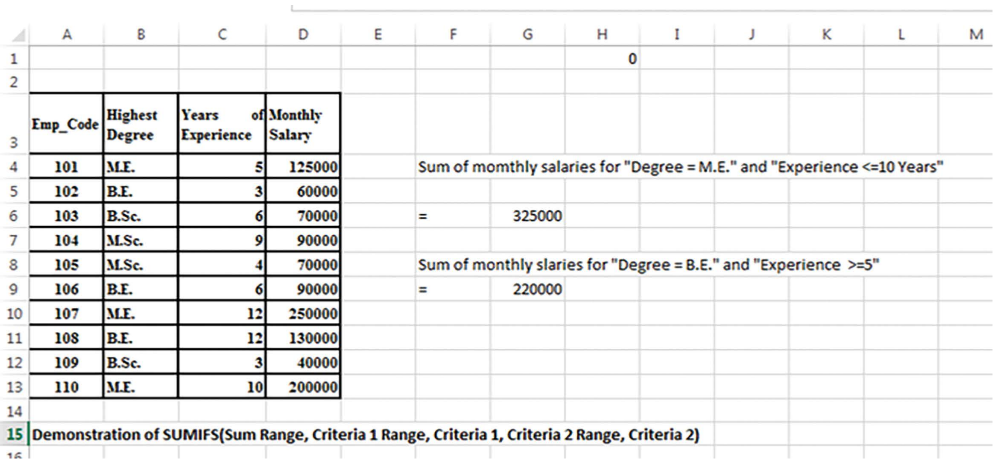
Consider an example as shown in Table 2.5.

Based on the data shown in Table 2.5, the following two situations are presented using the SUMIFS command.

Sum of the salaries of employees whose highest degree is ME and experience is less than or equal to 10 years.

Table 2.5 Details of Employees

Emp_Code	*Highest Degree*	*Years of Experience*	*Monthly Salary*
101	ME	5	125,000
102	BE	10	120,000
103	BSc	6	70,000
104	MSc	9	90,000
105	MSc	4	70,000
106	BE	6	90,000
107	ME	12	250,000
108	BE	12	130,000
109	BSc	3	40,000
110	ME	5	130,000



Emp_Code	Highest Degree	Years of Experience	Monthly Salary
101	M.E.	5	125000
102	B.E.	3	60000
103	B.Sc.	6	70000
104	M.Sc.	9	90000
105	M.Sc.	4	70000
106	B.E.	6	90000
107	M.E.	12	250000
108	B.E.	12	130000
109	B.Sc.	3	40000
110	M.E.	10	200000

0

Sum of momthly salaries for "Degree = M.E." and "Experience <=10 Years"

= 325000

Sum of monthly slaries for "Degree = B.E." and "Experience >=5"

= 220000

Demonstration of SUMIFS(Sum Range, Criteria 1 Range, Criteria 1, Criteria 2 Range, Criteria 2)

Figure 2.19 Screenshot of examples of SUMIFS command

Sum of the salaries of employees whose highest degree is BE and experience is greater than or equal to 5 years.

The data for Table 2.5 are copied starting from cell A3 to D13, as shown in Figure 2.19.

The formula to find the sum of the monthly salaries of employees for Degree = ME and Experience <= 10 years is as follows, which is entered in cell G6 in Figure 2.19.

= ***SUMIFS*(*D*4 : *D*13, *B*4 : *B*13, " = *M*.*E*.", C4 : C13, " < = 10"**

The formula to find the sum of the monthly salaries of employees for Degree = BE and Experience >=5 years is as given below, which is entered in cell G9 in Figure 2.19.

= ***SUMIFS*(*D*4 : *D*13, *B*4 : *B*13, " = *B*.*E*.", C4 : C13, " > = 5")**

A screenshot of the result of these commands is shown in Figure 2.19.

2.2.11 Absolute References to Cells While Copying Formulas

While copying a formula from a cell to another cell, it follows the relative positioning of cells which are involved in the formula. For example, if C1 = A1*B1 and this formula is copied to the cells from C2 to C5, then the respective changes in the formula will be as follows. The cell references are automatically changed when we copy the formula from C1 to the other cells from C2 to C5.

$$= \text{For cell C2} : C2 = A2 * B2$$

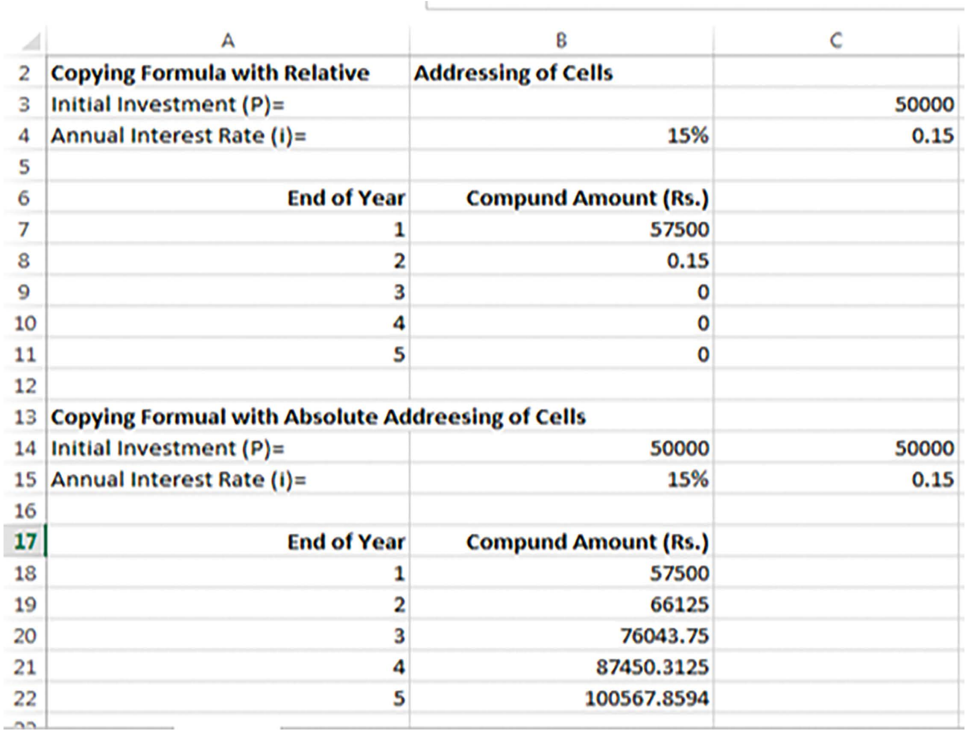
$$= \text{For cell C3} : C3 = A3 * B3$$

$$= \text{For cell C4} : C4 = A4 * B4$$

$$= \text{For cell C5} : C5 = A5 * B5$$

This may be advantageous in some situations, but it may cause serious problems in other situations.

To demonstrate this problem, consider the task of computing the compound amount (*F*) at the end of every year of an initial investment of ₹ 50,000 (*P*) for a specified period, say, 5 years at an annual interest rate *i*. These are summarised in the screenshot shown in Figure 2.20.The initial investment (*P*) is given in cell C3, and the annual interest rate (*i*) is given in cell C4.



	A	B	C
2	**Copying Formula with Relative**	**Addressing of Cells**	
3	**Initial Investment (P)=**		**50000**
4	**Annual Interest Rate (i)=**	**15%**	**0.15**
5			
6	**End of Year**	**Compund Amount (Rs.)**	
7	**1**	**57500**	
8	**2**	**0.15**	
9	**3**	**0**	
10	**4**	**0**	
11	**5**	**0**	
12			
13	**Copying Formual with Absolute Addreesing of Cells**		
14	**Initial Investment (P)=**	**50000**	**50000**
15	**Annual Interest Rate (i)=**	**15%**	**0.15**
16			
17	**End of Year**	**Compund Amount (Rs.)**	
18	**1**	**57500**	
19	**2**	**66125**	
20	**3**	**76043.75**	
21	**4**	**87450.3125**	
22	**5**	**100567.8594**	

Figure 2.20 Screenshot of working for compound amounts with relative addressing and absolute addressing of cells

The formula for the compound amount is $(F) = P \times (1+i)^n$

where
P is the initial investment
i is the annual interest rate
n is the period of investment

Under the heading, “Copying Formula with Relative Addressing of Cells”, the formula for the compound amount as given is entered in cell B7. Note that the power *m* of, say, *k* will be written in a formula like *k*^*m*.

$$= C3 * (1 + C4) \wedge A7$$

When this formula is copied from cell B8 to cell B11, the next four values after cell C3 vertically down will be substituted in place of C3 for cells B8, B9, B10, and B11, respectively. These values will be 0.15, 0, 0, and 0, which are wrong values, because the value of the initial investment remains constant as ₹ 50,000. Similarly, the values for annual interest rate substituted for cells B8, B9, B10, and B11 will be 0, 0, 0, and 0 from cells C5, C6, C7, and C8, respectively, which are wrong values, because the annual interest rate should be a constant and equals 15% (0.15). Hence, cells B8, B9, B10, and B11 contain wrong values.

This problem can be avoided by prefixing the $ symbol to the row and to the column, defining a cell which should have an absolute address in the formula when it is used, like the cell containing the initial investment and the cell containing the annual interest rate.

Wherever these cells are used in a formula, they should appear as follows:

Cell C3, which contains the initial investment: It should be typed as $C$3

Cell C4, which contains the annual interest rate: It should be typed as $C$4

Per these guidelines, in the lower part of the screenshot shown in Figure 2.15, the $ symbol is prefixed to the formula given in cell B18, as follows. Later, this formula is copied to the next four cells vertically down column B.

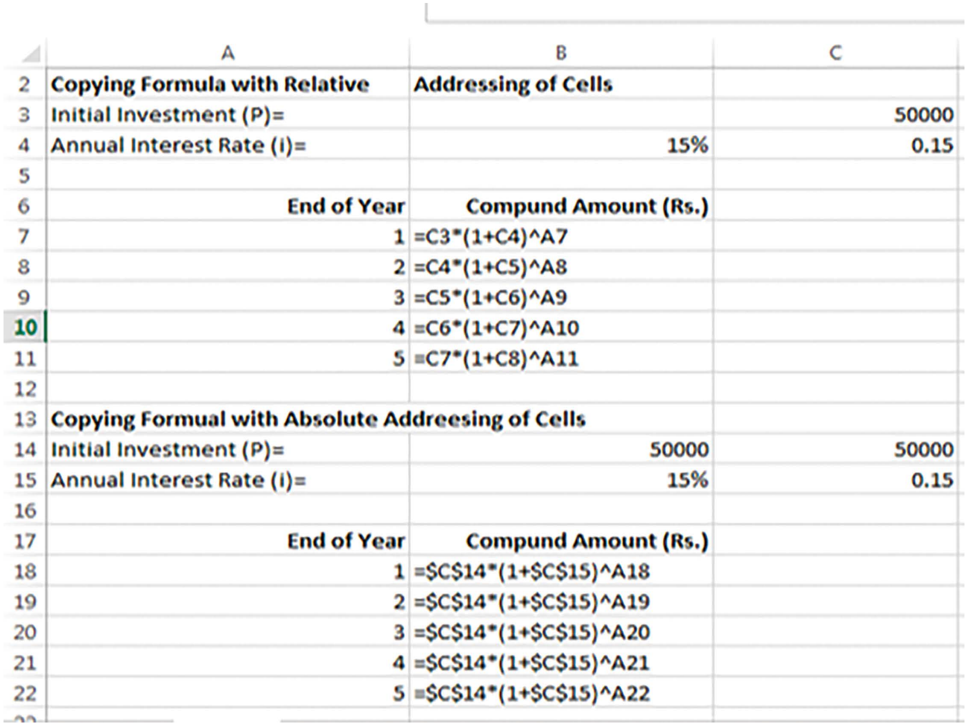
$$= \$C\$14 * (1 + \$C\$15) \wedge A18$$

The guidelines of the formulas for the working in the upper and lower parts of Figure 2.20 are shown in Figure 2.21.

2.2.12 COMBIN Command

In computing the probability of a distribution and computing test statistics of some tests, the investigator has to compute a number of combinations of a given set of numbers from 1 to *N* by taking, say, R at a time, which is ${}^{N}C_{R}$. If the numbers from 1 to 3 are considered, then ${}^{3}C_{2}$ will have three different combinations, which are as follows.

1, 2
1, 3
2, 3



	A	B	C
2	**Copying Formula with Relative**	**Addressing of Cells**	
3	Initial Investment (P)=		50000
4	Annual Interest Rate (i)=	15%	0.15
5			
6	**End of Year**	**Compund Amount (Rs.)**	
7	1	=C3*(1+C4)^A7	
8	2	=C4*(1+C5)^A8	
9	3	=C5*(1+C6)^A9	
10	4	=C6*(1+C7)^A10	
11	5	=C7*(1+C8)^A11	
12			
13	**Copying Formual with Absolute Addreesing of Cells**		
14	Initial Investment (P)=	50000	50000
15	Annual Interest Rate (i)=	15%	0.15
16			
17	**End of Year**	**Compund Amount (Rs.)**	
18	1	=$C$14*(1+$C$15)^A18	
19	2	=$C$14*(1+$C$15)^A19	
20	3	=$C$14*(1+$C$15)^A20	
21	4	=$C$14*(1+$C$15)^A21	
22	5	=$C$14*(1+$C$15)^A22	

Figure 2.21 Screenshot of guidelines for the formulas of working for compound amounts with relative addressing and absolute addressing of cells

The Excel command to find the number of combinations of N_{C_R} is:

$$= \text{COMBIN}(\text{Number, Number Chosen})$$

If N is 5 and R is 2, then N_{C_R} ; that is, 5_{C_2} is 10.
The Excel command for this case is as given.

$$= \text{COMBIN}(5,2), \text{ which gives 10 as the result}$$

This means that ten different pairs of combinations of numbers can be generated by taking two numbers at a time from the numbers from 1 to 5. This formula can be entered in any convenient cell of Excel as per the requirement of a problem that is solved.

2.2.13 FACT Command

The factorial of a given number can be obtained using the FACT command in Excel.
The syntax of the FACT command is as given.

$$= \text{FACT}(\text{Number})$$

The formula to find the factorial of 5 is as given.

= FACT(5)

This will give the result as 120 in the cell of Excel sheet where this formula is entered.

2.2.14 *Invoking the Data Analysis Button in Excel*

The Data Analysis button in Excel can be invoked using the following steps for Excel 2008, 2013, 2016, and 2019.

1. Open an Excel sheet.
2. Click File.
3. Click Options.
4. Click Add-Ins option in the dropdown menu of Step 3.
5. Click Excel Add-Ins in the Manage box at the bottom of the dropdown menu of Step 4 and click Go.
6. Click Analysis ToolPak in the dropdown menu of Step 5 and then click OK in the same dropdown menu.

Now, the Data Analysis button can be seen at the right side of the ribbon after clicking the DATA button in the main ribbon, as shown in Figure 2.22. Clicking the Data Analysis button in Figure 2.22 gives the display in Figure 2.23.

2.2.15 *MIN Function*

The MIN function finds the minimum among a given set of observations. The sequence of button clicks to obtain the minimum of the given set of observations is as given.

Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ MIN

A screenshot of this sequence of button clicks with the data of an example is shown in Figure 2.24.

In the dropdown menu of Figure 2.24, entering the range of cells B2:B7 against Number 1 and clicking OK gives the minimum of the observations in the range of cells B2:B7 in cell B10, where the cursor has already been positioned, as shown in Figure 2.25.

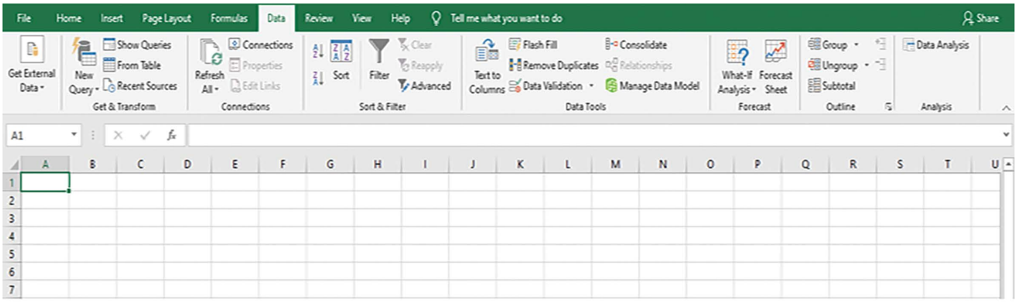


Figure 2.22 Screenshot showing Data Analysis button

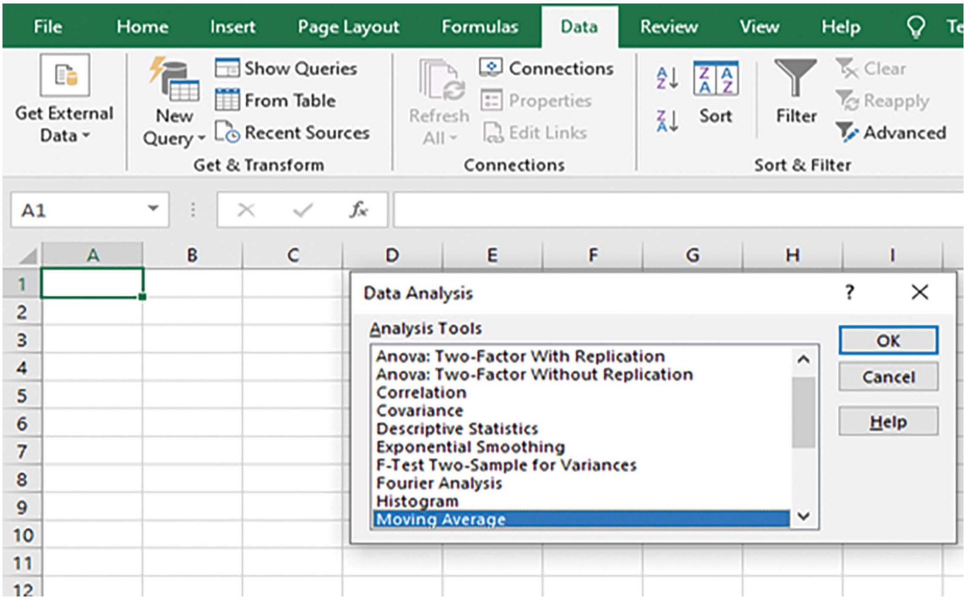
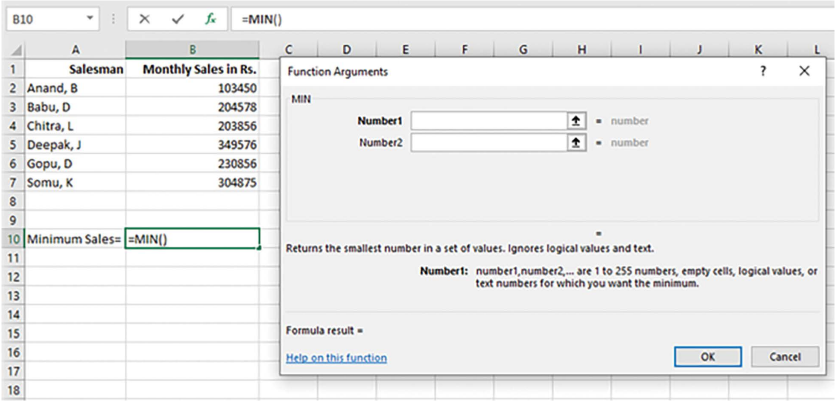


Figure 2.23 Screenshot of clicking Data Analysis button in the ribbon of Figure 2.22



	A	B
1	Salesman	Monthly Sales in Rs.
2	Anand, B	103450
3	Babu, D	204578
4	Chitra, L	203856
5	Deepak, J	349576
6	Gopu, D	230856
7	Somu, K	304875
8		
9		
10	Minimum Sales=	=MIN()
11		
12		
13		
14		
15		
16		
17		
18		

Returns the smallest number in a set of values. Ignores logical values and text.

Number1: number1,number2,... are 1 to 255 numbers, empty cells, logical values, or text numbers for which you want the minimum.

Formula result =

Help on this function

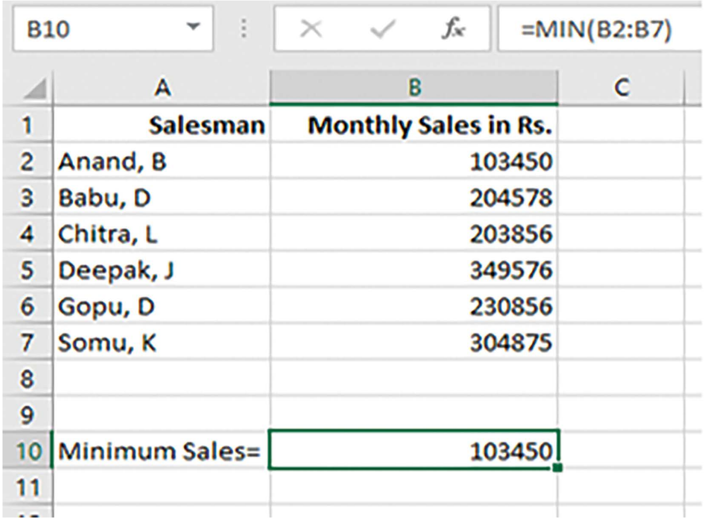
Figure 2.24 Screenshot for clicks of buttons, Home $\Longrightarrow$ Formulas $\Longrightarrow$ More Functions $\Longrightarrow$ Statistical $\Longrightarrow$ MIN

All these steps can be combined in the form of a formula, as shown, which is to be typed in cell B10 of Figure 2.24 to show the minimum monthly sales.

$$= \text{MIN}(\text{B2} : \text{B7})$$

2.2.16 *MINA Function*

The MINA function finds the minimum among a given set of observations, that is, numeric value, logical, and text. The TRUE of a logical is assumed to be 1 and the FALSE



B10 =MIN(B2:B7)

	A	B	C
1	Salesman	Monthly Sales in Rs.	
2	Anand, B	103450	
3	Babu, D	204578	
4	Chitra, L	203856	
5	Deepak, J	349576	
6	Gopu, D	230856	
7	Somu, K	304875	
8			
9			
10	Minimum Sales=	103450	
11			

Figure 2.25 Screenshot after filling the range of cells against Number 1 in the dropdown menu of Figure 2.24 and clicking OK

of a logical is assumed to be 0. Any other text is assumed to be 0. The sequence of button clicks to obtain the minimum of the given set of combinations of observations is as given.

Home $\Longrightarrow$ Formula $\Longrightarrow$ More Functions $\Longrightarrow$ Statistical $\Longrightarrow$ MINA

A screenshot of the sequence of button clicks with the data of an example is shown in Figure 2.26.

In the dropdown menu of Figure 2.26, entering the range of cells A2:A10 against Value 1 and clicking OK gives the minimum of the observations in the range of cells A2:A10 in cell B11, where the cursor has already been positioned as shown in Figure 2.27.

2.2.17 MAX Function

The MAX function finds the maximum among a given set of observations. The sequence of button clicks to obtain the maximum of the given set of observations is as given.

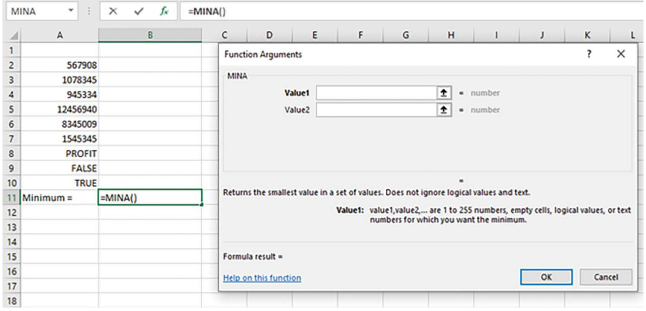
Home $\Longrightarrow$ Formula $\Longrightarrow$ More Functions $\Longrightarrow$ Statistical $\Longrightarrow$ MAX

A screenshot of the sequence of button clicks with the data of an example is shown in Figure 2.28.

In the dropdown menu of Figure 2.28, entering the range of cells B2:B7 against Number 1 and clicking OK gives the maximum of the observations in the range of cells B2:B7 in cell B10, where the cursor has already been positioned as shown in Figure 2.29.

All these steps can be combined in the form of a formula, as shown, which is to be typed in cell B10 of Figure 2.28 to show the maximum monthly sales.

$$= MAX(B2:B7)$$



MINA | =MINA()

	A	B
1		
2	567908	
3	1078345	
4	945334	
5	12456940	
6	8345009	
7	1545345	
8	PROFIT	
9	FALSE	
10	TRUE	
11	Minimum =	=MINA()

Function Arguments

MINA

Value1 = number

Value2 = number

=

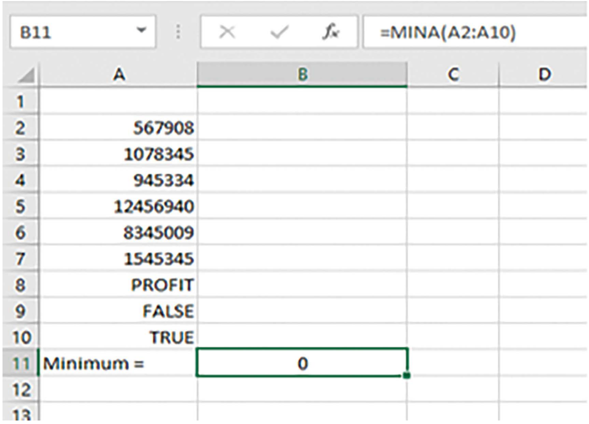
Returns the smallest value in a set of values. Does not ignore logical values and text.

Value1: value1,value2,... are 1 to 255 numbers, empty cells, logical values, or text numbers for which you want the minimum.

Formula result =

Help on this function

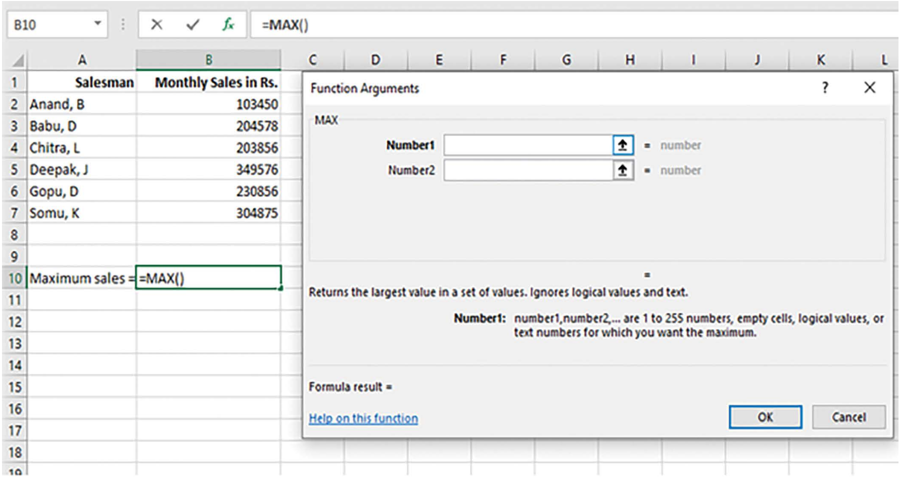
Figure 2.26 Screenshot for button clicks, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ MINA



B11 | =MINA(A2:A10)

	A	B
1		
2	567908	
3	1078345	
4	945334	
5	12456940	
6	8345009	
7	1545345	
8	PROFIT	
9	FALSE	
10	TRUE	
11	Minimum =	0

Figure 2.27 Screenshot after filling the range of cells against Number 1 in the dropdown menu of Figure 2.26 and clicking OK



B10 | =MAX()

	A	B
1	Salesman	Monthly Sales in Rs.
2	Anand, B	103450
3	Babu, D	204578
4	Chitra, L	203856
5	Deepak, J	349576
6	Gopu, D	230856
7	Somu, K	304875
8		
9		
10	Maximum sales =	=MAX()

Function Arguments

MAX

Number1 = number

Number2 = number

=

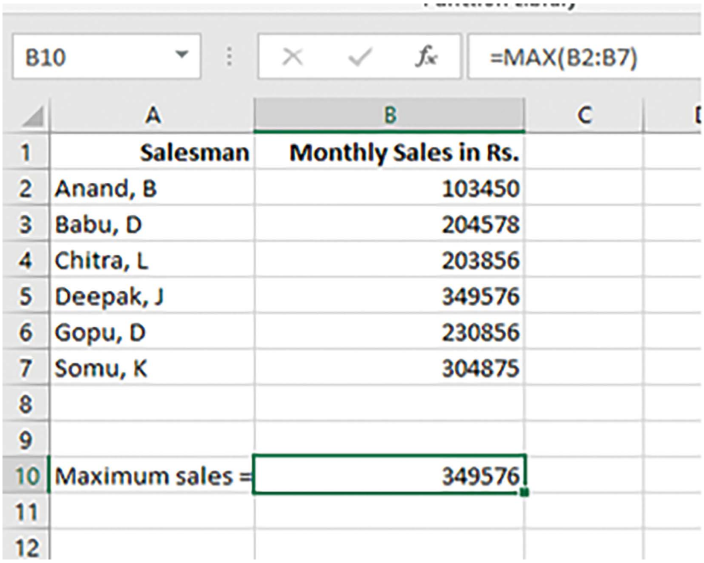
Returns the largest value in a set of values. Ignores logical values and text.

Number1: number1,number2,... are 1 to 255 numbers, empty cells, logical values, or text numbers for which you want the maximum.

Formula result =

Help on this function

Figure 2.28 Screenshot for button clicks, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ MAX



B10 =MAX(B2:B7)

	A	B	C
1	Salesman	Monthly Sales in Rs.	
2	Anand, B	103450	
3	Babu, D	204578	
4	Chitra, L	203856	
5	Deepak, J	349576	
6	Gopu, D	230856	
7	Somu, K	304875	
8			
9			
10	Maximum sales =	349576	
11			
12			

Figure 2.29 Screenshot after filling the range of cells against Number 1 in the dropdown menu of Figure 2.28 and clicking OK

2.2.18 MAXA Function

The MAXA function determines the maximum among a set of observations, including text, logical, and numeric values. It is assumed that a logical's TRUE condition is 1, and its FALSE condition is considered 0. Any other text is taken to be zero. The button-click order to get the maximum out of a given set of observations is listed in the following.

Home ⟹ Formula ⟹ More Functions ⟹ Statistical ⟹ MAXA

A screenshot of the sequence of button clicks with the data of an example is shown in Figure 2.30.

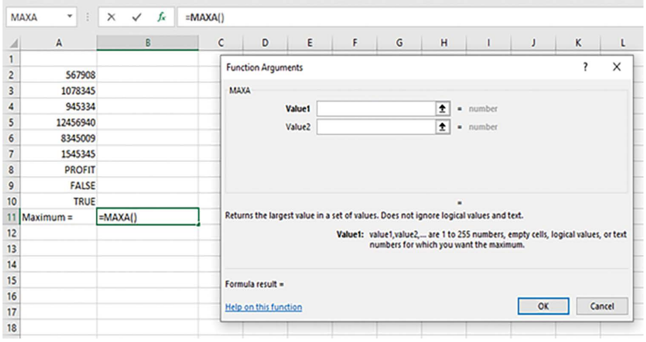
In the dropdown menu of Figure 2.30, entering the range of cells A2:A10 against Value 1 and clicking the OK button gives the maximum of the observations in the range of cells A2:A10 in cell B11, where the cursor has already been positioned, as shown in Figure 2.31.

All these steps can be combined in the form of a formula, as shown, which is to be typed in cell B11 of Figure 2.30 to show the maximum monthly sales.

$$= \text{MAXA}(\text{A2} : \text{A10})$$

2.2.19 MAXIFS Function

Excel 2016 and later versions have the MAXIFS function. The MAXIFS function determines the maximum values for one or more criteria applied to the data of the criterion range/criteria ranges, respectively, for a given data range [1, 2].



MAXA | =MAXA()

	A	B
1		
2	567908	
3	1078345	
4	945334	
5	12456940	
6	8345009	
7	1545345	
8	PROFIT	
9	FALSE	
10	TRUE	
11	Maximum =	=MAXA()

Function Arguments

MAXA

Value1 = number

Value2 = number

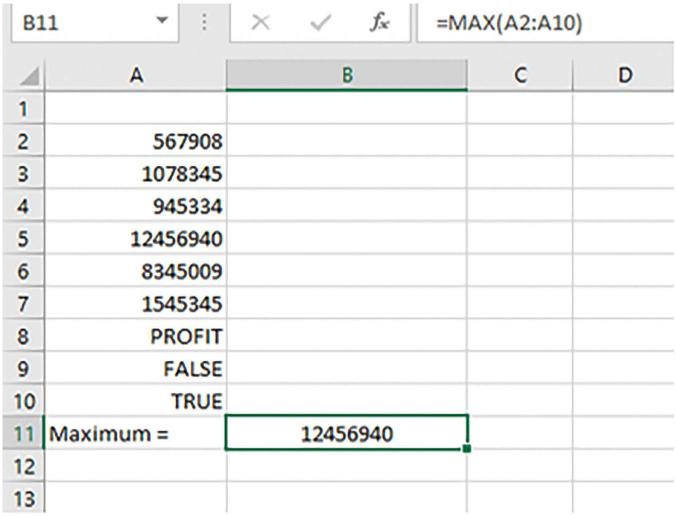
Returns the largest value in a set of values. Does not ignore logical values and text.

Value1: value1,value2,... are 1 to 255 numbers, empty cells, logical values, or text numbers for which you want the maximum.

Formula result =

Help on this function

Figure 2.30 Screenshot for button clicks, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ MAXA



B11 | =MAX(A2:A10)

	A	B	C	D
1				
2	567908			
3	1078345			
4	945334			
5	12456940			
6	8345009			
7	1545345			
8	PROFIT			
9	FALSE			
10	TRUE			
11	Maximum =	12456940		
12				
13				

Figure 2.31 Screenshot after filling the range of cells against Number 1 in the dropdown menu of Figure 2.30 and clicking OK

A screenshot of the problem is shown in Figure 2.32.

The screenshot in Figure 2.32 contains data for years from 1 to 10, incentive percentage for each year, and the sales in crores of rupees for each year. The objective of the question is to find the maximum of the sales subject to the following two criteria.

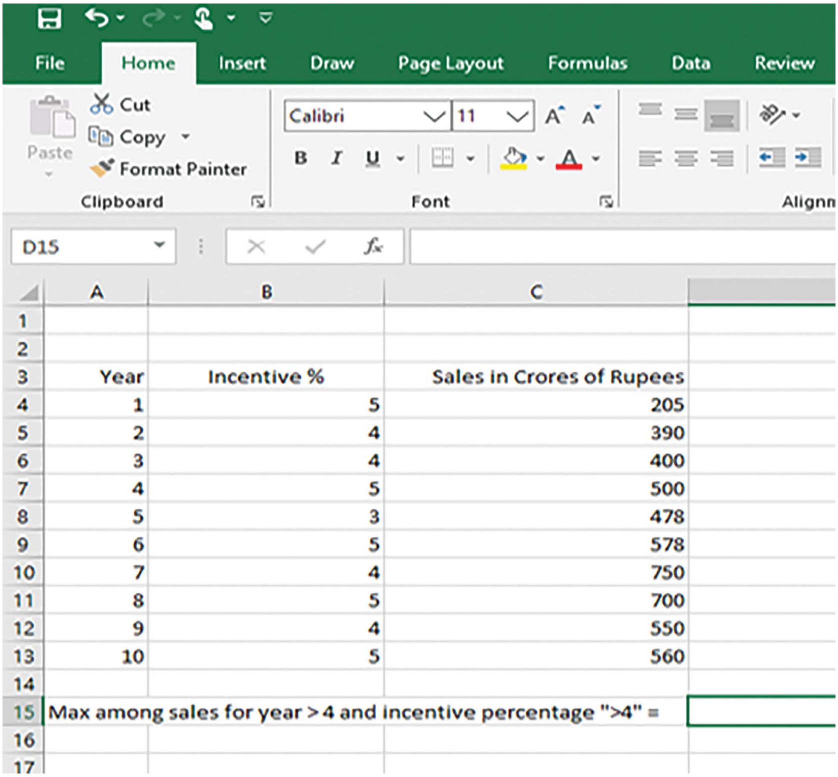
Criterion 1: Year > 4

Criterion 2: Incentive percentage = 4%

The sequence of button clicks required for the MAXIFS function is as given, which gives a screenshot along with suitable data, as shown in Figure 2.33.

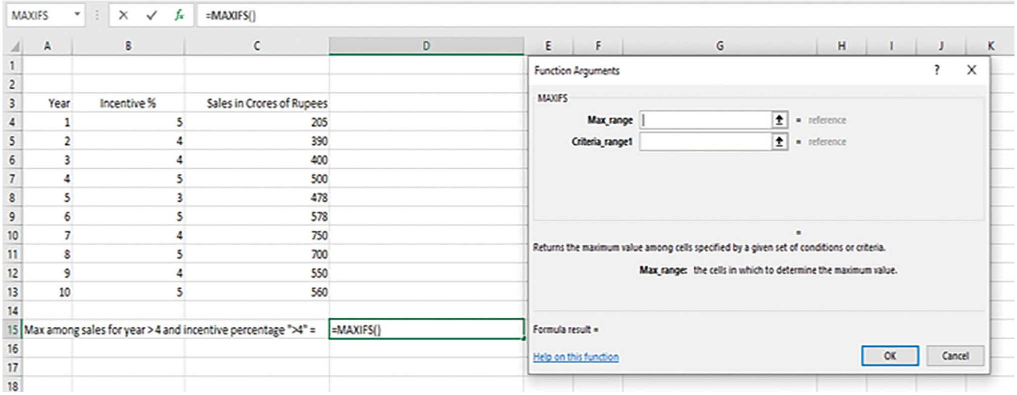
Home ⟹ Formula ⟹ More Functions ⟹ Statistical ⟹ MAXIFS

In the dropdown menu of Figure 2.33, fill the range of cells C4:C13 in the box against Max_Range, the range of cells A4:A13 in the box against as Criteria_Range1, >4 in the



	A	B	C
1			
2			
3	Year	Incentive %	Sales in Crores of Rupees
4	1	5	205
5	2	4	390
6	3	4	400
7	4	5	500
8	5	3	478
9	6	5	578
10	7	4	750
11	8	5	700
12	9	4	550
13	10	5	560
14			
15	Max among sales for year > 4 and incentive percentage ">4" =		
16			
17			

Figure 2.32 Screenshot of the data



	A	B	C	D
1				
2				
3	Year	Incentive %	Sales in Crores of Rupees	
4	1	5	205	
5	2	4	390	
6	3	4	400	
7	4	5	500	
8	5	3	478	
9	6	5	578	
10	7	4	750	
11	8	5	700	
12	9	4	550	
13	10	5	560	
14				
15	Max among sales for year > 4 and incentive percentage ">4" =			=MAXIFS()
16				
17				
18				

MAXIFS

Max_range = reference

Criteria_range1 = reference

Returns the maximum value among cells specified by a given set of conditions or criteria.

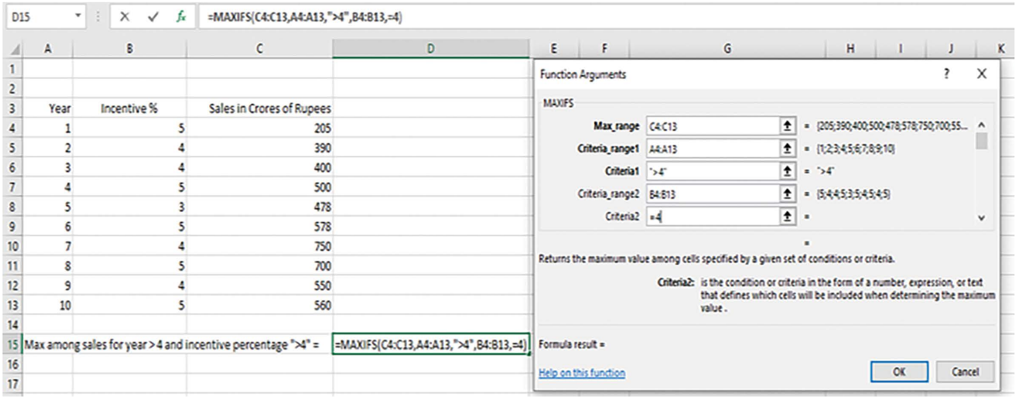
Max_range: the cells in which to determine the maximum value.

Formula result =

Help on this function

Figure 2.33 Screenshot for sequence of clicks of buttons, viz. Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ MAXIFS

box against Criteria1, the range of cells B4:B13 in the box against Criteria_Range2 and =4 in the box against Criteria2. Though Figure 2.33 contains only two boxes (Max_Range, Criteria_Range1) in its dropdown menu, the needed boxes, that is, Criteria1, Criteria_Range2, and Criteria2, will be included when data are entered, as shown in Figure 2.34.



D15 =MAXIFS(C4:C13,A4:A13,">4",B4:B13,=4)

	A	B	C	D
1				
2				
3	Year	Incentive %	Sales in Crores of Rupees	
4	1	5	205	
5	2	4	390	
6	3	4	400	
7	4	5	500	
8	5	3	478	
9	6	5	578	
10	7	4	750	
11	8	5	700	
12	9	4	550	
13	10	5	560	
14				
15	Max among sales for year >4 and incentive percentage ">4" =			=MAXIFS(C4:C13,A4:A13,">4",B4:B13,=4)
16				
17				

Function Arguments

MAXIFS

Max_range	C4:C13	= {205;390;400;500;478;578;750;700;55...
Criteria_range1	A4:A13	= {1;2;3;4;5;6;7;8;9;10}
Criteria1	">4"	= ">4"
Criteria_range2	B4:B13	= {5;4;4;5;3;5;4;5;4;5}
Criteria2	=4	=

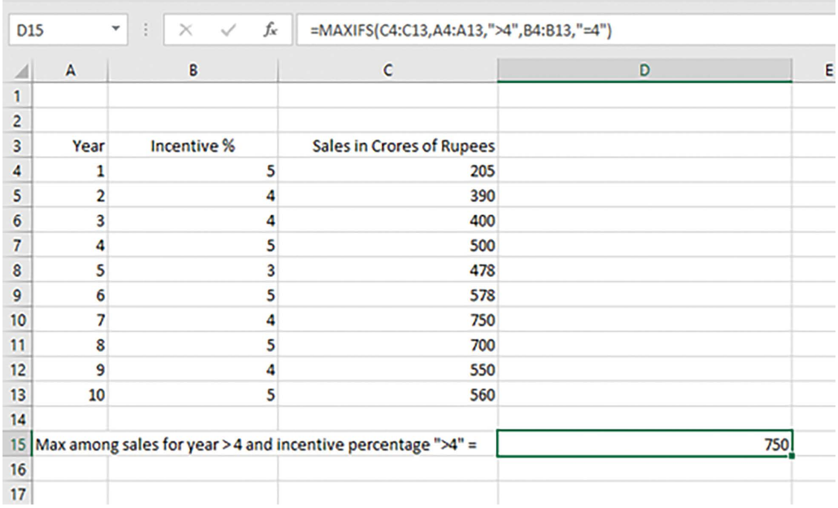
=

Returns the maximum value among cells specified by a given set of conditions or criteria.

Criteria2: is the condition or criteria in the form of a number, expression, or text that defines which cells will be included when determining the maximum value .

Formula result =

Figure 2.34 Screenshot after filling data in the dropdown menu of Figure 2.33



D15 =MAXIFS(C4:C13,A4:A13,">4",B4:B13,"=4")

	A	B	C	D
1				
2				
3	Year	Incentive %	Sales in Crores of Rupees	
4	1	5	205	
5	2	4	390	
6	3	4	400	
7	4	5	500	
8	5	3	478	
9	6	5	578	
10	7	4	750	
11	8	5	700	
12	9	4	550	
13	10	5	560	
14				
15	Max among sales for year >4 and incentive percentage ">4" =			750
16				
17				

Figure 2.35 Screenshot after clicking OK button in the dropdown menu of Figure 2.34

Clicking the OK button in the dropdown menu of Figure 2.34 gives the result shown in Figure 2.35.

The same can be achieved by typing the following formula in cell D15.

$$= MAXIFS(C4:C13, A4:A13, " > 4", B4:B13, " = 4")$$

2.2.20 MINIFS Function

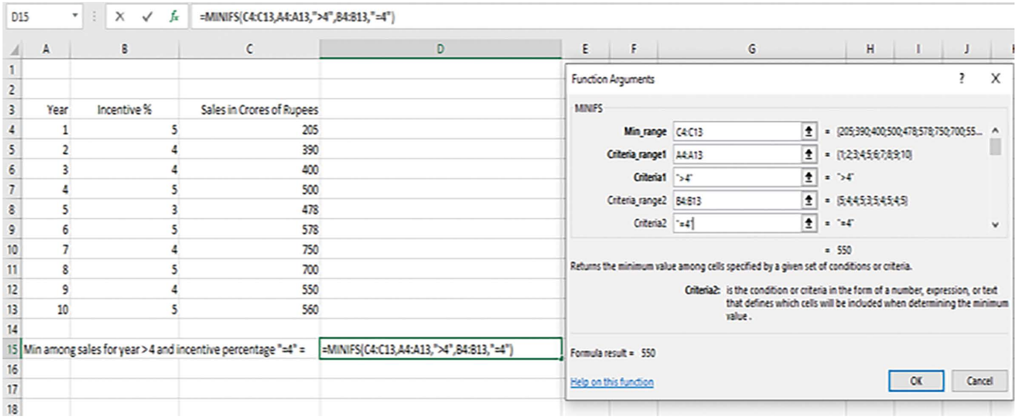
Excel 2016 and later versions support the MINIFS feature. When one or more criteria are applied to the data in the criterion range/criteria ranges, respectively, the MINIFS function determines the minimum of the values in the supplied data range [3]. Think about the information displayed in Figure 2.32.

Data for years 1 through 10 are shown in the screenshot in Figure 2.32, together with each year's incentive % and sales totals in crores of rupees. Finding the minimum sales subject to the next two conditions is the objective of the question.

Criterion 1: Year > 4

Criterion 2: Incentive percentage = 4%

The sequence of button clicks Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ MINIFS will give a screenshot similar to that in Figure 2.33 for finding the minimum with the difference that the first box in the dropdown menu is Min_Range instead of Max_Range. After filling the range of cells C4:C13 in the box against Min_Range, the range of cells A4:A13 in the box against Criteria_Range1, >4 in the box against Criteria1, the range of cells B4:B13 in the box against Criteria_Range2, and =4 in the box against Criteria2, the display will be as in Figure 2.36. Clicking the OK button in the dropdown menu of Figure 2.36 gives the result shown in Figure 2.37.



D15 =MINIFS(C4:C13,A4:A13,">4",B4:B13,"=4")

	A	B	C	D
1				
2				
3	Year	Incentive %	Sales in Crores of Rupees	
4	1	5	205	
5	2	4	390	
6	3	4	400	
7	4	5	500	
8	5	3	478	
9	6	5	578	
10	7	4	750	
11	8	5	700	
12	9	4	550	
13	10	5	560	
14				
15	Min among sales for year > 4 and incentive percentage "=4" =			=MINIFS(C4:C13,A4:A13,">4",B4:B13,"=4")

Function Arguments

MINFS

Min_range C4:C13 = {205;390;400;500;478;578;750;700;55...

Criteria_range1 A4:A13 = {1;2;3;4;5;6;7;8;9;10}

Criteria1 ">4" = ">4"

Criteria_range2 B4:B13 = {5;4;4;5;3;5;4;5;4;5}

Criteria2 "=4 = "=4"

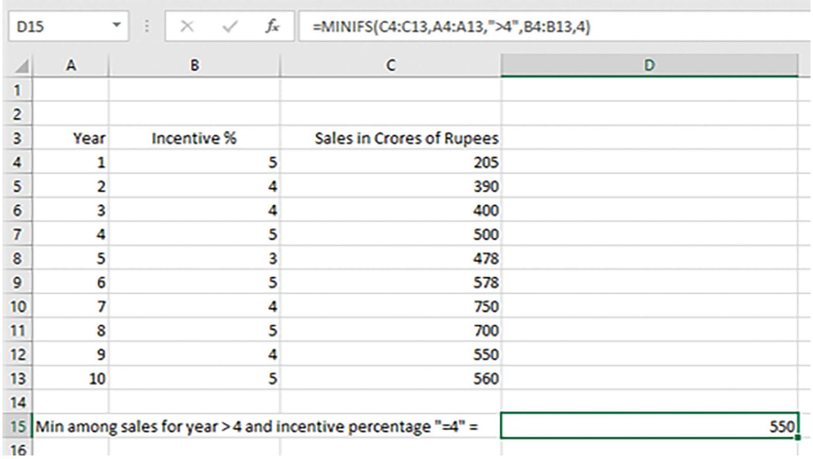
= 550

Returns the minimum value among cells specified by a given set of conditions or criteria.

Criteria2: is the condition or criteria in the form of a number, expression, or text that defines which cells will be included when determining the minimum value .

Formula result = 550

Figure 2.36 Screenshot after filling data for MINIFS function applied to problem in Figure 2.32



D15 =MINIFS(C4:C13,A4:A13,">4",B4:B13,4)

	A	B	C	D
1				
2				
3	Year	Incentive %	Sales in Crores of Rupees	
4	1	5	205	
5	2	4	390	
6	3	4	400	
7	4	5	500	
8	5	3	478	
9	6	5	578	
10	7	4	750	
11	8	5	700	
12	9	4	550	
13	10	5	560	
14				
15	Min among sales for year > 4 and incentive percentage "=4" =			550
16				

Figure 2.37 Screenshot after clicking OK button in the dropdown menu of Figure 2.36

The same can be achieved by typing the following formula in cell D15.

$$= \text{MINIFS}(\text{C4 : C13, A4 : A13,"} > 4\text{", B4 : B13, "} = 4\text{"})$$

2.2.21 ROUND Function

In many real-life applications, the values are converted to the respective nearest number with a desired number of decimal digits. If a mark, say, 25 out of 40, is converted to 15, then the corresponding Excel formula is as shown.

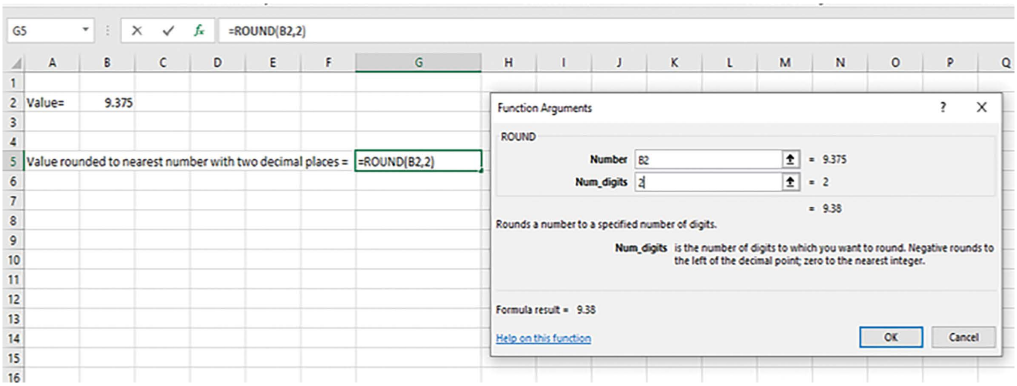
= (25/40)*15

The value of this is 9.375. If this is to be rounded to the nearest decimal number with two decimal digits, the required Excel sheet is shown in Figure 2.38 after clicking the Formulas ⟹ Math & Trig buttons. In Figure 2.38, the data against Number is the value in cell B2, and the data for Num_digits is 2, which is the number of decimal digits to be retained. Clicking OK in the dropdown menu of Figure 2.38 gives the rounded value in cell G5, as shown in Figure 2.39. If the number of decimal digits is assumed to be 0, then the given decimal number will be rounded to its nearest integer.

Case Study: For an elective subject, a faculty member administered two tests and gave the students an assignment. The maximum scores for Test 1, Test 2, and Assignment are, respectively, 40, 50, and 120. The total score for the internal exam is 40, which is broken down into three parts: 15 points for Test 1, 15 points for Test 2, and 10 points for the assignment (see Figure 2.40).

The actual marks earned on Test 1, Test 2, and Assignment are entered in Figure 2.40 and translated to a maximum of 15 marks, 15 marks, and 10 marks, respectively. Figure 2.41 shows the formulas for these conversions. The following ROUND formula is then used to convert the three components' converted marks back to their respective nearest integers. For Test 1, Test 2, and Assignment, as indicated in columns I, J, and K, respectively, use this formula on the first student before copying it to all subsequent students.

$$= \text{ROUND (Number, Num\_digits)}$$



G5 =ROUND(B2,2)

A2: Value= B2: 9.375

A5: Value rounded to nearest number with two decimal places = G5: =ROUND(B2,2)

Function Arguments

ROUND

Number B2 = 9.375

Num_digits 2 = 2

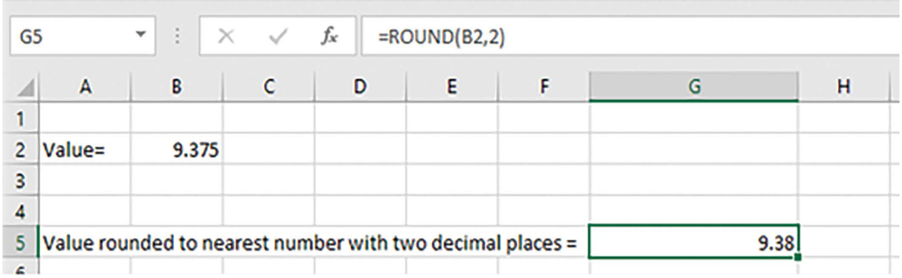
= 9.38

Rounds a number to a specified number of digits.

Num_digits is the number of digits to which you want to round. Negative rounds to the left of the decimal point; zero to the nearest integer.

Formula result = 9.38

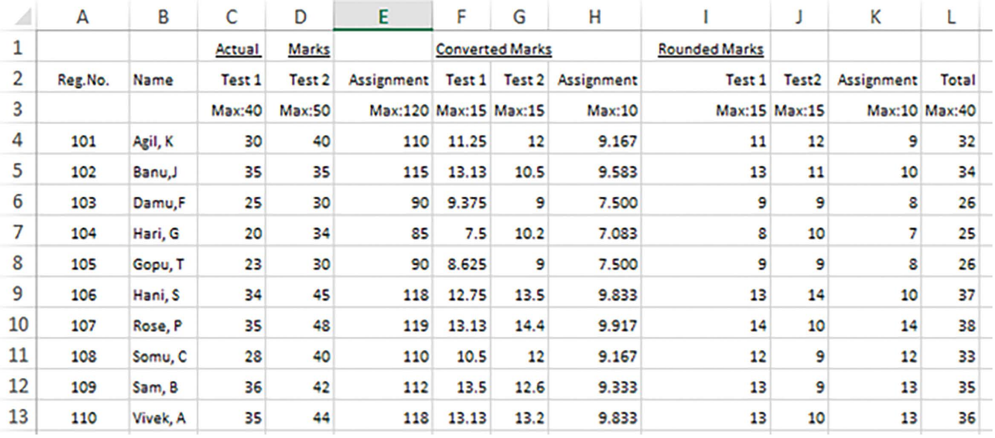
Figure 2.38 Screenshot for the sequence of clicks of buttons, viz. Formula ⟹ Math & Trig ⟹ ROUND and Filling Data



G5 =ROUND(B2,2)

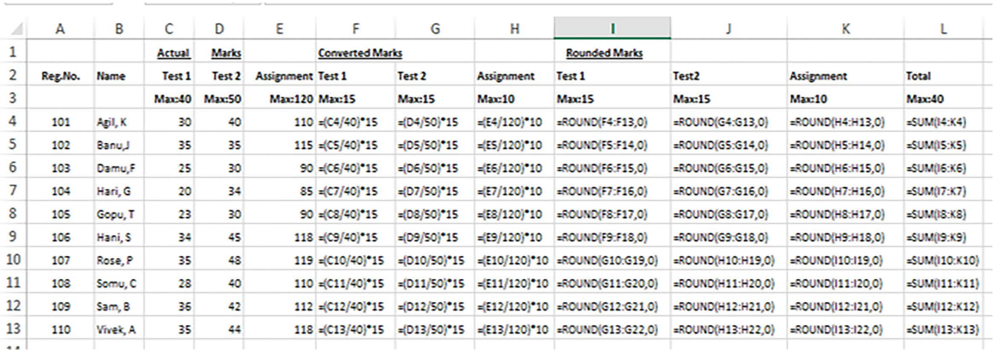
	A	B	C	D	E	F	G	H
1								
2	Value=	9.375						
3								
4								
5	Value rounded to nearest number with two decimal places =						9.38	

Figure 2.39 Screenshot after clicking OK in the dropdown menu of Figure 2.38



	A	B	C	D	E	F	G	H	I	J	K	L
1			Actual	Marks		Converted Marks			Rounded Marks			
2	Reg.No.	Name	Test 1	Test 2	Assignment	Test 1	Test 2	Assignment	Test 1	Test2	Assignment	Total
3			Max:40	Max:50	Max:120	Max:15	Max:15	Max:10	Max:15	Max:15	Max:10	Max:40
4	101	Agil, K	30	40	110	11.25	12	9.167	11	12	9	32
5	102	Banu,J	35	35	115	13.13	10.5	9.583	13	11	10	34
6	103	Damu,F	25	30	90	9.375	9	7.500	9	9	8	26
7	104	Hari, G	20	34	85	7.5	10.2	7.083	8	10	7	25
8	105	Gopu, T	23	30	90	8.625	9	7.500	9	9	8	26
9	106	Hani, S	34	45	118	12.75	13.5	9.833	13	14	10	37
10	107	Rose, P	35	48	119	13.13	14.4	9.917	14	10	14	38
11	108	Somu, C	28	40	110	10.5	12	9.167	12	9	12	33
12	109	Sam, B	36	42	112	13.5	12.6	9.333	13	9	13	35
13	110	Vivek, A	35	44	118	13.13	13.2	9.833	13	10	13	36

Figure 2.40 Workings of case study problem



	A	B	C	D	E	F	G	H	I	J	K	L
1			Actual	Marks		Converted Marks			Rounded Marks			
2	Reg.No.	Name	Test 1	Test 2	Assignment	Test 1	Test 2	Assignment	Test 1	Test2	Assignment	Total
3			Max:40	Max:50	Max:120	Max:15	Max:15	Max:10	Max:15	Max:15	Max:10	Max:40
4	101	Agil, K	30	40	110	=(C4/40)*15	=(D4/50)*15	=(E4/120)*10	=ROUND(F4:F13,0)	=ROUND(G4:G13,0)	=ROUND(H4:H13,0)	=SUM(I4:K4)
5	102	Banu,J	35	35	115	=(C5/40)*15	=(D5/50)*15	=(E5/120)*10	=ROUND(F5:F14,0)	=ROUND(G5:G14,0)	=ROUND(H5:H14,0)	=SUM(I5:K5)
6	103	Damu,F	25	30	90	=(C6/40)*15	=(D6/50)*15	=(E6/120)*10	=ROUND(F6:F15,0)	=ROUND(G6:G15,0)	=ROUND(H6:H15,0)	=SUM(I6:K6)
7	104	Hari, G	20	34	85	=(C7/40)*15	=(D7/50)*15	=(E7/120)*10	=ROUND(F7:F16,0)	=ROUND(G7:G16,0)	=ROUND(H7:H16,0)	=SUM(I7:K7)
8	105	Gopu, T	23	30	90	=(C8/40)*15	=(D8/50)*15	=(E8/120)*10	=ROUND(F8:F17,0)	=ROUND(G8:G17,0)	=ROUND(H8:H17,0)	=SUM(I8:K8)
9	106	Hani, S	34	45	118	=(C9/40)*15	=(D9/50)*15	=(E9/120)*10	=ROUND(F9:F18,0)	=ROUND(G9:G18,0)	=ROUND(H9:H18,0)	=SUM(I9:K9)
10	107	Rose, P	35	48	119	=(C10/40)*15	=(D10/50)*15	=(E10/120)*10	=ROUND(G10:G19,0)	=ROUND(H10:H19,0)	=ROUND(I10:I19,0)	=SUM(I10:K10)
11	108	Somu, C	28	40	110	=(C11/40)*15	=(D11/50)*15	=(E11/120)*10	=ROUND(G11:G20,0)	=ROUND(H11:H20,0)	=ROUND(I11:I20,0)	=SUM(I11:K11)
12	109	Sam, B	36	42	112	=(C12/40)*15	=(D12/50)*15	=(E12/120)*10	=ROUND(G12:G21,0)	=ROUND(H12:H21,0)	=ROUND(I12:I21,0)	=SUM(I12:K12)
13	110	Vivek, A	35	44	118	=(C13/40)*15	=(D13/50)*15	=(E13/120)*10	=ROUND(G13:G22,0)	=ROUND(H13:H22,0)	=ROUND(I13:I22,0)	=SUM(I13:K13)

Figure 2.41 Screenshot for formulas of the workings of the case problem

Then, copy the formula to all other cells in column L (cells L5 to L13) for the remaining students. Next, use the @SUM function to add the marks from Test 1, Test 2, and Assignment for the first candidate in cell L4 to get the total. Figure 2.41 displays the formulas for the working in Figure 2.40.

Note: The option available in the ribbon, as shown in Figure 2.42, can also be used to round a number to the nearest number. To the left of the icon are the left arrow with 0.00

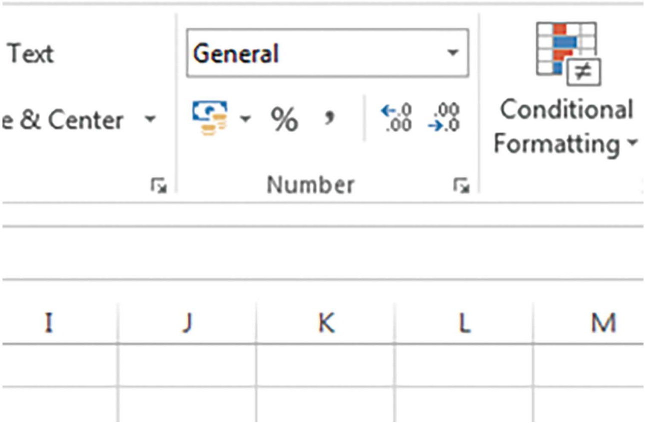


Figure 2.42 Screenshot of left arrow with 0.00 and the right arrow with 0.00 to the left of the icon conditional formatting

and the right arrow with 0.00. You can use conditional formatting to change the amount of decimal digits in a number by adding or subtracting more. Keep the cursor in any cell in that column and keep clicking the right arrow with 0.00 until the required number of decimal digits for the values in that range of cells is reached, for example, if the number of decimal digits for the values in a particular range of cells is to be reduced to 2.

If a value is to be made an integer, then the required number of decimal places is equal to 0. Such a modification is only for viewing the values. The actual value with decimal places in each of the cells in the range will be present in the memory. If two or more values in the range are added, then the original values with decimal places will be added, which will mislead users. *Therefore, adding numbers under this option should be avoided.*

2.2.22 INT Function

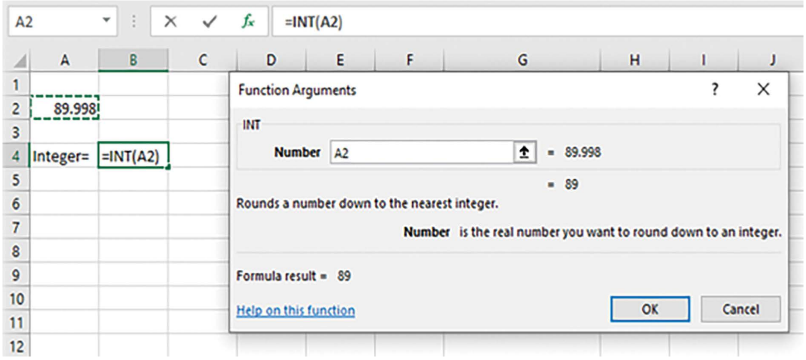
In many realities, the integer portion of a given number is obtained. The Excel formula for finding the integer portion of a given number is as given. Whatever the number of decimal places and decimal values, this function removes all the decimal digits of the given number.

= INT(Number)

The Number in the argument of INT function is a number with decimal places. If the number is 34.845, then the formula to find the integer portion of this number is given as follows, which gives the result as 34.

= INT(34.345)

The menu-driven approach to find the integer portion of a given number uses the sequence Home ⟹ Formulas ⟹ Math & Trig ⟹ INT, which gives the screenshot shown in Figure 2.43. After entering cell address A2, which contains the decimal number



=INT(A2)

	A	B
2	89.998	
4	Integer=	=INT(A2)

Function Arguments

INT

Number A2 = 89.998

= 89

Rounds a number down to the nearest integer.

Number is the real number you want to round down to an integer.

Formula result = 89

Figure 2.43 Screenshot of the sequence of buttons, viz. Home $\Longrightarrow$ Formulas $\Longrightarrow$ Math & Trig $\Longrightarrow$ INT with data filling in the dropdown menu

89.998, in the box against Number in the dropdown menu of Figure 2.43 and clicking the OK button in it, it will give the interger value of the decimal number 89.998 in cell B4 as 89.

The INT function can be used as a ROUND function by adding 0.5 to the decimal number. The corresponding formula to round a decimal number to the nearest number is as given. This addition of 0.5 to the number in the menu-driven approach of the INT function can be implemented by adding 0.5 to the cell address which contains the decimal number in the box against Number in Figure 2.43. The corresponding formula is as given.

$$= \mathrm{INT}(\mathrm{Number} + 0.5)$$

Specifically, it is as shown, which will give 90.

$$= \mathrm{INT}(\mathrm{A2} + 0.5)$$

2.2.23 ROUNDDOWN Function

The ROUNDDOWN function in Excel reduces the number of decimal places of a given decimal number to a desired number of decimal places without rounding to the previous nearest decimal number. Consider a number, say, 24.239. If the ROUNDDOWN function is used for this number with a provision to have only two decimal places, the result will be 24.23. This function is available under HOME $\Longrightarrow$ Formulas $\Longrightarrow$ Math & Trig.

The syntax of this function is as follows.

$$= \mathrm{ROUNDDOWN}\ (\mathrm{Number},\ \mathrm{Num\_digits})$$

The first argument in the formula is the number for which the number of decimal places is to be reduced. The second argument in the formula is the desired number of decimal places.

For the example given, the formula is ROUND DOWN (24.239, 2)

The result is 24.23.

This operation can be done using a menu-driven option, as explained for the ROUND function by clicking ROUNDDOWN after clicking Formula and Math & Trig, which will give a screenshot similar to the one shown in Figure 2.38.

The ROUNDDOWN function with zero decimal places is the same as the INT function. Both functions will truncate the given decimal number to obtain its integer value.

2.2.24 ROUNDUP Function

The ROUNDUP function in Excel reduces the number of decimal places of a given decimal number to a desired number of decimal places with rounding to the next highest decimal number. Consider a number, say, 24.231. If ROUNDUP function is used on this number with two decimal places, the result will be 24.24. This function is available under HOME ⟹ Formula ⟹ Math & Trig.

The syntax of this function is as follows.

$$= \text{ROUNDUP}(\text{Number, Num\_digits})$$

The first argument in the formula is the number for which the number of decimal places is to be increased. The second argument in the formula is the desired number of decimal places.

For the example given, the formula is: $= ROUNDUP(24.231, 2)$

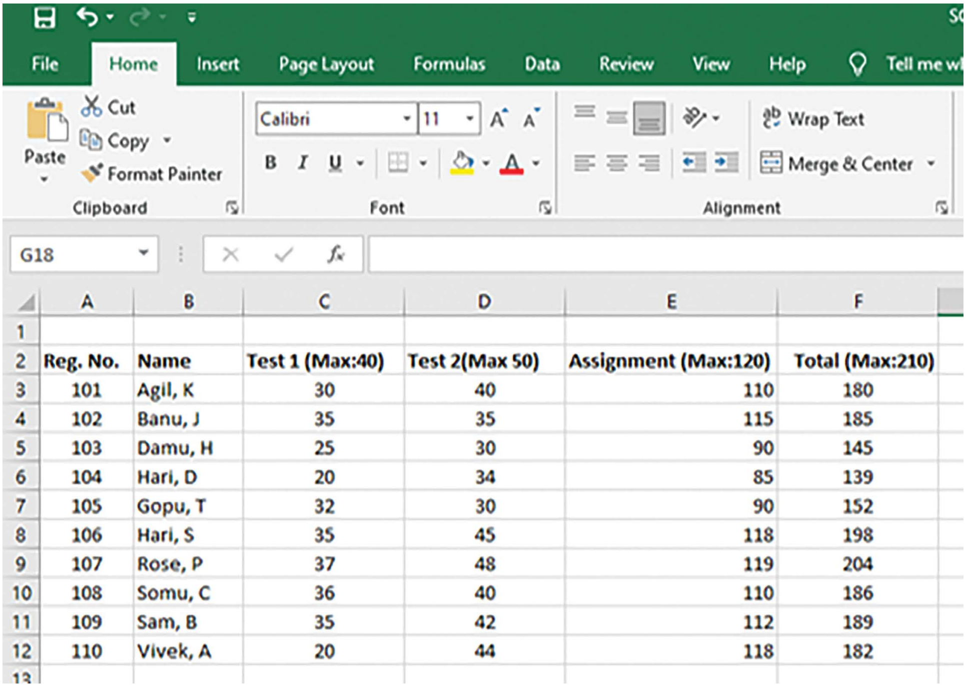
The result is 24.24.

This operation can be done using a menu-driven option, as explained for the ROUND function, by clicking ROUNDUP after clicking Formulas and Math & Trig, which will give a screenshot similar to the one shown in Figure 2.38.

2.2.25 SORT Function

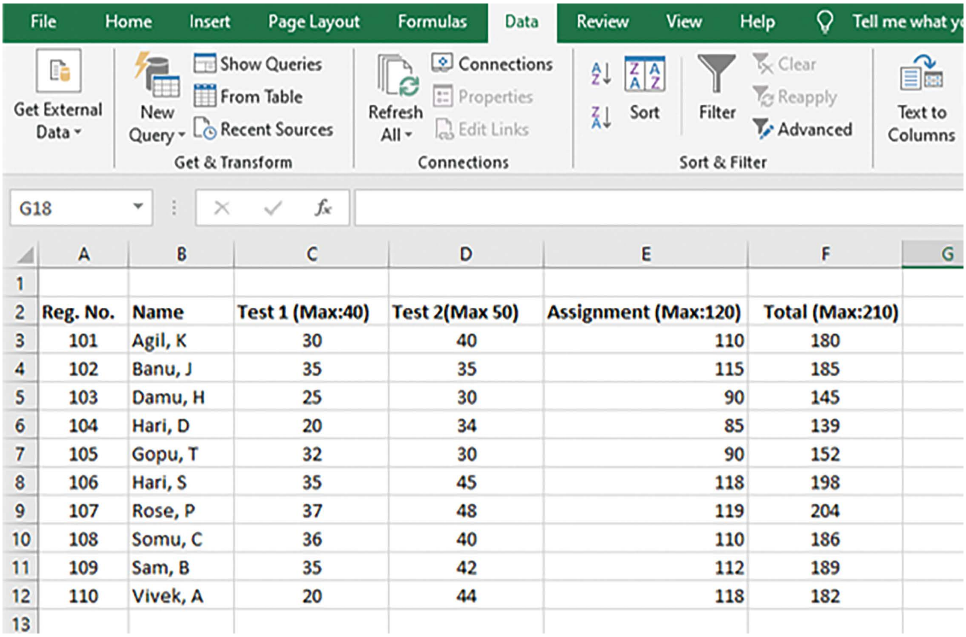
The SORT function makes it possible to reorder the contents of the cells in a given range in either the ascending or descending order of a specific column. Think about the information displayed in the Figure 2.44 screenshot. Follow these instructions to reorder the data in the range A3:F12 of Figure 2.44 in ascending order of the total mark listed in Column F.

1. Select the data range A2:F12, which includes the heading of each column in the range in Figure 2.44.
2. Click the DATA button shown in Figure 2.44, which will give a screenshot as shown in Figure 2.45.
3. Click the SORT button in the ribbon shown in Figure 2.45, which gives a screenshot as in Figure 2.46.
4. Open the options for Sort by in Figure 2.46, which gives a screenshot as shown in Figure 2.47.
5. In the dropdown menu of Figure 2.47, click the option Total (Max:210), click Smallest to Largest from the righthand side submenu, and click OK on the screen to obtain the sorted table per the ascending order of the total mark shown in Column F, which gives the final table as shown in Figure 2.48.



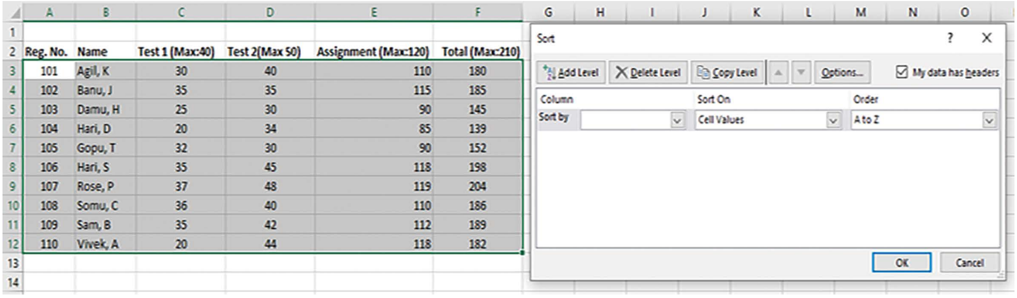
	A	B	C	D	E	F
1						
2	Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)
3	101	Agil, K	30	40	110	180
4	102	Banu, J	35	35	115	185
5	103	Damu, H	25	30	90	145
6	104	Hari, D	20	34	85	139
7	105	Gopu, T	32	30	90	152
8	106	Hari, S	35	45	118	198
9	107	Rose, P	37	48	119	204
10	108	Somu, C	36	40	110	186
11	109	Sam, B	35	42	112	189
12	110	Vivek, A	20	44	118	182
13						

Figure 2.44 Screenshot of data



	A	B	C	D	E	F	G
1							
2	Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)	
3	101	Agil, K	30	40	110	180	
4	102	Banu, J	35	35	115	185	
5	103	Damu, H	25	30	90	145	
6	104	Hari, D	20	34	85	139	
7	105	Gopu, T	32	30	90	152	
8	106	Hari, S	35	45	118	198	
9	107	Rose, P	37	48	119	204	
10	108	Somu, C	36	40	110	186	
11	109	Sam, B	35	42	112	189	
12	110	Vivek, A	20	44	118	182	
13							

Figure 2.45 Screenshot after clicking DATA button in the ribbon of Figure 2.44



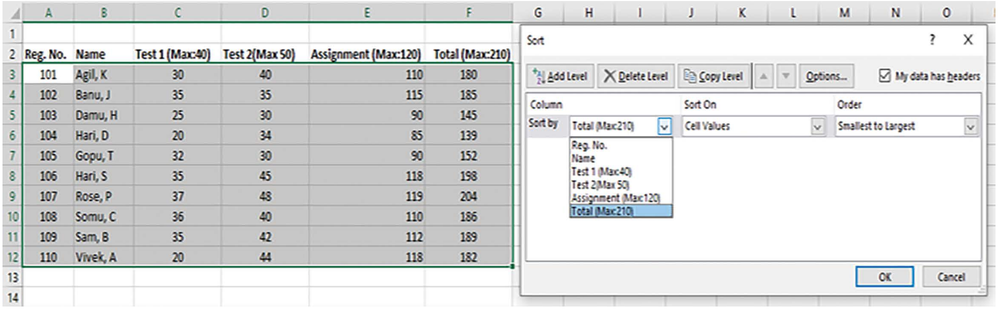
Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)
101	Agil, K	30	40	110	180
102	Banu, J	35	35	115	185
103	Damu, H	25	30	90	145
104	Hari, D	20	34	85	139
105	Gopu, T	32	30	90	152
106	Hari, S	35	45	118	198
107	Rose, P	37	48	119	204
108	Somu, C	36	40	110	186
109	Sam, B	35	42	112	189
110	Vivek, A	20	44	118	182

Sort

☑ My data has headers

	Column	Sort On	Order
Sort by		Cell Values	A to Z

Figure 2.46 Screenshot after SORT button in Figure 2.45



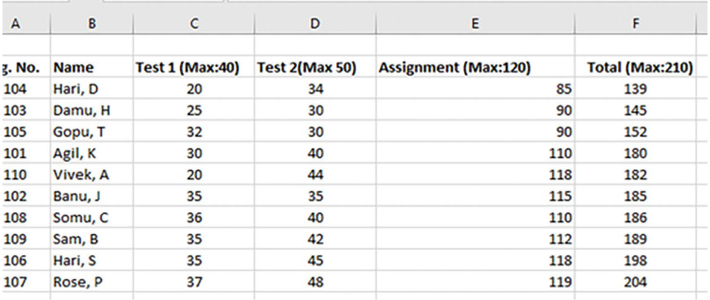
Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)
101	Agil, K	30	40	110	180
102	Banu, J	35	35	115	185
103	Damu, H	25	30	90	145
104	Hari, D	20	34	85	139
105	Gopu, T	32	30	90	152
106	Hari, S	35	45	118	198
107	Rose, P	37	48	119	204
108	Somu, C	36	40	110	186
109	Sam, B	35	42	112	189
110	Vivek, A	20	44	118	182

Sort

☑ My data has headers

	Column	Sort On	Order
Sort by	Total (Max:210)	Cell Values	Smallest to Largest

Figure 2.47 Screenshot after opening the option for sort by in Figure 2.46



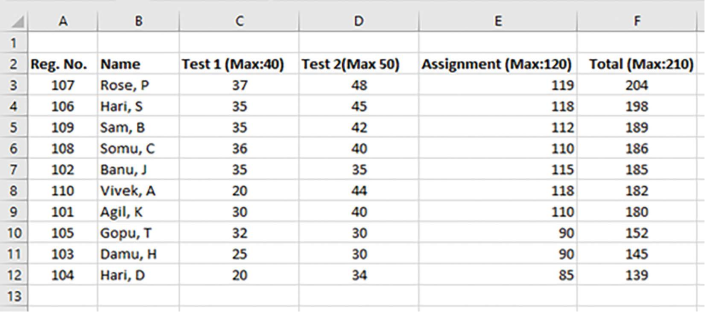
g. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)
104	Hari, D	20	34	85	139
103	Damu, H	25	30	90	145
105	Gopu, T	32	30	90	152
101	Agil, K	30	40	110	180
110	Vivek, A	20	44	118	182
102	Banu, J	35	35	115	185
108	Somu, C	36	40	110	186
109	Sam, B	35	42	112	189
106	Hari, S	35	45	118	198
107	Rose, P	37	48	119	204

Figure 2.48 Screenshot after clicking the option Total (Max:210) in the dropdown menu of Figure 2.47 and clicking OK button in it

If the data in the range A3:F12 are to be sorted in descending order of the total marks, the same sequence of steps is to be followed, but click Largest to Smallest in the dropdown menu of Figure 2.47, which will give the final table as in Figure 2.49.

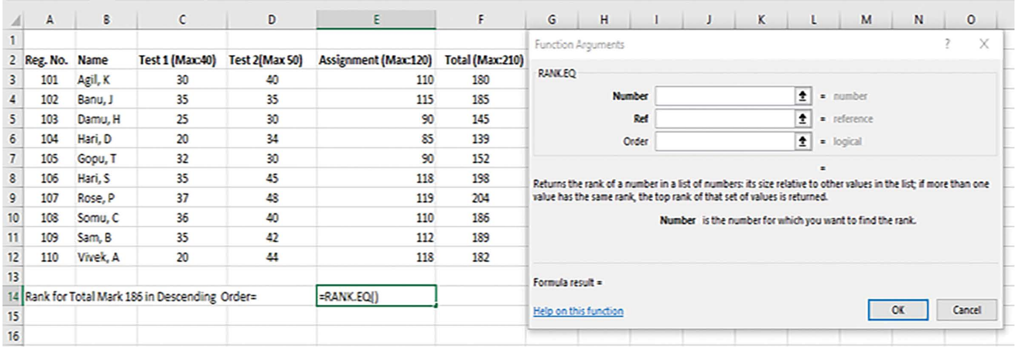
2.2.26 RANK.EQ Function

The RANK.EQ function obtains the rank of a number among a given set of numbers either in descending or ascending order.



	A	B	C	D	E	F
1						
2	Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)
3	107	Rose, P	37	48	119	204
4	106	Hari, S	35	45	118	198
5	109	Sam, B	35	42	112	189
6	108	Somu, C	36	40	110	186
7	102	Banu, J	35	35	115	185
8	110	Vivek, A	20	44	118	182
9	101	Agil, K	30	40	110	180
10	105	Gopu, T	32	30	90	152
11	103	Damu, H	25	30	90	145
12	104	Hari, D	20	34	85	139
13						

Figure 2.49 Screenshot of sorted table as per descending order of total mark



	A	B	C	D	E	F
1						
2	Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)
3	101	Agil, K	30	40	110	180
4	102	Banu, J	35	35	115	185
5	103	Damu, H	25	30	90	145
6	104	Hari, D	20	34	85	139
7	105	Gopu, T	32	30	90	152
8	106	Hari, S	35	45	118	198
9	107	Rose, P	37	48	119	204
10	108	Somu, C	36	40	110	186
11	109	Sam, B	35	42	112	189
12	110	Vivek, A	20	44	118	182
13						
14	Rank for Total Mark 186 in Descending Order=				=RANK.EQ()	
15						
16						

Function Arguments

RANK.EQ

Number = number

Ref = reference

Order = logical

=

Returns the rank of a number in a list of numbers: its size relative to other values in the list; if more than one value has the same rank, the top rank of that set of values is returned.

Number is the number for which you want to find the rank.

Formula result =

Figure 2.50 Screenshot for sequence of buttons, viz. Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ RANK.EQ Descending

The sequence of button clicks to obtain the dropdown menu of RANK.EQ is as follows.

HOME ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ RANK.EQ

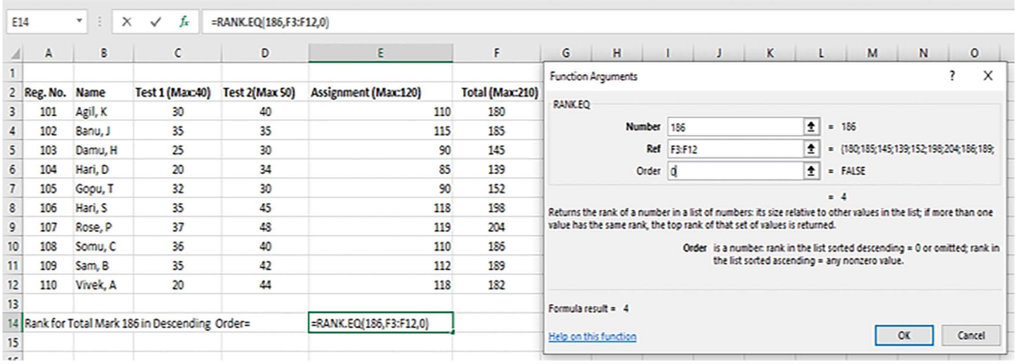
A screenshot for the sequence of button clicks with the data is shown in Figure 2.50. The explanations of the boxes in which data are to be fed in Figure 2.50 are as follows.

Number: This is the number for which the rank is required.

Ref: This is the range of cells containing data from which the rank of the given number is to be found.

Order: This is the order, which may be ascending or descending order of the ranks. In this box, if 0 is entered, the rank of the given number will be in descending order of the numbers in the given range of cells. In this box, if any non-zero number is entered, the rank of the given number will be in ascending order of the numbers in the given range of cells.

If the rank of the number 186 from the values available in the range F3:F12 per descending order of the ranks is required, the boxes in the dropdown menu for Number, Ref, and Order are filled with 186, F3:F12, and 0, respectively, as shown in Figure 2.51.



E14 =RANK.EQ(186,F3:F12,0)

	A	B	C	D	E	F
1						
2	Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)
3	101	Agil, K	30	40	110	180
4	102	Banu, J	35	35	115	185
5	103	Damu, H	25	30	90	145
6	104	Hari, D	20	34	85	139
7	105	Gopu, T	32	30	90	152
8	106	Hari, S	35	45	118	198
9	107	Rose, P	37	48	119	204
10	108	Somu, C	36	40	110	186
11	109	Sam, B	35	42	112	189
12	110	Vivek, A	20	44	118	182
13						
14	Rank for Total Mark 186 in Descending Order=				=RANK.EQ(186,F3:F12,0)	
15						

Function Arguments

RANK.EQ

Number 186 = 186

Ref F3:F12 = {180;185;145;139;152;198;204;186;189;

Order 0 = FALSE

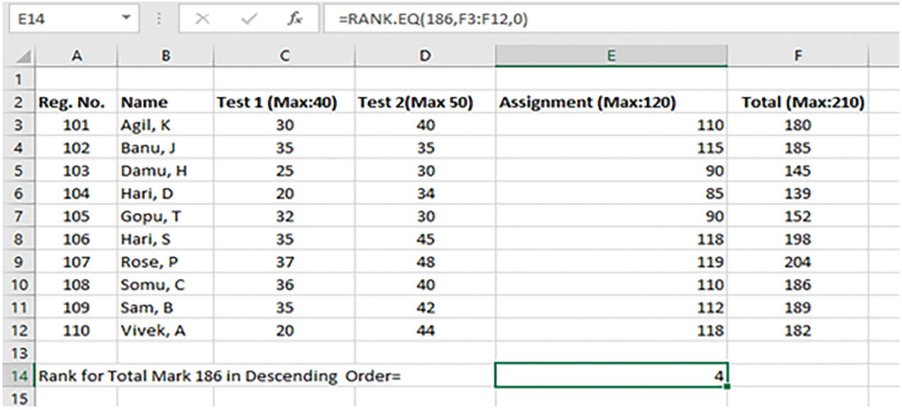
= 4

Returns the rank of a number in a list of numbers: its size relative to other values in the list; if more than one value has the same rank, the top rank of that set of values is returned.

Order is a number: rank in the list sorted descending = 0 or omitted; rank in the list sorted ascending = any nonzero value.

Formula result = 4

Figure 2.51 Screenshot after filling the data in the dropdown menu of Figure 2.50



E14 =RANK.EQ(186,F3:F12,0)

	A	B	C	D	E	F
1						
2	Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)
3	101	Agil, K	30	40	110	180
4	102	Banu, J	35	35	115	185
5	103	Damu, H	25	30	90	145
6	104	Hari, D	20	34	85	139
7	105	Gopu, T	32	30	90	152
8	106	Hari, S	35	45	118	198
9	107	Rose, P	37	48	119	204
10	108	Somu, C	36	40	110	186
11	109	Sam, B	35	42	112	189
12	110	Vivek, A	20	44	118	182
13						
14	Rank for Total Mark 186 in Descending Order=				4	
15						

Figure 2.52 Screenshot after clicking OK button in the dropdown menu of Figure 2.51

Clicking the OK button in the dropdown menu of Figure 2.51 gives the rank of the given number 186 in descending order of rank as 4 in cell E14, as shown in Figure 2.52.

The Excel formula to obtain this result is:

$$= RANK.EQ(Number, Ref, Order)$$

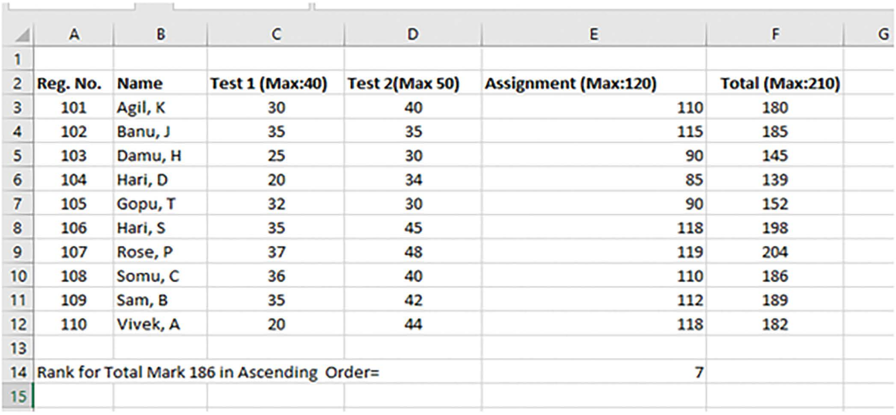
For the given problem, the formula is: $= \mathrm{RANK.EQ}(186, \mathrm{F3} : \mathrm{F12}, 0)$.

For the same problem, the rank of the number 186 from among the numbers in the range of cells F3:F12 per ascending order is given by the following formula, which gives the result in cell E14, as shown in Figure 2.53. The rank is 7.

$= \mathrm{RANK.EQ}(186, \mathrm{F3} : \mathrm{F12}, 1)$

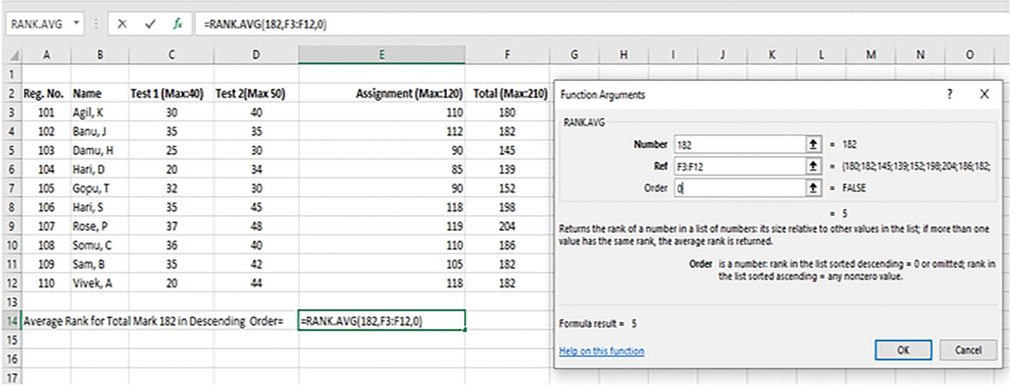
2.2.27 RANK.AVG Function

The RANK.AVG function determines a number's rank among a set of numbers, either in ascending or decreasing order. If multiple values share the same rank, the average of those ranks is calculated and used to represent the rank of those values.



	A	B	C	D	E	F	G
1							
2	Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)	
3	101	Agil, K	30	40	110	180	
4	102	Banu, J	35	35	115	185	
5	103	Damu, H	25	30	90	145	
6	104	Hari, D	20	34	85	139	
7	105	Gopu, T	32	30	90	152	
8	106	Hari, S	35	45	118	198	
9	107	Rose, P	37	48	119	204	
10	108	Somu, C	36	40	110	186	
11	109	Sam, B	35	42	112	189	
12	110	Vivek, A	20	44	118	182	
13							
14	Rank for Total Mark 186 in Ascending Order=				7		
15							

Figure 2.53 Screenshot after changing the data for order to any value other than 0 say 1 in Figure 2.51 to find the rank of 186 as per ascending order of numbers in the range F3:F12



RANK.AVG =RANK.AVG(182,F3:F12,0)

	A	B	C	D	E	F
1						
2	Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)
3	101	Agil, K	30	40	110	180
4	102	Banu, J	35	35	112	182
5	103	Damu, H	25	30	90	145
6	104	Hari, D	20	34	85	139
7	105	Gopu, T	32	30	90	152
8	106	Hari, S	35	45	118	198
9	107	Rose, P	37	48	119	204
10	108	Somu, C	36	40	110	186
11	109	Sam, B	35	42	105	182
12	110	Vivek, A	20	44	118	182
13						
14	Average Rank for Total Mark 182 in Descending Order=				=RANK.AVG(182,F3:F12,0)	
15						
16						
17						

Function Arguments

RANK.AVG

Number 182 = 182

Ref F3:F12 = {180;182;145;139;152;198;204;186;182;

Order 0 = FALSE

= 5

Returns the rank of a number in a list of numbers: its size relative to other values in the list; if more than one value has the same rank, the average rank is returned.

Order is a number: rank in the list sorted descending = 0 or omitted; rank in the list sorted ascending = any nonzero value.

Formula result = 5

Figure 2.54 Screenshot of the clicks of buttons, viz. Home => Formulas => More Functions => Statistical => RANK.AVG with data filled in dropdown menu

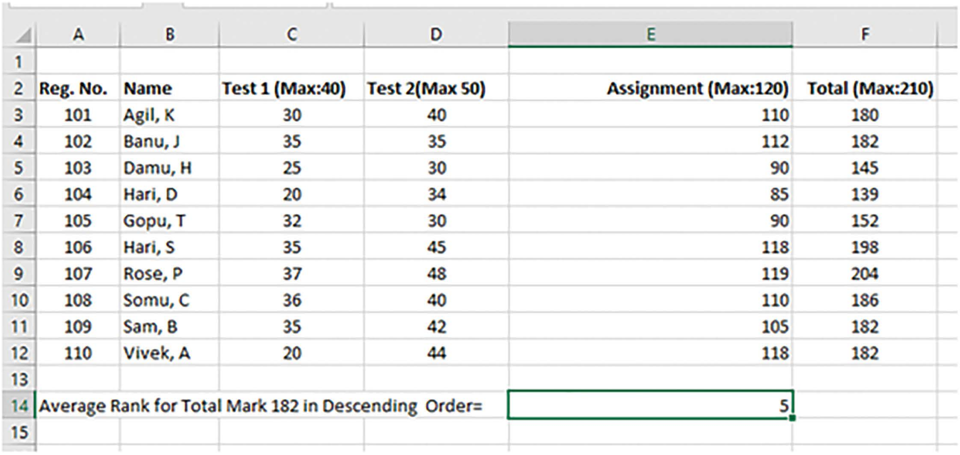
The sequence of button clicks to obtain the dropdown menu of RANK.AVG is as follows.

Home ⟹ Formula ⟹ More Functions ⟹ Statistical ⟹ RANK.AVG

A screenshot filled with data for this sequence of button clicks with the data to obtain the rank of 182 in descending order is shown in Figure 2.54. Clicking the OK button in the dropdown menu of Figure 2.54 gives the rank of the number 182 in descending order of the numbers in the range F3:F12 in cell E14, as shown in Figure 2.55.

The formula to obtain the rank of 182 in descending order is shown in the following, which gives the rank as 5.

$$= \text{RANK.AVG}(182, \text{F3}:\text{F12}, 0)$$



	A	B	C	D	E	F
1						
2	Reg. No.	Name	Test 1 (Max:40)	Test 2(Max 50)	Assignment (Max:120)	Total (Max:210)
3	101	Agil, K	30	40	110	180
4	102	Banu, J	35	35	112	182
5	103	Damu, H	25	30	90	145
6	104	Hari, D	20	34	85	139
7	105	Gopu, T	32	30	90	152
8	106	Hari, S	35	45	118	198
9	107	Rose, P	37	48	119	204
10	108	Somu, C	36	40	110	186
11	109	Sam, B	35	42	105	182
12	110	Vivek, A	20	44	118	182
13						
14	Average Rank for Total Mark 182 in Descending Order=				5	
15						

Figure 2.55 Screenshot after clicking OK button in the drop-down menu of Figure 2.54

The formula to obtain the rank of 182 in ascending order is shown in the following, which gives the rank as 6. This can be verified.

$$= \text{RANK.AVG}(182, \text{F3} : \text{F12}, 1)$$

2.2.28 Hide and Unhide Commands

The analyst can hide certain columns or rows in an Excel sheet to see the remaining columns or rows. The Hide/Unhide command is used to achieve this. Later, with the Unhide command, some of the hidden columns and rows may be revealed.

HIDE: The Hide command is illustrated using the data shown in Figure 2.56. This figure contains data for Test 1(Max: 40), Test 2 (Max:50), and Assignment (Max:120), which are shown in columns C, D, and E, respectively. The maximum of the total of these marks is 210, which is shown in Column F. Now, the weights of the Test 1, Test 2, and Assignment are fixed as 15, 15, and 10, respectively. So, the maximum of the revised total is 40.

The formula for Test 1 in Column G = INT[(Test 1 in Column C/40)*15 + 0.5].

The formula for Test 2 in Column H = INT[(Test 2 in Column D/50)*15 + 0.5].

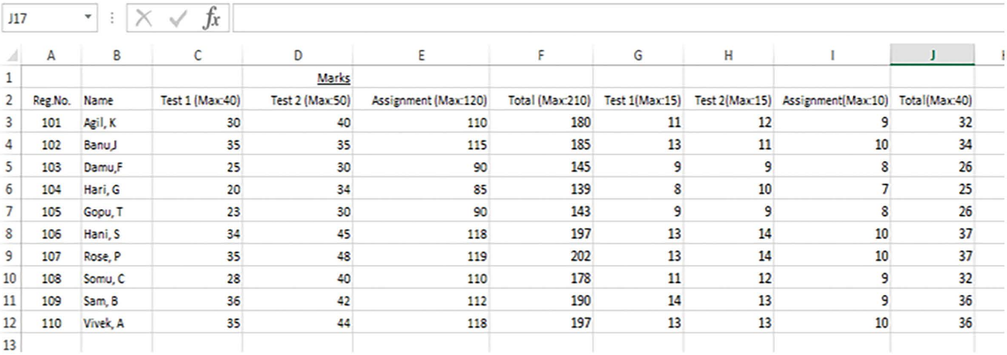
The formula for Assignment in Column I = INT[(Assignment in Column E/120)*10 + 0.5].

$$\text{The formula for Total in Column J} = \text{Test 1 in Column G} + \text{Test 2 in Column H} \\ + \text{Assignment in Column I.}$$

Now, to present the final view with Columns A, B, G, H, I, and J, columns C, D, E, and F must be hidden, which can be achieved using Hide Columns command.

Steps to Hide Columns

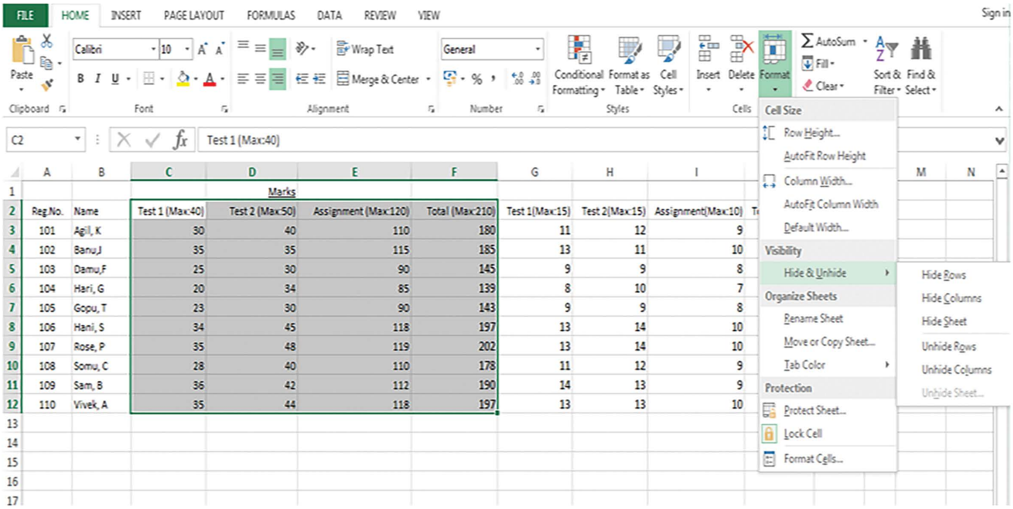
1. Select columns C, D, E, and F.
2. Click the Format button in the ribbon and then select the Hide & Unhide button under Visibility, whose screenshot is shown in Figure 2.57.



J17

	A	B	C	D	E	F	G	H	I	J
1				Marks						
2	Reg.No.	Name	Test 1 (Max:40)	Test 2 (Max:50)	Assignment (Max:120)	Total (Max:210)	Test 1(Max:15)	Test 2(Max:15)	Assignment(Max:10)	Total(Max:40)
3	101	Agil, K	30	40	110	180	11	12	9	32
4	102	Banu,J	35	35	115	185	13	11	10	34
5	103	Damu,F	25	30	90	145	9	9	8	26
6	104	Hari, G	20	34	85	139	8	10	7	25
7	105	Gopu, T	23	30	90	143	9	9	8	26
8	106	Hani, S	34	45	118	197	13	14	10	37
9	107	Rose, P	35	48	119	202	13	14	10	37
10	108	Somu, C	28	40	110	178	11	12	9	32
11	109	Sam, B	36	42	112	190	14	13	9	36
12	110	Vivek, A	35	44	118	197	13	13	10	36
13										

Figure 2.56 Screenshot of sample data to demonstrate Hide Columns command



C2 Test 1 (Max:40)

	A	B	C	D	E	F	G	H	I	T
1				Marks						
2	Reg.No.	Name	Test 1 (Max:40)	Test 2 (Max:50)	Assignment (Max:120)	Total (Max:210)	Test 1(Max:15)	Test 2(Max:15)	Assignment(Max:10)	T
3	101	Agil, K	30	40	110	180	11	12	9	
4	102	Banu,J	35	35	115	185	13	11	10	
5	103	Damu,F	25	30	90	145	9	9	8	
6	104	Hari, G	20	34	85	139	8	10	7	
7	105	Gopu, T	23	30	90	143	9	9	8	
8	106	Hani, S	34	45	118	197	13	14	10	
9	107	Rose, P	35	48	119	202	13	14	10	
10	108	Somu, C	28	40	110	178	11	12	9	
11	109	Sam, B	36	42	112	190	14	13	9	
12	110	Vivek, A	35	44	118	197	13	13	10	
13										
14										
15										
16										
17										

Figure 2.57 Screenshot after clicking Format, and Hide & Unhide under Visibility in the dropdown menu of Format button

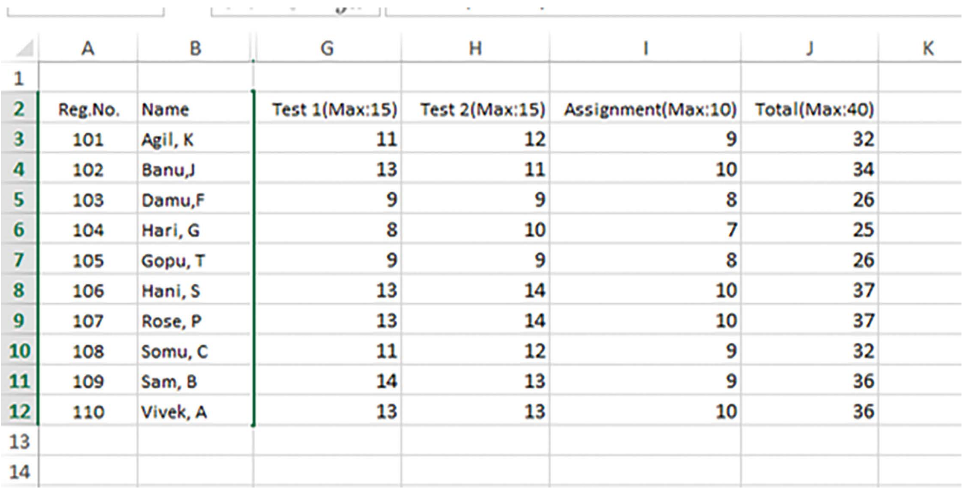
3. Then click the Hide Columns button in the rightmost dropdown menu of Figure 2.57, which gives the desired view with columns A, B, G, H, I, and J, as shown in Figure 2.58.

UNHIDE: The columns which have been hidden using the Hide Columns command can be unhidden using the Unhide Columns command.

The steps for the Unhide Columns command are:

1. Click the Format button in the ribbon and then click the Hide & Unhide button under Visibility.
2. Then click the Unhide Columns button in the rightmost dropdown menu, which will unhide columns C, D, E, and F, which have been hidden previously using the Hide Columns command. The corresponding view will be as shown in Figure 2.56.

Hide Rows and Unhide Rows can be done in similar ways.



	A	B	G	H	I	J	K
1							
2	Reg.No.	Name	Test 1(Max:15)	Test 2(Max:15)	Assignment(Max:10)	Total(Max:40)	
3	101	Agil, K	11	12	9	32	
4	102	Banu,J	13	11	10	34	
5	103	Damu,F	9	9	8	26	
6	104	Hari, G	8	10	7	25	
7	105	Gopu, T	9	9	8	26	
8	106	Hani, S	13	14	10	37	
9	107	Rose, P	13	14	10	37	
10	108	Somu, C	11	12	9	32	
11	109	Sam, B	14	13	9	36	
12	110	Vivek, A	13	13	10	36	
13							
14							

Figure 2.58 Screenshot after clicking Hide Columns in the rightmost dropdown menu of Figure 2.57

2.2.29 Filter Function

Based on one or more criteria applied to one or more columns, the filter function displays a subset of rows in a specific range of cells that contain data [4, 5]. Consider the information in Table 2.6, which details the quarterly sales of ten salespeople in lakhs of rupees. The sales manager is interested in the quarterly sales of salespeople whose mean yearly sales are greater than or equal to ₹ 25 lakhs and whose fourth-quarter sales are greater than or equal to that amount.

The mean annual sales of the salespeople can be obtained using the AVERAGE function, and it is ₹ 109 lakhs.

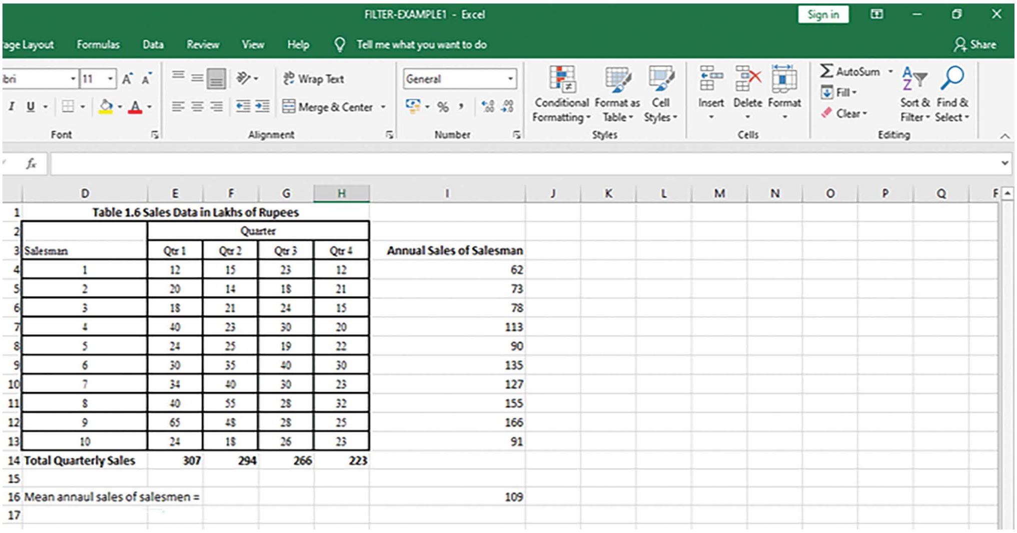
Steps of Filter Function

The steps of the Filter function are explained using the screenshot of the given problem in Figure 2.59.

1. Select the range of cells (D3:I13) containing the data to be filtered, including the column headings in the Excel sheet shown in Figure 2.59.
2. Click the Sort & Filter button, which can be seen at the top right corner in the Home menu of Figure 2.59, and then keep the cursor over Filter with Funnel Symbol in the dropdown menu, which gives a dropdown menu giving instructions, as shown in Figure 2.60.
3. Click the Filter button in the dropdown menu of Figure 2.60, which gives the output as shown in Figure 2.61, with a selection option at each of the column headings selected in Step 1.
4. Click the button at column I against the column heading, Total Sales of Salespeople in Figure 2.61, and then click Number Filters in its dropdown menu, which gives a display as in Figure 2.62.

Table 2.6 Sales Data in Lakhs of Rupees

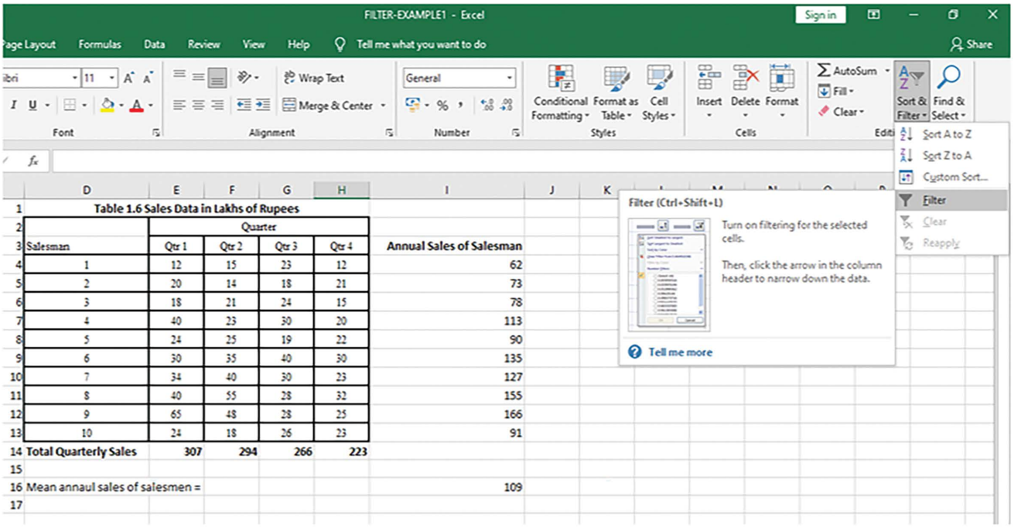
Salesperson	*Quarter*				*Annual Sales of Salespeople*
	Qtr 1	*Qtr 2*	*Qtr 3*	*Qtr 4*	
1	12	15	23	12	62
2	20	14	18	21	73
3	18	21	24	15	78
4	40	23	30	20	113
5	24	25	19	22	90
6	30	35	40	30	135
7	34	40	30	23	127
8	40	55	28	32	155
9	65	48	28	25	166
10	24	18	26	23	91
Total quarterly sales	307	294	266	223	



	D	E	F	G	H	I
1	Table 1.6 Sales Data in Lakhs of Rupees					
2		Quarter				
3	Salesman	Qtr 1	Qtr 2	Qtr 3	Qtr 4	Annual Sales of Salesman
4	1	12	15	23	12	62
5	2	20	14	18	21	73
6	3	18	21	24	15	78
7	4	40	23	30	20	113
8	5	24	25	19	22	90
9	6	30	35	40	30	135
10	7	34	40	30	23	127
11	8	40	55	28	32	155
12	9	65	48	28	25	166
13	10	24	18	26	23	91
14	Total Quarterly Sales	307	294	266	223	
15						
16	Mean annaul sales of salesmen =					109
17						

Figure 2.59 Screenshot of data

5. Click the Greater Than Or Equal To option in the second-level dropdown menu of Figure 2.62, which gives a display as in Figure 2.63.
6. Fill the mean annual sales of the salespeople (₹ 109 lakhs) as shown in the dropdown menu of Figure 2.63, and then click the OK button to show the filtered rows per the first criterion. The annual sales of the salespeople are greater than or equal to the mean annual sales of the salespeople, as shown in Figure 2.64.
7. Clicking button in column H against the column heading Qtr 4; clicking Number Filters in its dropdown menu, Greater Than Or Equal to in the dropdown menu of Number Filters; entering 25 in the box against Qtr 4 in the next dropdown menu; and finally clicking the OK button to perform second-level filtering of rows for Fourth quarter sales is greater than or equal to ₹ 25 lakhs will give the result shown in Figure 2.65.



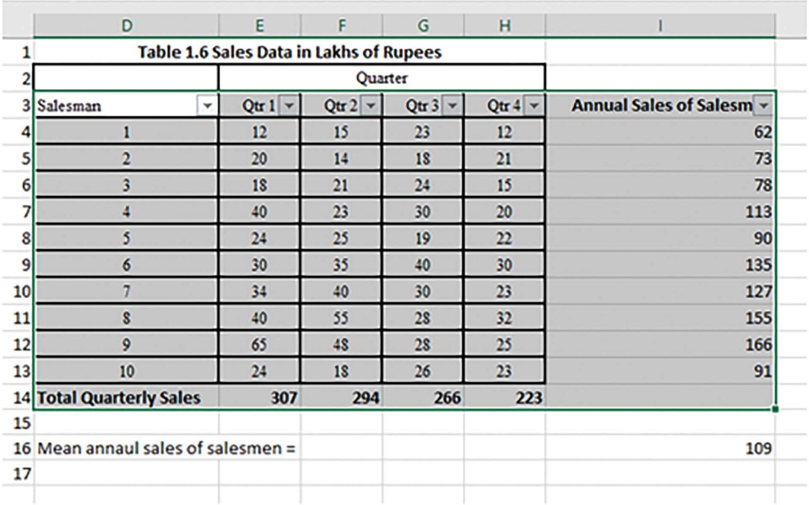
Filter (Ctrl+Shift+L)

Turn on filtering for the selected cells.

Then, click the arrow in the column header to narrow down the data.

	D	E	F	G	H	I
1	Table 1.6 Sales Data in Lakhs of Rupees					
2		Quarter				
3	Salesman	Qtr 1	Qtr 2	Qtr 3	Qtr 4	Annual Sales of Salesman
4	1	12	15	23	12	62
5	2	20	14	18	21	73
6	3	18	21	24	15	78
7	4	40	23	30	20	113
8	5	24	25	19	22	90
9	6	30	35	40	30	135
10	7	34	40	30	23	127
11	8	40	55	28	32	155
12	9	65	48	28	25	166
13	10	24	18	26	23	91
14	Total Quarterly Sales	307	294	266	223	
15						
16	Mean annaul sales of salesmen =					109
17						

Figure 2.60 Screenshot after clicking Sort & Filter button which can be seen at the top right corner in the Home menu of Figure 2.59 and then keeping the cursor over Filter with Funnel Symbol in the dropdown menu



	D	E	F	G	H	I
1	Table 1.6 Sales Data in Lakhs of Rupees					
2		Quarter				
3	Salesman	Qtr 1	Qtr 2	Qtr 3	Qtr 4	Annual Sales of Salesm
4	1	12	15	23	12	62
5	2	20	14	18	21	73
6	3	18	21	24	15	78
7	4	40	23	30	20	113
8	5	24	25	19	22	90
9	6	30	35	40	30	135
10	7	34	40	30	23	127
11	8	40	55	28	32	155
12	9	65	48	28	25	166
13	10	24	18	26	23	91
14	Total Quarterly Sales	307	294	266	223	
15						
16	Mean annaul sales of salesmen =					109
17						

Figure 2.61 Screenshot after clicking Filter button in the dropdown menu of Figure 2.60

Note: Alternatively, in Step 4, one can unclick the checkbox against each row which is not wanted in that column and then click the OK button to give the filtered rows. The same may be repeated for another column to remove unwanted rows from the rows obtained in the first filter. This can be continued for the necessary number of criteria (column criteria).

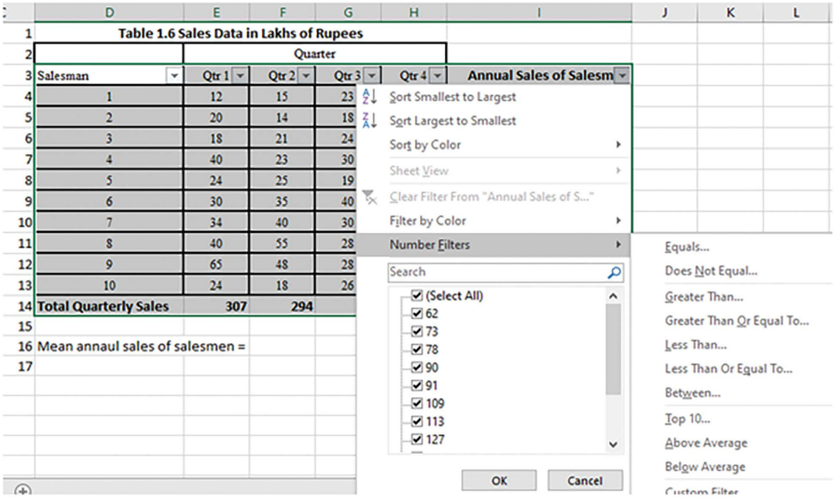


Table 1.6 Sales Data in Lakhs of Rupees

	Quarter		
Salesman	Qtr 1	Qtr 2	Qtr 3
1	12	15	23
2	20	14	18
3	18	21	24
4	40	23	30
5	24	25	19
6	30	35	40
7	34	40	30
8	40	55	28
9	65	48	28
10	24	18	26
Total Quarterly Sales	307	294	

Mean annaul sales of salesmen =

Figure 2.62 Screenshot after clicking button at the column F against column heading, "Total Sales of Salesman" and then clicking Number Filters in its dropdown menu.

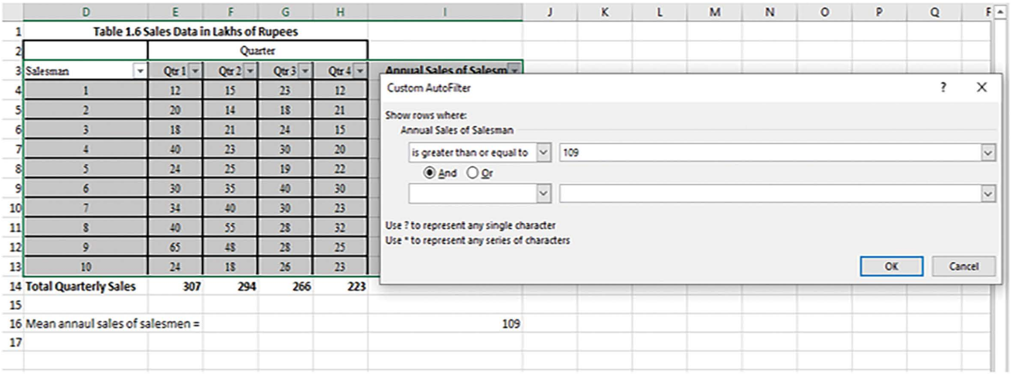


Table 1.6 Sales Data in Lakhs of Rupees

	Quarter			
Salesman	Qtr 1	Qtr 2	Qtr 3	Qtr 4
1	12	15	23	12
2	20	14	18	21
3	18	21	24	15
4	40	23	30	20
5	24	25	19	22
6	30	35	40	30
7	34	40	30	23
8	40	55	28	32
9	65	48	28	25
10	24	18	26	23
Total Quarterly Sales	307	294	266	223

Mean annaul sales of salesmen = 109

Custom AutoFilter

Show rows where:

Annual Sales of Salesman

is greater than or equal to 109

● And ○ Or

Use ? to represent any single character

Use * to represent any series of characters

Figure 2.63 Screenshot after clicking "Greater Than Or Equal To" option in the second-level drop-down menu of Figure 2.62

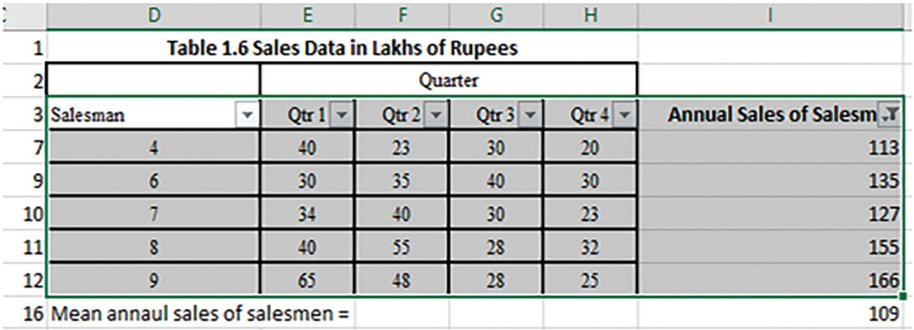
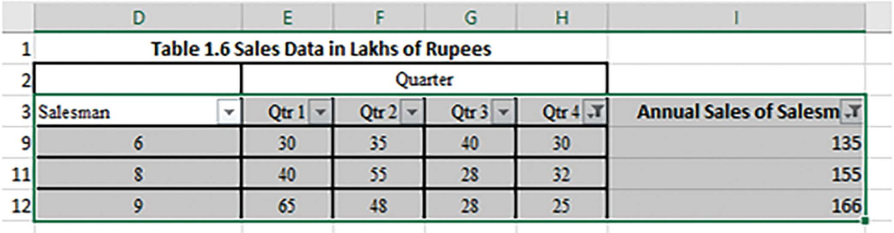


Table 1.6 Sales Data in Lakhs of Rupees

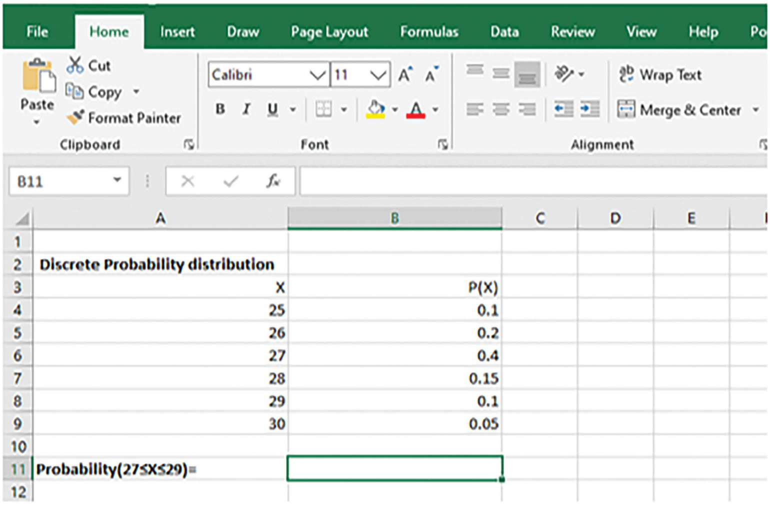
		Quarter				
3	Salesman	Qtr 1	Qtr 2	Qtr 3	Qtr 4	Annual Sales of Salesm
7	4	40	23	30	20	113
9	6	30	35	40	30	135
10	7	34	40	30	23	127
11	8	40	55	28	32	155
12	9	65	48	28	25	166
16	Mean annaul sales of salesmen =					109

Figure 2.64 Screenshot after filling the mean annual sales of the salesmen (₹ 109 lakhs) as shown in the dropdown menu of Figure 2.63 and clicking OK button in it



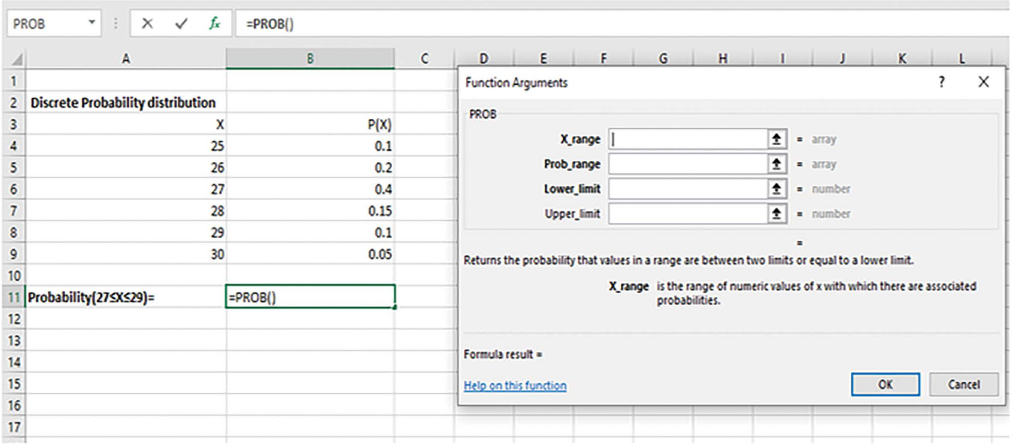
	D	E	F	G	H	I
1	Table 1.6 Sales Data in Lakhs of Rupees					
2		Quarter				
3	Salesman	Qtr 1	Qtr 2	Qtr 3	Qtr 4	Annual Sales of Salesm
9	6	30	35	40	30	135
11	8	40	55	28	32	155
12	9	65	48	28	25	166

Figure 2.65 Screenshot after clicking button in column E against the column heading Qtr 4, clicking Number Filters in its dropdown menu, “Greater Than Or Equal To” in the dropdown menu of Number Filters, Filling 25 in the box against Qtr 4 in the next dropdown menu and finally clicking OK button in it



	A	B	C	D	E
1					
2	Discrete Probability distribution				
3	X	P(X)			
4	25	0.1			
5	26	0.2			
6	27	0.4			
7	28	0.15			
8	29	0.1			
9	30	0.05			
10					
11	Probability(27≤X≤29)=				
12					

Figure 2.66 Screenshot of data of discrete probability distribution



	A	B	C
1			
2	Discrete Probability distribution		
3	X	P(X)	
4	25	0.1	
5	26	0.2	
6	27	0.4	
7	28	0.15	
8	29	0.1	
9	30	0.05	
10			
11	Probability(27≤X≤29)=	=PROB()	
12			
13			
14			
15			
16			
17			

Function Arguments

PROB

X_range = array

Prob_range = array

Lower_limit = number

Upper_limit = number

=

Returns the probability that values in a range are between two limits or equal to a lower limit.

X_range is the range of numeric values of x with which there are associated probabilities.

Formula result =

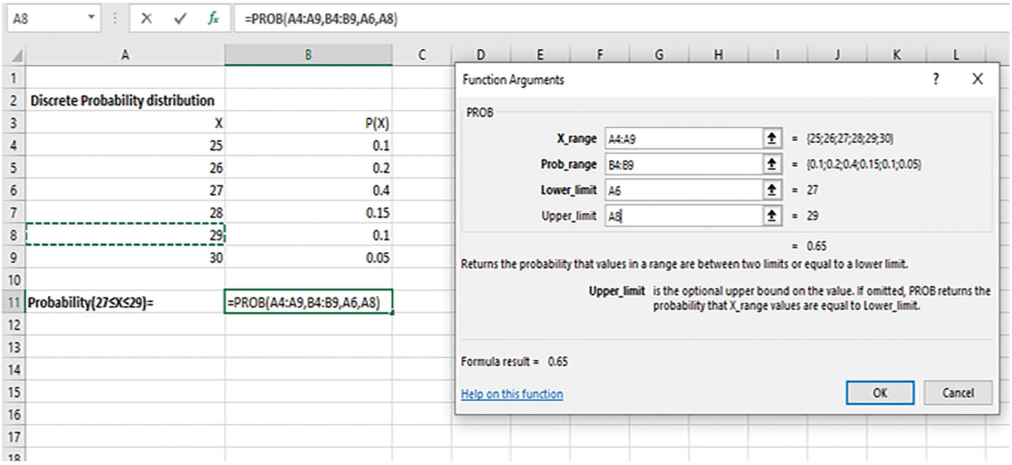
Help on this function

Figure 2.67 Screenshot for the sequence of clicks of buttons, viz. Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ PROB

2.2.30 PROB Function in Excel 2019

The sum of the probabilities for a range of values of a random variable inside a discrete probability distribution is provided by the PROB function in Excel 2019 [6].

Consider a discrete probability distribution, as shown in the screenshot of Figure 2.66. The sequence of button clicks Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ PROB will give a display as in Figure 2.67. The screenshot after filling the boxes against X_Range with cells A4:A9, Prob_Range with cells B4:B9, Lower_limit with A6, and Upper_limit with A8 is shown in Figure 2.68. Clicking the OK button in the drop-down menu of Figure 2.68 gives a screenshot as in Figure 2.69, which gives the sum of the probabilities for the values of *X*, that is, 27, 28, and 29, which is 0.65. If the value



=PROB(A4:A9,B4:B9,A6,A8)

	A	B	C
1			
2	Discrete Probability distribution		
3	X	P(X)	
4	25	0.1	
5	26	0.2	
6	27	0.4	
7	28	0.15	
8	29	0.1	
9	30	0.05	
10			
11	Probability(27≤X≤29)=	=PROB(A4:A9,B4:B9,A6,A8)	
12			
13			
14			
15			
16			
17			

Function Arguments

PROB

X_range A4:A9 = {25;26;27;28;29;30}

Prob_range B4:B9 = {0.1;0.2;0.4;0.15;0.1;0.05}

Lower_limit A6 = 27

Upper_limit A8 = 29

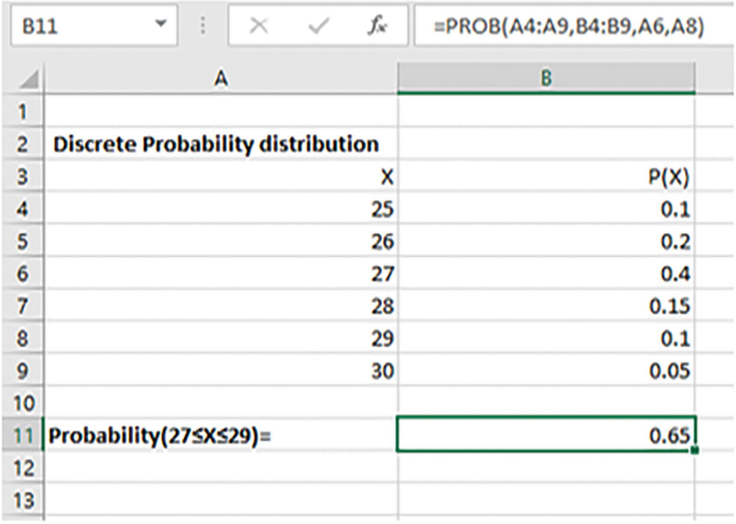
= 0.65

Returns the probability that values in a range are between two limits or equal to a lower limit.

Upper_limit is the optional upper bound on the value. If omitted, PROB returns the probability that X_range values are equal to Lower_limit.

Formula result = 0.65

Figure 2.68 Screenshot after filling data in the dropdown menu of Figure 2.67



=PROB(A4:A9,B4:B9,A6,A8)

	A	B
1		
2	Discrete Probability distribution	
3	X	P(X)
4	25	0.1
5	26	0.2
6	27	0.4
7	28	0.15
8	29	0.1
9	30	0.05
10		
11	Probability(27≤X≤29)=	0.65
12		
13		

Figure 2.69 Screenshot after clicking OK button in the dropdown menu of Figure 2.68

for the Lower_limit for *X* is the default, then the values for *X* from the beginning of the distribution will be considered. If the value for the Upper_limit for *X* is the default, then the values for *X* to its end will be considered.

Summary

- Excel is a handy tool for data analytics. It has the features of mapping even a semi-structured decision environment to support business managers in their decision-making activities.
- In Excel, the rows are numbered as 1, 2, 3, 4, 5, . . . 1048576, and the columns are labelled as A to XFD, which totals 16,384.
- The default column width of an Excel cell in an Excel sheet is 8.43, units and it can be changed.
- The Excel template can be formatted in the desired way using strings.
- The @SUM function helps to find the sum of numeric values in a range of cells.
- A formula in Excel can be copied if the same formula is used in a range of cells.
- The SUMIF function helps to sum the values of the cells in a range of cells based on a criterion.
- The SUMIFS command helps to find the sum of a range of cells for two criteria.
- Prefixing the $ symbol to the row and column defining a cell is required for a cell which should have absolute address in the formula when it is used in another cells.
- The COMBIN command gives the number of combinations when R numbers are selected from *N* numbers, where R is less than *N*.
- The FACT command gives the factorial of a given number.
- The Data Analysis button can be invoked in Excel using a procedure.
- The MIN function finds the minimum among a given set of observations.
- The MINA function finds the minimum among a given set of observations, that is, numeric value, logical, and text. The TRUE of a logical is assumed to be 1 and the FALSE of a logical is assumed to be 0. Any other text is assumed to be 0.
- The MAX function finds the maximum among a given set of observations.
- The MAXA function finds the maximum among a given set of observations, that is, numeric value, logical, and text. The TRUE of a logical is assumed to be 1 and the FALSE of a logical is assumed to be 0. Any other text is assumed to be 0.
- The MAXIFS function finds the maximum of the values in a given data range for one or more criteria applied to the data of the criterion range/criteria ranges, respectively.
- The MINIFS function finds the minimum of the values in a given data range for one or more criteria applied to the data of the criterion range/criteria ranges, respectively.
- The ROUND function rounds a given decimal number to the nearest number with a desired number of decimal digits.
- The INT function finds the truncated value of a given decimal number.
- The ROUNDDOWN function reduces the number of decimal places of a given decimal number to a desired number of decimal places with truncation to the previous nearest decimal number.
- The ROUNDDOWN function with zero decimal digits is the same as the INT function. Both functions will truncate the given decimal number to obtain its integer value.
- The ROUNDUP function in Excel rounds a given decimal number to the next nearest number with a desired number of decimal digits.
- The SORT function helps to rearrange the content of a range of cells in a given range in a desired order, that is, ascending or descending order of a particular column.

- The RANK.EQ function obtains the rank of a number among a given set of numbers either in descending or ascending order.
- The RANK.AVG function obtains the rank of a number among a given set of numbers either in descending or ascending order. If more than one value has the same rank, then the average of the ranks of those numbers is obtained and treated as the rank of those numbers.
- The Hide Columns/Hide Rows command will hide a set of columns/rows.
- The Unhide Columns/Unhide Rows command will unhide a set of columns/rows.
- The Filter function presents a subset of rows in a given range of cells which has data based on one or more criteria applied to one or more columns, respectively.
- The PROB function in Excel 2019 gives the sum of the probabilities of a range of values of a random variable of a discrete probability distribution.

Keywords

Excel is a handy tool for data analytics.

Absolute addressing of a cell requires prefixing the $ symbol to the row and the column defining that cell.

The Data Analysis button can be invoked in Excel using a procedure.

The MIN function finds the minimum among a given set of observations.

The MINA function finds the minimum among a given set of observations, that is, numeric value, logical, and text. The TRUE of a logical is assumed to be 1 and the FALSE of a logical is assumed to be 0. Any other text is assumed to be 0.

The MAX function finds the maximum among a given set of observations.

The MAXA function finds the maximum among a given set of observations, that is, numeric value, logical, and text. The TRUE of a logical is assumed to be 1 and the FALSE of a logical is assumed to be 0. Any other text is assumed to be 0.

The ROUND function rounds the given decimal number to the nearest number with a desired number of decimal digits.

The ROUNDDOWN function in Excel reduces the number of decimal places to a desired number of decimal digits without rounding.

The ROUNDUP function in Excel reduces the number of decimal places of a decimal number with a desired number of decimal digits with rounding.

The SORT function helps to rearrange the content of the cells in a given range in a desired order, that is, ascending or descending order of a particular column.

The RANK.EQ function obtains the rank of a number among a given set of numbers either in descending or ascending order.

The RANK.AVG function obtains the rank of a number among a given set of numbers either in descending or ascending order. If more than one value has the same rank, then the average of the ranks of those numbers is obtained and treated as the rank of those numbers.

The Filter function presents a subset of rows in a given range of cells which has data based on one or more criteria applied to one or more columns, respectively.

Review Questions

1. Give the specification of the Excel template.
2. Explain the arithmetic operators that are used in Excel.
3. How do you change the width of a column in Excel?

4. Illustrate the following in Excel using suitable examples.
   a. Adding two numbers.
   b. Subtracting one number from another number.
   c. Dividing one number by another number.
5. Explain the formatting in Excel for any example of your choice.
6. Illustrate the addition of values in a range of cells by including the cells in the range in a formula.
7. The components of internal marks in a subject of a PG course are shown in the following table for 10 students in an elective subject.
   a. Illustrate the use of the @SUM function to compute the total mark for the candidate with Reg. No 101.
   b. Illustrate the use of copying a formula to obtain the total marks of the remaining candidates shown in the table.

Reg. No.	*Name*	*Test 1*	*Test 2*	*Assignment 1*	*Assignment 2*	*Total*
		Max. Marks: 15	*Max. Marks: 15*	*Max. Marks: 10*	*Max. Marks: 10*	*Max. Marks: 50*
101	Anand, E.	12	13	8	9	
102	Babu, N.	11	12	7	8	
104	Chitra, K.	10	12	6	7	
108	Domnic, C.	13	14	8	7	
110	Fathima, S.	14	12	5	8	
115	Mohan, K.	12	12	6	7	
123	Nandhini, S.	13	14	8	7	
135	Peter, H.	9	12	5	9	
140	Rabinson, T.	8	12	9	8	
145	Sunil, K.	12	13	10	8	

8. Consider the internal and external marks of the following candidates. Answer the following questions using Excel.
   a. Find the total of the internal and external marks of each of the candidates.
   b. Find the sum of the internal and external marks for the following conditions using the SUMIFS function.

   The total mark is greater than or equal to 50 and the external mark is greater than or equal to 25.

   c. Find the average of the sum obtained in part b for the candidates satisfying the conditions in it.

Reg. No.	*Name*	*Internal Mark*	*External Mark*	*Total*
		Max. 50	*Max. 50*	*Max. 100*
101	Anand, E.	30	35	
102	Babu, N.	25	26	
104	Chitra, K.	14	40	

Reg. No.	*Name*	*Internal Mark*	*External Mark*	*Total*
		Max. 50	*Max. 50*	*Max. 100*
108	Domnic, D.	35	20	
110	Fathima, S.	40	45	
115	Mohan, S.	34	24	
123	Nandhini, S.	45	42	
135	Peter, H.	30	40	
140	Rabinson, T.	45	48	
145	Sunil, K.	42	40	

9. Distinguish between an absolute reference to a cell while copying a formula and relative reference to a cell while copying a formula using a suitable example for each.
10. Illustrate the COMBIN function using a suitable example in Excel.
11. Illustrate the FACT function using a suitable example in Excel.
12. Give the procedure to invoke the Data Analysis function in Excel.
13. Illustrate the MIN function and MINA function using suitable examples.
14. Illustrate the MAX function and MAXA function using suitable examples.
15. Illustrate the use of the MAXIFS function using an example.
16. Illustrate the use of the MINIFS function using an example.
17. What are the types of ROUND functions? Explain them with suitable examples.
18. What is the INT function? Compare it with the ROUNDDOWN function.
19. Illustrate the application of the SORT function to sort numbers in descending order using a suitable example.
20. Illustrate the application of the SORT function to sort numbers in ascending order using a suitable example.
21. Distinguish the RANK.EQ and RANK.AVG functions. Also illustrate them using suitable examples.
22. Illustrate Hide columns and Unhide columns using an example.
23. Illustrate Hide rows and Unhide rows using an example.
24. Consider the data shown in the following table. From that table, filter the rows for the following two criteria using Excel to obtain the list of candidates who passed the subject.

    Criterion 1: Total mark is greater than or equal to 50.
    and
    Criterion 2: External mark is greater than or equal to 30.

Reg. No.	*Name*	*Internal Marks*	*External Marks*	*Total*
		Max. Marks: 50	*Max. Marks: 50*	*Max. Marks: 100*
101	Anand, E.	30	35	
102	Babu, N.	25	26	
104	Chitra, K.	14	40	
108	Domnic, D.	35	20	
110	Fathima, S.	40	45	
115	Mohan, S.	34	24	
123	Nandhini, S.	45	42	

Reg. No.	*Name*	*Internal Marks*	*External Marks*	*Total*
		Max. Marks: 50	*Max. Marks: 50*	*Max. Marks: 100*
135	Peter, H.	30	40	
140	Rabinson, T.	45	48	
145	Sunil, K.	42	40	

25. Illustrate the application of the PROB function in Excel 2019 using an example.
26. The purchase price (*P*) of a machine is Rs 80,00,000. The life of the machine (*n*) is 10 years, and its scrap value (*S*) is Rs 5,00,000. The formula for the *depreciation amount* during the year *t* is given below.

$$D_t = \frac{P - S}{n}, \mathrm{t} = 1, 2, 3, \ldots., \mathrm{n}$$

The formula for the *book value* of the machine at the end of year *t* is given below.

$$B_t = B_{t-1} - D_t, t = 1,2,3,\ldots.,10$$

Construct an Excel sheet in the following format and compute D_t and B_t for t = 1, 2, 3, . . . , 10.

Year	*Deprecitaion for year t* (D_t)	*Book Vlaue at the end of year t* (B_t)
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

27. Fifty suppliers supply raw materials and subassemblies to a major vehicle OEM. The OEM's materials manager begins to suspect that most vendors frequently furnish raw materials after the deadlines. He therefore takes into account the most recent shipments of all the raw materials from all the suppliers for analysis. As a consultant, assist the materials manager in creating an Excel sheet that includes the following information, Supplier Code, Supplier Name, Raw Material Code, Due Date of Raw Material, Supply Data for Raw Material, Extent of Lateness if Applicable, and filters out the suppliers who supplied the raw materials after the given due dates.

References

1. https://corporatefinanceinstitute.com/resources/Excel/functions/maxifs-function-in-Excel/ [June 29, 2020].
2. https://corporatefinanceinstitute.com/resources/Excel/functions/maxifs-function-in-Excel/ [June 29, 2020].

3. https://Exceljet.net/Excel-functions/Excel-minifs-function [June 29, 2020].
4. https://support.microsoft.com/en-us/office/filter-function-f4f7cb66-82eb-4767-8f7c-4877ad80c759 [June 26, 2020].
5. www.Exceltip.com/Excel-365-functions/how-to-use-the-Excel-filter-function.html [June 26, 2020].
6. https://support.microsoft.com/en-us/office/prob-function-9ac30561-c81c-4259-8253-34f0a238fc49 [June 29, 2020].

3 Count, Frequency, and Histogram Functions

Learning Objectives

A complete reading of this chapter will help readers to

- Know the syntax of COUNT and implement it in Excel.
- Understand the difference between the COUNT function and COUNTA, with an illustration of the COUNTA function using an example.
- Implement the COUNTBLANK function, which is normally used in analysing survey data.
- Understand the syntax of the COUNTIF function while a criterion is involved in counting cells.
- Know the extension of the COUNTIF function in the form of COUNTIFS to handle situations with more than one criterion.
- Understand the use of the frequency function to form frequencies of a given set of data.
- Analyse data using histograms.

3.1 Introduction

The COUNT function is crucial in statistics. The count function counts the instances of observations of a specific variable, either with or without a condition. Additionally, the histogram and frequency functions take care of the initial data processing. The quantity of occurrences of a value in a given set of observations is known as its frequency.

The frequency analysis is extended by the histogram. The frequencies are plotted graphically against the data values, either in discrete form or in interval form. The values are taken in the X axis, and the frequencies are taken on the Y axis. This graph makes it simple to understand the pattern of the provided data.

3.2 Count Functions

The different count functions available in Excel are as listed.

1. COUNT
2. COUNTA
3. COUNTBLANK
4. COUNTIF
5. COUNTIFS

DOI: 10.4324/9781032671703-3

3.2.1 COUNT Function

The COUNT function counts the number of cells in a given range of cells of an Excel sheet that contains numbers.

The following sequence of buttons in Excel gives the screenshot shown in Figure 3.1.

Formulas ⟹ More Functions ⟹ Statistical

In Figure 3.1, clicking COUNT gives the screenshot shown in Figure 3.2. Now, one can perform any one of the following.

1. Enter the data (numbers or any other type) against Value 1, Value 2, and so on.
2. Enter a range of cells which contains data (numbers or any other type) against Value 1.

(Value 1, Value 2, . . . are 1 to 255 arguments that can contain or refer to a variety of different types of data, but only numbers are counted.)

Then clicking the OK button in the dropdown menu of Figure 3.2 will give the number of cells, which contains numbers only.

The Excel command for the COUNT function is as follows.

$$= \text{COUNT}(\text{range of cells containing data})$$

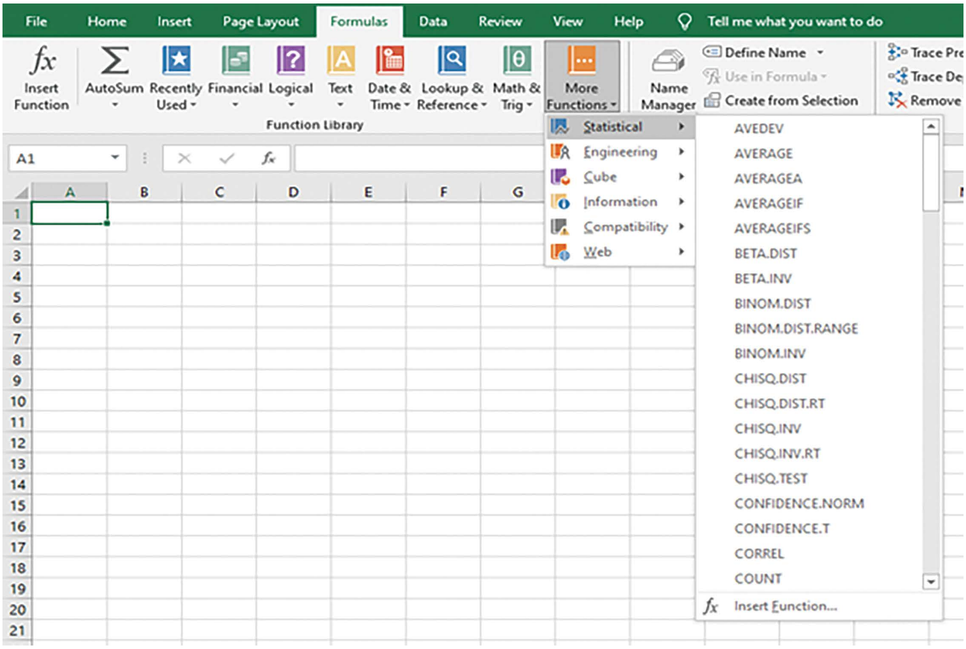
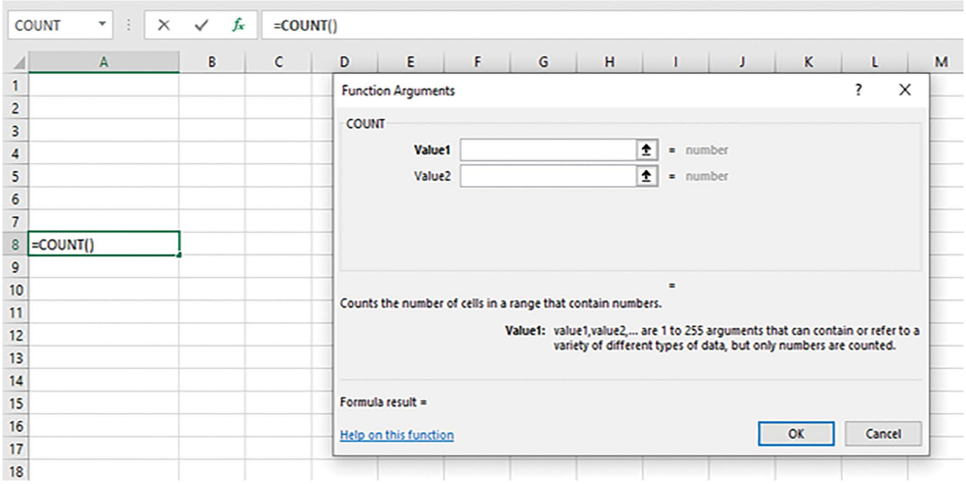


Figure 3.1 Screenshot of sequence of button clicks, Formulas ⟹ More Functions ⟹ Statistical



=COUNT()

Counts the number of cells in a range that contain numbers.

Value1: value1,value2,... are 1 to 255 arguments that can contain or refer to a variety of different types of data, but only numbers are counted.

Formula result =

Help on this function

Figure 3.2 Screenshot of clicking the COUNT function in the dropdown menu of Figure 3.1

Example 3.1

The profit (+)/loss (–)/no revenue of a company in crores of rupees for the past eight years are summarised in Table 3.1. Find the count of years during which either profit or loss is reported

Table 3.1 Data for Example 3.1

Year	*Profit*
1	220
2	230
3	210
4	–270
5	No revenue
6	No revenue
7	10
8	100

Solution

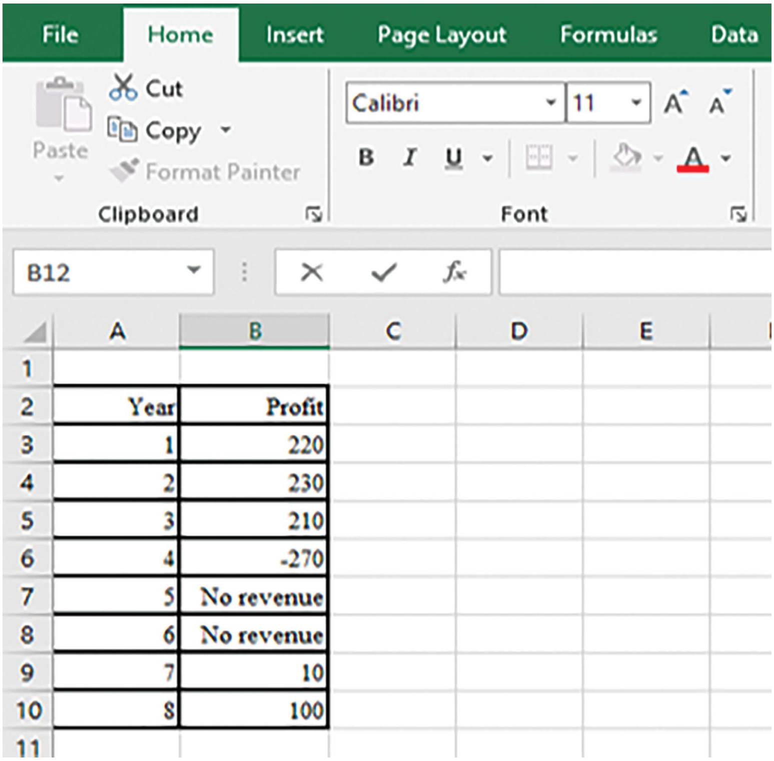
The data for Example 3.1 are shown in Table 3.2.

Step 1: A screenshot after inputting data into the Excel sheet is shown in Figure 3.3.
Step 2: A screenshot after clicking the sequence of buttons Formulas ⟹ More Functions ⟹ Statistical gives the display shown in Figure 3.4.
Step 3: Clicking the COUNT function in the dropdown menu of Figure 3.4 gives the display shown in Figure 3.5.
Step 4: Now, enter the range of cells from B3 to B10 against Value 1 in the dropdown menu of Figure 3.5 as shown in Figure 3.6, and then click the OK button in the same dropdown menu to show the result for the COUNT function, as shown in Figure 3.7.

The count of cells from B3 to B10 that contain numbers is 6, which is displayed in cell B12.

Table 3.2 Data for Example 3.1

Year	*Profit*
1	220
2	230
3	210
4	–270
5	No revenue
6	No revenue
7	10
8	100



	A	B
1		
2	Year	Profit
3	1	220
4	2	230
5	3	210
6	4	-270
7	5	No revenue
8	6	No revenue
9	7	10
10	8	100
11		

Figure 3.3 Screenshot after inputting data into the Excel sheet

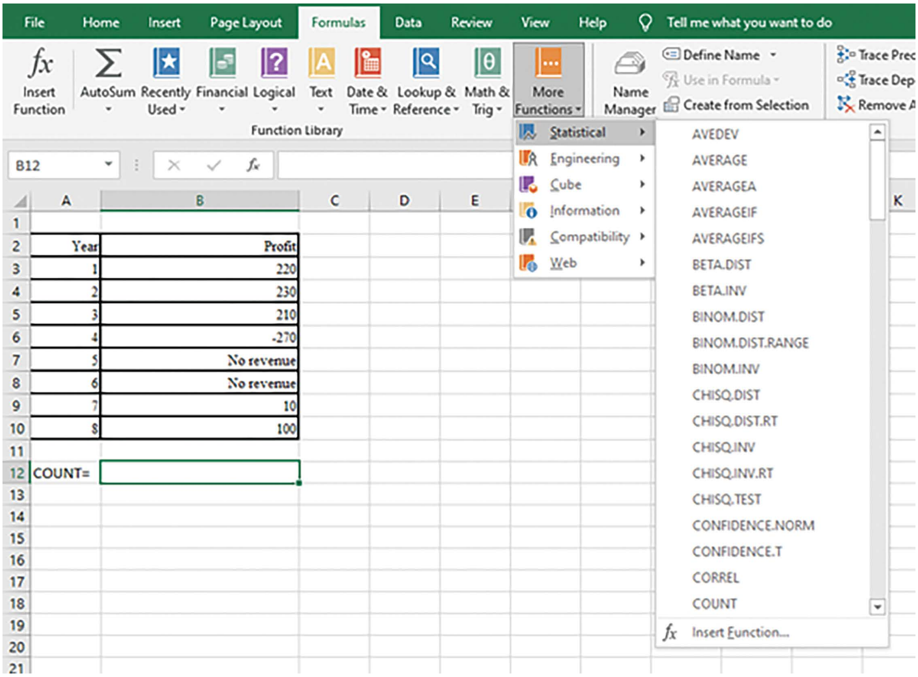
All these steps can be combined into a formula type command as shown.

= COUNT(B3 : B10)

This formula may be entered in any convenient cell, say, B12, to give the result.

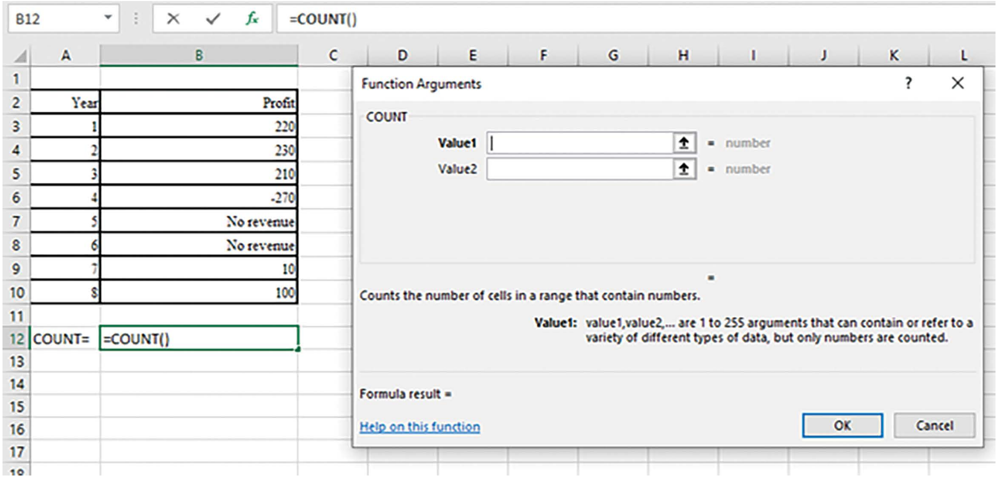
3.2.2 COUNTA Function

The COUNTA function counts the number of cells in a specified range of cells that are not empty. Data may not be accessible for a particular variable for certain respondents during the data collection process. The number of non-empty cells in such a result can be



	A	B
1		
2	Year	Profit
3	1	220
4	2	230
5	3	210
6	4	-270
7	5	No revenue
8	6	No revenue
9	7	10
10	8	100
11		
12	COUNT=	

Figure 3.4 Screenshot after clicking the sequence of buttons, Formulas ⟹ More Functions ⟹ Statistical



	A	B
1		
2	Year	Profit
3	1	220
4	2	230
5	3	210
6	4	-270
7	5	No revenue
8	6	No revenue
9	7	10
10	8	100
11		
12	COUNT=	=COUNT()

COUNT

Value1 = number

Value2 = number

=

Counts the number of cells in a range that contain numbers.

Value1: value1,value2,... are 1 to 255 arguments that can contain or refer to a variety of different types of data, but only numbers are counted.

Formula result =

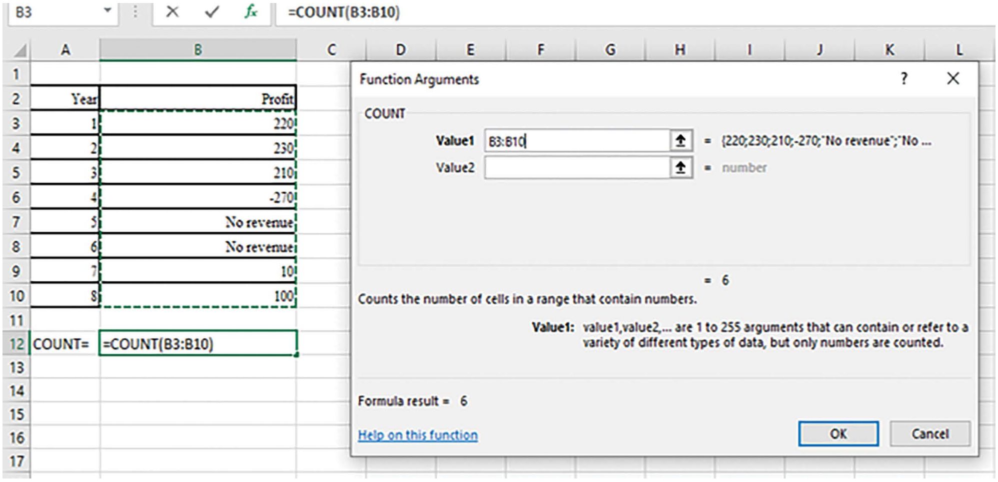
Help on this function

Figure 3.5 Screenshot after clicking COUNT function in the dropdown menu of Figure 3.4

determined with the use of this function. The following formula can be used to determine COUNTA's value.

=COUNTA(Range of cells to be considered) (Range of cells to be considered)

Example 3.2 shows this.



B3 =COUNT(B3:B10)

	A	B
1		
2	Year	Profit
3	1	220
4	2	230
5	3	210
6	4	-270
7	5	No revenue
8	6	No revenue
9	7	10
10	8	100
11		
12	COUNT=	=COUNT(B3:B10)

Function Arguments

COUNT

Value1 B3:B10 = {220;230;210;-270;"No revenue";"No ...

Value2 = number

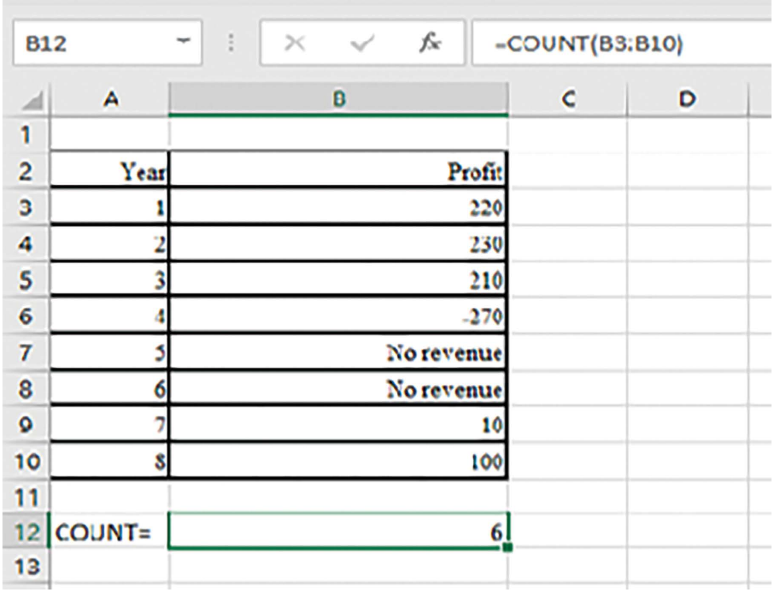
= 6

Counts the number of cells in a range that contain numbers.

Value1: value1,value2,... are 1 to 255 arguments that can contain or refer to a variety of different types of data, but only numbers are counted.

Formula result = 6

Figure 3.6 Screenshot after selecting the range of cells, that is, B3:B10 in the data box against Number 1 in the dropdown menu of Figure 3.5



B12 =COUNT(B3:B10)

	A	B
1		
2	Year	Profit
3	1	220
4	2	230
5	3	210
6	4	-270
7	5	No revenue
8	6	No revenue
9	7	10
10	8	100
11		
12	COUNT=	6

Figure 3.7 Screenshot after clicking OK in the dropdown menu shown in Figure 3.6

Example 3.2

According to a survey, Table 3.3 provides a summary of the power consumption data for businesses. Some businesses have not provided these details; thus, they are displayed as blanks. Find the number of businesses that have provided data for power usage using the COUNTA function.

Table 3.3 Power Consumption Data

Company	*Power Consumption (KWH)*
1	1,000
2	6,000
3	
4	2,900
5	
6	4,000
7	
8	9,000

Solution

The data for Example 3.2 are shown in Table 3.4.

The input of the data in Excel is shown in Figure 3.8. A screenshot of button clicks Formulas ⟹> More Functions ⟹> Statistical ⟹ COUNTA is shown in Figure 3.9. A screenshot after entering the range of cells B3:B10 in the cell against Value 1 is shown in Figure 3.10. A screenshot after clicking the OK button in the dropdown menu of Figure 3.10 is shown in Figure 3.11. The result of the COUNTA function applied to this example is 5.

If one wants to use a formula for COUNTA, then position the cursor in the desired location and enter the formula, which is as follows.

= COUNTA(B3 : B10)

3.2.3 COUNTBLANK Function

The COUNTBLANK function counts the number of empty or blank cells within a specified range of cells [1]. While data are being gathered, it’s possible that some respondents won’t have data for a particular variable of interest. These empty cells must be counted in order to determine how many respondents failed to supply information for that variable. Through the COUNTBALNK function, this is accomplished.

The formula to find the value of COUNTBLANK is as follows.

= COUNTBLANK(Range of cells to be considered)

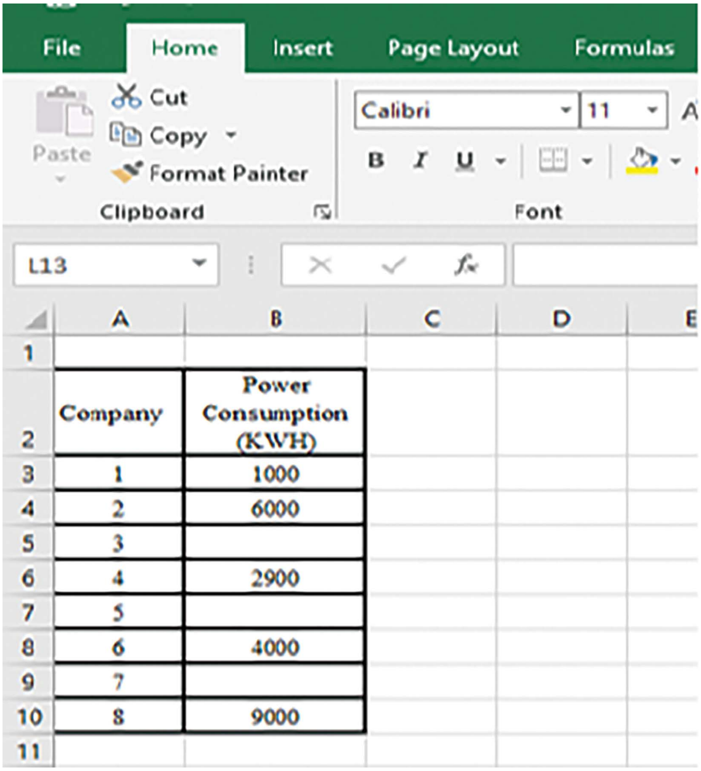
This is illustrated using an example as follows.

Example 3.3

In a survey, the power consumption data of companies are summarised in Table 3.5. Some companies have not furnished such data, so the cells are shown as blank. Using the COUNTBLANK function, find the number of companies which did not furnish data for power consumption.

Table 3.4 Data for Example 3.2

Company	*Power Consumption (KWH)*
1	1,000
2	6000
3	
4	2900
5	
6	4000
7	
8	9000



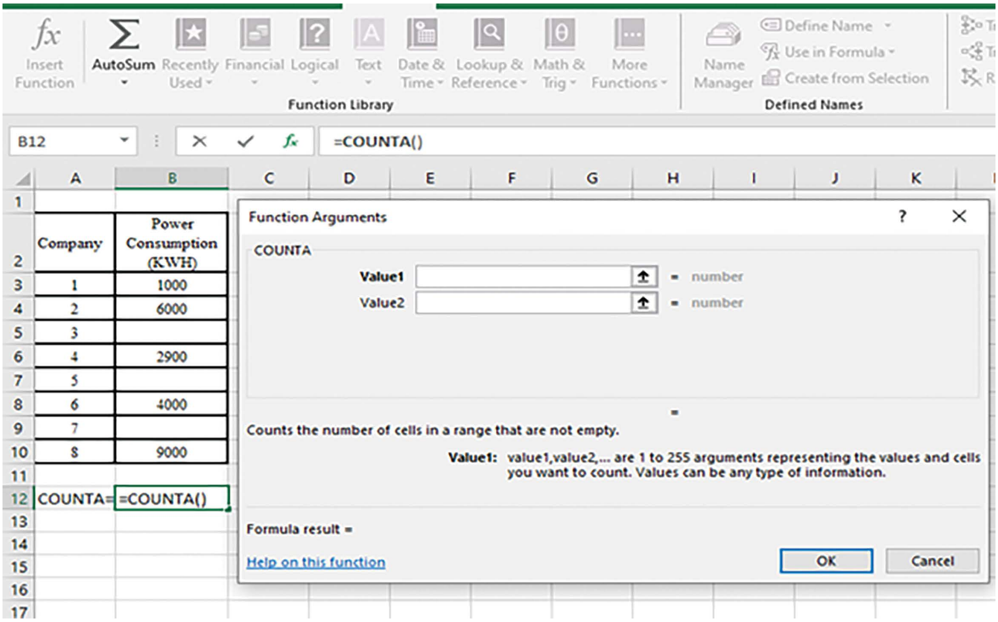
	A	B
1		
2	Company	Power Consumption (KWH)
3	1	1000
4	2	6000
5	3	
6	4	2900
7	5	
8	6	4000
9	7	
10	8	9000
11		

Figure 3.8 Screenshot of input of Example 3.2 in Excel sheet

Solution

The data for Example 3.3 are shown in Table 3.6.

The input of the data in Excel is shown in Figure 3.12. A screenshot of the button clicks Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTBLANK is shown in Figure 3.13. A screenshot after entering the range of cells B3:B10 under consideration in the cell against Range is shown in Figure 3.14. A screenshot after clicking the OK button in the dropdown menu of Figure 3.14 is shown in Figure 3.15. The result of the COUNTBLANK function applied to this example is 3.



=COUNTA()

	A	B
1		
2	Company	Power Consumption (KWH)
3	1	1000
4	2	6000
5	3	
6	4	2900
7	5	
8	6	4000
9	7	
10	8	9000
11		
12	COUNTA=	=COUNTA()

Function Arguments

COUNTA

Value1 = number

Value2 = number

=

Counts the number of cells in a range that are not empty.

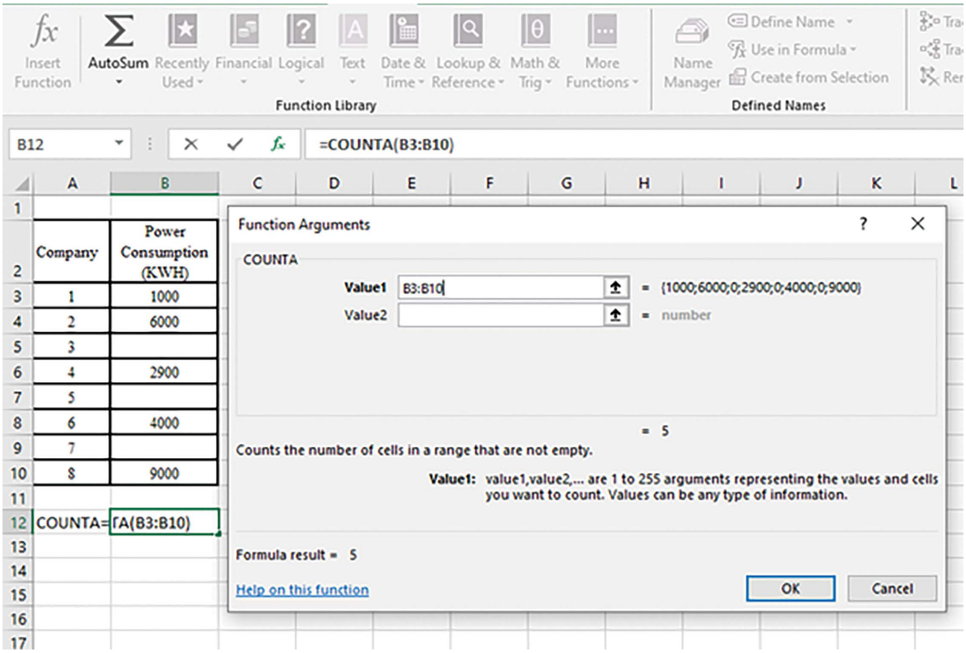
Value1: value1,value2,... are 1 to 255 arguments representing the values and cells you want to count. Values can be any type of information.

Formula result =

Help on this function

OK Cancel

Figure 3.9 Screenshot of button clicks, Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTA



=COUNTA(B3:B10)

	A	B
1		
2	Company	Power Consumption (KWH)
3	1	1000
4	2	6000
5	3	
6	4	2900
7	5	
8	6	4000
9	7	
10	8	9000
11		
12	COUNTA=	TA(B3:B10)

Function Arguments

COUNTA

Value1 B3:B10 = {1000;6000;0;2900;0;4000;0;9000}

Value2 = number

= 5

Counts the number of cells in a range that are not empty.

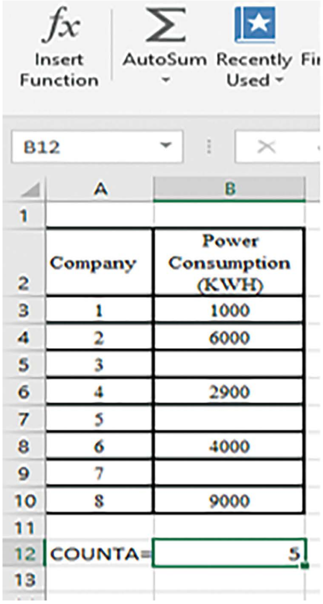
Value1: value1,value2,... are 1 to 255 arguments representing the values and cells you want to count. Values can be any type of information.

Formula result = 5

Help on this function

OK Cancel

Figure 3.10 Screenshot after the range of cells is entered in Value 1



B12

	A	B
1		
2	Company	Power Consumption (KWH)
3	1	1000
4	2	6000
5	3	
6	4	2900
7	5	
8	6	4000
9	7	
10	8	9000
11		
12	COUNTA=	5
13		

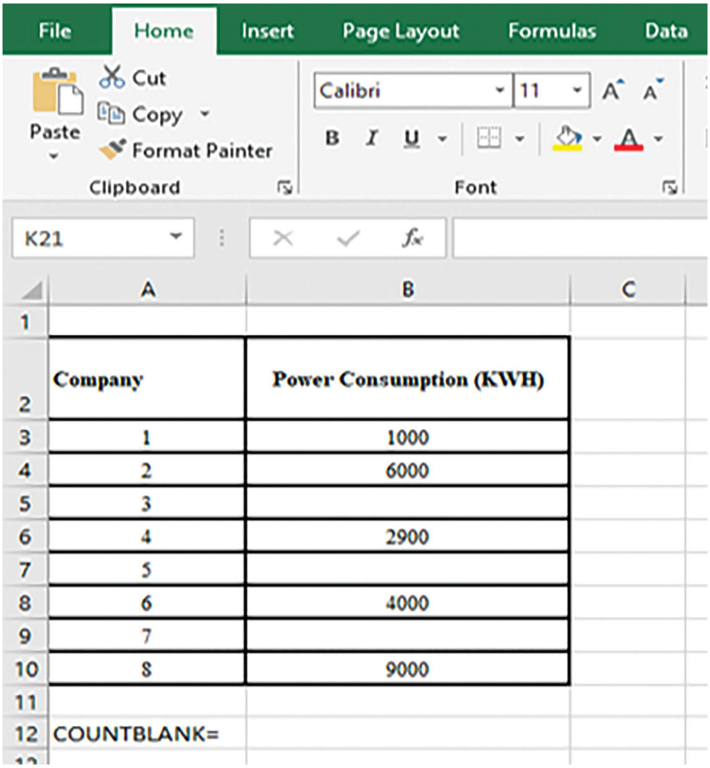
Figure 3.11 Screenshot after clicking the OK button in the dropdown menu of Figure 3.10

Table 3.5 Power Consumption Data

Company	*Power Consumption (KWH)*
1	1,000
2	6,000
3	
4	2,900
5	
6	4,000
7	
8	9,000

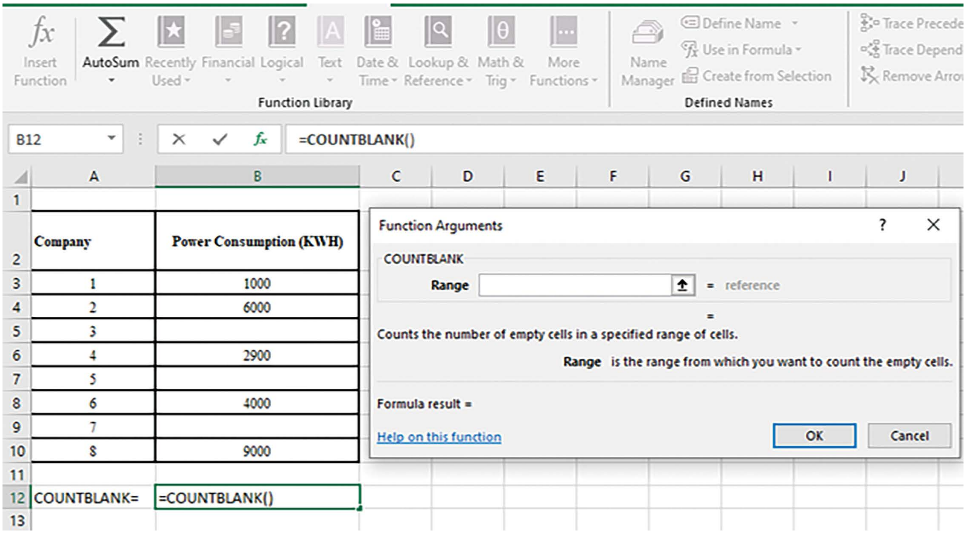
Table 3.6 Data for Example 3.3

Company	*Power Consumption (KWH)*
1	1,000
2	6,000
3	
4	2,900
5	
6	4,000
7	
8	9,000



	A	B
1		
2	**Company**	**Power Consumption (KWH)**
3	1	1000
4	2	6000
5	3	
6	4	2900
7	5	
8	6	4000
9	7	
10	8	9000
11		
12	COUNTBLANK=	

Figure 3.12 Screenshot of input of Example 3.3 in Excel sheet



=COUNTBLANK()

	A	B
1		
2	**Company**	**Power Consumption (KWH)**
3	1	1000
4	2	6000
5	3	
6	4	2900
7	5	
8	6	4000
9	7	
10	8	9000
11		
12	COUNTBLANK=	=COUNTBLANK()

Function Arguments

COUNTBLANK

Range = reference

=

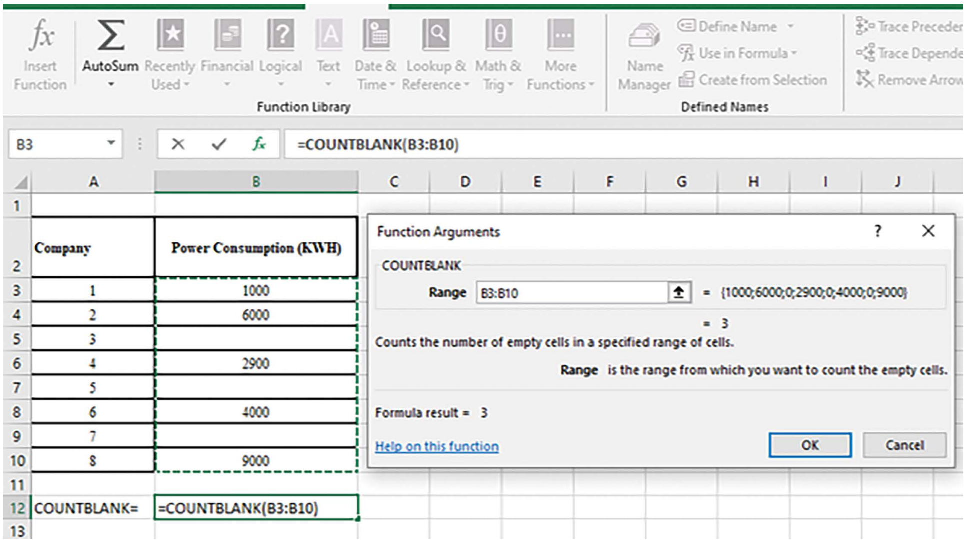
Counts the number of empty cells in a specified range of cells.

Range is the range from which you want to count the empty cells.

Formula result =

Help on this function

Figure 3.13 Screenshot of button clicks, Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTBLANK



B3 =COUNTBLANK(B3:B10)

	A	B
1		
2	Company	Power Consumption (KWH)
3	1	1000
4	2	6000
5	3	
6	4	2900
7	5	
8	6	4000
9	7	
10	8	9000
11		
12	COUNTBLANK=	=COUNTBLANK(B3:B10)
13		

Function Arguments

COUNTBLANK

Range B3:B10 = {1000;6000;0;2900;0;4000;0;9000}

= 3

Counts the number of empty cells in a specified range of cells.

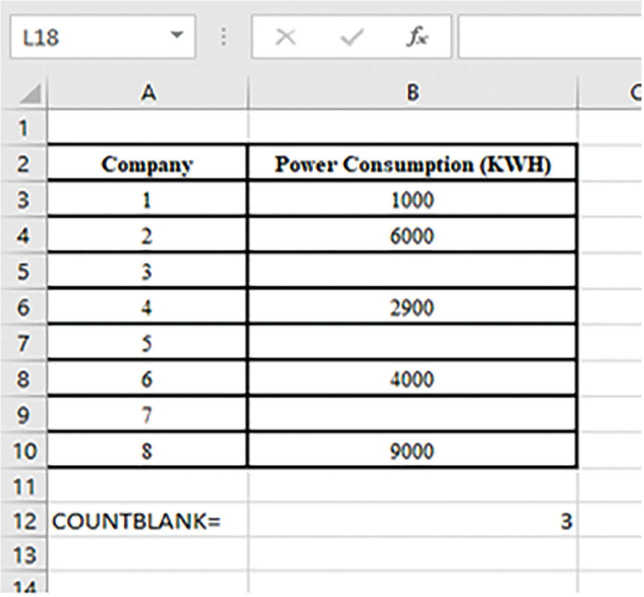
Range is the range from which you want to count the empty cells.

Formula result = 3

Help on this function

OK Cancel

Figure 3.14 Screenshot after entering the range of cells in the box against Range



L18

	A	B
1		
2	Company	Power Consumption (KWH)
3	1	1000
4	2	6000
5	3	
6	4	2900
7	5	
8	6	4000
9	7	
10	8	9000
11		
12	COUNTBLANK=	3
13		

Figure 3.15 Screenshot after clicking the OK button in the dropdown menu of Figure 3.14

If one wants to use a formula for COUNTBLANK, position the cursor in the desired location and then enter the formula, which is as follows.

$$= \text{COUNTBLANK}(\text{B3} : \text{B10})$$

3.2.4 COUNTIF Function

The COUNTIF function counts the number of cells that satisfy a certain criterion inside a specified range of cells. The criterion may be either <, <=, =, >, or >= [2, 3].

Consider the data of a company about 52 weeks' highest share prices for different years ending March as shown in Table 3.7. Here, the objective may be to find the number of years during which the 52-week highest share price is more than, say, ₹ 200. The answer is 3, which can be obtained using the COUNTIF function.

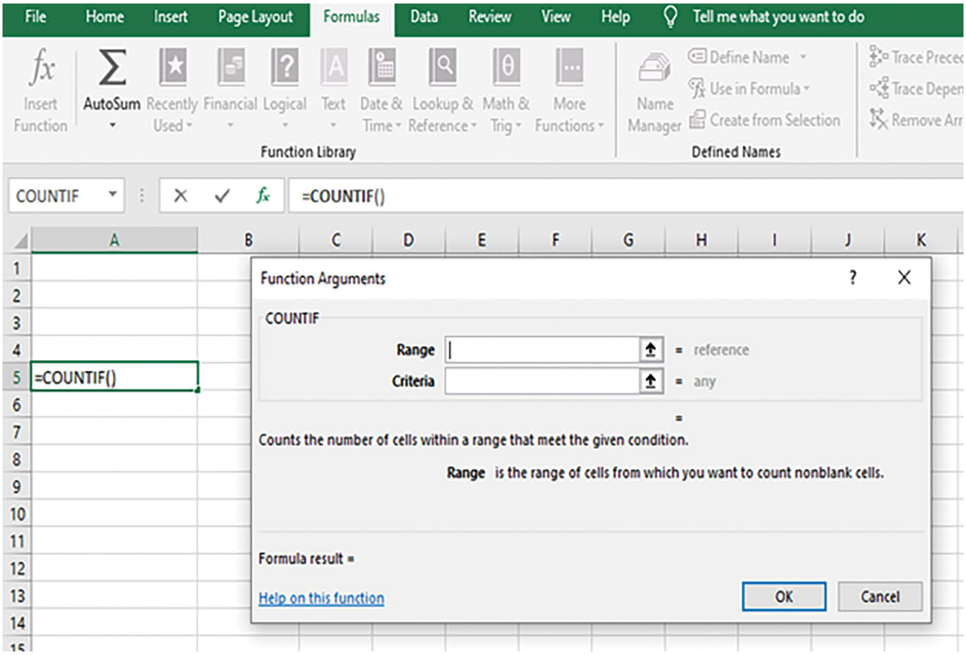
The formula to obtain the result is as follows.

= COUNTIF(Range of cells containing data, Criterion used to count the cells in the specified range)

The sequence of button clicks Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTIF gives the screenshot in Figure 3.16.

Table 3.7 Sample Data

Year Ending March	*52-Week Highest Price*
2013	190
2014	200
2015	180
2016	300
2017	250
2018	350



=COUNTIF()

Function Arguments

COUNTIF

Range | = reference

Criteria | = any

=

Counts the number of cells within a range that meet the given condition.

Range is the range of cells from which you want to count nonblank cells.

Formula result =

Help on this function

Figure 3.16 Screenshot after clicking sequence of buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTIF

Example 3.4

Table 3.8 lists the 52-week highest share prices for Alpha Engineering Company over the previous six years. Utilise the COUNTIF function to get the number of years where the 52-week highest share price for that company exceeded ₹ 200.

Solution

The data for Example 3.4 are shown in Table 3.9.

The input of the data for Example 3.4 in an Excel sheet is shown in Figure 3.17. The clicks of the buttons Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTIF give a display, as shown in Figure 3.18. In Figure 3.18, Range means the range of cells which contain the data for which one is interested in counting the cells satisfying the criterion that the value in each cell is more than 200. Criteria means the condition (<, <=, =, >, or >=), which is used to count the cells in the specified range of the Excel sheet. The entry of the range of cells B3:B8 against the range and >200 against the criteria in the dropdown menu of Figure 3.18 are shown in Figure 3.19. Then clicking the OK button in the dropdown menu of Figure 3.19 gives the result shown in Figure 3.20. The result of applying COUNTIF with the specified condition to the range of cells specified is shown in Figure 3.20, and the result is 3.

3.2.5 COUNTIFS Function

The COUNTIFS function is comparable to the COUNTIF function with the exception that it contains multiple criteria. Take a look at Table 3.10, which provides information on the age and weight of ten employees of a company.

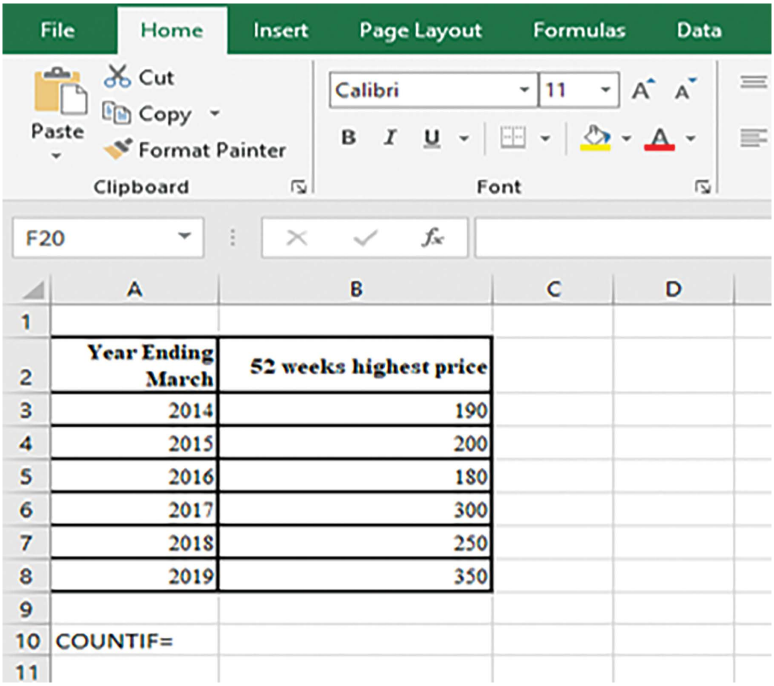
The investigator may be interested in counting the number of employees with age more than 40 and weight less than 55, which can be obtained using the COUTNTIFS function.

Table 3.8 Data for Example 3.4

Year Ending March	*52-Week Highest Price*
2014	190
2015	200
2016	180
2017	300
2018	250
2019	350

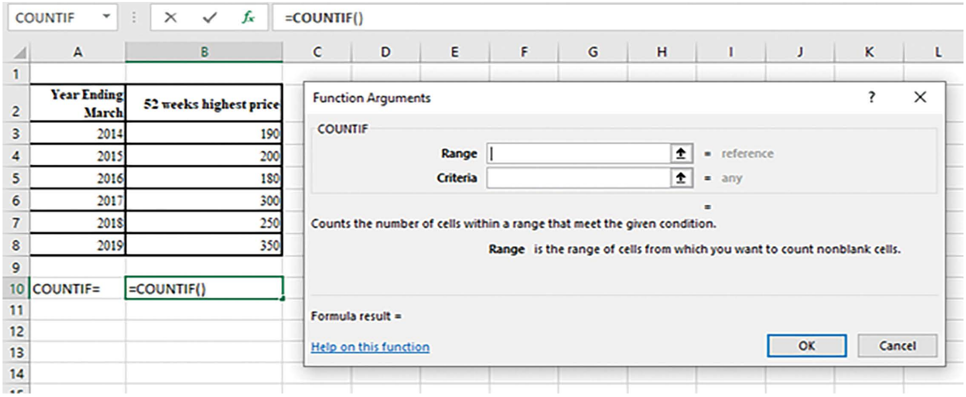
Table 3.9 Reproduction of Data for Example 3.4

Year Ending March	*52-week Highest Price*
2014	190
2015	200
2016	180
2017	300
2018	250
2019	350



	A	B
1		
2	Year Ending March	52 weeks highest price
3	2014	190
4	2015	200
5	2016	180
6	2017	300
7	2018	250
8	2019	350
9		
10	COUNTIF=	
11		

Figure 3.17 Screenshot of input of Example 3.4 in Excel sheet



	A	B
1		
2	Year Ending March	52 weeks highest price
3	2014	190
4	2015	200
5	2016	180
6	2017	300
7	2018	250
8	2019	350
9		
10	COUNTIF=	=COUNTIF()

COUNTIF

Range = reference

Criteria = any

=

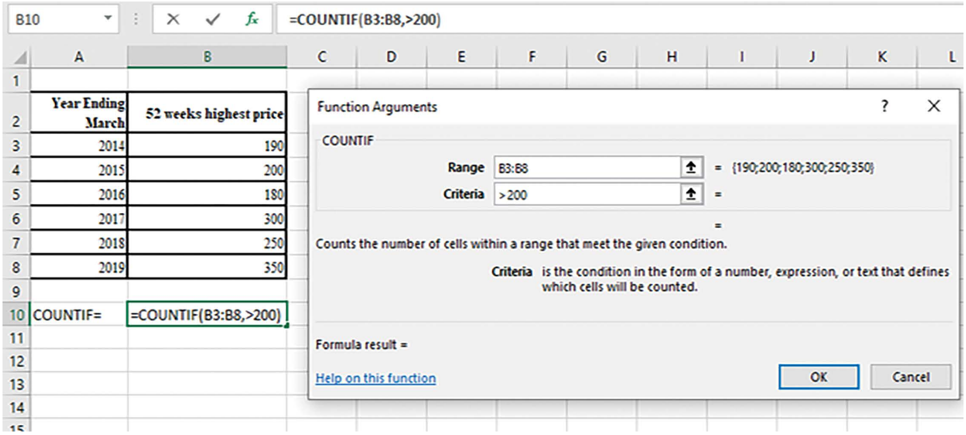
Counts the number of cells within a range that meet the given condition.

Range is the range of cells from which you want to count nonblank cells.

Formula result =

Figure 3.18 Screenshot for the sequence of clicks, Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTIF

The button clicks Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTIFS give a display, as shown in Figure 3.21. The dropdown menu of Figure 3.21 contains Criteria_Range1 and Criteria1. This is extendable for a number of criteria used by the investigator. The range of cells which contain data for Criterion 1 must be entered in the box against Criteria1_Range, and the condition for Criterion1 must be entered in the box against Criteria1. After entering these details in the dropdown menu of Figure 3.21 for



B10 =COUNTIF(B3:B8,>200)

	A	B
1		
2	Year Ending March	52 weeks highest price
3	2014	190
4	2015	200
5	2016	180
6	2017	300
7	2018	250
8	2019	350
9		
10	COUNTIF=	=COUNTIF(B3:B8,>200)

Function Arguments

COUNTIF

Range B3:B8 = {190;200;180;300;250;350}

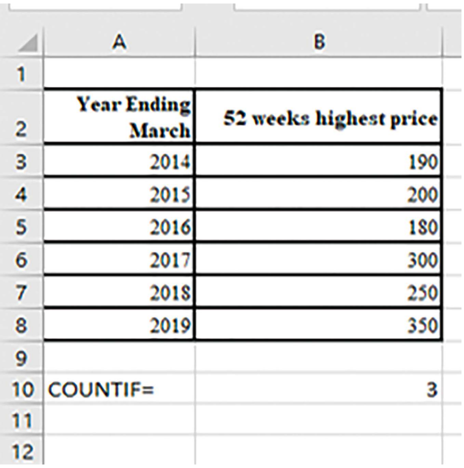
Criteria >200 =

Counts the number of cells within a range that meet the given condition.

Criteria is the condition in the form of a number, expression, or text that defines which cells will be counted.

Formula result =

Figure 3.19 Screenshot after entering the range of cells and criterion against Range and Criteria, respectively, in the dropdown menu of Figure 3.18

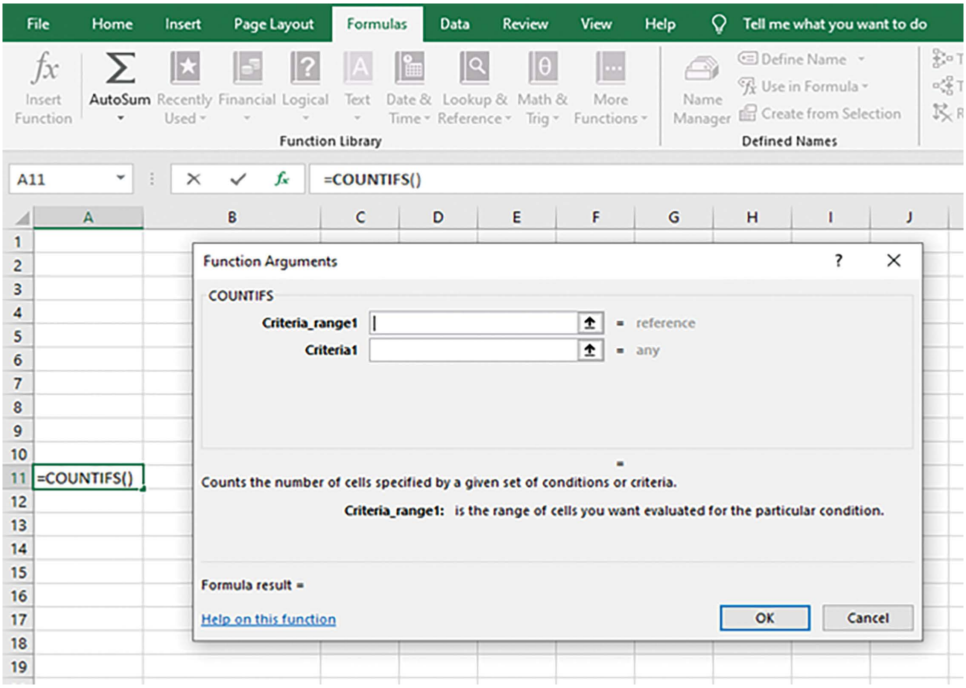


	A	B
1		
2	Year Ending March	52 weeks highest price
3	2014	190
4	2015	200
5	2016	180
6	2017	300
7	2018	250
8	2019	350
9		
10	COUNTIF=	3
11		
12		

Figure 3.20 Screenshot after clicking the OK button in the dropdown menu of Figure 3.19

Table 3.10 Age and Weight of Employees

Employee Code	*Age (Years)*	*Weight (Kg)*
1	23	55
2	34	60
3	54	65
4	21	50
5	33	64
6	47	52
7	57	70
8	28	65
9	45	73
10	40	80



A11 =COUNTIFS()

11 =COUNTIFS()

Function Arguments

COUNTIFS

Criteria_range1 = reference

Criteria1 = any

=

Counts the number of cells specified by a given set of conditions or criteria.

Criteria_range1: is the range of cells you want evaluated for the particular condition.

Formula result =

Figure 3.21 Screenshot of the clicks, Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTIFS

the required number of criteria, clicking the OK button in its dropdown menu will give the desired result for the COUNTIFS function.

Example 3.5

Consider an example which contains data on the age and weight of ten employees in an organisation, as shown in Table 3.11.

Find the number of employees with age more than 40 and weight less than 50 using the COUTNTIFS function.

Solution

The data for Example 3.5 are shown in Table 3.12.

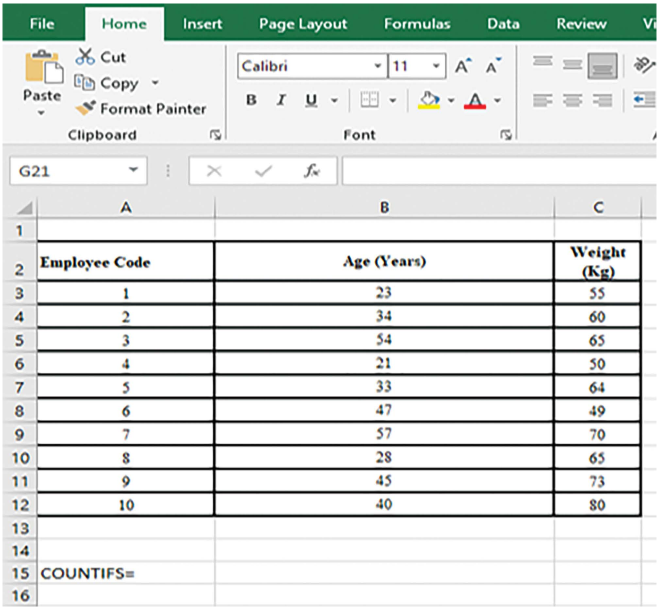
The input of Example 3.5 in an Excel sheet is shown in Figure 3.22. The button clicks Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTIFS give a display, as shown in Figure 3.23. The dropdown menu of Figure 3.23 contains Criteria range 1 and Criteria1. The display after entering the range of cells B3:B12 in the box against Criteria_Range1 and >40 in the box against Criteria1 is shown in Figure 3.24. The display after entering the range of cells C3:C12 in the box against Criteria_Range2 and <50 in the box against Criteria2 is shown in Figure 3.25. Then clicking the OK button in the dropdown menu of Figure 3.25 gives the result in cell B15, which is 1, as shown in Figure 3.26.

Table 3.11 Age and Weight of Employees

Employee Code	*Age (Years)*	*Weight (Kg)*
1	23	55
2	34	60
3	54	65
4	21	50
5	33	64
6	47	49
7	57	70
8	28	65
9	45	73
10	40	80

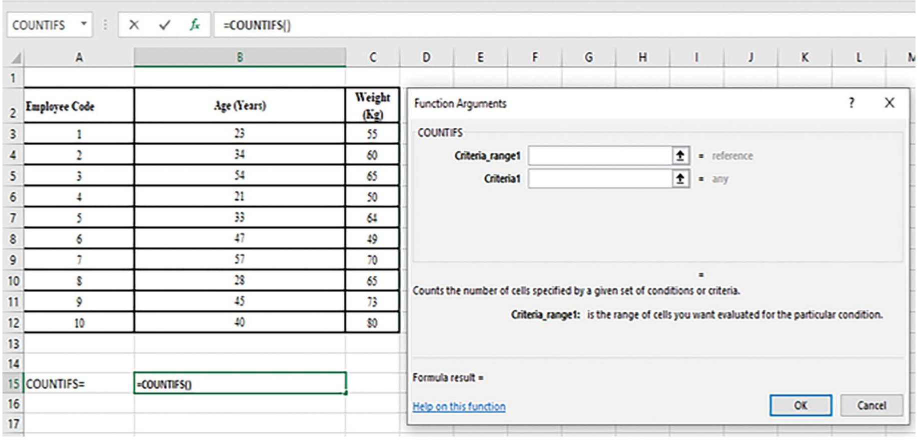
Table 3.12 Data for Example 3.5

Employee Code	*Age (Years)*	*Weight (Kg)*
1	23	55
2	34	60
3	54	65
4	21	50
5	33	64
6	47	49
7	57	70
8	28	65
9	45	73
10	40	80



	A	B	C
1			
2	Employee Code	Age (Years)	Weight (Kg)
3	1	23	55
4	2	34	60
5	3	54	65
6	4	21	50
7	5	33	64
8	6	47	49
9	7	57	70
10	8	28	65
11	9	45	73
12	10	40	80
13			
14			
15	COUNTIFS=		
16			

Figure 3.22 Screenshot of input of Example 3.5



COUNTIFS   =COUNTIFS()

Employee Code	Age (Years)	Weight (Kg)
1	23	55
2	34	60
3	54	65
4	21	50
5	33	64
6	47	49
7	57	70
8	28	65
9	45	73
10	40	80

COUNTIFS= =COUNTIFS()

Function Arguments

COUNTIFS

Criteria_range1 = reference

Criteria1 = any

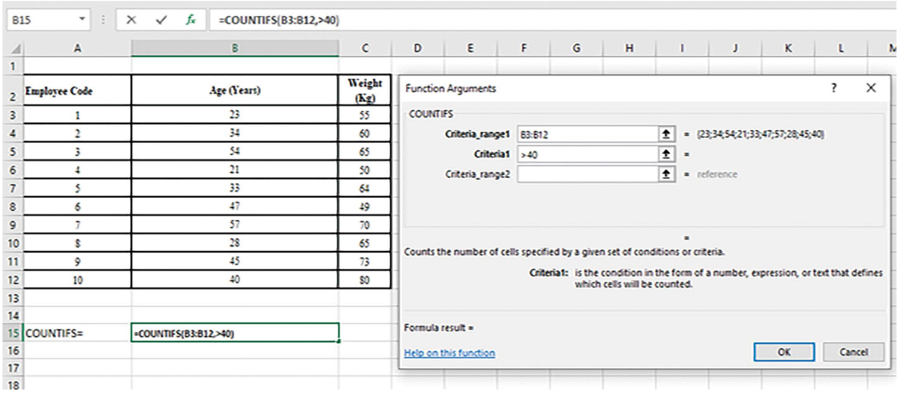
Counts the number of cells specified by a given set of conditions or criteria.

Criteria_range1: is the range of cells you want evaluated for the particular condition.

Formula result =

Help on this function

Figure 3.23 Screenshot for clicking buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ COUNTIFS



B15   =COUNTIFS(B3:B12,>40)

Employee Code	Age (Years)	Weight (Kg)
1	23	55
2	34	60
3	54	65
4	21	50
5	33	64
6	47	49
7	57	70
8	28	65
9	45	73
10	40	80

COUNTIFS= =COUNTIFS(B3:B12,>40)

Function Arguments

COUNTIFS

Criteria_range1 B3:B12 = {23;34;54;21;33;47;57;28;45;40}

Criteria1 >40 =

Criteria_range2 = reference

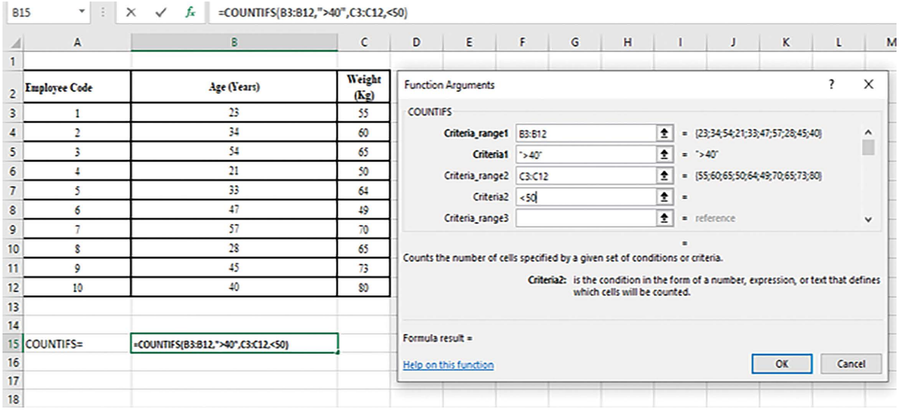
Counts the number of cells specified by a given set of conditions or criteria.

Criteria1: is the condition in the form of a number, expression, or text that defines which cells will be counted.

Formula result =

Help on this function

Figure 3.24 Screenshot after entering Cells B3:B12 in Criteria_Range1 and >40 in Criteria1 in the dropdown menu of Figure 3.23



B15   =COUNTIFS(B3:B12,">40",C3:C12,<50)

Employee Code	Age (Years)	Weight (Kg)
1	23	55
2	34	60
3	54	65
4	21	50
5	33	64
6	47	49
7	57	70
8	28	65
9	45	73
10	40	80

COUNTIFS= =COUNTIFS(B3:B12,">40",C3:C12,<50)

Function Arguments

COUNTIFS

Criteria_range1 B3:B12 = {23;34;54;21;33;47;57;28;45;40}

Criteria1 ">40" = ">40"

Criteria_range2 C3:C12 = {55;60;65;50;64;49;70;65;73;80}

Criteria2 <50 =

Criteria_range3 = reference

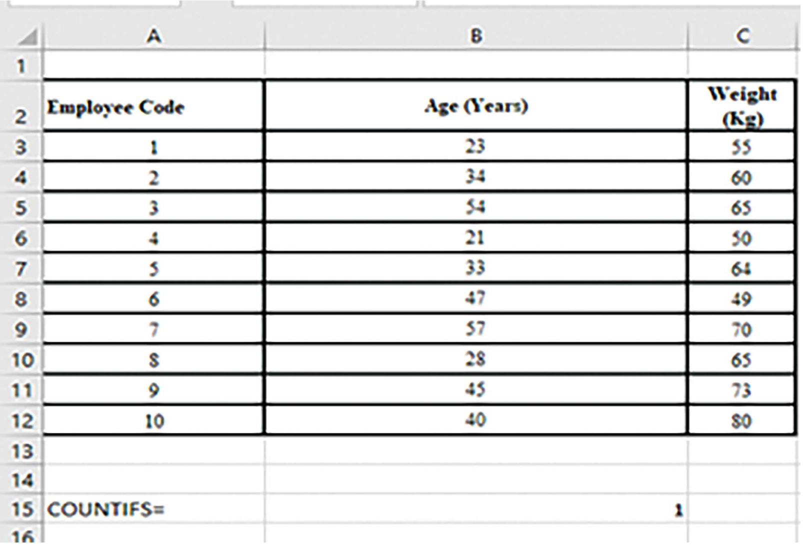
Counts the number of cells specified by a given set of conditions or criteria.

Criteria2: is the condition in the form of a number, expression, or text that defines which cells will be counted.

Formula result =

Help on this function

Figure 3.25 Screenshot after entering Cells C3:C12 in Criteria_Range2 and <50 in Criteria2 in the dropdown menu of Figure 3.24



	A	B	C
1			
2	Employee Code	Age (Years)	Weight (Kg)
3	1	23	55
4	2	34	60
5	3	54	65
6	4	21	50
7	5	33	64
8	6	47	49
9	7	57	70
10	8	28	65
11	9	45	73
12	10	40	80
13			
14			
15	COUNTIFS=	1	
16			

Figure 3.26 Results after clicking the OK button in the dropdown menu of Figure 3.25

The same result can be obtained using the following formula.

= COUNTIFS(B3 : B12, " > 40", C3 : C12, " < 50")

3.3 Frequency

The frequency of a number or a range of numbers in a given collection of data refers to how frequently they appear [4]. Think about the information in Table.3.13 about the employees' ages in a company. Table 3.14 displays the frequency distribution of these values.

The display for clicking the sequence of buttons Formula ⟹ More Functions ⟹ Statistical ⟹ FREQUENCY is shown in Figure 3.27.

In the dropdown menu of Figure 3.27, Data_array represents the array of data from which the frequencies for the elements of the data are to be formed, and Bin_array represents an array of reference to intervals to which the given data are to be grouped.

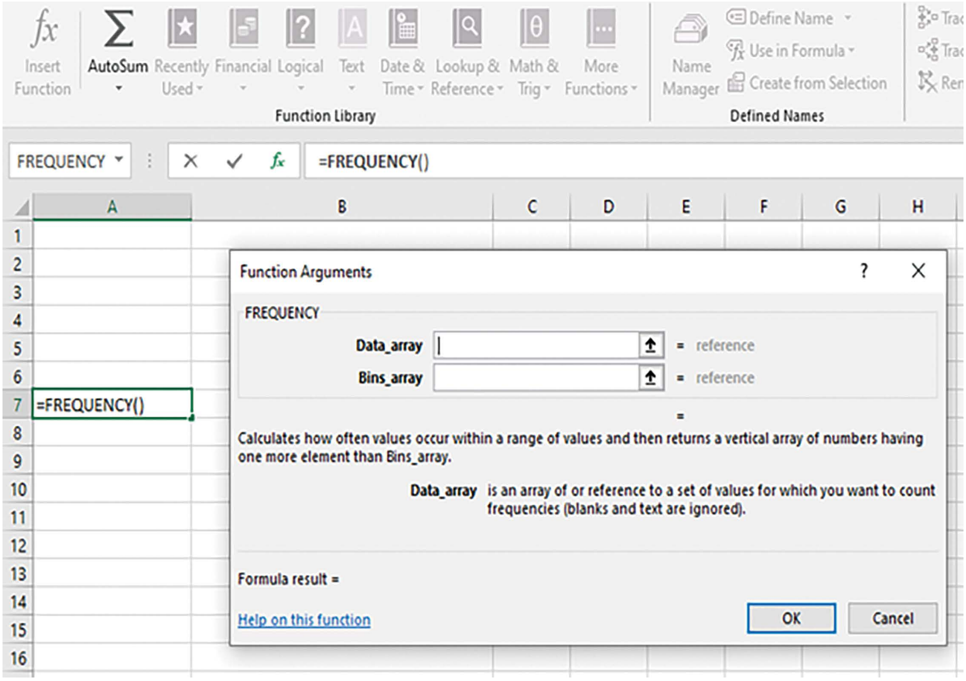
Step 1: Input the data for the problem in an Excel sheet.
Step 2: Find the minimum of the data using the MIN function [=MIN(Range of cells containing data)].
Step 3: Find the maximum of the data using the MAX function [=MAX(Range of cells containing data)].
Step 4: Decide and enter the starting and end of class intervals (Start of interval and End of interval) of the data in two different columns.
Step 5: Select the range of cells where the frequencies are to be displayed.
Step 6: Simultaneously press the SHIFT, CTRL, and ENTER keys.
Step 7: Click the sequence of buttons Formula ⟹ More Functions ⟹ Statistical ⟹ FREQUENCY.

Table 3.13 Data on Age of Employees

Employee Code	*Age in Years*
1	26
2	27
3	30
4	27
5	28
6	27
7	29
8	28
9	30
10	27

Table 3.14 Frequency Distribution of Data on Age of Employees

Age in Years	*Frequency*
26	1
27	4
28	2
29	1
30	2



=FREQUENCY()

Data_array = reference

Bins_array = reference

Calculates how often values occur within a range of values and then returns a vertical array of numbers having one more element than Bins_array.

Data_array is an array of or reference to a set of values for which you want to count frequencies (blanks and text are ignored).

Formula result =

Figure 3.27 Screenshot of display for clicking the sequence of buttons, Formula ⟹ More Functions ⟹ Statistical ⟹ FREQUENCY

Step 8: Enter the range of cells in the cell against Data_array in the dropdown menu of Step 7.
Step 9: Enter the range of cells containing the end of the intervals in the cell against Bins_array in the dropdown menu of Step 7.
Step 10: Simultaneously press the SHIFT, CTRL, and ENTER keys to give the frequencies in the range of cells selected in Step 5.

Example 3.6

The daily demand values of an item are shown in Table 3.15. Form the frequency distribution of these data by assuming a class interval of 2 using the frequency function of Excel.

Solution

The data for Example 3.6 are shown in Table 3.16.

Step 1: Input the data for the problem in an Excel sheet, as shown in Figure 3.28.
Step 2: Find the minimum of the data using the MIN function [=MIN(Range of cells containing data)], as shown in Figure 3.28.

Table 3.15 Demand Values

Day	*Demand*	*Day*	*Demand*
1	101	12	102
2	110	13	105
3	105	14	103
4	104	15	102
5	101	16	109
6	102	17	107
7	105	18	109
8	110	19	110
9	112	20	101
10	108	21	104
11	104	22	106

Table 3.16 Data for Example 3.6

Day	*Demand*	*Day*	*Demand*
1	101	12	102
2	110	13	105
3	105	14	103
4	104	15	102
5	101	16	109
6	102	17	107
7	105	18	109
8	110	19	110
9	112	20	101
10	108	21	104
11	104	22	106

Step 3: Find the maximum of the data using the MAX function [=MAX(Range of cells containing data)], as shown in Figure 3.28.

Step 4: Decide on and enter the end of class intervals (Start of interval and End of interval) of the data, as shown in Figure 3.28.

The execution of the following steps gives a display, as shown in Figure 3.29.

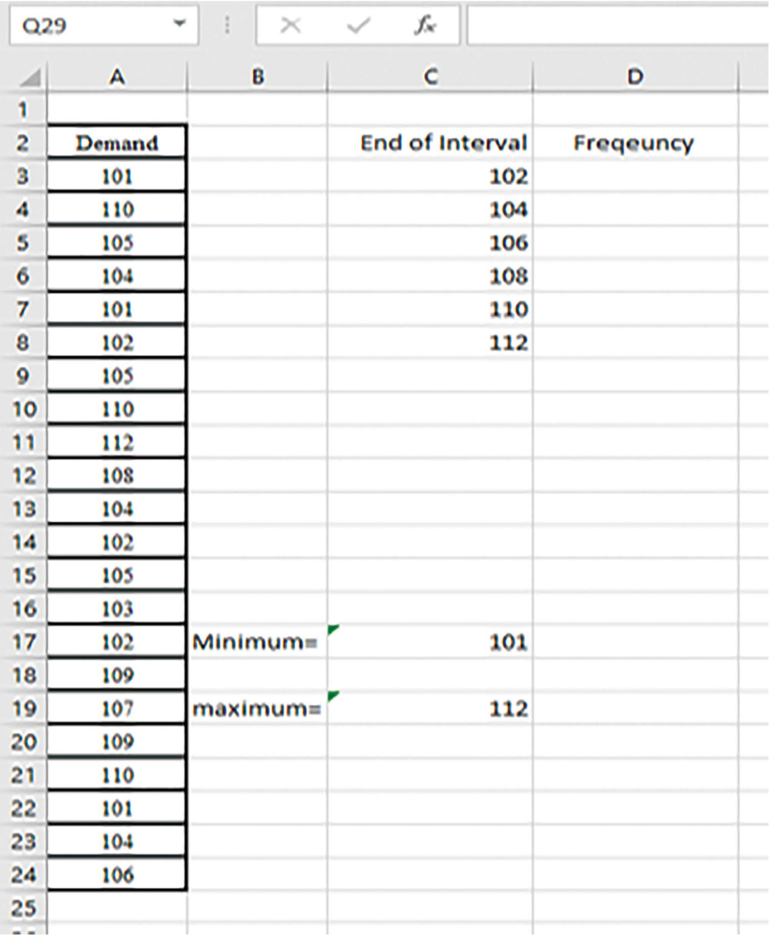
Step 5: Select the range of cells where the frequencies are to be displayed.

Step 6: Simultaneously press the SHIFT, CTRL, and ENTER keys.

Step 7: Click the sequence of buttons Formula ⟹ More Functions ⟹ Statistical ⟹ FREQUENCY.

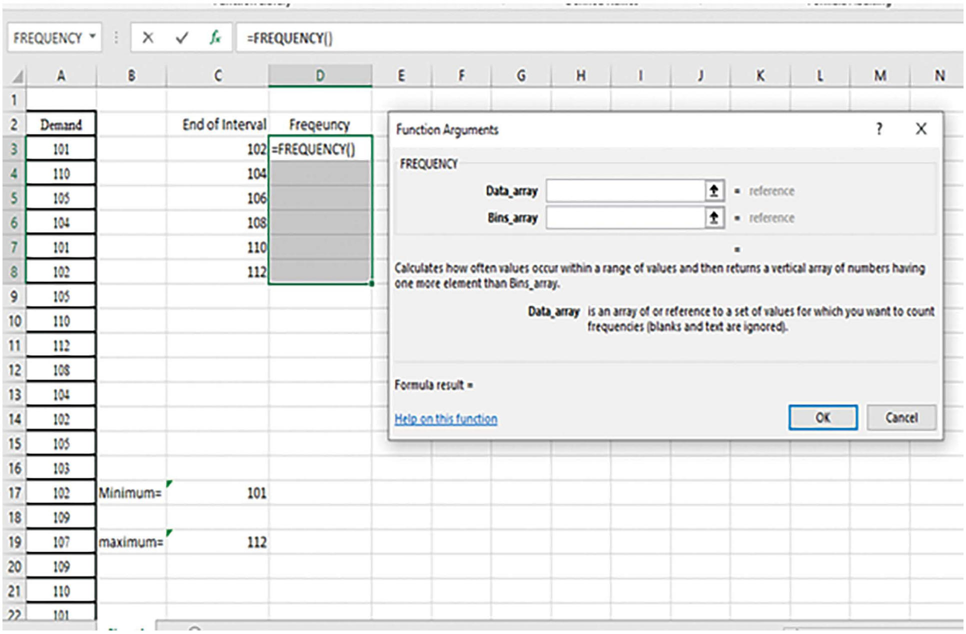
The execution of the following steps gives a display, as shown in Figure 3.30.

Step 8: Enter the range of cells in the cell against Data_array in the dropdown menu of Figure 3.29, as shown in Figure 3.30.



	A	B	C	D
1				
2	Demand		End of Interval	Freqeuncy
3	101		102	
4	110		104	
5	105		106	
6	104		108	
7	101		110	
8	102		112	
9	105			
10	110			
11	112			
12	108			
13	104			
14	102			
15	105			
16	103			
17	102	Minimum=	101	
18	109			
19	107	maximum=	112	
20	109			
21	110			
22	101			
23	104			
24	106			
25				

Figure 3.28 Screenshot of input of given data along with the upper limits of class intervals whose width is 2



FREQUENCY | =FREQUENCY()

	A	B	C	D
1				
2	Demand		End of Interval	Freqeuncy
3	101		102	=FREQUENCY()
4	110		104	
5	105		106	
6	104		108	
7	101		110	
8	102		112	
9	105			
10	110			
11	112			
12	108			
13	104			
14	102			
15	105			
16	103			
17	102	Minimum=	101	
18	109			
19	107	maximum=	112	
20	109			
21	110			
22	101			

Function Arguments

FREQUENCY

Data_array = reference

Bins_array = reference

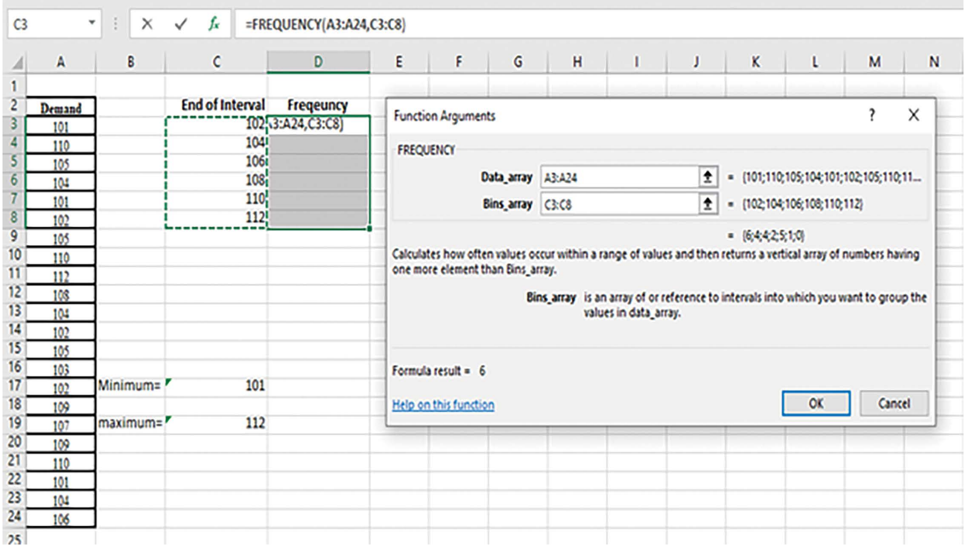
Calculates how often values occur within a range of values and then returns a vertical array of numbers having one more element than Bins_array.

Data_array is an array of or reference to a set of values for which you want to count frequencies (blanks and text are ignored).

Formula result =

Help on this function

Figure 3.29 Screenshot of display after executing Steps 5, 6, and 7



C3 | =FREQUENCY(A3:A24,C3:C8)

	A	B	C	D
1				
2	Demand		End of Interval	Freqeuncy
3	101		102	3:A24,C3:C8)
4	110		104	
5	105		106	
6	104		108	
7	101		110	
8	102		112	
9	105			
10	110			
11	112			
12	108			
13	104			
14	102			
15	105			
16	103			
17	102	Minimum=	101	
18	109			
19	107	maximum=	112	
20	109			
21	110			
22	101			
23	104			
24	106			
25				

Function Arguments

FREQUENCY

Data_array A3:A24 = {101;110;105;104;101;102;105;110;11...

Bins_array C3:C8 = {102;104;106;108;110;112}

= {6;4;4;2;5;1;0}

Calculates how often values occur within a range of values and then returns a vertical array of numbers having one more element than Bins_array.

Bins_array is an array of or reference to intervals into which you want to group the values in data_array.

Formula result = 6

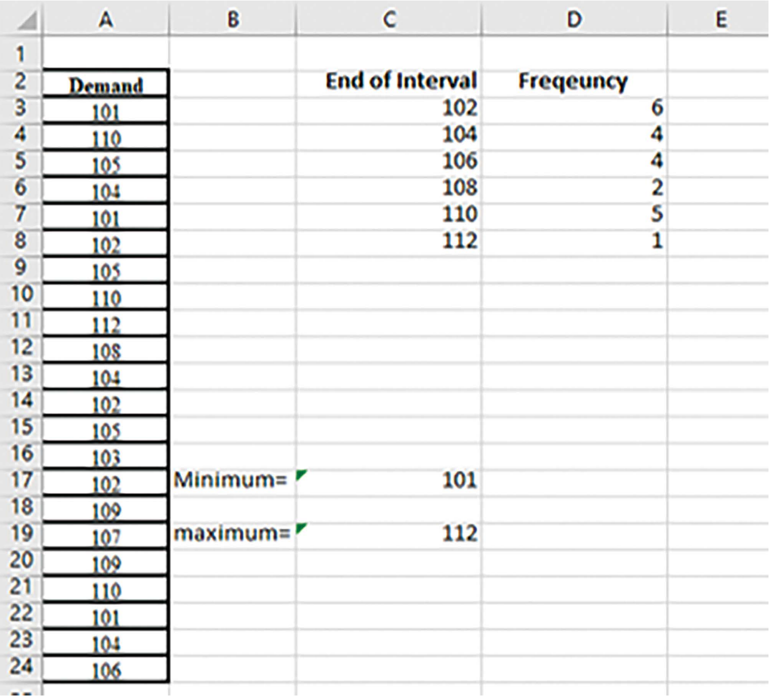
Help on this function

Figure 3.30 Screenshot of display after executing Steps 8 and 9

Step 9: Enter the range of cells containing the end of the intervals in the cell against Bins_array in the dropdown menu of Figure 3.29, as shown in Figure 3.30.

The execution of the following step gives a display as shown in Figure 3.31.

Step 10: Simultaneously press the SHIFT, CTRL, and ENTER keys to give the frequencies in the range of cells selected, as shown in Figure 3.31.



	A	B	C	D	E
1					
2	Demand		End of Interval	Freqeuncy	
3	101		102	6	
4	110		104	4	
5	105		106	4	
6	104		108	2	
7	101		110	5	
8	102		112	1	
9	105				
10	110				
11	112				
12	108				
13	104				
14	102				
15	105				
16	103				
17	102	Minimum=	101		
18	109				
19	107	maximum=	112		
20	109				
21	110				
22	101				
23	104				
24	106				

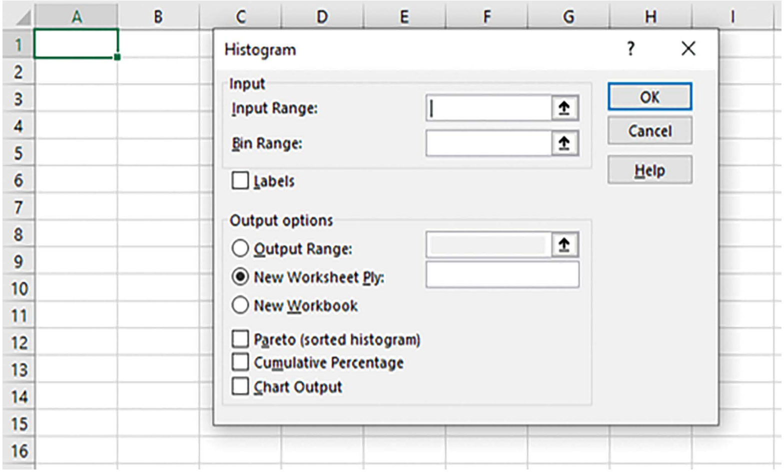
Figure 3.31 Screenshot of display after execution Step 10 to give the frequency distribution in cells D3:D8

3.4 Histograms

Similar to the frequency function is the histogram function. In a given set of data, this determines the frequency of the observations. After selecting the Data and Data Analysis buttons, this becomes available [5].

The steps of using the Histogram function in Excel are as follows.

Step1: Enter the data in a range of cells.
Step 2: Find the minimum of the data using the MIN function.
Step 3: Find the maximum of the data using the MAX function.
Step 4: Decide on and enter the start and end of the intervals in another range of cells with proper column headings.
Step 5: Click the Data button, the Data Analysis button, and then Histogram in the dropdown menu. The display after this is shown in Figure 3.32.
Step 6: Enter the range of cells containing the data including the column heading in the box against, "Input Range" in the dropdown menu of Figure 3.32.
Step 7: Enter the range of cells containing the end of intervals including the column heading in the box against, "Bin Range" in the dropdown menu of Figure 3.32.
Step 8: Click the checkbox for the labels.
Step 9: Click the Output Range and enter the range of cells where the output is to be displayed with a provision for headings.
Step 10: Click the OK button in the dropdown menu of Figure 3.32.



Histogram

Input

Input Range:

Bin Range:

□ Labels

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

□ Chart Output

Figure 3.32 Screenshot of display after clicking buttons, Data ⟹ Data Analysis ⟹ Histogram

Step 11: Select the output range, click the Insert button in the ribbon at the top, and click the Column button under Charts in the ribbon.

Step 12: Click the desired bar chart icon to give the bar chart.

Example 3.7

The age distribution of the employees in a company is shown in Table 3.17. Form the frequency distribution using the Histogram function and then draw a bar chart for the frequencies.

Solution

The data for Example 3.7 are shown in Table 3.18.

Input the age distribution of the employees along with start of interval and end of interval data in an Excel sheet as shown in Figure 3.33 using the following steps.

Step1: Enter the data in a range of cells.

Step 2: Find the minimum of the data using the MIN function.

Step 3: Find the maximum of the data using the MAX function.

Step 4: Decide on and enter the start and end of the intervals in another range of cells with proper column headings.

The button clicks Data in the top Ribbon ⟹ Data Analysis in the Top Ribbon give a display, as in Figure 3.34. Clicking the Histogram button in the dropdown menu of Figure 3.34 gives a display, as in Figure 3.35. In Figure 3.35, perform Steps 6 to 9 to show the display as in Figure 3.36.

Step 6: Enter the range of cells containing the data including the column heading in the box against, “Input Range” in the dropdown menu of Figure 3.35.

Table 3.17 Age Distribution of Employees

Employee Code	*Age in Years*
1	26
2	27
3	30
4	27
5	28
6	27
7	25
8	28
9	30
10	27
11	32
12	34
13	31
14	38
15	39
16	40
17	33
18	34
19	45
20	37

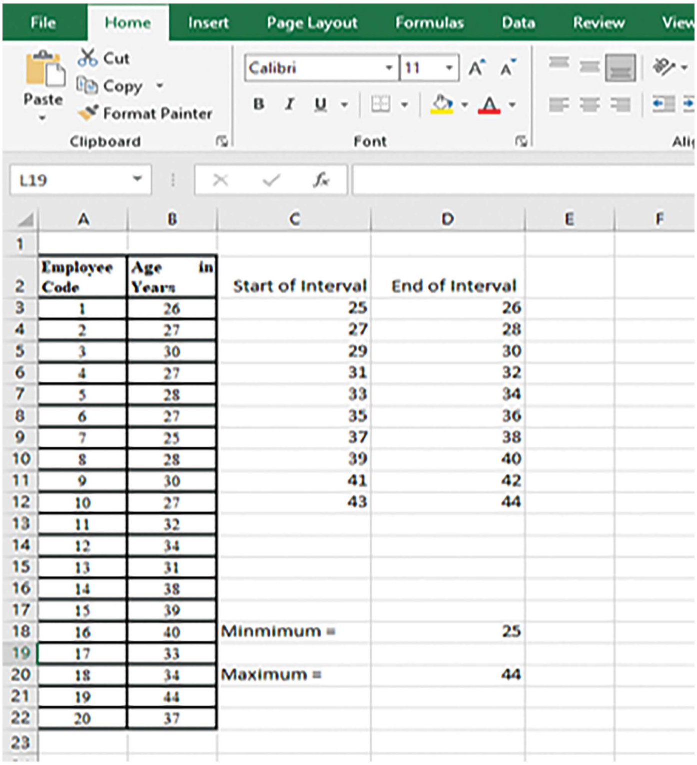
Table 3.18 Data for Example 3.7

Employee Code	*Age in Years*
1	26
2	27
3	30
4	27
5	28
6	27
7	25
8	28
9	30
10	27
11	32
12	34
13	31
14	38
15	39
16	40
17	33
18	34
19	45
20	37

Step 7: Enter the range of cells containing end of intervals including the column heading in the box against, "Bin Range" in the dropdown menu of Figure 3.35.

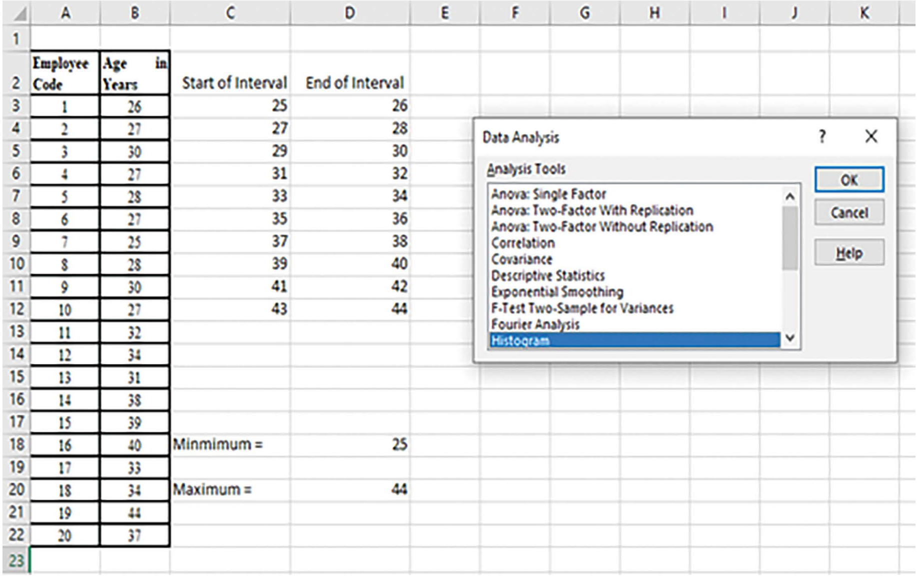
Step 8: Click the checkbox for the labels.

Step 9: Click the Output Range and enter the range of cells where the output is to be displayed. Here, include one extra cell for the frequency with regard to more than the end value of the last class interval.



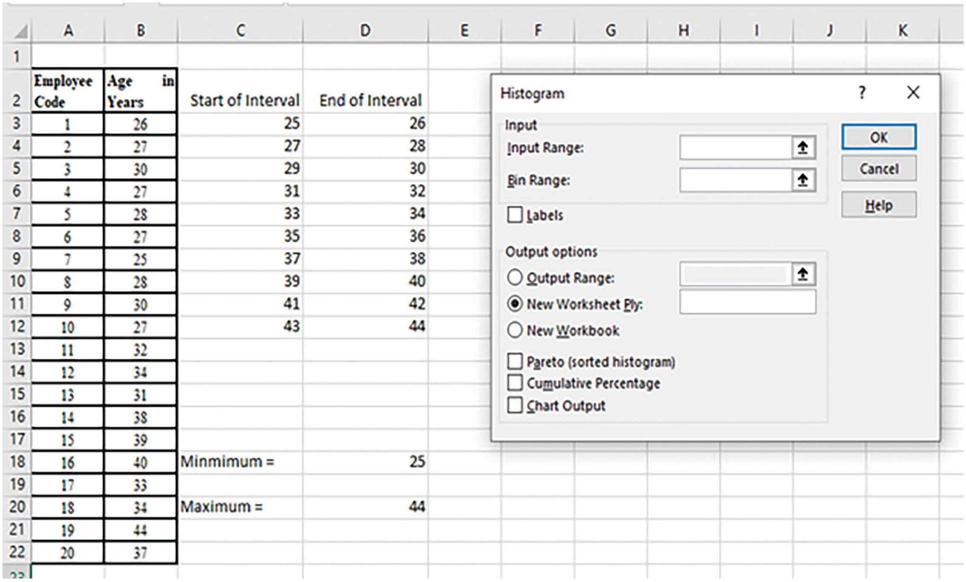
	A	B	C	D	E	F
1						
2	Employee Code	Age in Years	Start of Interval	End of Interval		
3	1	26	25	26		
4	2	27	27	28		
5	3	30	29	30		
6	4	27	31	32		
7	5	28	33	34		
8	6	27	35	36		
9	7	25	37	38		
10	8	28	39	40		
11	9	30	41	42		
12	10	27	43	44		
13	11	32				
14	12	34				
15	13	31				
16	14	38				
17	15	39				
18	16	40	Minmimum =	25		
19	17	33				
20	18	34	Maximum =	44		
21	19	44				
22	20	37				
23						

Figure 3.33 Screenshot of input of age distribution of employees along with the start and end of interval data



	A	B	C	D	E	F	G	H	I	J	K
1											
2	Employee Code	Age in Years	Start of Interval	End of Interval							
3	1	26	25	26							
4	2	27	27	28							
5	3	30	29	30							
6	4	27	31	32							
7	5	28	33	34							
8	6	27	35	36							
9	7	25	37	38							
10	8	28	39	40							
11	9	30	41	42							
12	10	27	43	44							
13	11	32									
14	12	34									
15	13	31									
16	14	38									
17	15	39									
18	16	40	Minmimum =	25							
19	17	33									
20	18	34	Maximum =	44							
21	19	44									
22	20	37									
23											

Figure 3.34 Screenshot of display of button clicks, Data in the Top Ribbon ⟹ Data Analysis in the Top Ribbon



	A	B	C	D
1				
2	Employee Code	Age in Years	Start of Interval	End of Interval
3	1	26	25	26
4	2	27	27	28
5	3	30	29	30
6	4	27	31	32
7	5	28	33	34
8	6	27	35	36
9	7	25	37	38
10	8	28	39	40
11	9	30	41	42
12	10	27	43	44
13	11	32		
14	12	34		
15	13	31		
16	14	38		
17	15	39		
18	16	40	Minmimum =	25
19	17	33		
20	18	34	Maximum =	44
21	19	44		
22	20	37		

Histogram

Input

Input Range:

Bin Range:

☐ Labels

Output options

○ Output Range:

● New Worksheet Ply:

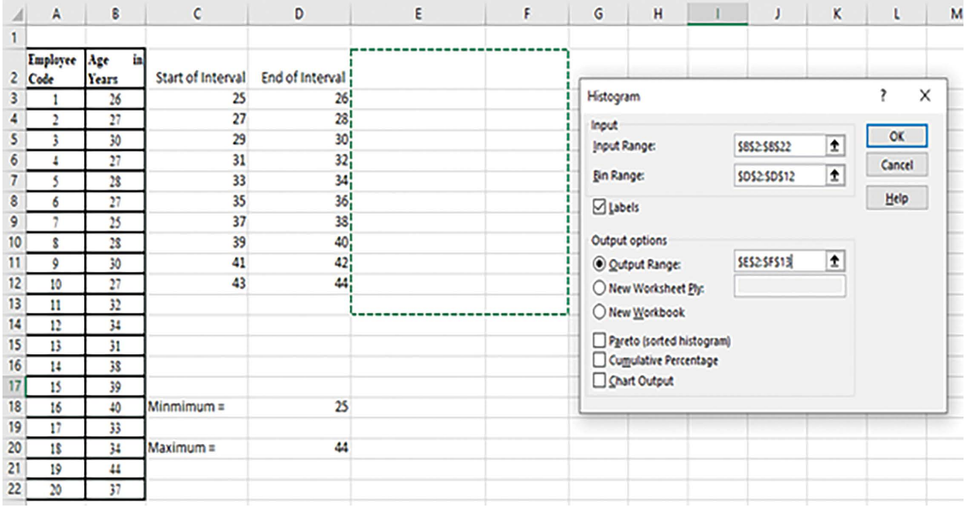
○ New Workbook

☐ Pareto (sorted histogram)

☐ Cumulative Percentage

☐ Chart Output

Figure 3.35 Screenshot of display of Histogram in the resultant dropdown menu of Figure 3.34



	A	B	C	D
1				
2	Employee Code	Age in Years	Start of Interval	End of Interval
3	1	26	25	26
4	2	27	27	28
5	3	30	29	30
6	4	27	31	32
7	5	28	33	34
8	6	27	35	36
9	7	25	37	38
10	8	28	39	40
11	9	30	41	42
12	10	27	43	44
13	11	32		
14	12	34		
15	13	31		
16	14	38		
17	15	39		
18	16	40	Minmimum =	25
19	17	33		
20	18	34	Maximum =	44
21	19	44		
22	20	37		

Histogram

Input

Input Range: $B$2:$B$22

Bin Range: $D$2:$D$12

☑ Labels

Output options

● Output Range: $E$2:$F$13

○ New Worksheet Ply:

○ New Workbook

☐ Pareto (sorted histogram)

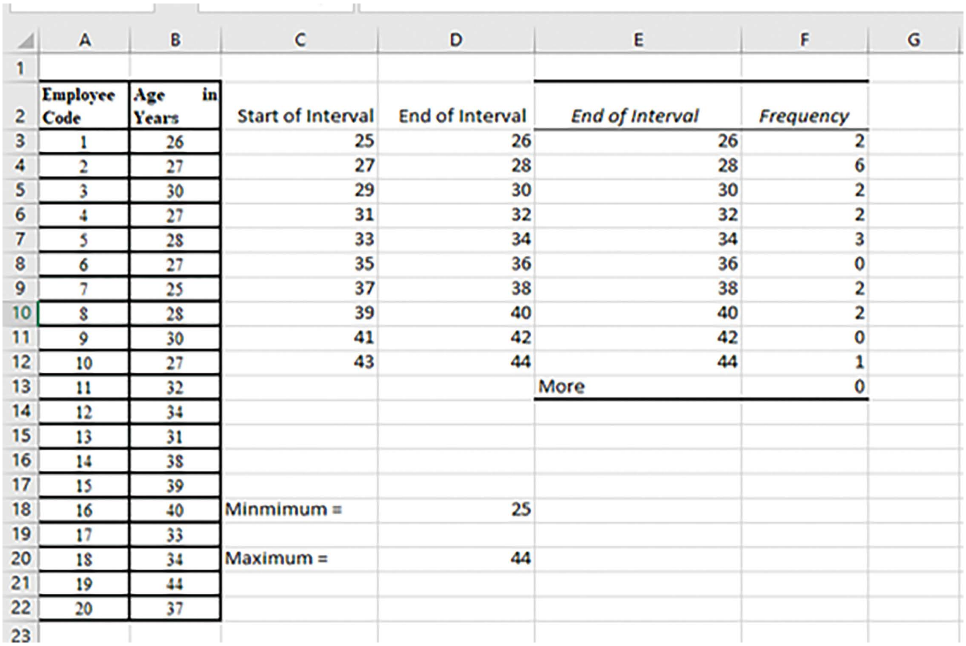
☐ Cumulative Percentage

☐ Chart Output

Figure 3.36 Screenshot after executing Steps from 6 to 9

Step 10: Clicking the OK button in the dropdown menu of Figure 3.36 gives the result of the frequencies along with the end of the intervals in the specified output range, as shown in Figure 3.37.

The execution of the following steps gives a display, as in Figure 3.38, which contains a bar chart for the frequencies shown in Figure 3.37.



	A	B	C	D	E	F	G
1							
2	Employee Code	Age in Years	Start of Interval	End of Interval	*End of Interval*	*Frequency*	
3	1	26	25	26	26	2	
4	2	27	27	28	28	6	
5	3	30	29	30	30	2	
6	4	27	31	32	32	2	
7	5	28	33	34	34	3	
8	6	27	35	36	36	0	
9	7	25	37	38	38	2	
10	8	28	39	40	40	2	
11	9	30	41	42	42	0	
12	10	27	43	44	44	1	
13	11	32			More	0	
14	12	34					
15	13	31					
16	14	38					
17	15	39					
18	16	40	Minmimum =	25			
19	17	33					
20	18	34	Maximum =	44			
21	19	44					
22	20	37					
23							

Figure 3.37 Screenshot after clicking the OK button in the dropdown menu of Figure 3.36

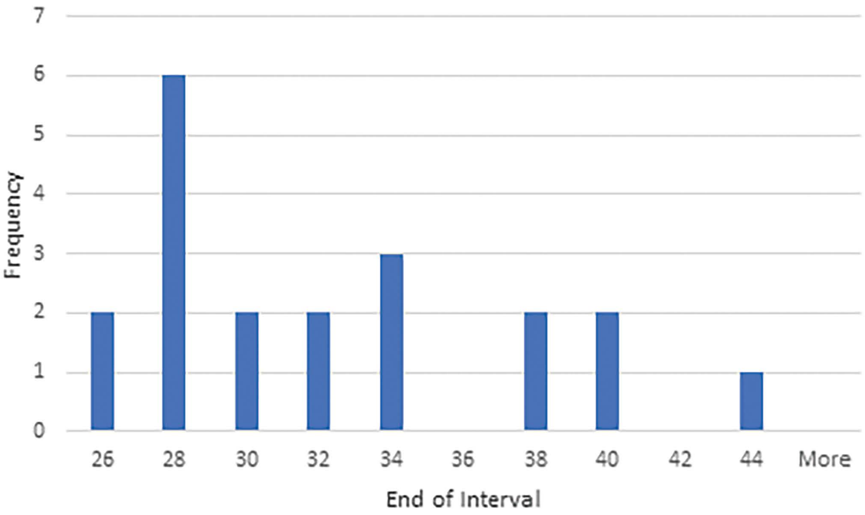


Figure 3.38 Screenshot of bar chart for the frequencies obtained in Figure 3.37

Step 11: Select the output range, that is, End of Interval and Frequency, including the headings in Column E and Column F, respectively, in Figure 3.37, and then click the Insert button in the ribbon at the top and the Column button under Charts in the ribbon.

Step 12: Click the desired bar chart icon (first at the top) to show the bar chart as shown in Figure 3.38, in which the *X* axis and *Y* axis are labelled.

Summary

- The COUNT function counts the number of cells in a given range of cells of an Excel sheet that contains numbers.
- The Excel command for the COUNT function is:

$$= \text{COUNT}\left(\text{Range of cells containing data}\right)$$

- The COUNTA function determines the number of cells that are not empty in a given range of cells.
- The formula to find the value of COUNTA is:

$$= \text{COUNTA}\left(\text{range of cells to be considered}\right)$$

- The COUNTBLANK function determines the number of cells that are blank (empty) in a given range of cells.
- The formula to find the value of COUNTBLANK is:

$$= \text{COUNTBLANK}\left(\text{Range of cells to be considridred}\right)$$

- The COUNTIF function finds the number of cells within a given range of cells satisfying a given criterion.
- The COUNTIFS function is similar to the COUNTIF function, with the difference that it has more than one criterion.
- The formula to obtain the result of the COUNTIF function is:

$$= \text{COUNTIF}\begin{pmatrix}\text{Range of cells containing data, Criterion used}\\ \text{to count the cells in the specified range}\end{pmatrix}$$

- The formula to obtain the result of the COUNTIFS function is: =COUNTIFS(Criteria_Range1, Criteria1, Criteria_Range2, Criteria2, etc.)
- Frequency is the number of occurrences of a number or a range of numbers in a given set of data.
- The sequence of button clicks for the FREQUENCY function is Formula ⟹ More Functions ⟹ Statistical ⟹ FREQUENCY
- Enter the range of cells in the cell against Data_array in the dropdown menu of FREQUENCY function, enter the range of cells containing the end of the intervals in the cell against Bins_array in the same dropdown menu, and at the end simultaneously press the SHIFT, CTRL, and ENTER keys to find the frequencies in the selected range of cells.
- The Histogram function is also like the frequency function, which finds the frequency of the observations in a given set of data. This is available after clicking the Data and Data Analysis buttons.

Keywords

- The COUNT function counts the number of cells in a given range of cells of an Excel sheet that contains numbers.
- The COUNTA function determines the number of cells that are not empty in a given range of cells.
- The COUNTBLANK function determines the number of cells that are blank (empty) in a given range of cells.

- COUNTIF function finds the number of cells within a given range of cells satisfying a given criterion.
- The COUNTIFS function is similar to the COUNTIF function, with the difference that it has more than one criterion.
- Frequency is the number of occurrences of a number or a range of numbers in a given set of data.
- The Histogram function is also like the frequency function, which finds the frequency of the observations in a given set of data. This is available after clicking the Data and Data Analysis buttons.

Review Questions

1. a. What is purpose of the COUNT function in Excel?
   b. List the sequence of button clicks before clicking the COUNT button.
   c. Give the formula for the COUNT function.
2. Consider the data given in the following table. In the second row of the table, the empty cells represent absent status of the candidates for the test. Write the formula to COUNT the cells with test marks in the following table.

Roll Number	01	02	03	04	05	06	07	08	09	10
Test Mark Max:25	20	22		18	23		24	25	17	16

3. Illustrate the COUNTBLANK function in Excel using your own example.
4. a. Write the syntax of the COUNTIF function in Excel.
   b. Consider the data given in the following table about the earnings per share (EPS) of companies.
   Write the formula for the COUNTIF function in Excel to obtain the number of companies with an EPS higher than 50.

Company Code	1011	102	103	104	105	106	107	108	109	110
EPS	20	70	30	40	22	60	120	25	35	65

5. a. Write the syntax of the COUNTIFS function in Excel.
   b. Consider the internal and external marks of the following candidates. Answer the following questions using Excel.
   i. Find the total of the internal and external marks of each of the candidates.
   ii. Find the count for the number of candidates satisfying the following conditions using COUTNTIFS function.
   "The total mark is greater than or equal to 50 and the external mark is greater than or equal to 25".

Reg. No.	*Name*	*Internal Mark*	*External Mark*	*Total*
		Max. 50	*Max. 50*	*Max. 100*
101	Anand, E	30	35	
102	Babu, N	25	26	
104	Chitra, K	14	40	

Reg. No.	*Name*	*Internal Mark*	*External Mark*	*Total*
		Max. 50	*Max. 50*	*Max. 100*
108	Domnic, D	35	20	
110	Fathima, S	40	45	
115	Mohan, S	34	24	
123	Nandhini, S	45	42	
135	Peter, H	30	40	
140	Rabinson, T	45	48	
145	Sunil, K	42	40	

6. a. Give the syntax of the FREQUENCY function in Excel.

   b. Give the applications of the FREQUENCY function.

7. Give the steps of obtaining the frequency of a given set of interval data using Excel.
8. The daily prices of a share are shown in the following table. Form the frequency distribution of these data by assuming a class interval of 2 using the frequency function in Excel.

Day	*Price (₹)*	*Day*	*Price (₹)*
1	500	12	650
2	550	13	675
3	575	14	700
4	500	15	600
5	600	16	725
6	650	17	700
7	625	18	675
8	700	19	600
9	750	20	725
10	525	21	800
11	650	22	725

9. Give the steps of constructing a histogram for the given interval data.

Employee Code	*Monthly Income in Thousands of ₹*
1	30
2	31
3	34
4	31
5	32
6	31
7	29
8	32
9	34
10	31
11	36
12	38
13	35
14	42
15	43
16	44
17	37

Employee Code	*Monthly Income in Thousands of ₹*
18	38
19	49
20	41

10. The monthly income distribution of the employees in a company is shown in the previous table. Form the frequency distribution using the Histogram function, and then draw a bar chart for the frequencies.
11. The monthly income of 40 employees in an industrial park ranges from ₹ 25,000 to ₹ 1,250,000. Create a histogram using Excel to illustrate the frequencies of the employees' monthly income with a class interval of ₹ 25,000 by assuming appropriate data (monthly income) for this problem.

References

1. https://support.microsoft.com/en-us/office/countif-function-e0de10c6-f885-4e71-abb4-1f464816df34 [July 2, 2020].
2. https://support.microsoft.com/en-us/office/countblank-function-6a92d772-675c-4bee-b346-24af6bd3ac22 [July 2, 2020].
3. www.Excel-easy.com/examples/frequency.html [June 25, 2020].
4. www.Exceluser.com/formulas/countifs-frequency-distributions.htm [June, 25, 2020].
5. www.Excel-easy.com/examples/histogram.html [July 2, 2020].

4 Average Functions

Learning Objectives

The study of this chapter will enable readers to:

- Understand the syntax of the AVERAGE function with an illustration.
- Find the difference between the AVERAGE function and the AVEDEV function.
- Analyse the use of the AVERAGEA function for business applications
- Apply the AVERAGEIFS function with more than one criterion.
- Analyse the weighted average of a given set of data.
- Understand the computation of the mean of grouped data.

4.1 Introduction

Finding the average, variance, standard deviation, and numerous other measurements are just a few of the many procedures involved in the data processing endeavour [1]. The various functions linked to the average of a given set of data are presented in this chapter.

The different average functions are listed in the following.

1. AVERAGE
2. AVEDEV
3. AVERAGEA
4. AVERAGEIF
5. AVERAGEIFS

4.2 AVERAGE Function

The AVERAGE function determines the arithmetic mean of a set of observations, which is computed using the following formula [2].

$$AVERAGE = \frac{\sum_{i=1}^{n} X_i}{n}$$

where
n is the total number of observations
X_i is the i^{th} observation, where i = 1, 2, 3, . . ., n

DOI: 10.4324/9781032671703-4

Clicking the following sequence of buttons in Excel gives the screenshot in Figure 4.1.

Formulas ⇨ More Functions ⇨ Statistical

In Figure 4.1, clicking the option AVERAGE gives the screenshot in Figure 4.2. Now, one should enter the cells one after another which contain the data against Number 1, Number 2, Number 3, and so on. If the data are arranged in a range of cells continuously, then that range of cells is to be entered against Number 1. After defining the addresses of the data, clicking the OK button in the dropdown menu of Figure 4.2 will display the average of the data in the desired cell where the cursor has been positioned before the start of this procedure. Clicking OK will return the average of its arguments, which can be numbers, names, arrays, or references that contain numbers. Number 1, Number 2, Number 3, and so on can go up to Number 255.

Instead of following the menu-driven option, one can enter a formula in a desired cell, say, A11, to get the average of a given set of observations as follows.

= AVERAGE(Range of cells containing data)

If the data are stored from cell A1 to cell A10, then the formula at cell A11 to compute their average is as follows.

= AVERAGE(A1 : A10)

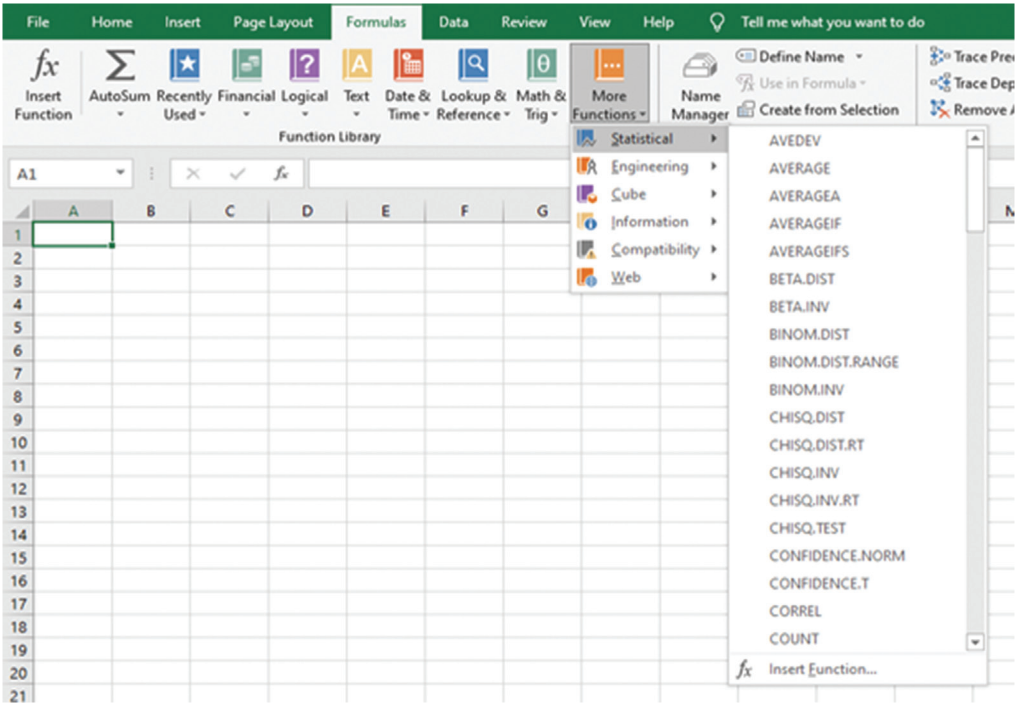
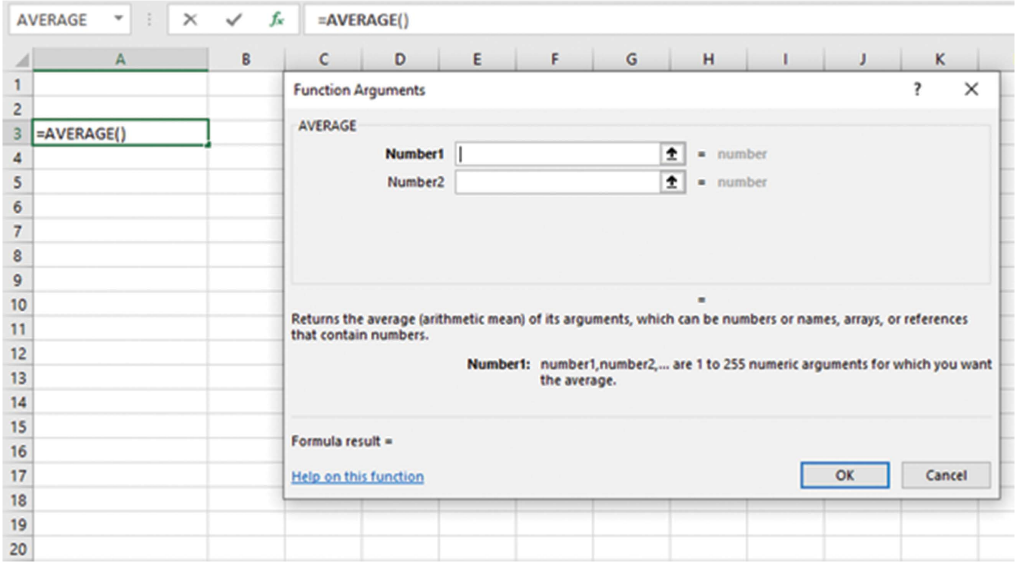


Figure 4.1 Screenshot of sequence of button clicks, Formulas ⇨ More Formulas ⇨ Statistical



=AVERAGE()

Function Arguments

AVERAGE

Number1 = number

Number2 = number

Returns the average (arithmetic mean) of its arguments, which can be numbers or names, arrays, or references that contain numbers.

Number1: number1,number2,... are 1 to 255 numeric arguments for which you want the average.

Formula result =

Figure 4.2 Screenshot of click, Formulas ⟹ More Functions ⟹ Statistical and AVERAGE

Out of the range of data from cell A1 to cell A10, if the average of the data at cells A1, A5, and A9 is required, then the corresponding formula is as follows.

$$= \text{AVERAGE}(\text{A1, A5, A9})$$

Example 4.1

The demand values in thousands of units of a product for the past eight years are summarised in Table 4.1.

1. Find the mean demand of the product using the AVERAGE function.
2. Find the mean of the demand values of odd years.

Solution

The data for Example 4.1 are shown in Table 4.2.

1. Compute the mean demand using the AVERAGE function.

   Step 1: A screenshot after inputting the data into the Excel sheet is shown in Figure 4.3.

   Step 2: A screenshot after clicking the sequence of buttons Formulas ⟹ More Function ⟹ Statistical gives the display shown in Figure 4.4.

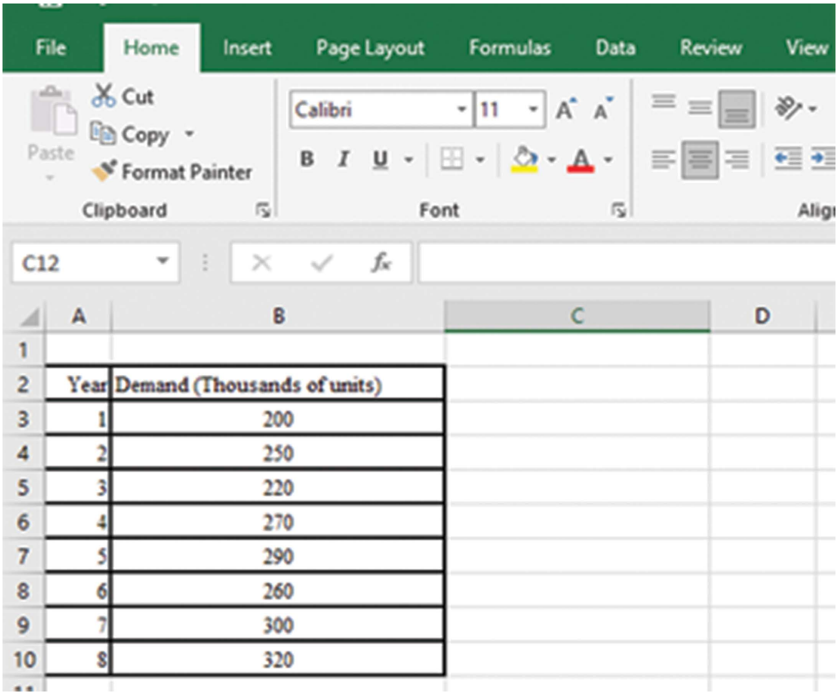
   Step 3: Click the AVERAGE function in the dropdown menu of Figure 4.4, which gives the display shown in Figure 4.5.

Table 4.1 Demand Values

Year	*Demand (Thousands of Units)*
1	200
2	250
3	220
4	270
5	290
6	260
7	300
8	320

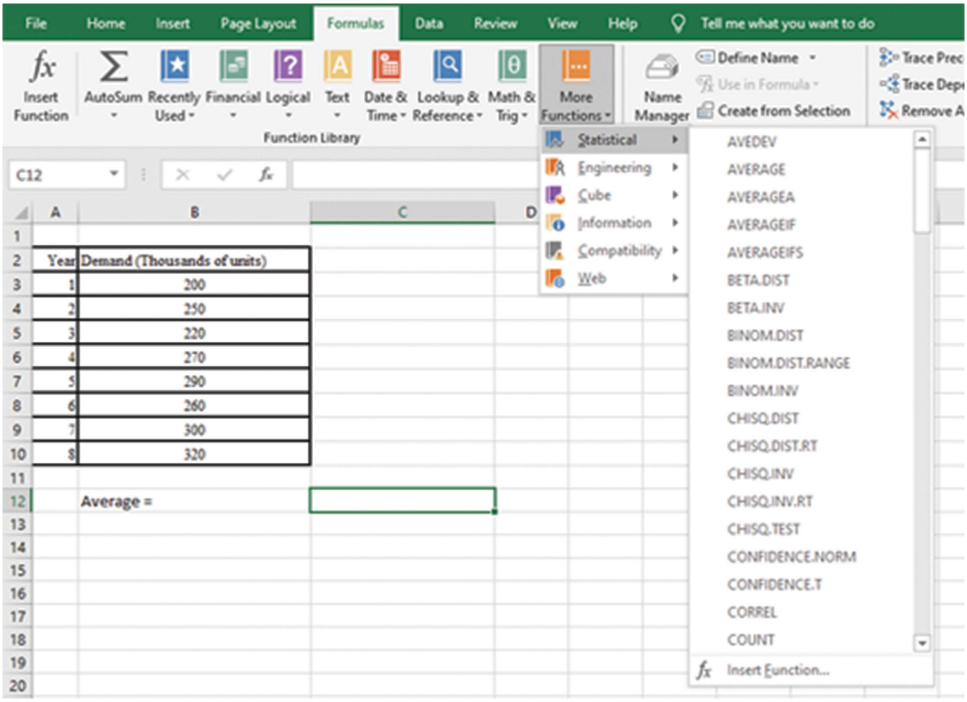
Table 4.2 Data for Example 4.1

Year	*Demand (Thousands of Units)*
1	200
2	250
3	220
4	270
5	290
6	260
7	300
8	320



Year	Demand (Thousands of units)
1	200
2	250
3	220
4	270
5	290
6	260
7	300
8	320

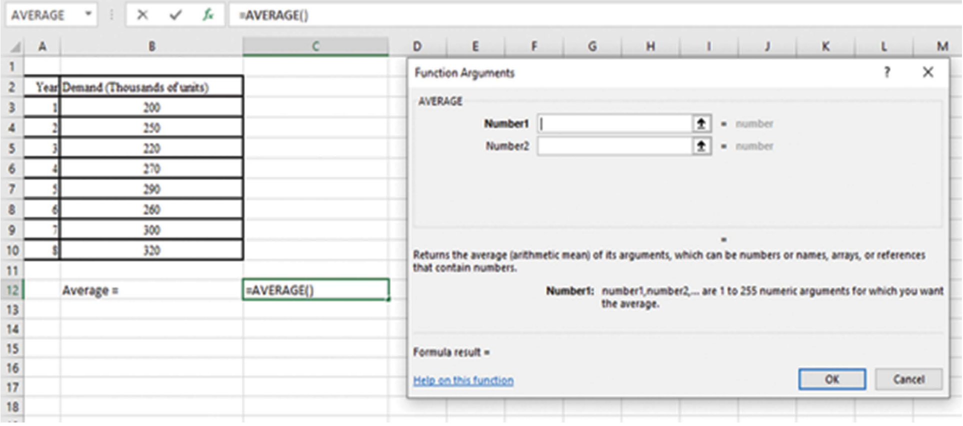
Figure 4.3 Screenshot after inputting data in the Excel sheet



Year	Demand (Thousands of units)
1	200
2	250
3	220
4	270
5	290
6	260
7	300
8	320

Average =

Figure 4.4 Screenshot after clicking the sequence of buttons, Formulas ⟹ More Functions ⟹ Statistical



Year	Demand (Thousands of units)
1	200
2	250
3	220
4	270
5	290
6	260
7	300
8	320

Average = =AVERAGE()

AVERAGE

Number1 = number

Number2 = number

=

Returns the average (arithmetic mean) of its arguments, which can be numbers or names, arrays, or references that contain numbers.

Number1: number1,number2,... are 1 to 255 numeric arguments for which you want the average.

Formula result =

Help on this function

Figure 4.5 Screenshot after clicking the AVERAGE function in the dropdown menu of Figure 4.4

Step 4: From Figure 4.5, it is clear that one can enter the series of data one after another in the cells Number 1, Number 2, . . . Number 255. The arguments can be numbers, names, arrays, or references that contain numbers.

Alternatively, one can enter the range of cells containing data against Number 1, as shown in Figure 4.6, which will give the result for the mean in the cell where the cursor is already positioned.

In Figure 4.6, the cursor is positioned in cell C12 to view the formula simultaneously.

Step 5: Click OK in the dropdown menu shown in Figure 4.6 to give the result of the AVERAGE function, as shown in Figure 4.7.

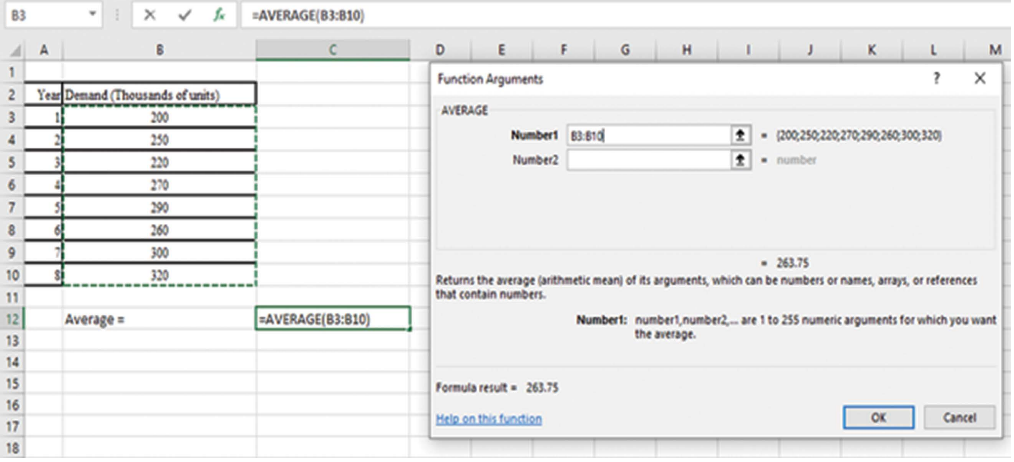
The mean demand of the data in the range of cells from B3:B10 is 263.75 thousand units.

All these steps can be combined into a formula-type command as follows.

$$= \text{AVERAGE}(\text{B3} : \text{B10})$$

This formula may be entered in any convenient cell, say, C12, to give the result.

2. Mean of the demand values of odd years.



B3 =AVERAGE(B3:B10)

Year	Demand (Thousands of units)
1	200
2	250
3	220
4	270
5	290
6	260
7	300
8	320

Average = =AVERAGE(B3:B10)

Function Arguments

AVERAGE

Number1 B3:B10 = {200;250;220;270;290;260;300;320}

Number2 = number

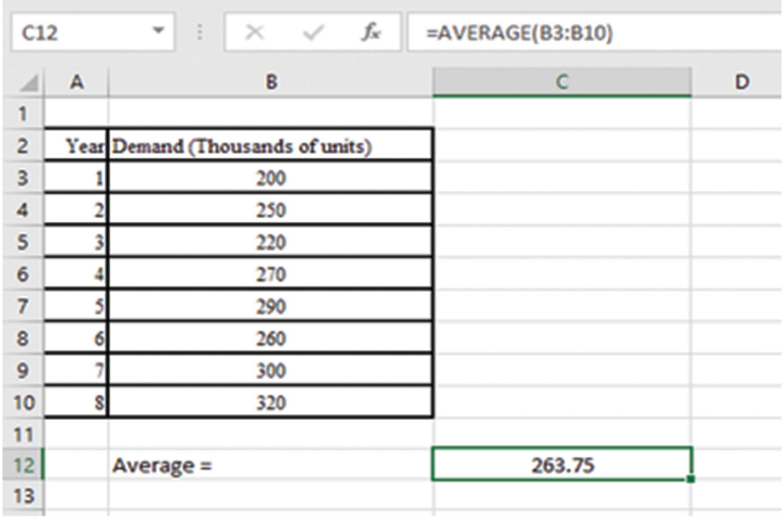
= 263.75

Returns the average (arithmetic mean) of its arguments, which can be numbers or names, arrays, or references that contain numbers.

Number1: number1,number2,... are 1 to 255 numeric arguments for which you want the average.

Formula result = 263.75

Figure 4.6 Screenshot after entering the range of cells B3:B10 in the data box against Number 1

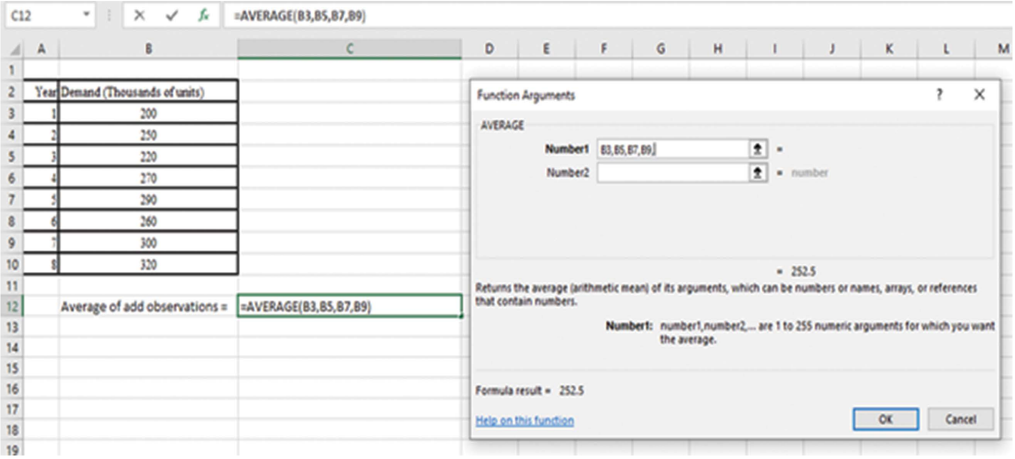


C12 =AVERAGE(B3:B10)

Year	Demand (Thousands of units)
1	200
2	250
3	220
4	270
5	290
6	260
7	300
8	320

Average = 263.75

Figure 4.7 Screenshot after clicking OK in the dropdown menu shown in Figure 4.6



C12

=AVERAGE(B3,B5,B7,B9)

Year	Demand (Thousands of units)
1	200
2	250
3	220
4	270
5	290
6	260
7	300
8	320

Average of add observations = =AVERAGE(B3,B5,B7,B9)

Function Arguments

AVERAGE

Number1 B3,B5,B7,B9 =

Number2 = number

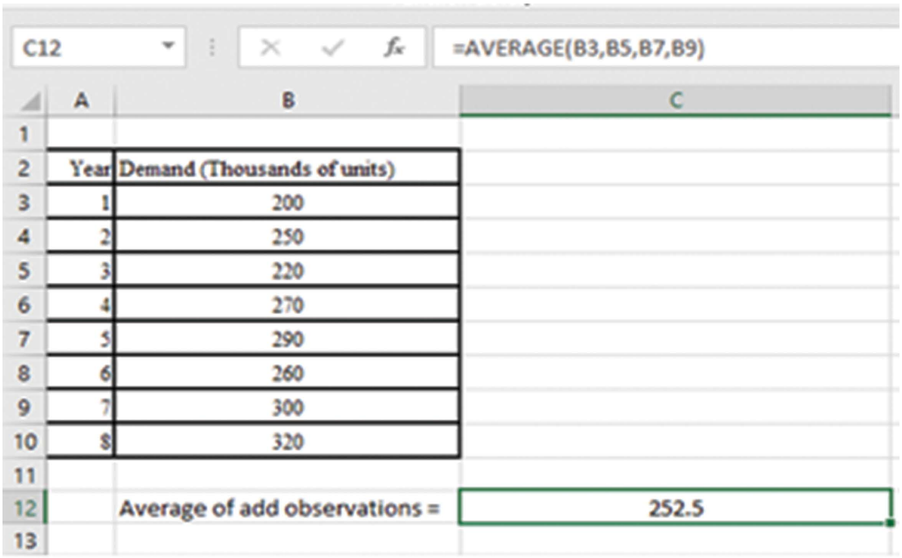
= 252.5

Returns the average (arithmetic mean) of its arguments, which can be numbers or names, arrays, or references that contain numbers.

Number1: number1,number2,... are 1 to 255 numeric arguments for which you want the average.

Formula result = 252.5

Figure 4.8 Screenshot of selection of cells to compute mean demand of cells of odd years of Example 4.1



C12

=AVERAGE(B3,B5,B7,B9)

Year	Demand (Thousands of units)
1	200
2	250
3	220
4	270
5	290
6	260
7	300
8	320

Average of add observations = 252.5

Figure 4.9 Screenshot of result after clicking the OK button in the dropdown menu of Figure 4.8

The selection of cells for the computation of the mean demand of odd years is shown in cell Number 1 of Figure 4.8.

Clicking the OK button in the dropdown menu of Figure 4.8 gives the result shown in Figure 4.9.

All the steps may be combined through the following formula.

= AVERAGE(B3, B5, B7, B9)

This formula may be entered in any convenient cell, say, C12, to give the result.

4.3 AVEDEV Function

The AVEDEV function calculates the average of the absolute deviations of observations from their arithmetic mean (Xi, i = 1, 2, 3, . . . , n, where n is the number of data). Alternatively, this could be referred to as the mean absolute deviation (MAD) of a set of observations.

The formula for the arithmetic mean is as follows.

$$\text{Arithmetic mean, } \bar{\mathrm{X}} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{X_i}}{\mathrm{n}}$$

where
n is the number of observations
X_i is the i^{th} observation for i = 1, 2,3, . . ., n
$\bar{X}$ is the arithmetic mean

The formula for AVEDEV is as follows.

$$\text{AVEDEV} = \frac{\sum_{i=1}^{n} \left| \bar{X} - X_i \right|}{n}$$

where
n is the number of cells containing data
X_i is the i^{th} observation for i = 1, 2,3, . . ., n
$\bar{X}$ is the arithmetic mean

Clicking the following sequence of buttons in Excel gives the screenshot in Figure 4.10.

Formulas ⟹ More Functions ⟹ Statistical

In Figure 4.10, clicking the option AVRDEV gives the screenshot in Figure 4.11. Now, one should enter the cells one after another, which contain the data against Number 1, Number 2, Number 3, and so on. If the data are arranged in a range of cells continuously, then that range of cells is to be entered against Number 1. After defining the addresses of the data, clicking the OK button in the dropdown menu of Figure 4.11 will display the average of the absolute deviations of data points from their mean.

Clicking OK will return the average of its arguments, which can be numbers, names, arrays, or references that contain numbers. Number 1, Number 2, Number 3, and so on, up to Number 255, are the arguments for which the mean absolute deviation is to be computed.

Instead of following the menu-driven option, one can enter a formula in the desired cell, say, A11, to get the AVEDEV of a given set of observations as follows, if the data are stored from cell A1 to cell A10.

$$= \text{AVEDEV}(\text{A1} : \text{A10})$$

Out of the range of data from cell A1 to cell A10, if the average deviation of the data at cells A2, A5, and A10 is required, then the corresponding formula is as follows.

$$= \text{AVEDEV}(\text{A2, A5, A10})$$

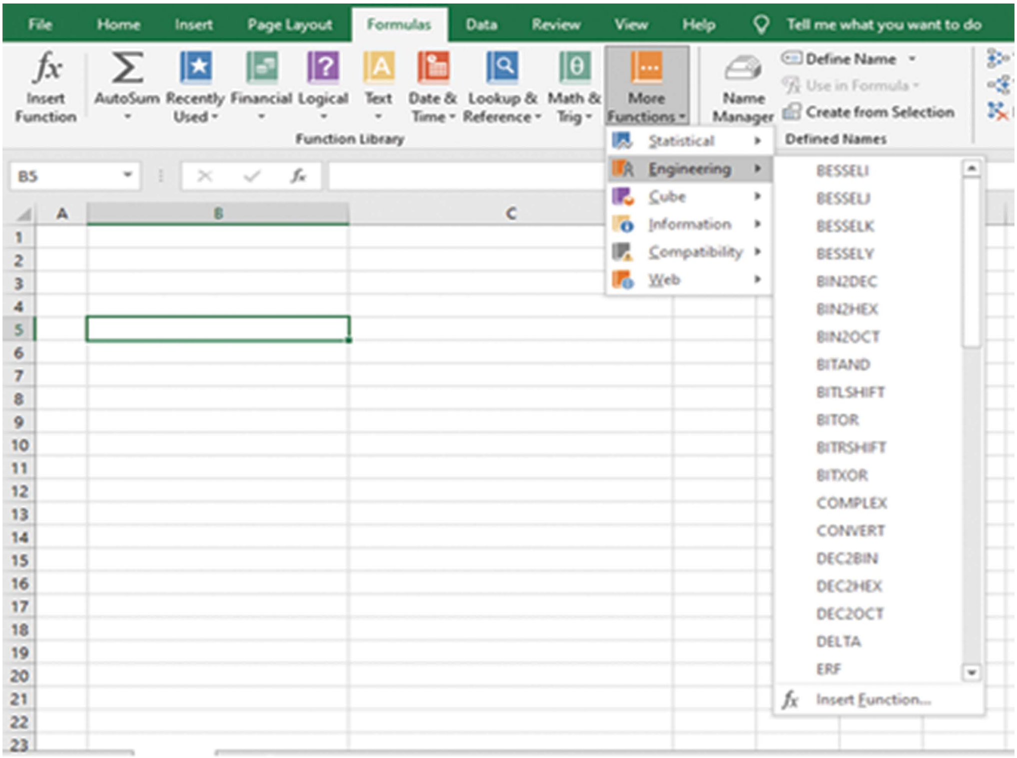
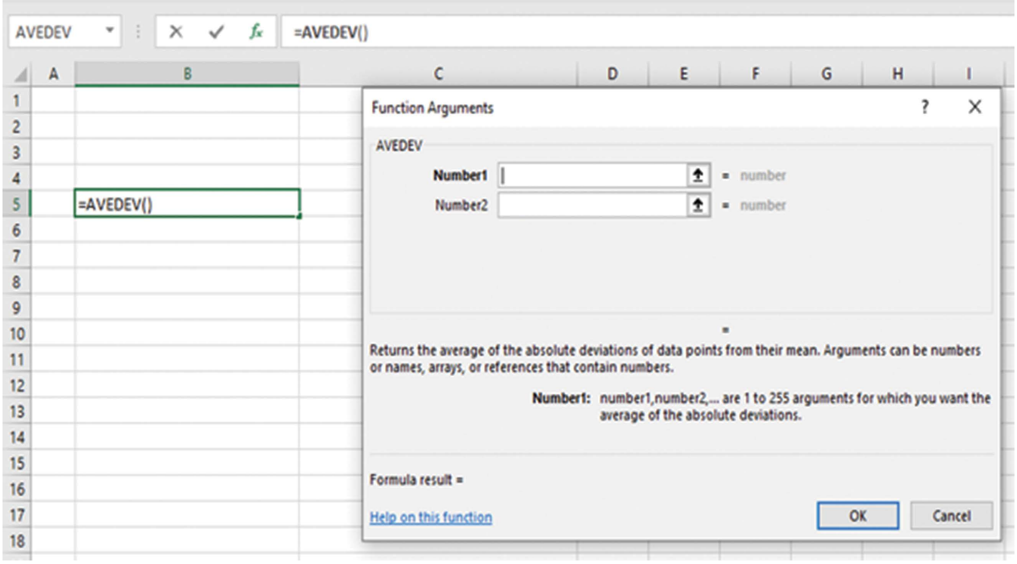


Figure 4.10 Screenshot of sequence of button clicks, Formulas $\Longrightarrow$ More Formulas $\Longrightarrow$ Statistical



=AVEDEV()

Function Arguments

AVEDEV

Number1 = number

Number2 = number

Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

Number1: number1,number2,... are 1 to 255 arguments for which you want the average of the absolute deviations.

Formula result =

Help on this function

OK Cancel

Figure 4.11 Screenshot after clicking the option AVEDEV in the dropdown menu of Figure 4.10

Example 4.2

Determine the mean absolute deviation of the salaries of the employees, which are shown in Table 4.3.

Solution

The data for Example 4.2 are shown in Table 4.4.

$$\text{Arithmetic mean of salaries} = \frac{\sum_{i=1}^{7} X_I}{7}$$

where X_i is the salary of the i^{th} employee, i = 1, 2, 3, . . ., 7.

The AVEDEV function gives the mean of the absolute deviations of the observations (X_i, i = 1, 2, 3, . . . , 7) from the arithmetic mean of the observations, which is given by the following equation.

The formula for AVEDEV is as follows.

$$AVEDEV = \frac{\sum_{i=1}^{7} \left|\bar{X} - X_i\right|}{7}$$

where
X_i is the i^{th} observation for i = 1, 2,3, . . ., 7
$\bar{X}$ is the arithmetic mean

Table 4.3 Salaries of Employees

Employee No.	*Monthly Salary (₹)*
1	13000
2	16000
3	9500
4	14500
5	20500
6	19500
7	7500

Table 4.4 Data for Example 4.2

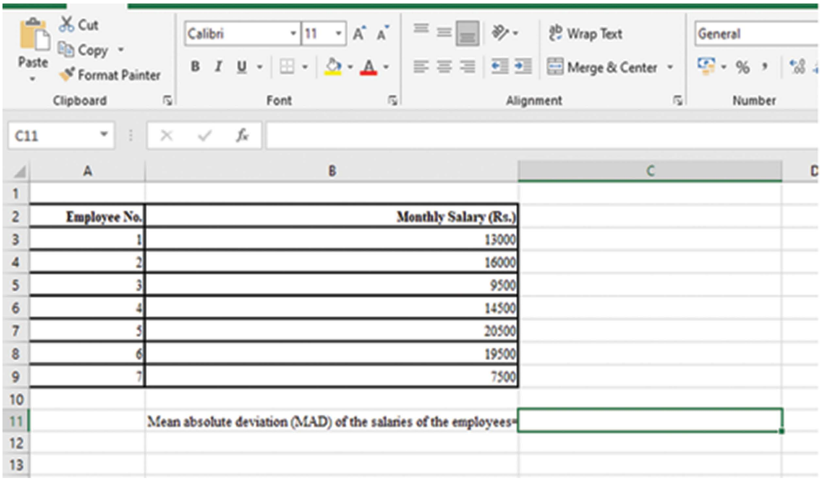
Employee No.	*Monthly Salary (₹)*
1	13000
2	16000
3	9500
4	14500
5	20500
6	19500
7	7500

Step 1: A screenshot after inputting the data into the Excel sheet is shown in Figure 4.12.

Step 2: A screenshot after clicking the sequence of buttons Formulas ⟹ More Functions ⟹ Statistical gives the display shown in Figure 4.13.

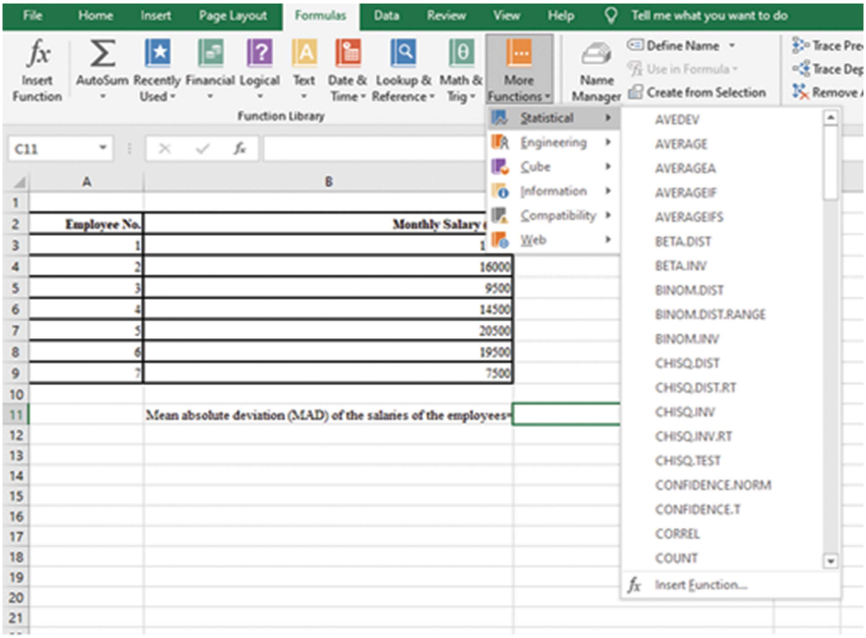
Step 3: Click the AVEDEV function in the dropdown menu of Figure 4.13, which gives the display shown in Figure 4.14.

Step 4: From Figure 4.14, it is clear that one can enter the series of data one after another in the cells Number 1, Number 2, . . . Number 255. The arguments can be numbers,



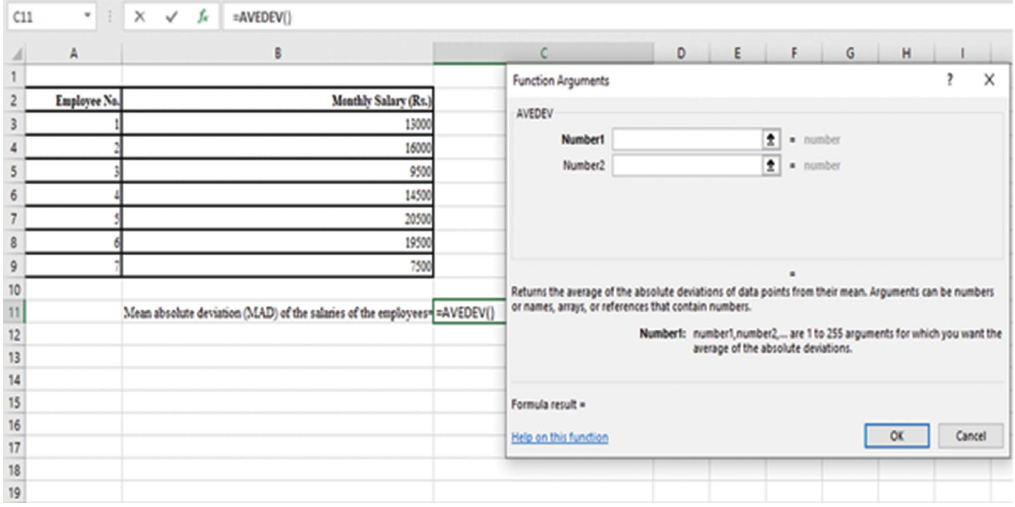
	A	B	C
1			
2	Employee No.	Monthly Salary (Rs.)	
3	1	13000	
4	2	16000	
5	3	9500	
6	4	14500	
7	5	20500	
8	6	19500	
9	7	7500	
10			
11		Mean absolute deviation (MAD) of the salaries of the employees=	
12			
13			

Figure 4.12 Screenshot after inputting data into the Excel sheet



	A	B
1		
2	Employee No.	Monthly Salary (
3	1	1
4	2	16000
5	3	9500
6	4	14500
7	5	20500
8	6	19500
9	7	7500
10		
11		Mean absolute deviation (MAD) of the salaries of the employees=

Figure 4.13 Screenshot after clicking the sequence of buttons, Formulas ⟹ More Functions ⟹ Statistical



C11 =AVEDEV()

Employee No.	Monthly Salary (Rs.)
1	13000
2	16000
3	9500
4	14500
5	20500
6	19500
7	7500

Mean absolute deviation (MAD) of the salaries of the employees= =AVEDEV()

Function Arguments

AVEDEV

Number1 = number

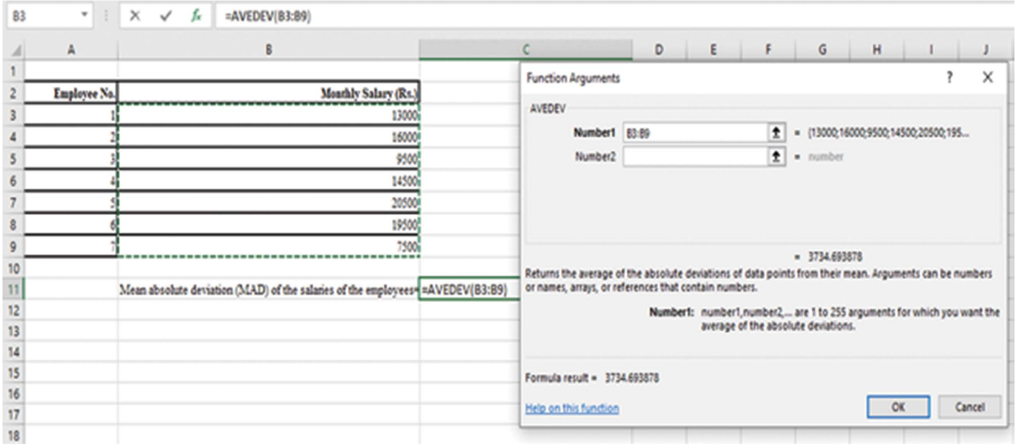
Number2 = number

Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

Number1: number1,number2,... are 1 to 255 arguments for which you want the average of the absolute deviations.

Formula result =

Figure 4.14 Screenshot after clicking AVEDEV function



B3 =AVEDEV(B3:B9)

Employee No.	Monthly Salary (Rs.)
1	13000
2	16000
3	9500
4	14500
5	20500
6	19500
7	7500

Mean absolute deviation (MAD) of the salaries of the employees= =AVEDEV(B3:B9)

Function Arguments

AVEDEV

Number1 B3:B9 = {13000;16000;9500;14500;20500;195...

Number2 = number

= 3734.693878

Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

Number1: number1,number2,... are 1 to 255 arguments for which you want the average of the absolute deviations.

Formula result = 3734.693878

Figure 4.15 Screenshot after entering the range of cells in the box against Number 1 in the dropdown menu of Figure 4.14

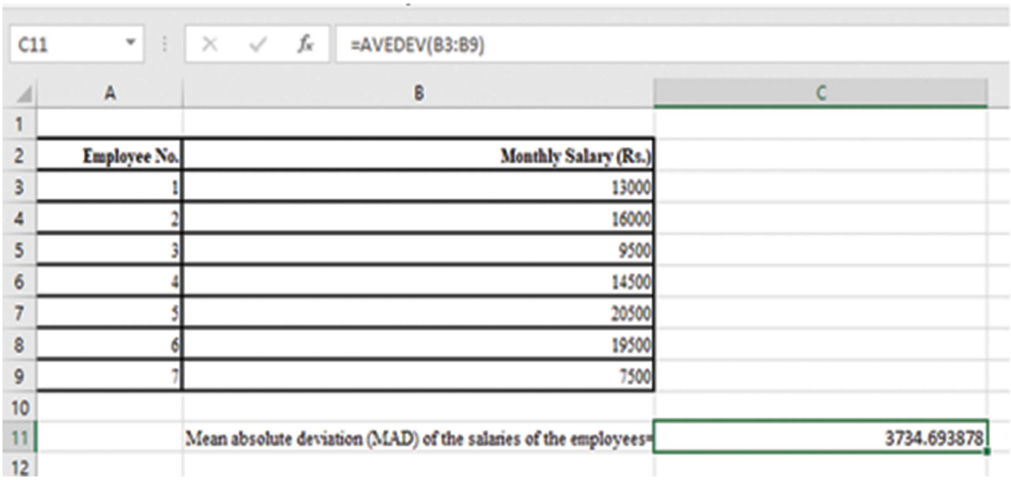
names, arrays, or references that contain numbers. Enter the range of cells containing the data in the box against Number 1 in the dropdown menu of Figure 4.15. In Figure 4.15, the cursor is positioned in cell B11 to view the formula.

Step 5: Click the OK button in the dropdown menu of Figure 4.15 to show the mean absolute deviation of the salaries of employees, as shown in Figure 4.16. The average deviation of the salaries of the employees from the mean salary is 3734.693878.

All these steps can be combined using the following formula.

$$= \text{AVEDEV}(\text{B3} : \text{B9})$$

Entering this formula in any convenient cell, say, B11, will give the desired result.



C11 =AVEDEV(B3:B9)

Employee No.	Monthly Salary (Rs.)
1	13000
2	16000
3	9500
4	14500
5	20500
6	19500
7	7500

Mean absolute deviation (MAD) of the salaries of the employees= 3734.693878

Figure 4.16 Screenshot for the formula = AVEDEV(B3:B9) with the result for AVEDEV in cell B11

Example 4.3

Take into account the employee deductions made on a monthly basis, as detailed in Table 4.5. Calculate the mean absolute deviation of the employee monthly deductions for the employees as given in parts a and b using the AVEDEV() function.

a. Employees 3, 5, 7, and 9.
b. Employees 4, 6, and 8.

Solution

The data for Example 4.3 are shown in Table 4.6.

a. After copying the data for the problem into an Excel sheet, the formula to find the AVEDEV of the data given in cells B3, B5, B7, and B9 can be issued as follows.

$$= \text{AVEDEV}\left(\text{B3, B5, B7, B9}\right)$$

This can be done by directly entering the formula in the formula bar at the top or entering the required cells separated by commas in the box against Number 1 in the dropdown menu after issuing the sequence of clicks Formulas ⟹ More Functions ⟹ Statistical ⟹ AVEDEV, as shown in Figure 4.17.

Next, click the OK button in the dropdown menu of Figure 4.17 to show the result in Figure 4.18. The value of the mean absolute deviation of the data shown in cells B3, B5, B7, and B9 from their mean is 3875. The same result can be obtained using the following formula directly in a convenient cell, say, C12.

$$= \text{AVEDEV}\left(\text{B3, B5, B7, B9}\right)$$

b. The formula to find the AVEDEV of the data in cells B4, B6, and B8 is shown in the following.

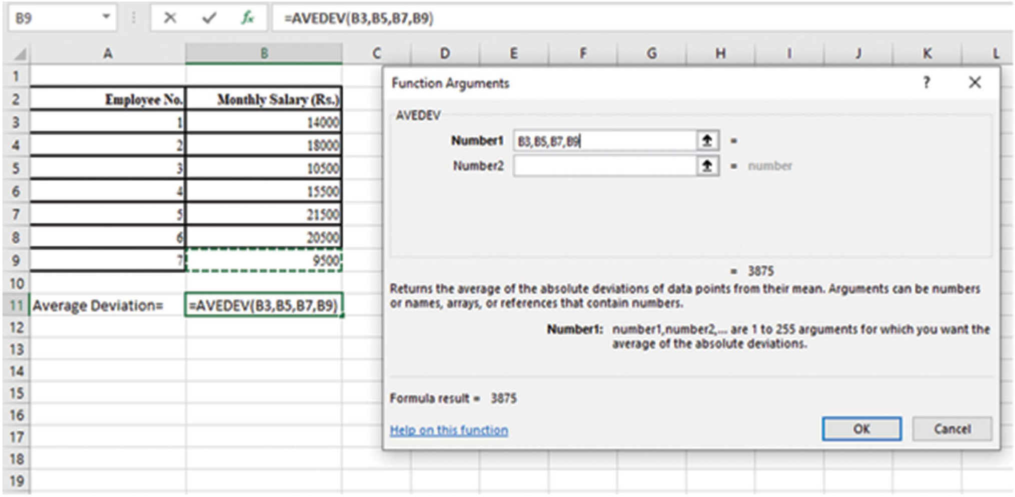
$$= \text{AVEDEV}\left(\text{B4, B6, B8}\right)$$

Table 4.5 Total Monthly Deductions of Employees

Employee No.	*Monthly Deduction (₹)*
1	14000
2	18000
3	10500
4	15500
5	21500
6	20500
7	9500

Table 4.6 Data for Example 4.3

Employee No.	*Monthly Deduction (₹)*
1	14000
2	18000
3	10500
4	15500
5	21500
6	20500
7	9500



=AVEDEV(B3,B5,B7,B9)

	A	B
1		
2	Employee No.	Monthly Salary (Rs.)
3	1	14000
4	2	18000
5	3	10500
6	4	15500
7	5	21500
8	6	20500
9	7	9500
10		
11	Average Deviation=	=AVEDEV(B3,B5,B7,B9)

Number1 B3,B5,B7,B9 =

= 3875

Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

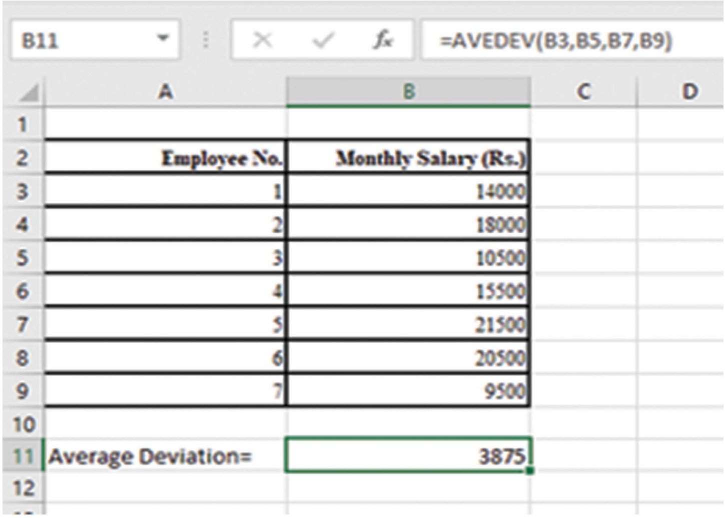
Number1: number1,number2,... are 1 to 255 arguments for which you want the average of the absolute deviations.

Formula result = 3875

Figure 4.17 Screenshot of the AVEDEV function and the result of data in Cells B3, B5, B7, and B9

This can be done by directly entering the formula in the formula bar at the top or entering the required cells separated by commas in the box against Number 1 in the dropdown menu after issuing the sequence of clicks Formulas ⟹ More Functions ⟹ Statistical ⟹ AVEDEV, as shown in Figure 4.19.

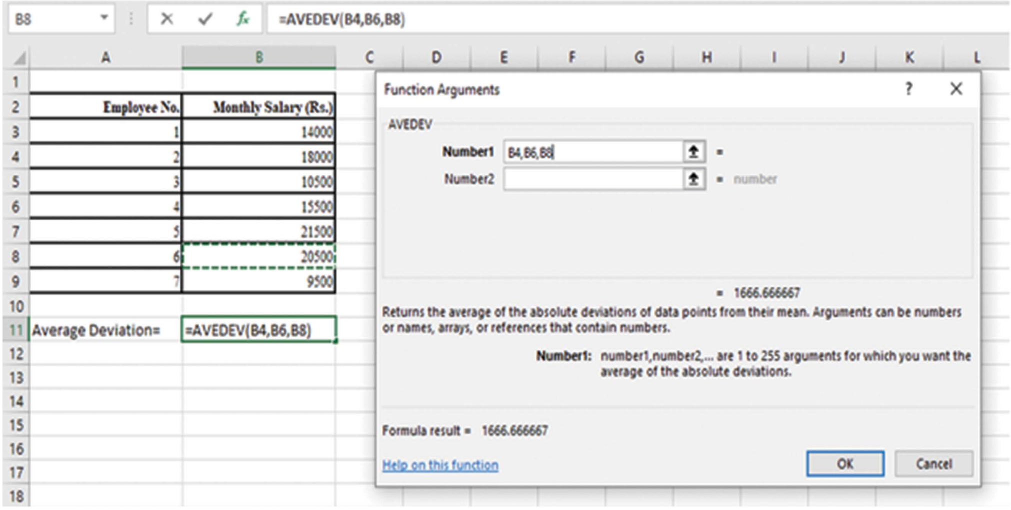
Next, click the OK button in the dropdown menu of Figure 4.19 to show the result in Figure 4.20. The value of the average deviation of the data shown in cells B4, B6, and B8 from their mean is 1666.666667.



B11 =AVEDEV(B3,B5,B7,B9)

	A	B
1		
2	Employee No.	Monthly Salary (Rs.)
3	1	14000
4	2	18000
5	3	10500
6	4	15500
7	5	21500
8	6	20500
9	7	9500
10		
11	Average Deviation=	3875
12		

Figure 4.18 Screenshot of result for AVEDEV for the data in cells B3, B5, B7, and B9



B8 =AVEDEV(B4,B6,B8)

	A	B
1		
2	Employee No.	Monthly Salary (Rs.)
3	1	14000
4	2	18000
5	3	10500
6	4	15500
7	5	21500
8	6	20500
9	7	9500
10		
11	Average Deviation=	=AVEDEV(B4,B6,B8)

Function Arguments

AVEDEV

Number1 B4,B6,B8 =

Number2 = number

= 1666.666667

Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

Number1: number1,number2,... are 1 to 255 arguments for which you want the average of the absolute deviations.

Formula result = 1666.666667

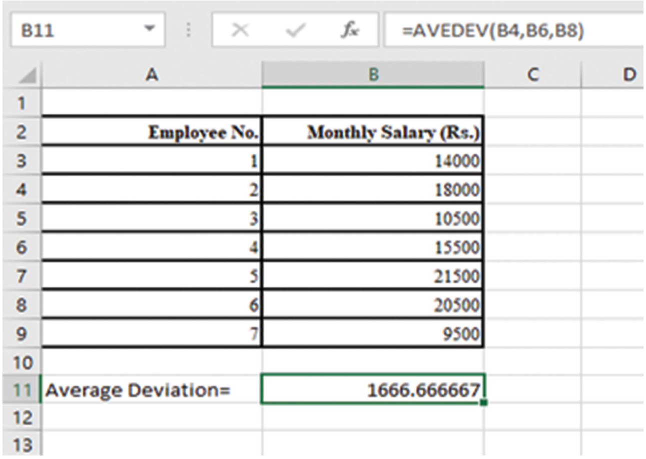
Figure 4.19 Screenshot after clicking the required cells B4, B6, and B8, separated by commas in the box against Number 1 of the dropdown menu

The same result can be obtained by using the following formula in a convenient cell, say, C12.

= AVEDEV(B4, B6, B8)

4.4 AVERAGEA Function

The AVERAGEA function calculates the average of a specified set of data. Sales of a product could be the desired data for which the average needs to be computed. The



B11 =AVEDEV(B4,B6,B8)

	A	B
2	Employee No.	Monthly Salary (Rs.)
3	1	14000
4	2	18000
5	3	10500
6	4	15500
7	5	21500
8	6	20500
9	7	9500
11	Average Deviation=	1666.666667

Figure 4.20 Screenshot of result for AVEDEV for the data in cells B4, B6, and B8

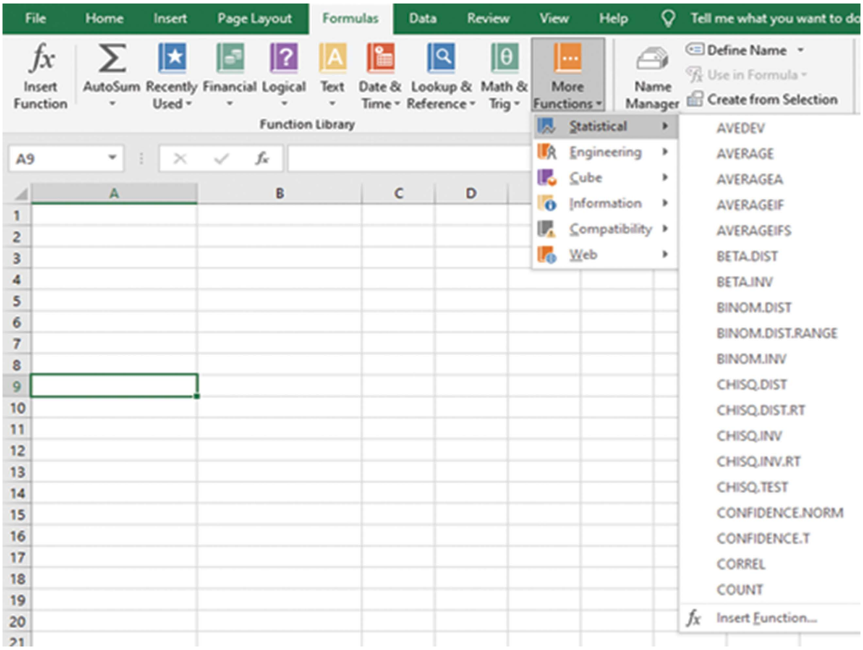


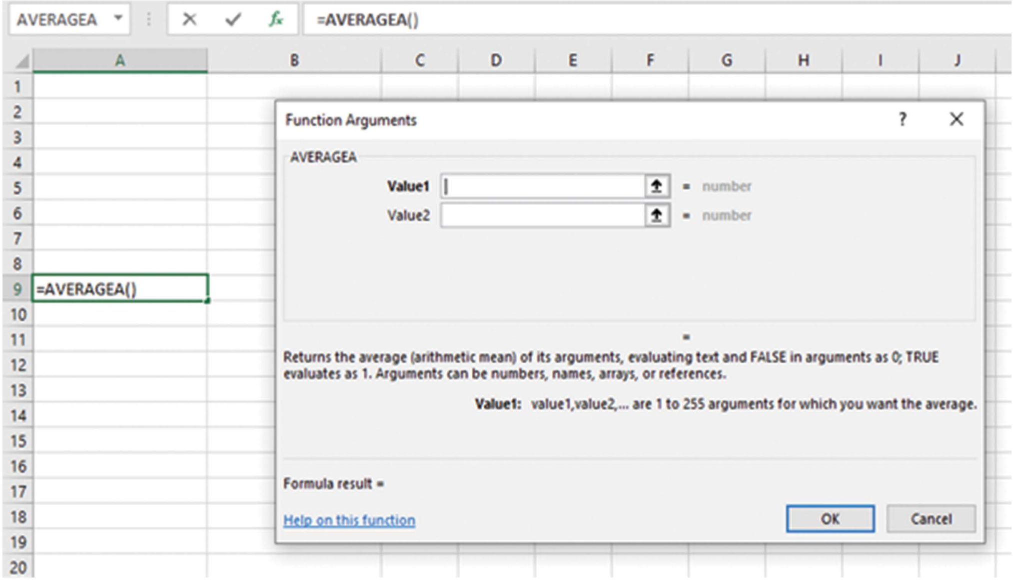
Figure 4.21 Screenshot of sequence of button clicks, Formulas ⟹ More Formulas ⟹ Statistical

criterion could be PROFIT or LOSS in terms of TRUE or FALSE during the previous ten years, respectively. In this case, TRUE is evaluated as 1 and FALSE as 0.

Clicking the following sequence of buttons in Excel gives the screenshot in Figure 4.21.

Home ⟹ Formulas ⟹ More Functions ⟹ Statistical

In Figure 4.21, clicking the option AVERAGEA gives the screenshot in Figure 4.22. Now one should enter the cells which contain the data one after another against Value



Returns the average (arithmetic mean) of its arguments, evaluating text and FALSE in arguments as 0; TRUE evaluates as 1. Arguments can be numbers, names, arrays, or references.

Value1: value1,value2,... are 1 to 255 arguments for which you want the average.

Figure 4.22 Screenshot after clicking the option AVEREAGEA in the dropdown menu of Figure 4.21

1, Value 2, Value 3, and so on, up to Value 255. The arguments can be numbers, names, arrays, or references that contain numbers. Value 1, Value 2, and so on, up to Value 255, are the arguments for which the average is to be computed. If the logical is FALSE, it is evaluated as 0, and if it is TRUE, it is evaluated as 1. If the data are arranged in a range of cells continuously, then that range of the cells is to be entered against Number 1. After defining the addresses of the data, clicking the OK button in the dropdown menu of Figure 4.22 will display the average of the data in a given range of cells in the Excel sheet for a given criterion or set of criteria.

Instead of following the menu-driven option, one can enter a formula in a desired cell, say, A11, to get the AVERAGEA of a given set of observations, as follows, if the data are stored from cell A1 to cell A10.

$$= \text{AVERAGEA}(\text{A1} : \text{A10})$$

Out of the range of data from cell A1 to cell A10, if the average of the data at cells A3, A4, A6, A8, and A10 is required, then the corresponding formula is as follows.

$$= \text{AVERAGEA}(\text{A3, A4, A6, A8, A10})$$

Example 4.4

Dividend status is indicated as Paid or Not Paid and is characterised as either being TRUE or FALSE. Table 4.7 displays information on a company in the industry for

Table 4.7 Dividend Status of Companies

Year Ending March	*Dividend (Paid: TRUE/Not Paid: FALSE)*
2015	FALSE
2016	TRUE
2017	TRUE
2018	FALSE
2019	TRUE
2020	FALSE

various years, ending in March. Find the average of the dividend status for the years from 2015 to 2020.

Solution

The data for Example 4.4 are shown in Table 4.8.

The input of the data for Example 4.4 in an Excel sheet is shown in Figure 4.23.

Clicking the sequence of buttons Formulas $\Longrightarrow$ More Functions $\Longrightarrow$ Statistical $\Longrightarrow$ AVERAGEA gives a screenshot, as shown in Figure 4.24. Then inputting the range of cells B3:B8 in the cell against Value 1 in the dropdown menu of Figure 4.24 gives the screenshot in Figure 4.25. Following this, clicking the OK button in the dropdown menu of Figure 4.25 gives the output in cell B10, as shown in Figure 4.26.

The entire sequence of operations for the AVERAGEA function can be executed by entering the formula in cell B10 as follows, which gives 0.5 as the average of TRUE and FALSE data in the range of cells from B3 to B8.

$$\text{Formula} := \text{AVERAGEA}(\text{B3} : \text{B8})$$

The average dividend status of the company is 0.5.

Example 4.5

Consider the data for Example 4.4 and find the average of TRUE and FALSE, which are present in cells B3, B5, and B7, using the AVERAGEA function.

Solution

The data for the problem are shown in Table 4.9.

Clicking Formulas $\Longrightarrow$ More Functions $\Longrightarrow$ Statistical $\Longrightarrow$ AVERAGEA and entering cells B3, B5, and B7 in the cell against Value 1 of the corresponding dropdown menu gives a display as shown in Figure 4.27. Then clicking the OK button in the dropdown menu of Figure 4.27 gives the result of AVERAGEA, as shown in Figure 4.28.

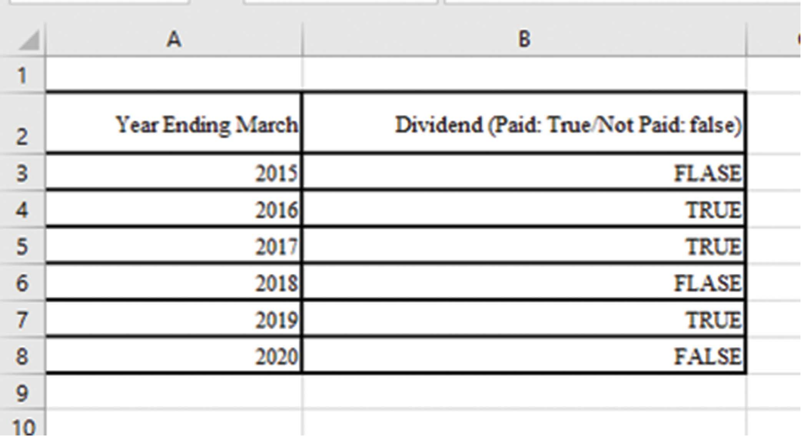
The same sequence of operations can be executed using the following formula to get the average of the data in cells B3, B5, and B7.

$$\text{Formula} := \text{AVERAGEA}(\text{B3, B5, B7})$$

The average dividend status of the company is 0.666666667.

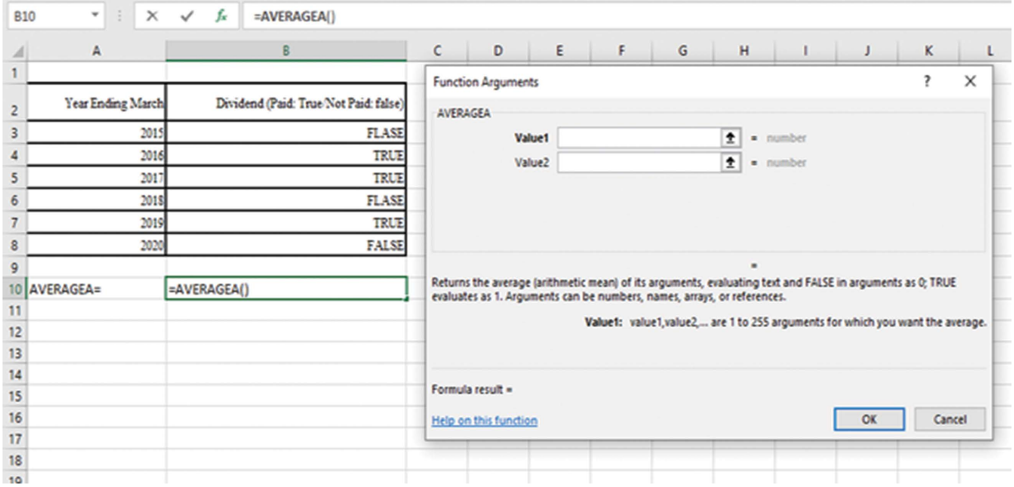
Table 4.8 Data for Example 4.4

Year Ending March	*Dividend (Paid: TRUE/Not Paid: FALSE)*
2015	FALSE
2016	TRUE
2017	TRUE
2018	FALSE
2019	TRUE
2020	FALSE



Year Ending March	Dividend (Paid: True/Not Paid: false)
2015	FLASE
2016	TRUE
2017	TRUE
2018	FLASE
2019	TRUE
2020	FALSE

Figure 4.23 Screenshot of input of data for Example 4.4 in Excel sheet



B10 =AVERAGEA()

Year Ending March	Dividend (Paid: True/Not Paid: false)
2015	FLASE
2016	TRUE
2017	TRUE
2018	FLASE
2019	TRUE
2020	FALSE
AVERAGEA=	=AVERAGEA()

Function Arguments

AVERAGEA

Value1 = number

Value2 = number

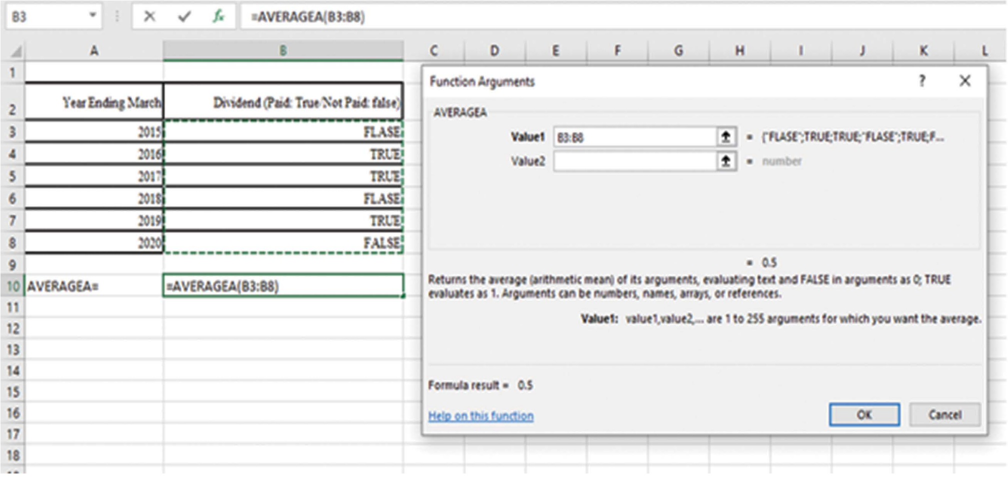
Returns the average (arithmetic mean) of its arguments, evaluating text and FALSE in arguments as 0; TRUE evaluates as 1. Arguments can be numbers, names, arrays, or references.

Value1: value1,value2,... are 1 to 255 arguments for which you want the average.

Formula result =

Help on this function

Figure 4.24 Screenshot of clicks of the sequence of buttons, Formulas ⟹More Functions ⟹ Statistical ⟹ AVERAGEA



B3 =AVERAGEA(B3:B8)

	A	B
2	Year Ending March	Dividend (Paid: True/Not Paid: false)
3	2015	FLASE
4	2016	TRUE
5	2017	TRUE
6	2018	FLASE
7	2019	TRUE
8	2020	FALSE
10	AVERAGEA=	=AVERAGEA(B3:B8)

Function Arguments

AVERAGEA

Value1 B3:B8 = {"FLASE";TRUE;TRUE;"FLASE";TRUE;F...

Value2 = number

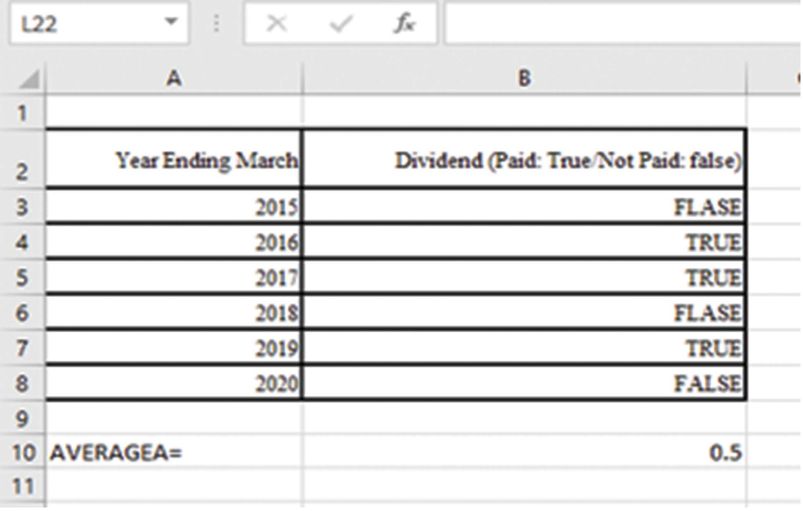
= 0.5

Returns the average (arithmetic mean) of its arguments, evaluating text and FALSE in arguments as 0; TRUE evaluates as 1. Arguments can be numbers, names, arrays, or references.

Value1: value1,value2,... are 1 to 255 arguments for which you want the average.

Formula result = 0.5

Figure 4.25 Screenshot of input of range of cells (B3:B8) in the cell against Value 1 in the dropdown menu of Figure 4.24



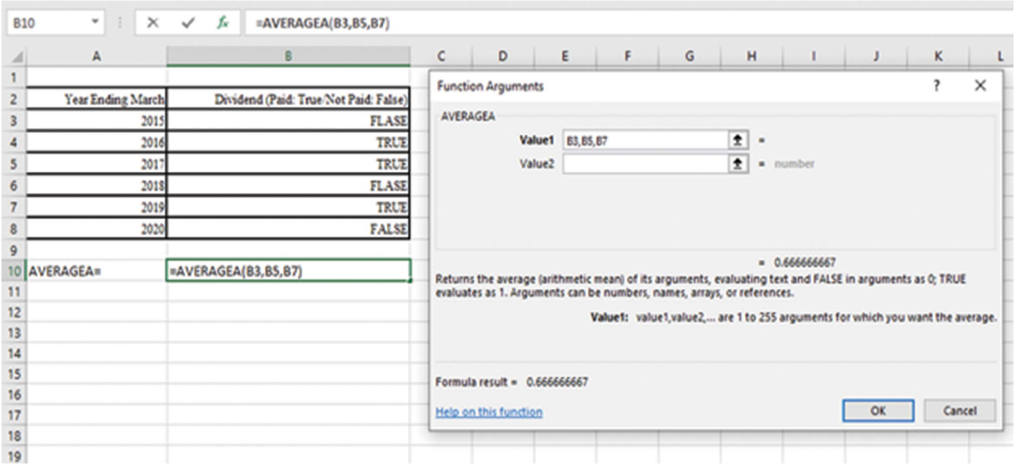
L22

	A	B
2	Year Ending March	Dividend (Paid: True/Not Paid: false)
3	2015	FLASE
4	2016	TRUE
5	2017	TRUE
6	2018	FLASE
7	2019	TRUE
8	2020	FALSE
10	AVERAGEA=	0.5

Figure 4.26 Screenshot after clicking the OK button in the dropdown menu of Figure 4.25

Table 4.9 Data for Example 4.5

Year Ending March	*Dividend (Paid: TRUE/Not Paid: FALSE)*
2015	FALSE
2016	TRUE
2017	TRUE
2018	FALSE
2019	TRUE
2020	FALSE



B10 =AVERAGEA(B3,B5,B7)

	A	B
1		
2	Year Ending March	Dividend (Paid: True/Not Paid: False)
3	2015	FLASE
4	2016	TRUE
5	2017	TRUE
6	2018	FLASE
7	2019	TRUE
8	2020	FALSE
9		
10	AVERAGEA=	=AVERAGEA(B3,B5,B7)

Function Arguments

AVERAGEA

Value1 B3,B5,B7 =

Value2 = number

= 0.666666667

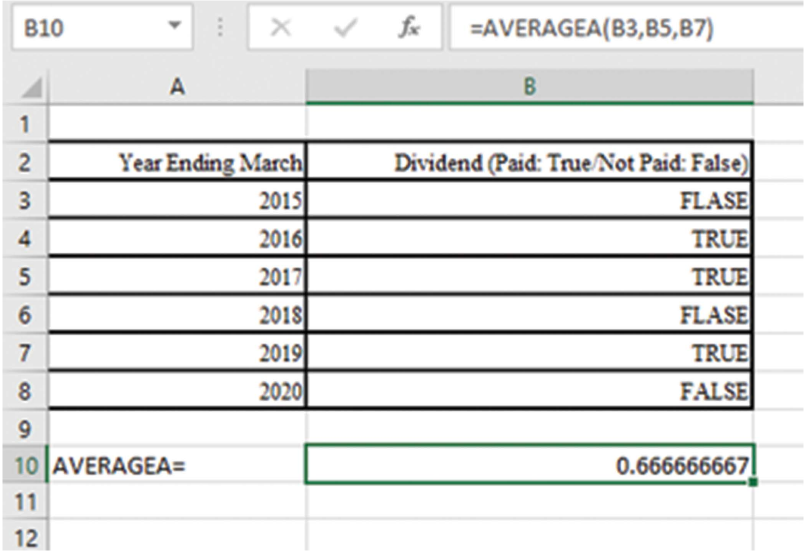
Returns the average (arithmetic mean) of its arguments, evaluating text and FALSE in arguments as 0; TRUE evaluates as 1. Arguments can be numbers, names, arrays, or references.

Value1: value1,value2,... are 1 to 255 arguments for which you want the average.

Formula result = 0.666666667

Help on this function

Figure 4.27 Screenshot of button clicks, Formulas ⟹ More Functions ⟹ Statistical ⟹ AVERAGEA and entry of Cells B3, B5, and B7 in the cell against Value 1 of the corresponding dropdown menu



B10 =AVERAGEA(B3,B5,B7)

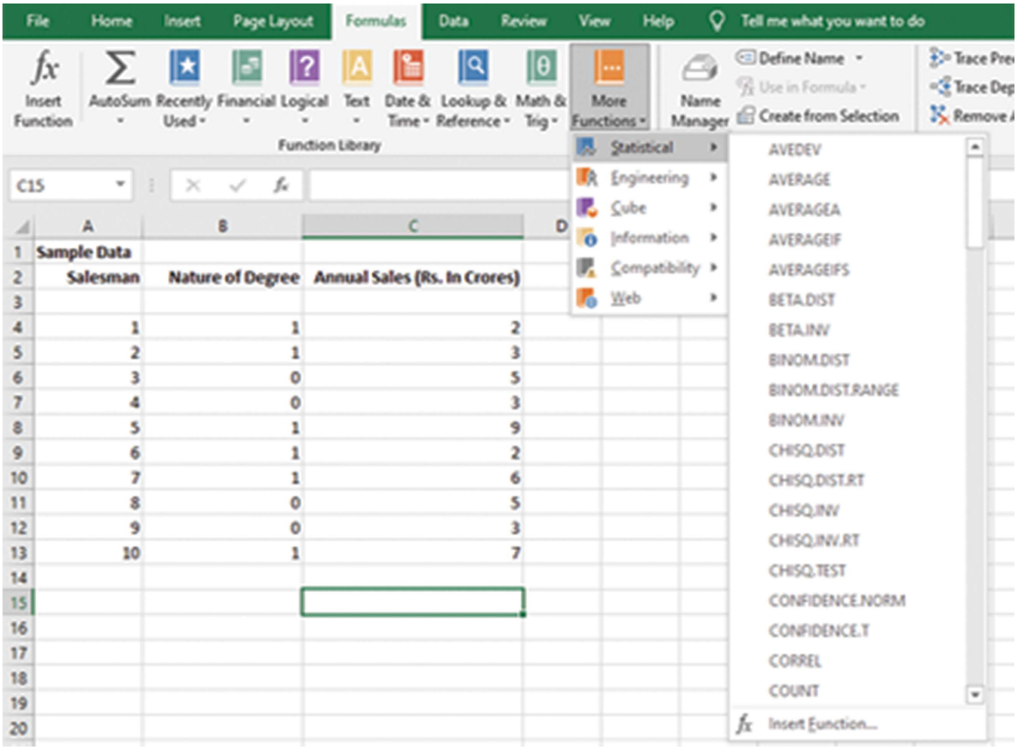
	A	B
1		
2	Year Ending March	Dividend (Paid: True/Not Paid: False)
3	2015	FLASE
4	2016	TRUE
5	2017	TRUE
6	2018	FLASE
7	2019	TRUE
8	2020	FALSE
9		
10	AVERAGEA=	0.666666667
11		
12		

Figure 4.28 Screenshot after clicking the OK button in the dropdown menu of Figure 4.27

4.5 AVERAGEIF Function

For a specific criterion, the AVERAGEIF function calculates the average of a series of observations that are kept in a range of cells.

Consider the salesforce population of a company consisting of salespeople with engineering degrees (Code 1) and non-engineering degrees (Code 0). The objective may be to find the average annual sales made by the salespeople with engineering degrees as well as those made by the salespeople with non-engineering degrees. Here, the data on the



Sample Data

Salesman	Nature of Degree	Annual Sales (Rs. In Crores)
1	1	2
2	1	3
3	0	5
4	0	3
5	1	9
6	1	2
7	1	6
8	0	5
9	0	3
10	1	7

Figure 4.29 Screenshot of sequence of button clicks, Formulas ⟹ More Formulas ⟹ Statistical with sample data

annual sales form the data to find the average, and the nature of degrees of the salespeople forms the criterion (1 for engineering and 0 for non-engineering).

After entering a sample set of data, clicking the following sequence of buttons in Excel gives the screenshot in Figure 4.29.

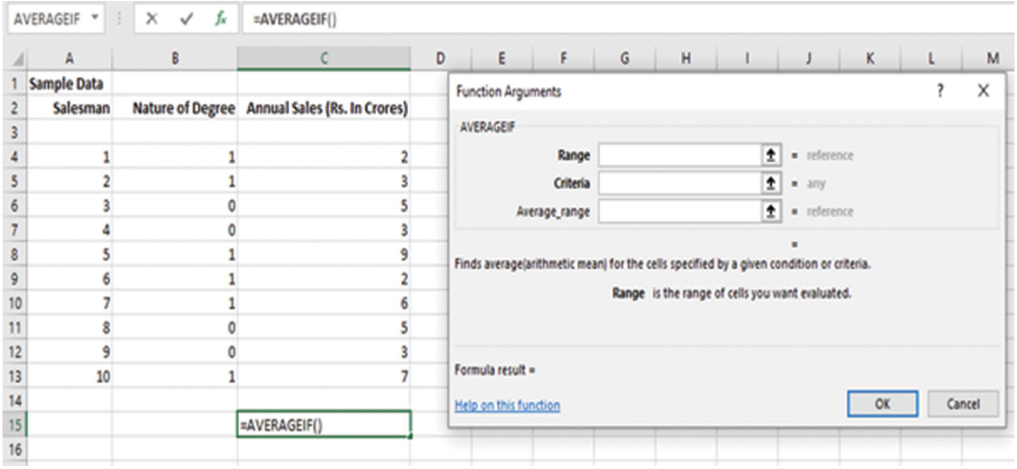
Home ⟹ Formulas ⟹ More Functions ⟹ Statistical

In Figure 4.29, clicking the option AVERAGEIF gives the screenshot in Figure 4.30. The different arguments in the dropdown menu of this screenshot are listed in the following.

Range
Criteria
Average_ range

The range is a range of cells which contains the data based on which criterion will be defined. The criterion filters the data for which the average is required. The average range is the range of cells, which contains the data for which the average is to be computed based on the defined criterion.

Clicking OK in the dropdown menu of Figure 4.30 will return the average (arithmetic mean) of the data stored in the range of cells based on certain defined criterion, say, the annual sales made by salespeople with engineering degrees (Code 1).



AVERAGEIF =AVERAGEIF()

	A	B	C
1	Sample Data		
2	Salesman	Nature of Degree	Annual Sales (Rs. In Crores)
3			
4	1	1	2
5	2	1	3
6	3	0	5
7	4	0	3
8	5	1	9
9	6	1	2
10	7	1	6
11	8	0	5
12	9	0	3
13	10	1	7
14			
15			=AVERAGEIF()
16			

Function Arguments

AVERAGEIF

Range = reference

Criteria = any

Average_range = reference

=

Finds average(arithmetic mean) for the cells specified by a given condition or criteria.

Range is the range of cells you want evaluated.

Formula result =

Figure 4.30 Screenshot after clicking the option AVERAGEIF from the dropdown menu of Figure 4.29

Instead of following the menu-driven option, one can enter a formula in the desired cell, say, C15, to get the AVERAGEIF of a given set of observations as follows, if the data on the annual sales are stored from cells C4 to C13.

= AVERAGEIF(Range of cells containing criterion data, Criterion,

Range of cells containing data for which average is required)

= AVERAGEIF(B4 : B13, 1, C4 : C13)

One can verify that the required average is 4.833333333.

Example 4.6

Take a look at Table 4.10 for information on earnings per share (EPS), dividend percentage, and 52-week high share price for various companies for years ending in March. Using Excel's AVERAGEIF function, get the median of the 52-week high share prices on the condition that EPS is 20.

Solution

The data for Example 4.6 are shown in Table 4.11.

A screenshot after inputting the given data in an Excel sheet is shown in Figure 4.31. Then the button clicks Formulas ⟹ More Functions ⟹ Statistical ⟹ VERAGEIF give a screenshot as shown in Figure 4.32.

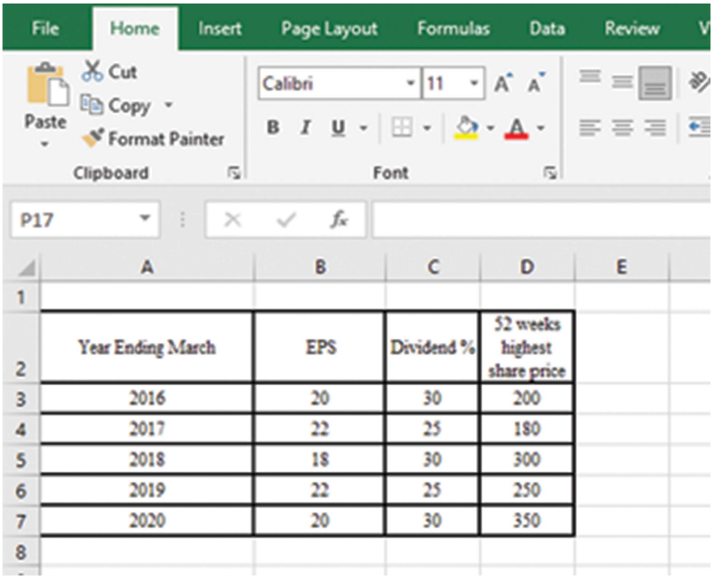
Then entering the range of cells B3:B7 in the cell against Range, 20 as the criterion value in the cell against Criteria, and the range of cells D3:D7 in the cell against Average Range in the dropdown menu of Figure 4.32 will be as shown in Figure 4.33. Then

Table 4.10 Data for EPS, Dividend %, and 52-Week Highest Share Price

Year Ending March	*EPS*	*Dividend %*	*52-Week Highest Share Price*
2016	20	30	200
2017	22	25	180
2018	18	30	300
2019	22	25	250
2020	20	30	350

Table 4.11 Data for Example 4.6

Year Ending March	*EPS*	*Dividend %*	*52-Week Highest Share Price*
2016	20	30	200
2017	22	25	180
2018	18	30	300
2019	22	25	250
2020	20	30	350



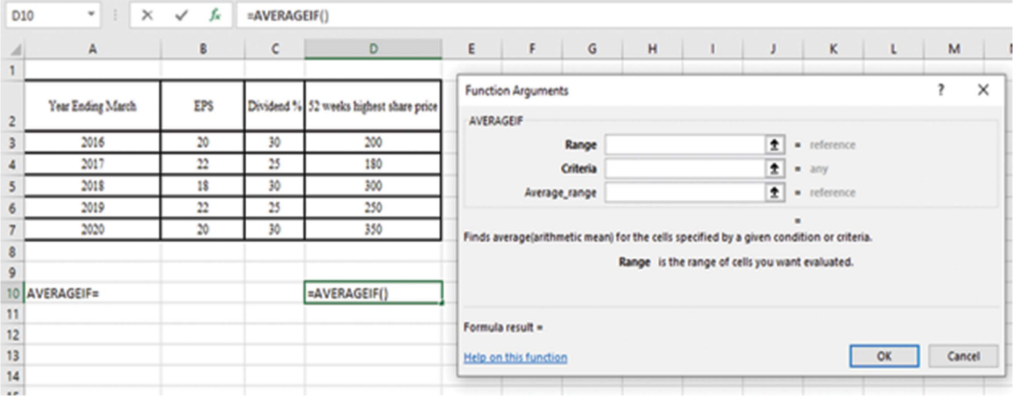
Year Ending March	EPS	Dividend %	52 weeks highest share price
2016	20	30	200
2017	22	25	180
2018	18	30	300
2019	22	25	250
2020	20	30	350

Figure 4.31 Screenshot after inputting data for Example 4.6 in an Excel sheet

clicking the OK button in the dropdown menu of Figure 4.33 gives the result shown in Figure 4.34, which is ₹ 275.

The entire sequence of operations can be combined in a formula as follows in a convenient cell, say, D10, to obtain the required average.

Formula : = AVERAGEIF(B3 : B7, 20, D3 : D7)



D10 =AVERAGEIF()

	A	B	C	D
2	Year Ending March	EPS	Dividend %	52 weeks highest share price
3	2016	20	30	200
4	2017	22	25	180
5	2018	18	30	300
6	2019	22	25	250
7	2020	20	30	350
10	AVERAGEIF=			=AVERAGEIF()

Function Arguments

AVERAGEIF

Range = reference

Criteria = any

Average_range = reference

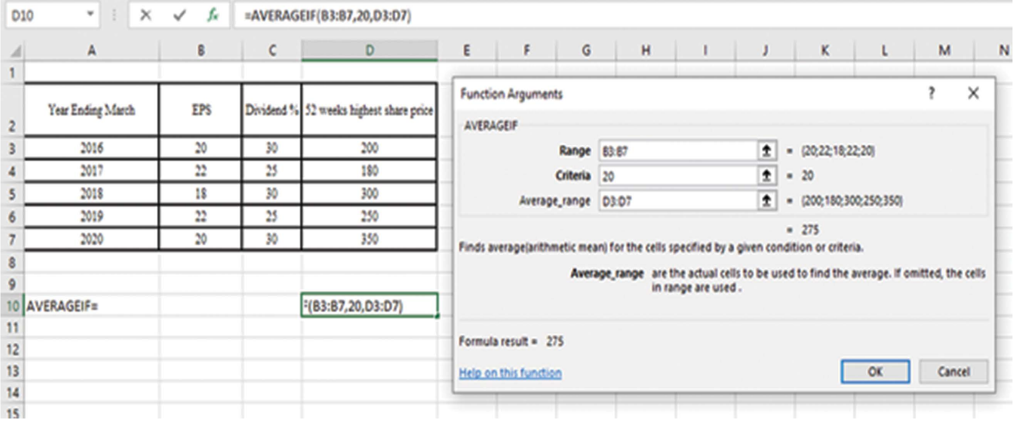
=

Finds average(arithmetic mean) for the cells specified by a given condition or criteria.

Range is the range of cells you want evaluated.

Formula result =

Figure 4.32 Screenshot after clicking buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ AVERAGEIF function



D10 =AVERAGEIF(B3:B7,20,D3:D7)

	A	B	C	D
2	Year Ending March	EPS	Dividend %	52 weeks highest share price
3	2016	20	30	200
4	2017	22	25	180
5	2018	18	30	300
6	2019	22	25	250
7	2020	20	30	350
10	AVERAGEIF=			F(B3:B7,20,D3:D7)

Function Arguments

AVERAGEIF

Range B3:B7 = {20;22;18;22;20}

Criteria 20 = 20

Average_range D3:D7 = {200;180;300;250;350}

= 275

Finds average(arithmetic mean) for the cells specified by a given condition or criteria.

Average_range are the actual cells to be used to find the average. If omitted, the cells in range are used.

Formula result = 275

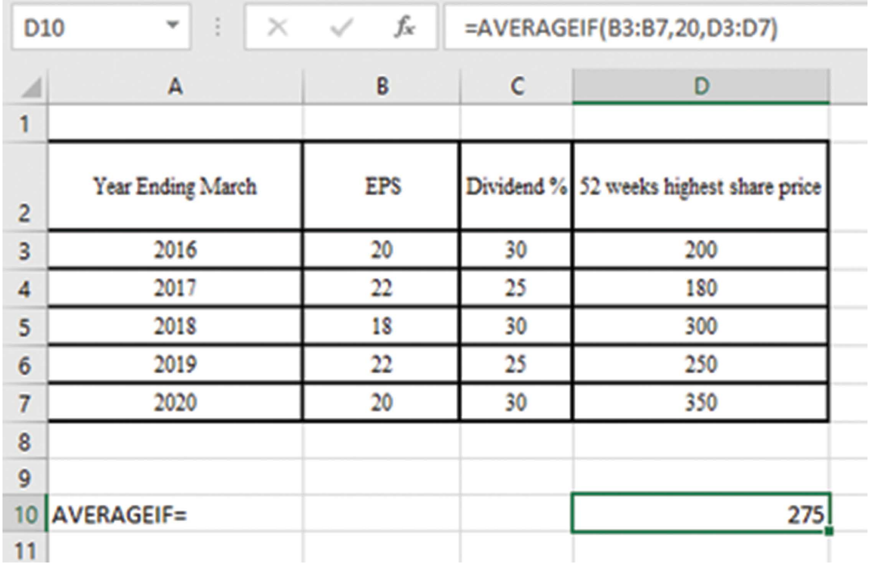
Figure 4.33 Screenshot after entering the range of Cells B3:B7 in the cell against Range, 20 as the criterion value in the cell against Criteria, and the range of cells D3:D7 in the cell against Average Range

The average of the 52-week highest share prices with respect to an EPS value of 20 is ₹ 275.

4.6 AVERAGEIFS Function

The AVERAGEIFS function finds the average of a set of observations stored in a range of cells for two criteria [3].

Consider the sales population of a company consisting of salespeople with engineering degrees (Code 1: 1) and non-engineering degrees (Code 2: 0). The ages of the salespeople differ significantly. The objective may be to find the average annual sales made by the salespeople with the nature of the degree as the first criterion and the age of the



Year Ending March	EPS	Dividend %	52 weeks highest share price
2016	20	30	200
2017	22	25	180
2018	18	30	300
2019	22	25	250
2020	20	30	350

AVERAGEIF= 275

Figure 4.34 Screenshot after clicking the OK button in the dropdown menu of Figure 4.33

salespeople as the second criterion. Here, the data on the annual sales form the data to find the average for the two stated criteria.

Clicking the following sequence of buttons in Excel gives the screenshot in Figure 4.35.

Formulas ⟹ More Functions ⟹ Statistical

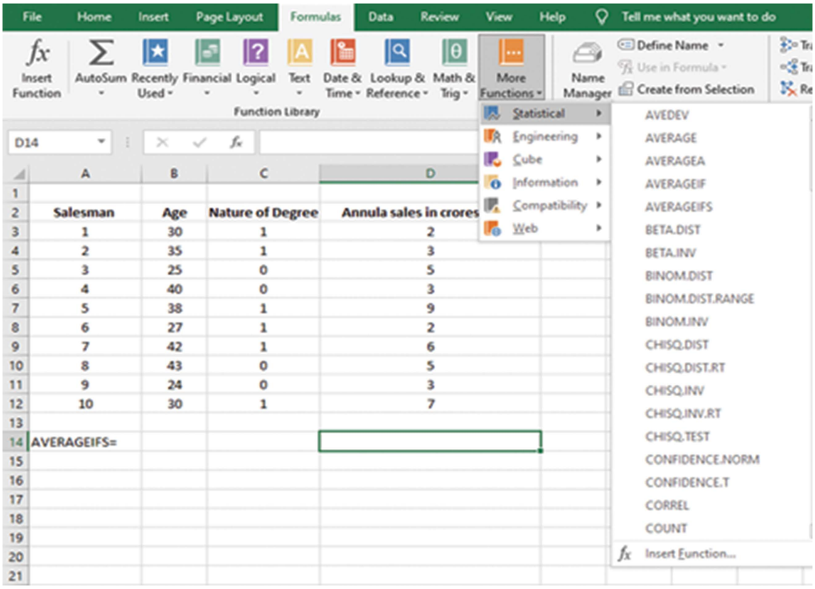
In Figure 4.35, clicking the option AVERAGEIFS gives the screenshot in Figure 4.36. The different arguments of this screenshot are listed in the following.

Average_Range
Criteria_Range1
Criteria1
Criteria_Range2
Criteria2

Note: There can be 127 criteria ranges and associated criteria.

The average range is a range of cells which contain the data for which the average is to be found for two criteria. Criteria_Range1 is the range of cells containing the data of Criterion1, Criteria1 is the condition of criterion 1, Criteria_Range2 is the range of cells containing the data of Criterion 2, and Criteria2 is the condition of criterion 2. The screenshot after entering the range of cells and conditions is shown in Figure 4.37.

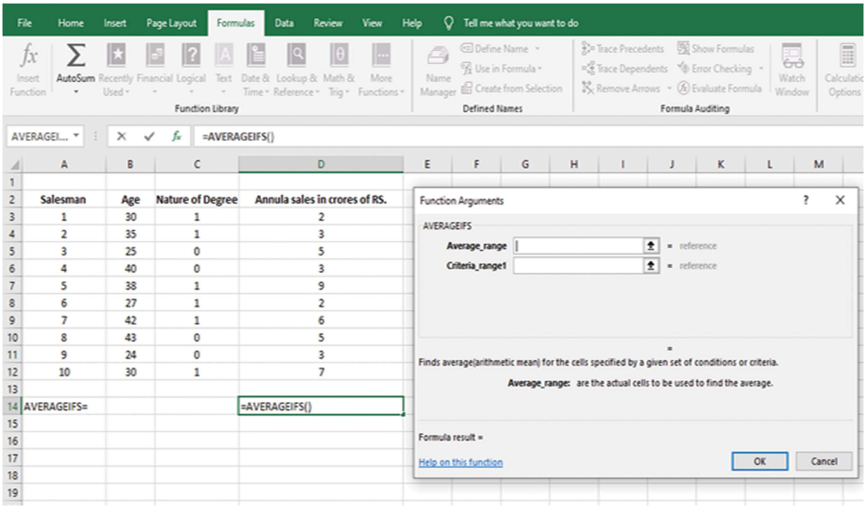
Clicking the OK button in the dropdown menu of Figure 4.37 will return the average (arithmetic mean) of the data stored in the range of cells against Average_Range when the nature of degree is 0 (Criterion1) and the age is >30 (Criteria2). The corresponding screenshot is shown in Figure 4.38. The average result is 4 crores of rupees.



D14

	A	B	C	D
1				
2	Salesman	Age	Nature of Degree	Annula sales in crores
3	1	30	1	2
4	2	35	1	3
5	3	25	0	5
6	4	40	0	3
7	5	38	1	9
8	6	27	1	2
9	7	42	1	6
10	8	43	0	5
11	9	24	0	3
12	10	30	1	7
13				
14	AVERAGEIFS=			

Figure 4.35 Screenshot of sequence of button clicks, Formulas ⟹ More Formulas ⟹ Statistical



AVERAGEI... =AVERAGEIFS()

	A	B	C	D
1				
2	Salesman	Age	Nature of Degree	Annula sales in crores of RS.
3	1	30	1	2
4	2	35	1	3
5	3	25	0	5
6	4	40	0	3
7	5	38	1	9
8	6	27	1	2
9	7	42	1	6
10	8	43	0	5
11	9	24	0	3
12	10	30	1	7
13				
14	AVERAGEIFS=			=AVERAGEIFS()

Function Arguments

AVERAGEIFS

Average_range = reference

Criteria_range1 = reference

=

Finds average(arithmetic mean) for the cells specified by a given set of conditions or criteria.

Average_range: are the actual cells to be used to find the average.

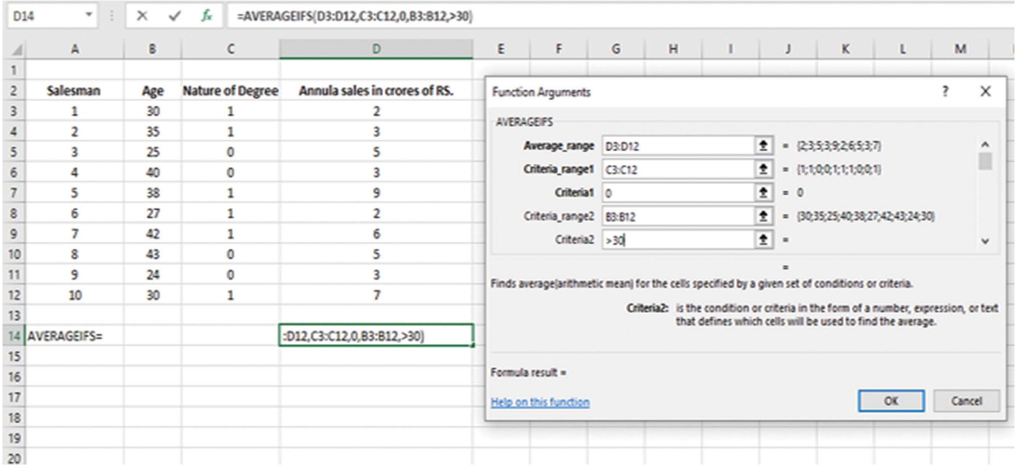
Formula result =

Help on this function

Figure 4.36 Screenshot after clicking the option AVERAGEIFS from the dropdown menu of Figure 4.35

Instead of following menu-driven option, one can enter a formula in a desired cell, say, C15, to get the AVERAGEIFS of a given set of observations as follows, if the data are stored from cell B4 to cell D13.

= AVERAGEIFS (Range of cells containing data for which the average is required, Range of cells containing data of Criterion 1, Criterion 1 data, Range of cells containing data of Criterion 2, Criterion 2 data)



D14 =AVERAGEIFS(D3:D12,C3:C12,0,B3:B12,>30)

	A	B	C	D
1				
2	Salesman	Age	Nature of Degree	Annula sales in crores of RS.
3	1	30	1	2
4	2	35	1	3
5	3	25	0	5
6	4	40	0	3
7	5	38	1	9
8	6	27	1	2
9	7	42	1	6
10	8	43	0	5
11	9	24	0	3
12	10	30	1	7
13				
14	AVERAGEIFS=			:D12,C3:C12,0,B3:B12,>30)

Function Arguments

AVERAGEIFS

Average_range D3:D12 = {2;3;5;3;9;2;6;5;3;7}

Criteria_range1 C3:C12 = {1;1;0;0;1;1;1;0;0;1}

Criteria1 0 = 0

Criteria_range2 B3:B12 = {30;35;25;40;38;27;42;43;24;30}

Criteria2 >30 =

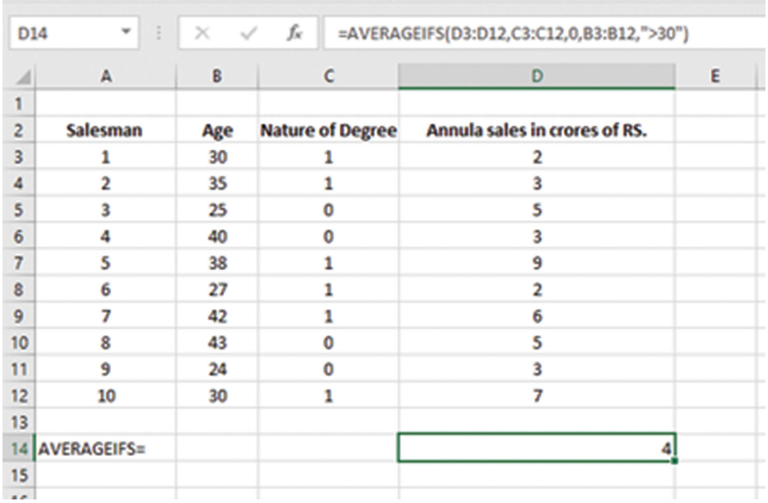
=

Finds average(arithmetic mean) for the cells specified by a given set of conditions or criteria.

Criteria2: is the condition or criteria in the form of a number, expression, or text that defines which cells will be used to find the average.

Formula result =

Figure 4.37 Screenshot after entering the range of cells against Average-range, range of cells for Criterion1, Criterion1 data, range of cells for Criterion2, and Criterion2 data



D14 =AVERAGEIFS(D3:D12,C3:C12,0,B3:B12,">30")

	A	B	C	D
1				
2	Salesman	Age	Nature of Degree	Annula sales in crores of RS.
3	1	30	1	2
4	2	35	1	3
5	3	25	0	5
6	4	40	0	3
7	5	38	1	9
8	6	27	1	2
9	7	42	1	6
10	8	43	0	5
11	9	24	0	3
12	10	30	1	7
13				
14	AVERAGEIFS=			4
15				

Figure 4.38 Screenshot after clicking OK in the dropdown menu of Figure 4.37 to show average per AVERAGIFS function

$$= \text{AVERAGEIFS}(\text{D3}:\text{D12}, \text{C3}:\text{C12}, 0, \text{B3}:\text{B12}, > 30)$$

One can verify that the required average is 4.

Example 4.7

Take a look at Table 4.12 for information on various companies' EPS, dividend percentage, and 52-week high share price for various years ending in March. Using Excel's

Table 4.12 EPS, Dividend %, and 52-Week Highest Share Price

Year Ending March	*EPS*	*Dividend %*	*52-Week Highest Share Price*
2016	20	21	200
2017	22	18	180
2018	18	30	300
2019	22	25	250
2020	20	22	350

AVERAGEIF function, get the mean of the 52-week high share prices for the condition that EPS is 20 and another condition that dividend percent is >20.

Solution

The data for Example 4.7 are shown in Table 4.13.

The screenshot after inputting the data in an Excel sheet is shown in Figure 4.39. Then button clicks Formulas ⟹ More Functions ⟹ Statistical ⟹ AVERAGEIFS give a screenshot, as shown in Figure 4.40.

The entries of the range of cells (D3:D7) for Average_Range, range of cells (B3:B7) for Criterion_Range1, Criteria1 data (20), Range of cells (C3:C7) for Criteria_Range2, and Criteria2 data (>20) are shown in Figure 4.41. Then, clicking the OK button in the dropdown menu of Figure 4.41 gives the result on the required average using the AVERAGEIFS function, as shown in Figure 4.42, which is ₹ 275.

The whole sequence of operations can be combined in a formula as follows to get the required average.

$$\text{Formula} := \text{AVERAGEIFS}(\text{D3} : \text{D7}, \text{B3} : \text{B7}, 20, \text{C3} : \text{C7}, ">20")$$

4.6.1 Weighted Average

In reality, a weighted average will be determined for the observations that pertain to an entity of study. As stated in Table 4.14, take into consideration a range of values for an interest variable and the accompanying weights.

The formula to compute the weighted average of the values of the variable X_i, where i varies from 1, 2, 3, . . . , n is as follows.

$$X_{wt} = \frac{\sum_{i=1}^{n} X_i W_i}{\sum_{i=1}^{n} W_i}$$

where

n is the number of observations

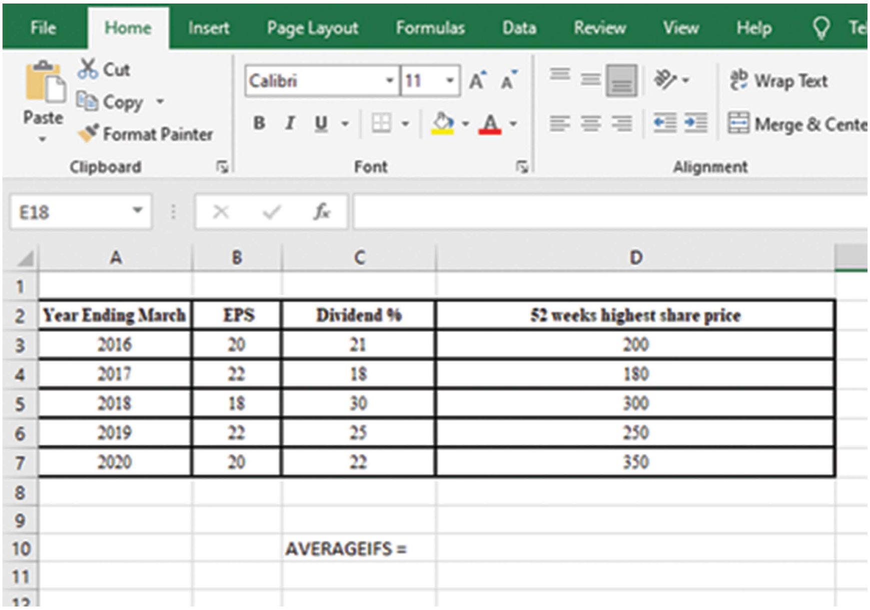
X_i is the value of the i^{th} observation

W_i is the weight of the i^{th} observation

One can verify the fact that the weighted average of the data given in Table 4.14 is 22.

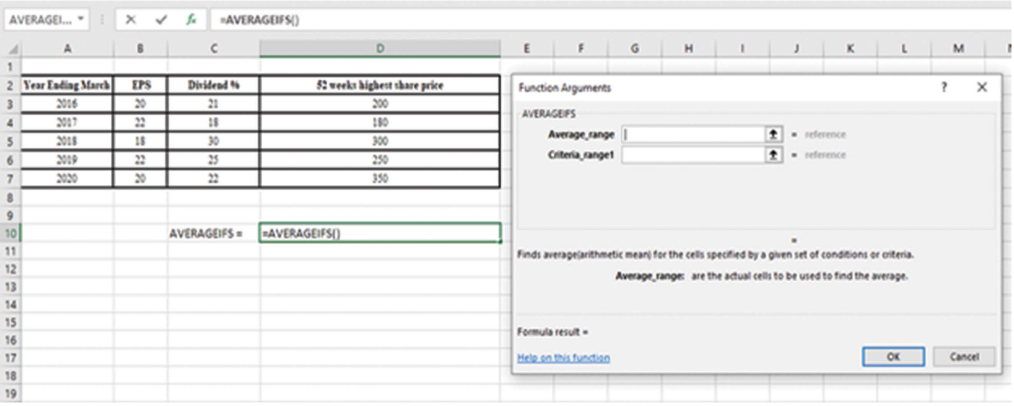
Table 4.13 Data for Example 4.7

Year Ending March	*EPS*	*Dividend %*	*52-Week Highest Share Price*
2016	20	21	200
2017	22	18	180
2018	18	30	300
2019	22	25	250
2020	20	22	350



	A	B	C	D
1				
2	Year Ending March	EPS	Dividend %	52 weeks highest share price
3	2016	20	21	200
4	2017	22	18	180
5	2018	18	30	300
6	2019	22	25	250
7	2020	20	22	350
8				
9				
10			AVERAGEIFS =	
11				

Figure 4.39 Screenshot after inputting the data in an Excel sheet



=AVERAGEIFS()

	A	B	C	D
1				
2	Year Ending March	EPS	Dividend %	52 weeks highest share price
3	2016	20	21	200
4	2017	22	18	180
5	2018	18	30	300
6	2019	22	25	250
7	2020	20	22	350
8				
9				
10			AVERAGEIFS =	=AVERAGEIFS()

Function Arguments

AVERAGEIFS

Average_range = reference

Criteria_range1 = reference

=

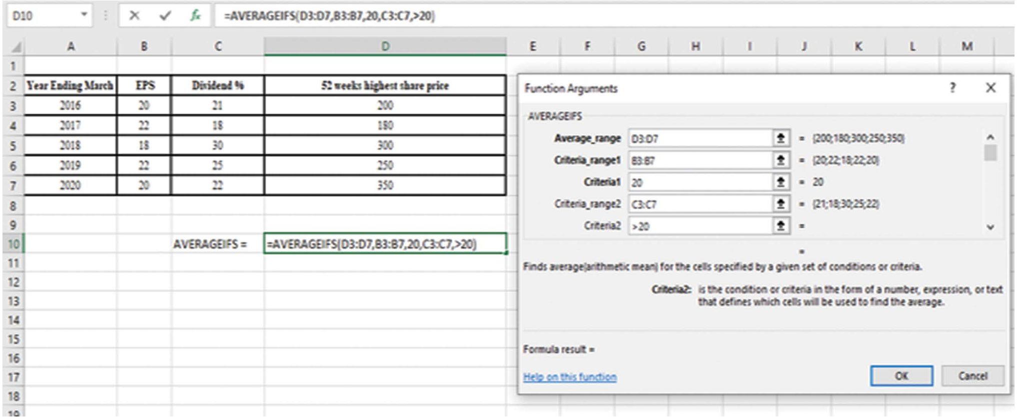
Finds average(arithmetic mean) for the cells specified by a given set of conditions or criteria.

Average_range: are the actual cells to be used to find the average.

Formula result =

Help on this function

Figure 4.40 Screenshot after clicking buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ AVERAGEIFS



D10

=AVERAGEIFS(D3:D7,B3:B7,20,C3:C7,>20)

Year Ending March	EPS	Dividend %	52 weeks highest share price
2016	20	21	200
2017	22	18	180
2018	18	30	300
2019	22	25	250
2020	20	22	350

AVERAGEIFS = =AVERAGEIFS(D3:D7,B3:B7,20,C3:C7,>20)

Function Arguments

AVERAGEIFS

Average_range	D3:D7	= {200;180;300;250;350}
Criteria_range1	B3:B7	= {20;22;18;22;20}
Criteria1	20	= 20
Criteria_range2	C3:C7	= {21;18;30;25;22}
Criteria2	>20	=

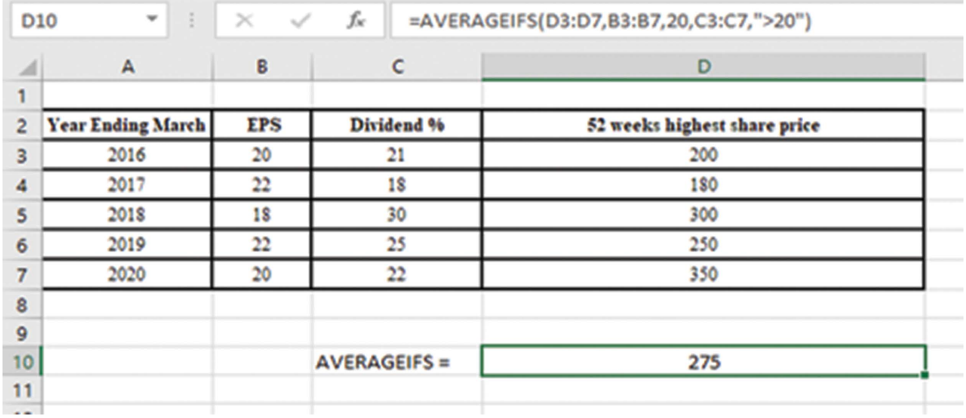
=

Finds average(arithmetic mean) for the cells specified by a given set of conditions or criteria.

Criteria2: is the condition or criteria in the form of a number, expression, or text that defines which cells will be used to find the average.

Formula result =

Figure 4.41 Screenshot after entering the range of cells against average-range, range of cells for Criterion1, Criterion1 data, range of cells for Criterion2, and Criterion2 data



D10

=AVERAGEIFS(D3:D7,B3:B7,20,C3:C7,">20")

Year Ending March	EPS	Dividend %	52 weeks highest share price
2016	20	21	200
2017	22	18	180
2018	18	30	300
2019	22	25	250
2020	20	22	350

AVERAGEIFS = 275

Figure 4.42 Screenshot after clicking OK in the dropdown menu of Figure 4.41 to show average per AVERAGIFS function

Table 4.14 Data for Variable and Weight

Variable (X_i)	Weight (W_i)
30	0.5
20	0.3
10	0.1

Example 4.8

The Beta Corporation has implemented a method of performance evaluation that rates each manager based on five key criteria. For each factor, the employees are graded on a 0–10 scale. The scheme of weights for the factors is as shown in Table 4.15.

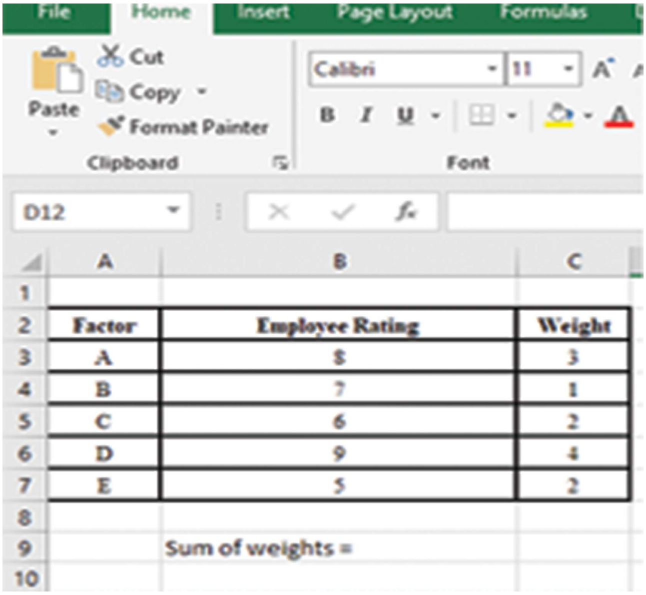
If an employee is rated with 8, 7, 6, 9, and 5 points for factors A, B, C, D, and E, respectively, find the overall weighted rating of that employee.

Table 4.15 Data for Example 4.8

Factor	Weight
A	3
B	1
C	2
D	4
E	2

Table 4.16 Data for Example 4.8

Factor	*Employee Rating*	*Weight*
A	8	3
B	7	1
C	6	2
D	9	4
E	5	2



Factor	Employee Rating	Weight
A	8	3
B	7	1
C	6	2
D	9	4
E	5	2

Sum of weights =

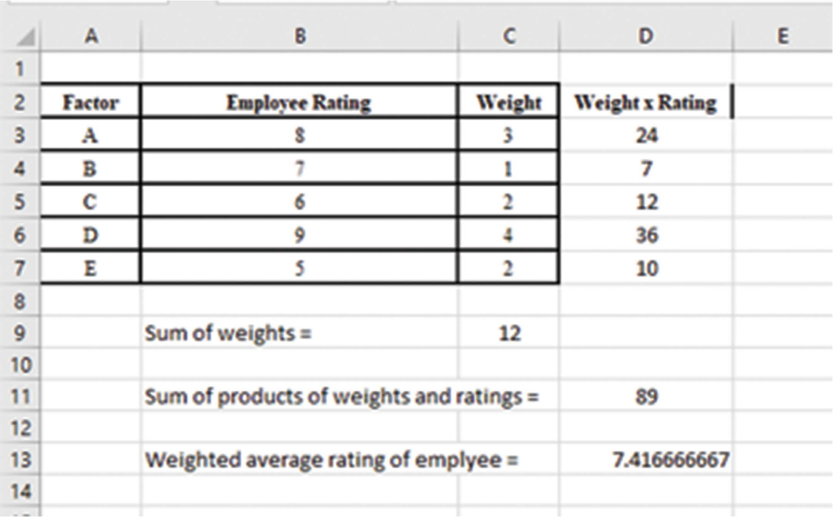
Figure 4.43 Screenshot of input of data in an Excel sheet for Example 4.8

Solution

The data for Example 4.8 along with the ratings of the employees are shown in Table 4.16.

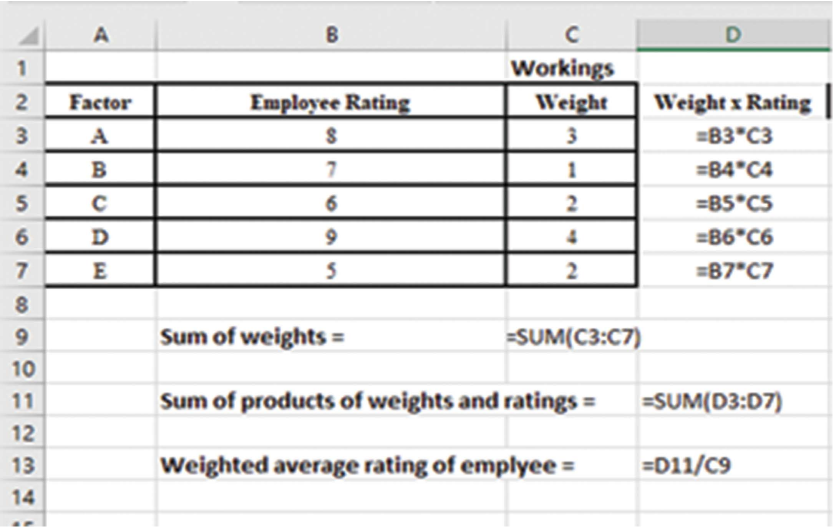
The input of the data in an Excel sheet is shown in Figure 4.43.

The working of the weighted average of the employee rating are shown in Figure 4.44. The corresponding formulas are displayed in Figure 4.45. The weighted average of the employee rating is 7.416667.



	A	B	C	D	E
1					
2	Factor	Employee Rating	Weight	Weight x Rating	
3	A	8	3	24	
4	B	7	1	7	
5	C	6	2	12	
6	D	9	4	36	
7	E	5	2	10	
8					
9		Sum of weights =	12		
10					
11		Sum of products of weights and ratings =		89	
12					
13		Weighted average rating of emplyee =		7.416666667	
14					

Figure 4.44 Screenshot of working of weighted average of employee rating



	A	B	C	D
1			Workings	
2	Factor	Employee Rating	Weight	Weight x Rating
3	A	8	3	=B3*C3
4	B	7	1	=B4*C4
5	C	6	2	=B5*C5
6	D	9	4	=B6*C6
7	E	5	2	=B7*C7
8				
9		Sum of weights =	=SUM(C3:C7)	
10				
11		Sum of products of weights and ratings =		=SUM(D3:D7)
12				
13		Weighted average rating of emplyee =		=D11/C9
14				

Figure 4.45 Screenshot of formulas for working of weighted average of employee rating

4.6.2 Arithmetic Mean of Grouped Data With Frequencies

If the data are in the form of interval data along with frequencies, as shown in Table 4.17, then an Excel sheet can be designed by the investigators themselves.

The mid-point of each class interval is computed using the following formula, as shown in Table 4.18.

$$X\,mid_i = \frac{(X_i + X_{i+1})}{2}$$

Table 4.17 Interval Data With Frequencies

Data Item Interval	*Frequency (f_i)*
X_0 to X_1	f_1
X_1 to X_2	f_2
X_2 to X_3	f_3
.	.
.	.
X_{i-1} to X_i	f_i
.	.
X_{n-1} to X_n.	f_n

Table 4.18 Interval Data With Mid-Points of Class Intervals

Data Item Interval	*Midpoint of Class Interval*	*Frequency (f_i)*
X_0 to X_1	$Y_1 = \frac{(X_0 + X_1)}{2}$	f_1
X_1 to X_2	$Y_2 = \frac{(X_1 + X_2)}{2}$	f_2
X_2 to X_3	$Y_3 = \frac{(X_2 + X_3)}{2}$	f_3
.	.	.
.	.	.
X_{i-1} to X_i	$Y_i = \frac{(X_{i-1} + X_i)}{2}$	f_i
.	.	.
X_{n-1} to X_n.	$Y_n = \frac{(X_{n-1} + X_n)}{2}$	f_n
Total frequency		$N = f_1 + f_2 + f_3 + \ldots + f_i +,,, + f_n$

The sum of all the frequencies (N) of the data items is computed using the following formula.

$N = \sum_{i=1}^{n} f_i$, f_i is the frequency of the ith class interval and n is the number of class intervals

The arithmetic mean of the grouped data items is computed using the following formula.

$$Arithmetic mean, \overline{Y} = \left[\frac{\sum_{i=1}^{n} (f_i Y_i)}{\sum_{i=1}^{n} f_i} \right] \times C$$

where
f_i is the frequency of the i^{th} class interval
Y_i is the mid-point of the i^{th} class interval
C is the width of the class interval
n is the total number of class intervals
$\bar{Y}$ is the arithmetic mean

Example 4.9

The lowest and highest monthly salary offered to employees of a computer company are ₹ 40,000 and ₹ 3,60,000, respectively. As indicated in Table 4.19, the frequency distribution of the firm's employees' monthly salaries is provided. Determine the arithmetic mean of the monthly salaries of the employees.

Solution

The data for Example 4.9 are shown in modified form in Table 4.20.

The arithmetic mean of the grouped data items is computed using the following formula.

$$Arithmetic mean, \bar{Y} = \left[\frac{\sum_{i=1}^{n}(f_i Y_i)}{\sum_{i=1}^{n} f_i}\right] \times C$$

Table 4.19 Frequency Distribution of Monthly Salaries of Employees

Monthly Salary (₹)	*Number of Employees*
40000 to 80000	150
80000 to 120000	250
120000 to 160000	275
160000 to 200000	350
200000 to 240000	200
240000 to 280000	175
280000 to 320000	100
320000 to 360000	50

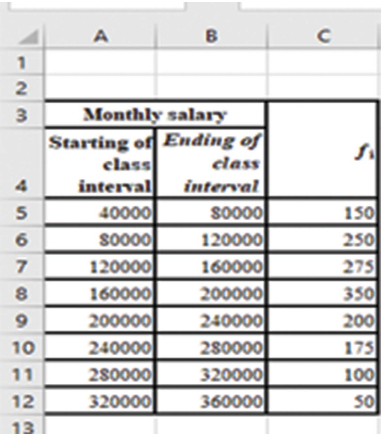
Table 4.20 Data for Example 4.9 in Modified Form

Monthly Salary		f_i
Start of Class Interval	*End of Class Interval*	
40000	80000	150
80000	120000	250
120000	160000	275
160000	200000	350
200000	240000	200
240000	280000	175
280000	320000	100
320000	360000	50

where

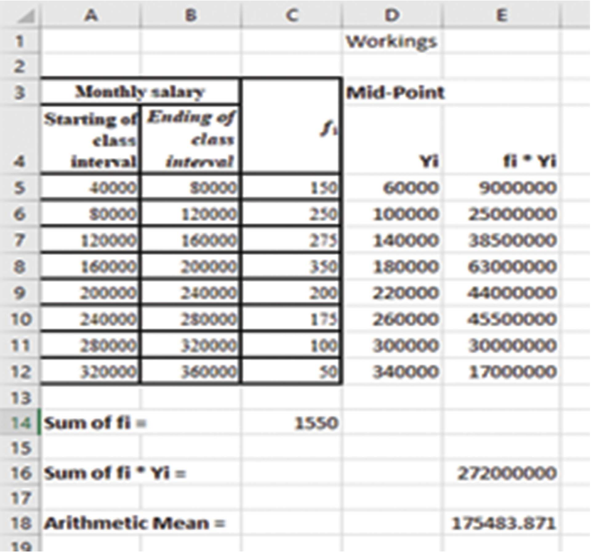
f_i is the frequency of the i^{th} class interval
Y_i is the mid-point of the i^{th} class interval
C is the width of the class interval
n is the total number of class intervals
$\bar{Y}$ is the arithmetic mean of the grouped data items

The screenshot of the input for the example in an Excel sheet is shown in Figure 4.46. The screenshot of an Excel worksheet to obtain the arithmetic mean of this example with grouped data is shown in Figure 4.47. The formulas for the Excel work to obtain the arithmetic mean of this example with grouped data are shown in Figure 4.48. The arithmetic mean is ₹ 1,75,483.871.



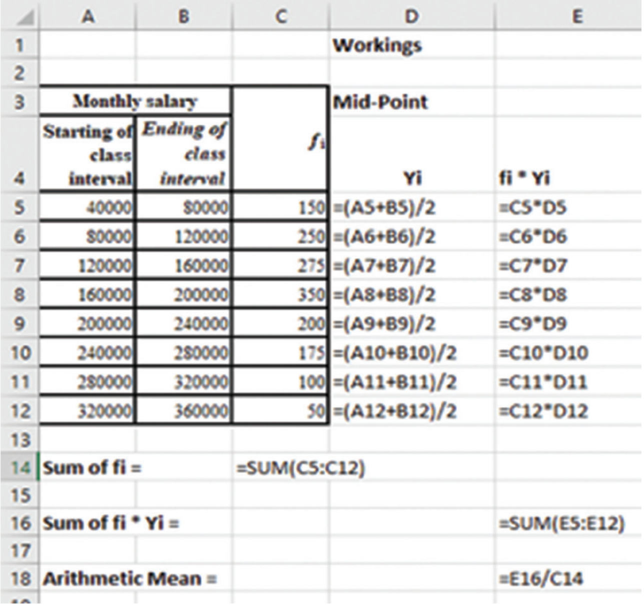
	A	B	C
1			
2			
3	**Monthly salary**		
4	**Starting of class interval**	***Ending of class interval***	f_i
5	40000	80000	150
6	80000	120000	250
7	120000	160000	275
8	160000	200000	350
9	200000	240000	200
10	240000	280000	175
11	280000	320000	100
12	320000	360000	50
13			

Figure 4.46 Screenshot of input of Example 4.9



	A	B	C	D	E
1				Workings	
2					
3	**Monthly salary**			**Mid-Point**	
4	**Starting of class interval**	***Ending of class interval***	f_i	**Yi**	**fi * Yi**
5	40000	80000	150	60000	9000000
6	80000	120000	250	100000	25000000
7	120000	160000	275	140000	38500000
8	160000	200000	350	180000	63000000
9	200000	240000	200	220000	44000000
10	240000	280000	175	260000	45500000
11	280000	320000	100	300000	30000000
12	320000	360000	50	340000	17000000
13					
14	**Sum of fi =**		**1550**		
15					
16	**Sum of fi * Yi =**				**272000000**
17					
18	**Arithmetic Mean =**				**175483.871**
19					

Figure 4.47 Screenshot of working of arithmetic mean of Example 4.9 with grouped data



	A	B	C	D	E
1				Workings	
2					
3	Monthly salary			Mid-Point	
4	Starting of class interval	*Ending of class interval*	f_i	Yi	fi * Yi
5	40000	80000	150	=(A5+B5)/2	=C5*D5
6	80000	120000	250	=(A6+B6)/2	=C6*D6
7	120000	160000	275	=(A7+B7)/2	=C7*D7
8	160000	200000	350	=(A8+B8)/2	=C8*D8
9	200000	240000	200	=(A9+B9)/2	=C9*D9
10	240000	280000	175	=(A10+B10)/2	=C10*D10
11	280000	320000	100	=(A11+B11)/2	=C11*D11
12	320000	360000	50	=(A12+B12)/2	=C12*D12
13					
14	Sum of fi =		=SUM(C5:C12)		
15					
16	Sum of fi * Yi =				=SUM(E5:E12)
17					
18	Arithmetic Mean =				=E16/C14

Figure 4.48 Screenshot of formulas for working of arithmetic mean of Example 4.9 with grouped data

Summary

- The Excel formula for average is: =AVERAGE(Range of cells) or = AVERAGE(Selected cells)
- The Excel formula for average deviation is: =AVEDEV(Range of cells) or = AVEDEV(Selected cells)
- The Excel formula for the average with numeric/logic (TRUE/FALSE) is: =AVERAGEA(Range of cells) or = AVERAGEA(Selected cells)
- The formula for AVERAGEIF function to find the average of a set of numeric observations stored in a range of cells for a given criterion is: =AVERAGEIF(Range, Criteria, Average Range)
- The AVERAGEIFS function finds the average of a set of numeric data for two criteria, whose formula is:

  =AVERGAEIFS (Range of cells containing data for which their average is required, Range of cells containing data of Criterion 1, Criterion 1 data, Range of cells containing data of Criterion 2, Criterion 2 data, etc.)

  - The formula to compute the weighted average of the values of the variable X_i, where i varies from 1, 2, 3, . . . , n is as follows.

$$X_{wt} = \frac{\sum_{i=1}^{n} X_i W_i}{\sum_{i=1}^{n} W_i}$$

- The arithmetic mean of the grouped data items is computed using the following formula.

$$Arithmetic mean, Y = \frac{\sum_{i=1}^{n} f_i Y_i}{\sum_{i=1}^{n} f_i} \times C$$

Keywords

The AVERAGE function determines the arithmetic mean of a set of observations.

The AVEDEV function finds the average of the absolute deviations of observations (X_i, i = 1, 2, 3, . . ., *n*, where *n* is the number of observations) from the arithmetic mean. Alternatively, this may be called the mean absolute deviation of a set of observations.

The AVERAGEA function finds the average of a desired set of data, that is, numeric, Text, False as 0 and TRUE as 1.

The AVERAGEIF function finds the average of a set of numeric observations stored in a range of cells for a criterion.

The AVERAGEIFS function finds the average of a set of observations stored in a range of cells for two criteria.

Weighted average is computed for the data in reality with certain weights.

Arithmetic mean of grouped data uses mid-points of intervals and the corresponding frequencies to estimate it.

Review Questions

1. Give the syntax of the AVERAGE function in Excel.
2. Illustrate the use of the AVERAGE function through button clicks from HOME using an example.
3. The quarterly sales of a product for the last two years are summarised in the following table. Find the mean quarterly sales of the product using the AVERAGE function in Excel.

Quarter	*Quarterly Sales in Crores of Rupees*
1	200
2	220
3	150
4	180
5	290
6	310
7	220
8	300

4. a. Give the mathematical formula of the AVEDEV function in Excel.
   b. Illustrate the use of the AVEDEV function through button clicks from HOME using an example.

5. Take a look at the information in the following table, which lists the number of degrees each person in a particular department of an IT company has earned. Using the AVEDEV function in Excel, get the average deviation of the number of degrees the employees hold.

Employee Code	*Employee Name*	*Number of Degrees Held*
1	Arthi, N	3
2	Balu, K	1
3	Baskaran, G	3
4	Domnic, H	4
5	Elango, J	2
6	Fathima, J	1
7	Gopi, K	3
8	Hendry, D	2
9	Jeyam, D	4
10	Lenin, K	3
11	Sachin, G	3
12	Yogesh, B	2

6. a. What is the purpose of the AVERAGEA function in Excel?
   b. Give the syntax of the AVERAGEA function in Excel.

7. Numerous research initiatives are carried out in a research institution. As stated in the accompanying table, the outcome of each of these initiatives may be either success (with a value of 1) or failure (with a value of 0). Explain how to use Excel's AVERAGEA function to determine the average of these numbers.

Research Project	*Outcome*
1	Success (1)
2	Success (1)
3	Failure (0)
4	Success (1)
5	Failure (0)
6	Success (1)
7	Failure (0)
8	Success (1)

8. a. Explain the purpose of the AVERAGEIF function in Excel.
   b. Give the syntax of the AVERAGEIF function in Excel.

9. Take a look at the data from various engineering college years as displayed in the following table. Find the average CTC for the condition Pass percentage is more than 90% and Placement percentage is more than 95% using Excel's AVERAGEIF function.

Data for Pass Percentage, Placement Percentage and CTC

Year Ending March	*Pass Percentage*	*Placement Percentage*	*Highest CTC (Cost to the Company) Adjusted for Inflation in Lakhs of ₹*
2014	90	95	9
2015	92	93	8
2016	95	99	10
2017	88	89	7
2018	98	100	11
2019	89	89	9
2020	97	98	10

10. What is weighted average? Illustrate the working of the weighted average using an example.
11. Alpha Business School has the following grading system for the subjects offered in that school.

Grade	*Weight*
A	4
B	3
C	2
D	1
E	0

In a semester, a student has to take eight subjects. The grade point average (GPA) of the student in that semester is computed using the following formula.

$$GPA = \frac{\sum_{i=1}^{8} \text{Grade equivalent of subject} \, i \times \text{Number of credits of subject} \, i}{\sum_{i=1}^{8} \text{Number of credits of subject} \, i}$$

The Subject Code and Grade Obtained of eight subjects are shown in the next table.

Subject Code	101	103	105	110	112	120	135	150
Grade Obtained	A	C	B	E	B	A	D	A

Find the GPA of the students in the semester using Excel.

12. The age range for employees in a company is 20 years old at the lowest and 60 years old at the highest. The following table provides the frequency distribution of the employee population's ages. Determine the arithmetic mean of the age of the employees using Excel.

Distribution of Employee Ages

Age Interval (Year)	*Number of Employees*
20–25	175
25–30	225
30–35	300
35–40	325
40–45	225
45–50	150
50–55	100
55–60	50

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. https://Exceljet.net/Excel-functions/Excel-average-function [June 25, 2020].
3. https://Exceljet.net/Excel-functions/Excel-averageifs-function [July 3, 2020].

5 Median and Mode

Learning Objectives

After reading this chapter, you will be able to

- Know the concept of median, which is the middlemost observation of a given set of data.
- Understand the method of computing the median of ungrouped data.
- Determine the median of grouped data.
- Understand the concept of mode of a given set of data (observations), which has the maximum frequency.
- Compute the mode of ungrouped data.
- Determine the mode of grouped data.
- Analyse percentile of a given set of data.
- Understand quartile and its classification.

5.1 Introduction

This chapter covers all the essential details and examples of the median, mode, percentile, and quartile. The median for ungrouped data and the median for grouped data are further distinguished in the section on the median. The mode portion is further divided into two categories: mode for grouped data and mode for ungrouped data. Excel worksheets are used to illustrate each of these.

If the data repeat, it is possible to create a frequency distribution for the observations of the data. Excel automatically creates this type of frequency distribution before determining the median and mode of the provided set of data; thus the investigator is not required to do so. However, if the total frequency is high, it will be difficult to input every instance in the Excel sheet. The investigator can use Excel to tackle this issue by creating a worksheet that is appropriate.

5.2 Median

A given set of data, such as the sales revenues of a corporation over the past several years, can be determined using the median function, which finds the middlemost observation in the set [1].

The actual data could either be ungrouped or grouped. As a result, this section is divided into subsections for the median of ungrouped and grouped data.

DOI: 10.4324/9781032671703-5

5.2.1 Median of Ungrouped Data Using Median Function

There are some real-life situations in which the instances of a variable mostly may not repeat. Hence, there may not be frequencies for such data. This sets an example of ungrouped data. Such data are arranged per their increasing order.

The formula to obtain the median of this type of data is as follows.

$$Median = X_p, \; if \;\; N \; is \; odd$$

$$= \frac{\left(X_{p_1} + X_{p_2}\right)}{2}, \; if \;\; N \; is \; even$$

where

N is the total number of observations of the data.

X_i is the i^{th} observation of a given data, where i = 1, 2, 3, . . . , N

p is the middlemost point of N observations.

$$p = \frac{(N-1)}{2} + 1, \; if \;\; N \; is \; odd$$

$$p_1 = \frac{(N)}{2} and \, p_2 = p_1 + 1, if \; N \, is \, even$$

The selection of the buttons in the sequence Formulas ⟹ More Functions ⟹ Statistical gives a display as shown in Figure 5.1 [2]. The screenshot after clicking the median option in the dropdown menu of Figure 5.1 is shown in Figure 5.2. If one copies the range of cells containing the data in the box against Number 1 in the dropdown menu of Figure 5.2 and clicks the OK button in that dropdown menu, the result of the median will be displayed in the cell where the cursor has been positioned originally in the Excel sheet. Number 1, Number 2, Number 3, . . . , and so on can be from 1 to 255 numbers, names, arrays, or references which contain the data for which the median estimate is required.

All these operations can be combined in a formula as follows.

$$\text{Formula} := \text{MEDIAN}(\text{Range of cells containing data})$$

If there are too many number of observations, then one can construct one's own Excel sheet, which will be explained at the end of this subsection.

Example 5.1

The sales revenue in crores of rupees of a company during the past 15 years is summarised in Table 5.1. Find the median sales revenue of the company using Excel.

Solution

The data for Example 5.1 are shown in Table 5.2.

The screenshot after inputting the data for this problem in an Excel sheet is shown in Figure 5.3.

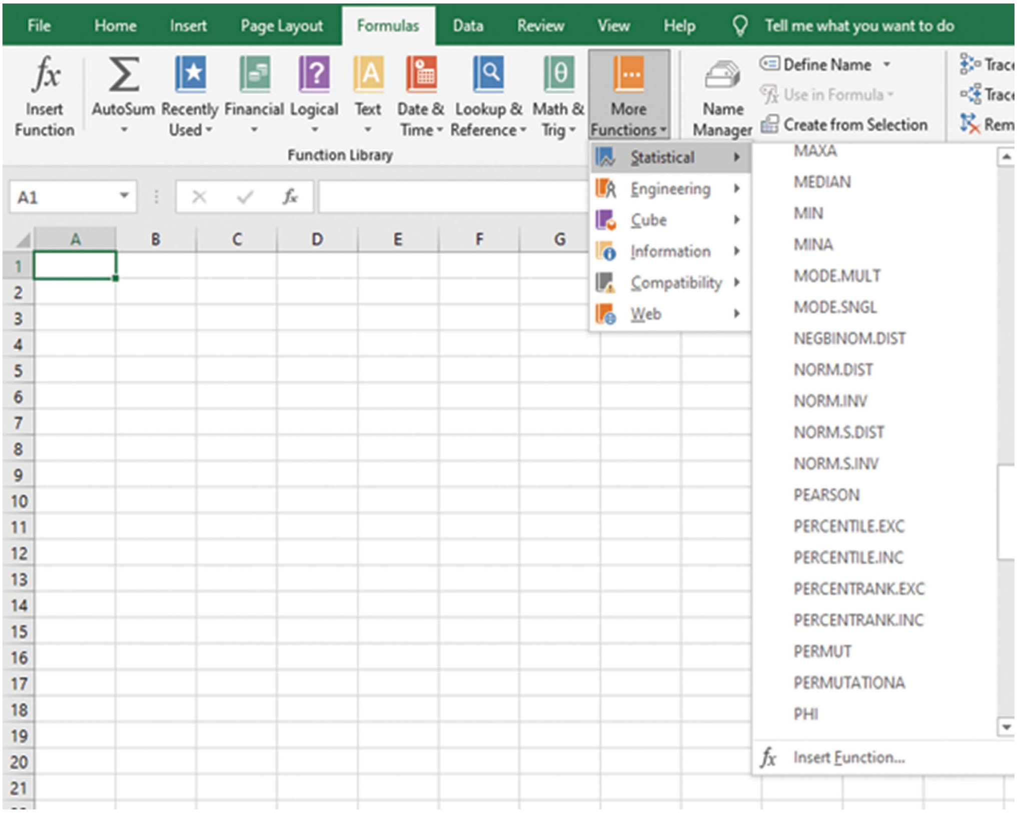
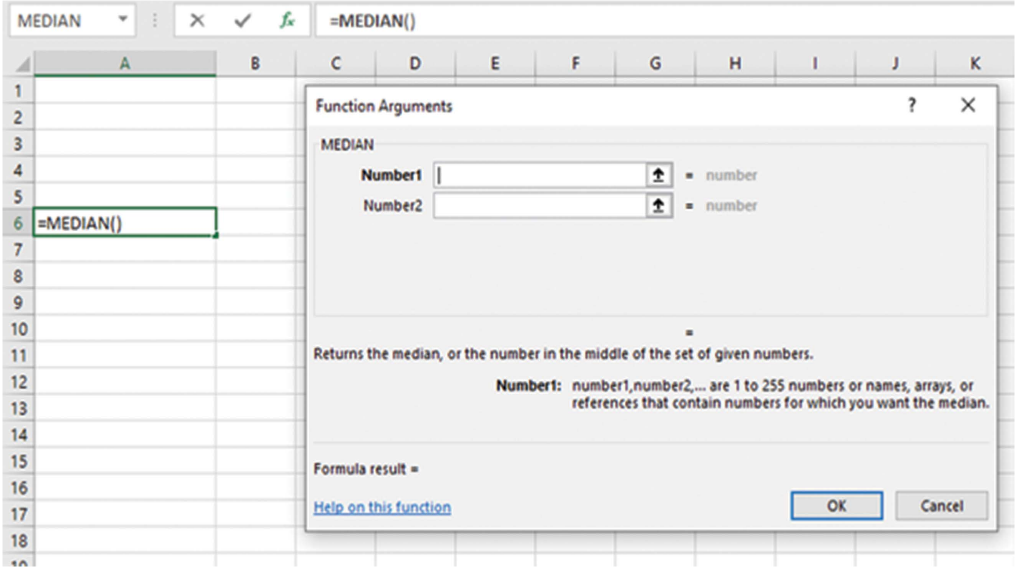


Figure 5.1 Screenshot after clicking buttons, Formulas ⟹ More Functions ⟹ Statistical


MEDIAN

=MEDIAN()

Function Arguments

MEDIAN

Number1 = number

Number2 = number

Returns the median, or the number in the middle of the set of given numbers.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the median.

Formula result =

Help on this function

OK Cancel

Figure 5.2 Screenshot after clicking the Median option in the dropdown menu of Figure 5.1

Table 5.1 Sales Revenues of Companies

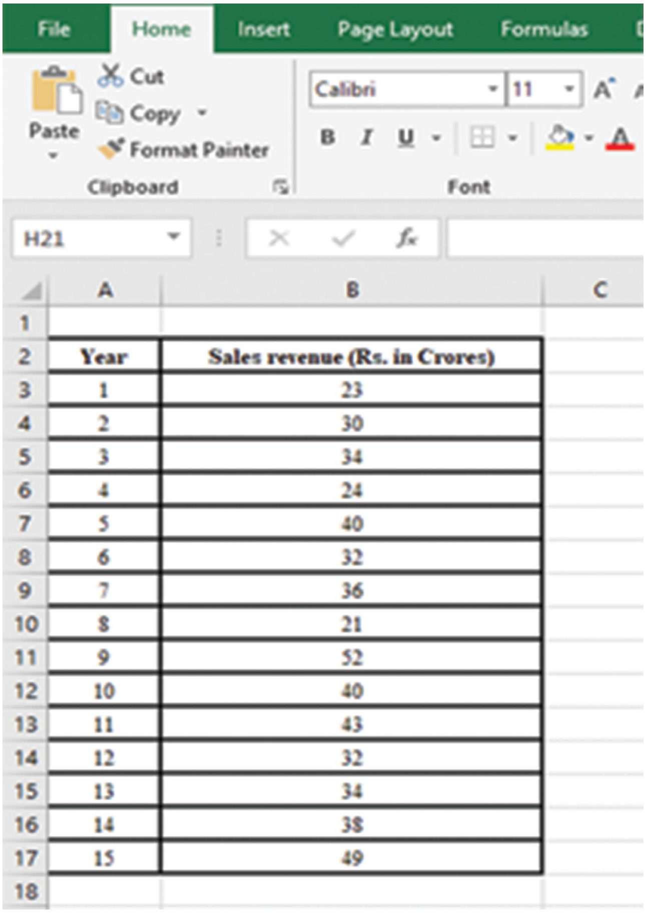
Year	*Sales Revenue (₹ in Crores)*
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49

Table 5.2 Data for Example 5.1

Year	*Sales Revenue (₹ in Crores)*
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49

The steps of finding the median of the given ungrouped data are presented in the following.

1. Clicking the buttons in the sequence Formulas ⟹ More Functions ⟹ Statistical ⟹ gives a display as in Figure 5.4.
2. Then, clicking the option Median in the dropdown menu of Figure 5.4 gives the display in Figure 5.5.
3. The display after copying the range of cells from B3 to B17 containing the data is shown in Figure 5.6.
4. Clicking OK in the dropdown menu of Figure 5.6 gives the result for the median, as in Figure 5.7, which is 34.



Year	Sales revenue (Rs. in Crores)
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49

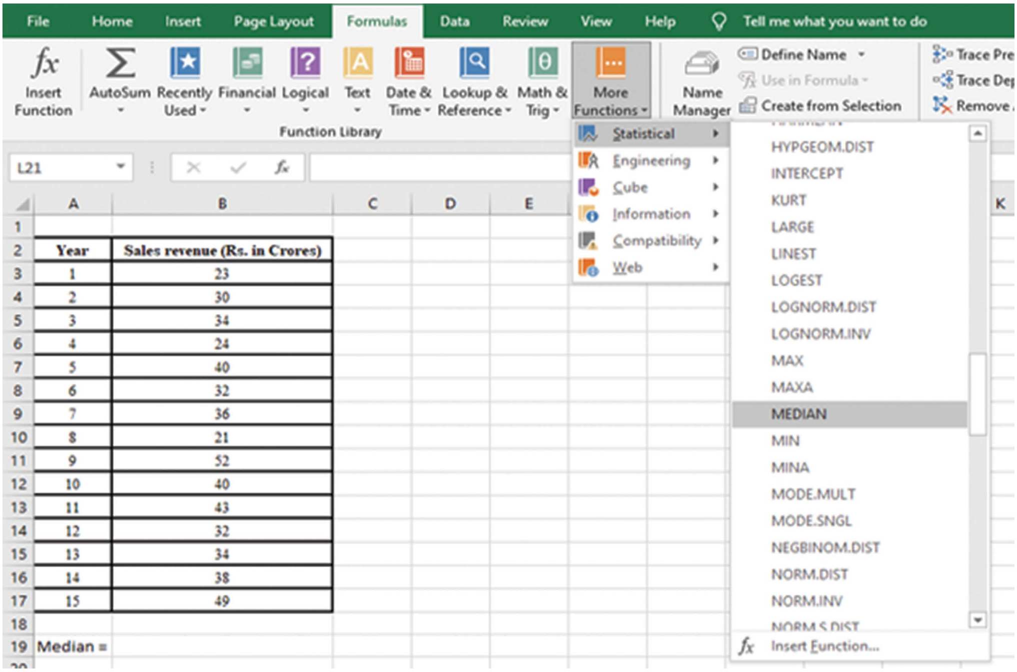
Figure 5.3 Screenshot after inputting data for Example 5.1

All these operations can be combined by entering the following formula in cell B19 to obtain the result for the median.

Formula : = MEDIAN(B3 : B17)

5.2.2 Median of Grouped Data With Frequencies Using Excel Sheets

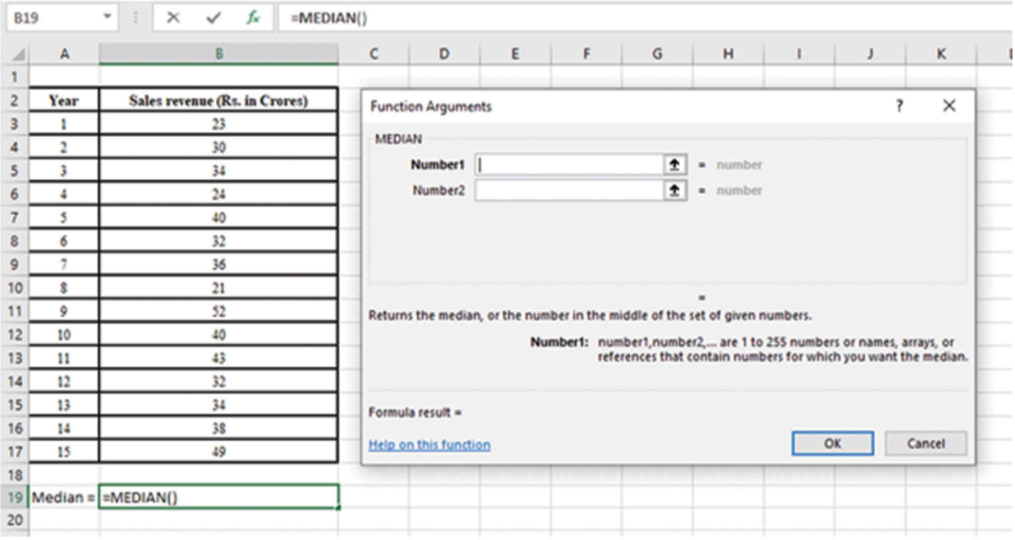
The investigators themselves can create an Excel sheet if the data are presented as interval data with frequencies, as illustrated in Table 5.3. The class intervals in the data could be open ended, which would indicate that neither the first-class interval's lower nor last class interval's higher limits would exist.



Year	Sales revenue (Rs. in Crores)
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49

Median =

Figure 5.4 Screenshot of button clicks in the sequence, Formulas ⟹ More Functions ⟹ Statistical



B19 =MEDIAN()

Year	Sales revenue (Rs. in Crores)
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49

Median = =MEDIAN()

Function Arguments

MEDIAN

Number1 = number

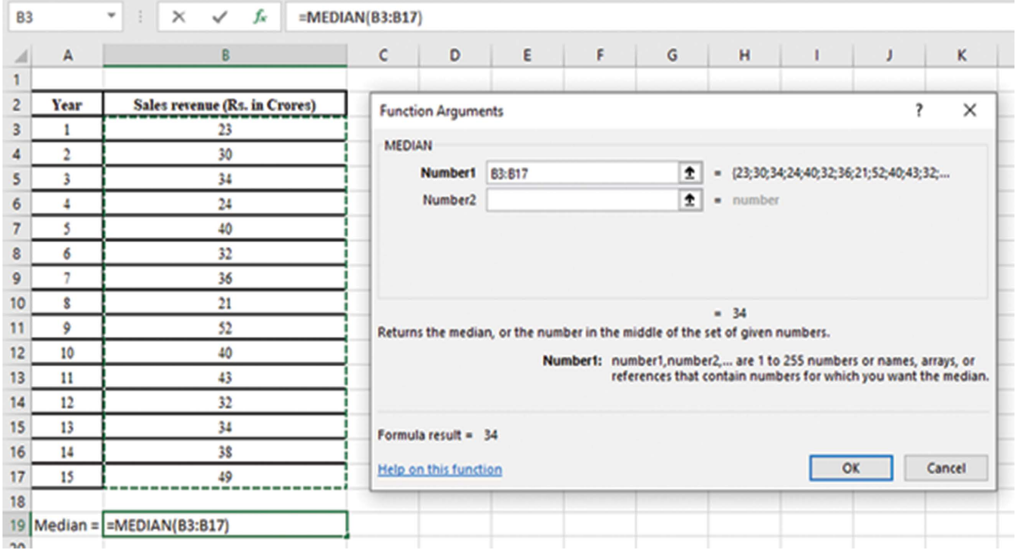
Number2 = number

Returns the median, or the number in the middle of the set of given numbers.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the median.

Formula result =

Figure 5.5 Screenshot of Median option in the dropdown menu of Figure 5.4



B3 =MEDIAN(B3:B17)

Year	Sales revenue (Rs. in Crores)
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49
Median =	=MEDIAN(B3:B17)

Function Arguments

MEDIAN

Number1 B3:B17 = {23;30;34;24;40;32;36;21;52;40;43;32;...

Number2 = number

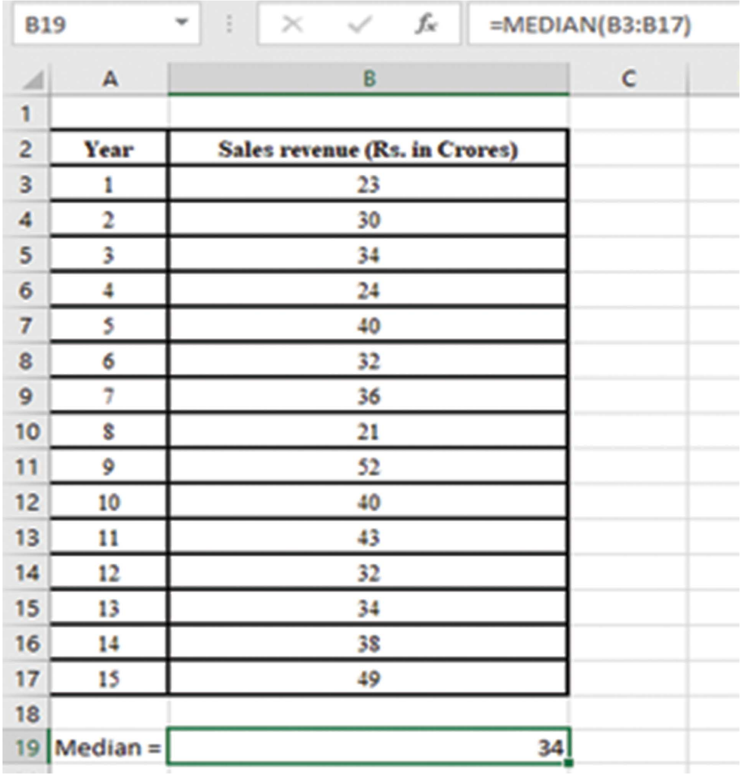
= 34

Returns the median, or the number in the middle of the set of given numbers.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the median.

Formula result = 34

Figure 5.6 Screenshot after copying the range of Cells from B3 to B17 in the box against Number 1 of the dropdown menu of Figure 5.5



B19 =MEDIAN(B3:B17)

Year	Sales revenue (Rs. in Crores)
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49
Median =	34

Figure 5.7 Screenshot after clicking OK in the dropdown menu of Figure 5.6 to show the result for the median

Table 5.3 Interval Data With Frequencies

Data Item Interval	*Frequency* (f_i)
Less than X_1	f_1
X_2–X_3	f_2
X_3–X_4	f_3
–	–
–	–
X_i–X_{i+1}	f_i
–	–
X_{n-1} to X_n	f_{n-1}
More than X_n	f_n

Table 5.4 Interval Data With Cumulative Frequencies

Data Item Interval	*Frequency* (f_i)	*Cumulative Frequency* (F_i)
Less than X_1	f_1	$F_1 = f_1$
X_2–X_3	f_2	$F_2 = f_1 + f_2$
X_3–X_4	f_3	$F_3 = f_1 + f_2 + f_3$
–	–	–
–	–	–
X_i–X_{i+1}	f_i	$F_{\mathrm{i}} = f_1 + f_2 + f_3 + \dots + f_i$
–	–	–
X_{n-1}–X_n	f_{n-1}	$F_{n\text{-}1} = f_1 + f_2 + f_3 + \dots + f_i + \dots + f_{n-1}$
More than X_n	f_n	$F_n = f_1 + f_2 + f_3 + \dots + f_i + \dots + f_{n-1} + f_n = N$

The cumulative frequency values of the data items can be computed as shown in Table 5.4. In the last row of Table 5.4, *N* represents the total frequency of the data items, which is given by the following formula.

$$N = F_n = \sum_{i=1}^{n} f_i$$

where
n is the total number of class intervals
f_i is the frequency of the i^{th} class interval
F_n is the cumulative frequency of the n^{th} class interval

The median of the data items is the value of the data item, which corresponds to 50% of the total frequency, which is *N*/2, where *N* is the total frequency.

The formula for the median of the grouped data is as follows.

$$\text{Median} = X_k + \left(\frac{\dfrac{N}{2} - F_k}{f_k} \right) \times C$$

where
N is the total frequency
k is the interval (class) corresponding to 50% of the total frequency
X_k is the lower limit of the data item of the k^{th} interval (median class)
f_k is the frequency of the k^{th} interval (median class)
F_k is the cumulative frequency of the previous interval with respect to the median class
C is the length of the class interval

Example 5.2

Based on a survey, the distribution of the number of years of usage of a particular brand of washing machine by its first buyers is shown in Table 5.5.

Find the median of the number of years of usage of the washing machine by its first buyers.

Solution

The data for Example 5.2 are shown in Table 5.6.

The formula for the median of the grouped data is as follows.

$$\text{Median} = X_k + \left(\frac{\frac{N}{2} - F_k}{f_k} \right) \times C$$

Table 5.5 Distribution of Number of Years of Usage of Washing Machine by Its First Buyers

No. of Years of Usage	*No. of Respondents*
Less than 2	10
2–4	15
4–6	20
6–8	25
8–10	20
10–12	12
More than 12	10

Table 5.6 Data for Example 5.2

No. of Years of Usage	*No. of Respondents*
Less than 2	10
2–4	15
4–6	20
6–8	25
8–10	20
10–12	12
More than 12	10

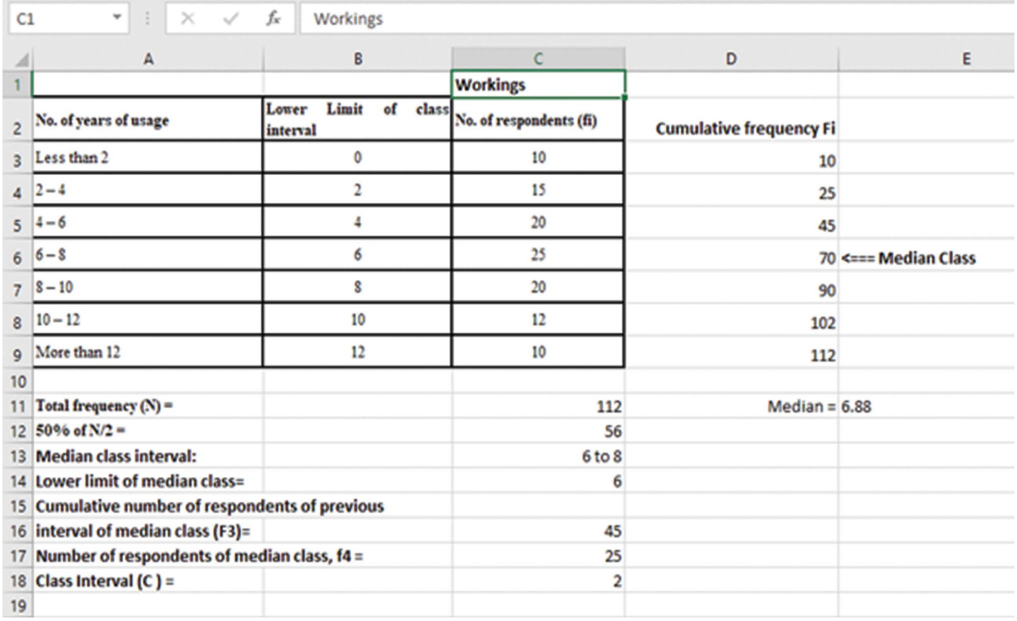
where
N is the total frequency
k is the interval (class) corresponding to 50% of the total frequency
X_k is the lower limit of the data item of the k^{th} interval (median class)
f_k is the frequency of the k^{th} interval (median class)
F_k is the cumulative frequency of the previous interval with respect to the median class
C is the length of the class interval

The screenshot of the input of the given problem and the working to obtain the median of the grouped data are shown in Figure 5.8, and the corresponding formulas are shown in Figure 5.9. The median of the number of years of usage of the washing machine by its first buyers is 6.88 years.

5.3 Mode

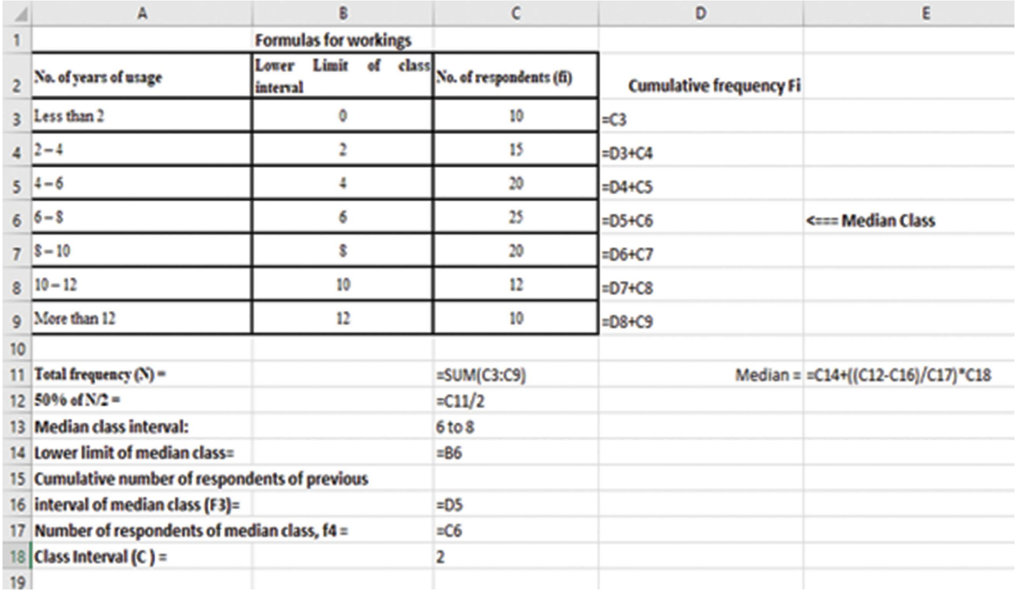
The element of a set of data (observations) that occurs most frequently is known as the mode [1]. This is a type of central tendency measure. Take the annual sales revenues of a company in an industrial estate as an example. More than one company may have the same yearly sales revenue, which results in the construction of a frequency distribution for the annual sales revenue. Finding the annual sales income whose number of occurrences is highest when compared to the occurrences of the other annual sales revenues is the objective at this point. The yearly sales income that occurs the most frequently is referred to as the mode of the annual sales revenue.

Consider the annual sales revenues of the firms in an industrial estate, as shown in Table 5.7.



	A	B	C	D	E
1			Workings		
2	No. of years of usage	Lower Limit of class interval	No. of respondents (fi)	Cumulative frequency Fi	
3	Less than 2	0	10	10	
4	2 – 4	2	15	25	
5	4 – 6	4	20	45	
6	6 – 8	6	25	70	<=== Median Class
7	8 – 10	8	20	90	
8	10 – 12	10	12	102	
9	More than 12	12	10	112	
10					
11	Total frequency (N) =		112	Median =	6.88
12	50% of N/2 =		56		
13	Median class interval:		6 to 8		
14	Lower limit of median class=		6		
15	Cumulative number of respondents of previous				
16	interval of median class (F3)=		45		
17	Number of respondents of median class, f4 =		25		
18	Class Interval (C) =		2		
19					

Figure 5.8 Screenshot of working to obtain median of number of years of usage of washing machine



	A	B	C	D	E
1		Formulas for workings			
2	No. of years of usage	Lower Limit of class interval	No. of respondents (fi)	Cumulative frequency Fi	
3	Less than 2	0	10	=C3	
4	2 – 4	2	15	=D3+C4	
5	4 – 6	4	20	=D4+C5	
6	6 – 8	6	25	=D5+C6	<=== Median Class
7	8 – 10	8	20	=D6+C7	
8	10 – 12	10	12	=D7+C8	
9	More than 12	12	10	=D8+C9	
10					
11	Total frequency (N) =		=SUM(C3:C9)	Median =	=C14+((C12-C16)/C17)*C18
12	50% of N/2 =		=C11/2		
13	Median class interval:		6 to 8		
14	Lower limit of median class=		=B6		
15	Cumulative number of respondents of previous				
16	interval of median class (F3)=		=D5		
17	Number of respondents of median class, f4 =		=C6		
18	Class Interval (C) =		2		
19					

Figure 5.9 Screenshot of formulas for the working to obtain median of number of years of usage of washing machine

Table 5.7 Data on Annual Sales of Firms

Annual Sales Revenue (Crores of Rupees; X_i)	*Number of Firms (f_i)*
500	12
600	16
700	24
800	36
900	28
1,000	22
1,100	10

Maximum frequency

From Table 5.7, the maximum frequency is for the annual sales revenue of ₹ 800 crores. Hence, the mode of the annual sales revenues of the firms is ₹ 800 crores. This is given by the following formula.

Mode = X_i for which f_i is maximum

The classifications of mode are as listed.

- Mode of ungrouped data
- Mode of grouped data

5.3.1 Mode of Ungrouped Data Using Mode Function

This section presents the determination of the mode of ungroup data. The determination of the mode of a given set of data can be obtained using Excel.

The mode may be single mode or multi-mode, which are explained in the following.

- Single mode means that there is only one unique observation, which has the highest frequency in the data set.
- Multi-mode means that more than one observation will have the maximum frequency in a given set of data.

There are two types of commands in Excel to find the mode(s): single mode and multi-mode of a given set of values, which are as listed in the following [3].

= MODE.SNGL (Range of cells containing the data items for which mode is to be found)

= MODE.MULT (Range of cells containing the data items for which several modes, if they exist, are to be found)

The screenshot after clicking the sequence of buttons Formulas ⟹ More Functions ⟹ Statistical is shown in Figure 5.10. Clicking the option MODE.SNGL from Figure 5.10 gives a display as shown in Figure 5.11. If one copies the range of cells containing the data in the box against Number 1 in the dropdown menu of Figure 5.11 and clicks the OK button in its dropdown menu, the result of the mode will be displayed in the cell where the cursor has been positioned originally in the Excel sheet. Number 1, Number 2, Number 3, . . . , and so on can be from 1 to 255 numbers, names, arrays, or references which contain the data for which the mode estimate is required.

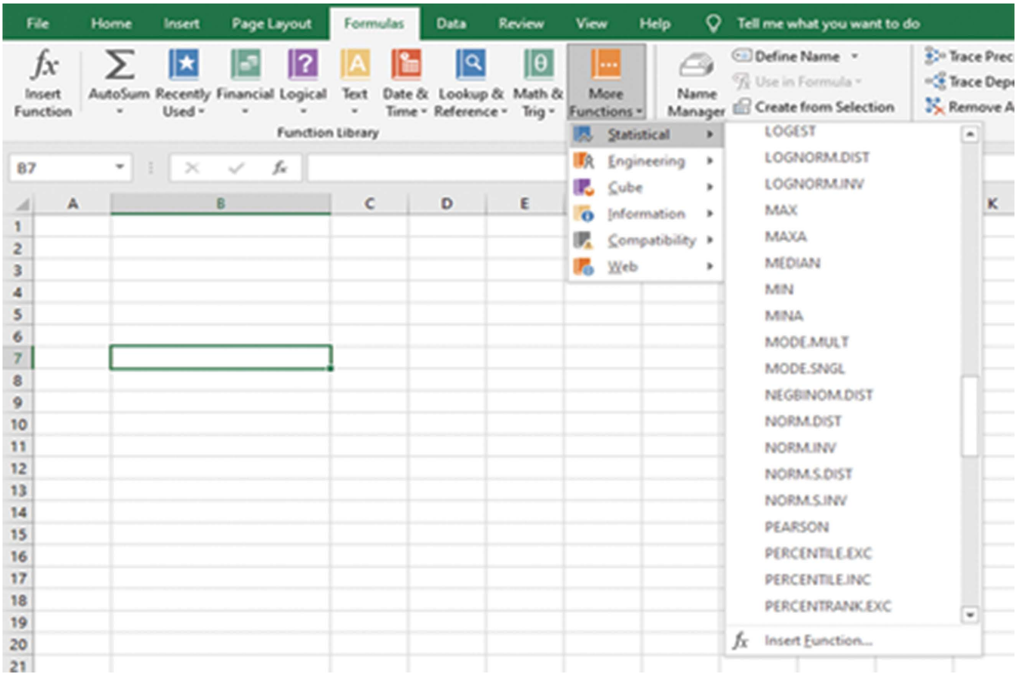
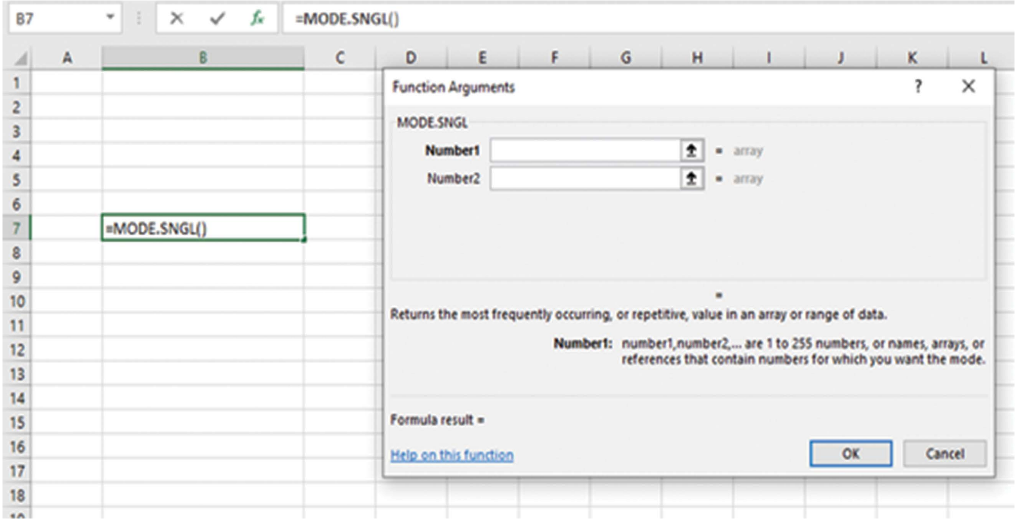


Figure 5.10 Screenshot of clicking the sequence of buttons, Formulas ⟹ More Functions ⟹ Statistical



B7

=MODE.SNGL()

Function Arguments

MODE.SNGL

Number1 = array

Number2 = array

Returns the most frequently occurring, or repetitive, value in an array or range of data.

Number1: number1,number2,... are 1 to 255 numbers, or names, arrays, or references that contain numbers for which you want the mode.

Formula result =

Figure 5.11 Screenshot after clicking the option MODE.SNGL from the dropdown menu of Figure 5.10

The data shown in Table 5.7 are with single mode. Hence, all these operations can be combined in a formula as follows to compute the single mode of the given set of data items.

Formula : = MODE.SNGL (Range of cells containing the data items for which mode is to be found)

A given set of data may be multi-mode. In such a case, the steps to obtain the multi-modes are as follows.

Step 1: Select a vertical column with a sufficient number of cells to display multiple modes if they exist.
Step 2: Click the sequence of buttons Formulas $\Longrightarrow$ More Functions $\Longrightarrow$ Statistical $\Longrightarrow$ MODE.MULT to show the display in Figure 5.12.
Step 3: Copy the range of cells which contains the data in the box against Number 1.
Step 4: Simultaneously press the Shift key, Ctrl key, and Enter key to show the display of multiple modes in the vertical column selected in Step 1, if they exist.

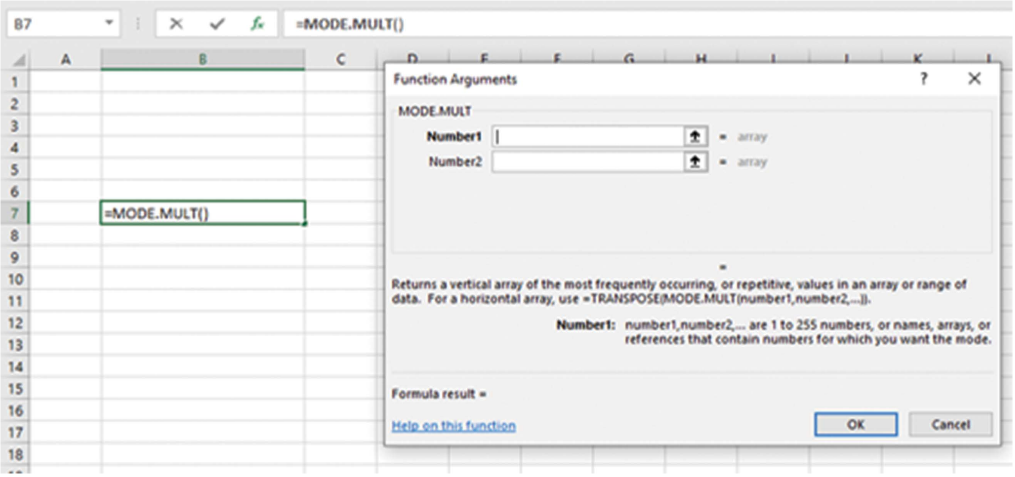
Example 5.3

The annual sales revenues of the firms functioning in an industrial estate are summarised in Table 5.8. Find the mode of the annual sales revenues of the firms using Excel.

Solution

The data for Example 5.3 are shown in Table 5.9.

- The screenshot after inputting the annual sales revenues of the firms in an Excel sheet is shown in Figure 5.13.



Returns a vertical array of the most frequently occurring, or repetitive, values in an array or range of data. For a horizontal array, use =TRANSPOSE(MODE.MULT(number1,number2,...)).

Number1: number1,number2,... are 1 to 255 numbers, or names, arrays, or references that contain numbers for which you want the mode.

Figure 5.12 Screenshot after clicking buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ MODE.MULT

Table 5.8 Annual Sales Revenues of Firms

Firms	*Annual Revenue (Crores of Rupees)*
1	25
2	30
3	28
4	30
5	29
6	28
7	40
8	25
9	45
10	28
11	30
12	45
13	29
14	28
15	25
16	45
17	30
18	28
19	40
20	28

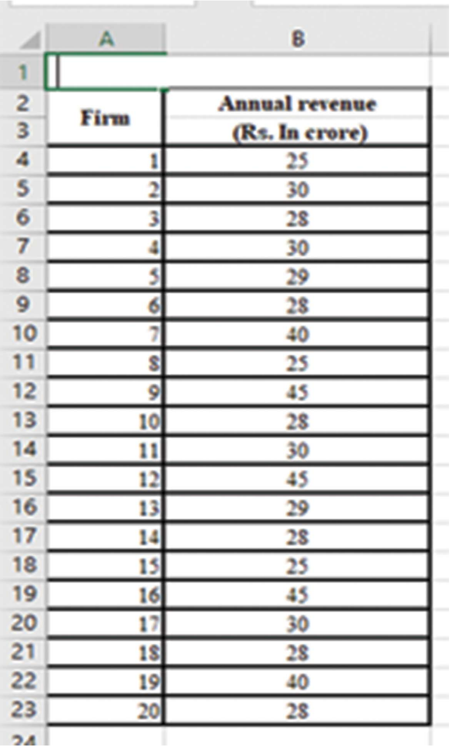
Table 5.9 Data for Example 5.3

Firms	*Annual Revenue (Crores of Rupees)*
1	25
2	30
3	28
4	30

(*Continued*)

Table 5.9 (Continued)

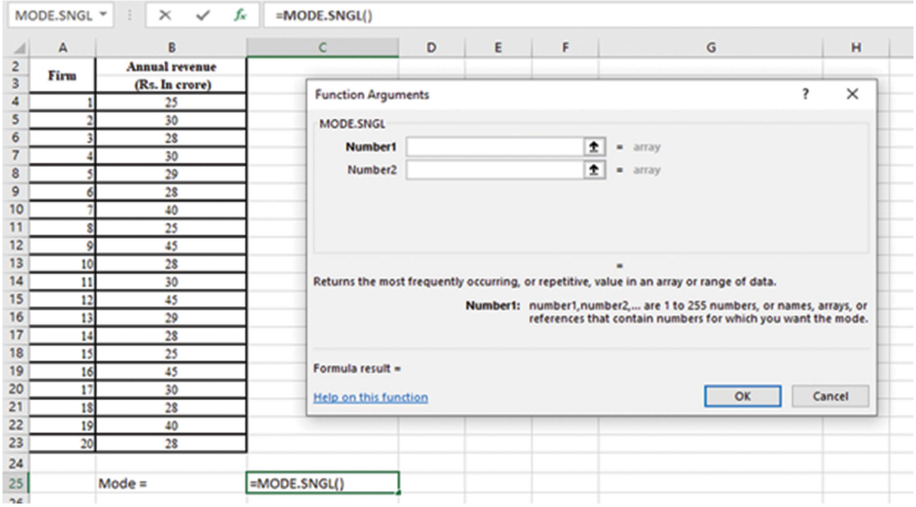
Firms	*Annual Revenue (Crores of Rupees)*
5	29
6	28
7	40
8	25
9	45
10	28
11	30
12	45
13	29
14	28
15	25
16	45
17	30
18	28
19	40
20	28



Firm	Annual revenue (Rs. In crore)
1	25
2	30
3	28
4	30
5	29
6	28
7	40
8	25
9	45
10	28
11	30
12	45
13	29
14	28
15	25
16	45
17	30
18	28
19	40
20	28

Figure 5.13 Screenshot after inputting annual sales revenues of firms in an Excel sheet

- The screenshot after clicking the sequence of buttons Formulas ⟹ More Functions ⟹ Statistical ⟹ MODE.SNGL is shown in Figure 5.14.
- The screenshot after copying the range of cells (B3:B22) in the box against Number 1 in the dropdown menu of Figure 5.14 is shown in Figure 5.15.



=MODE.SNGL()

Firm	Annual revenue (Rs. In crore)
1	25
2	30
3	28
4	30
5	29
6	28
7	40
8	25
9	45
10	28
11	30
12	45
13	29
14	28
15	25
16	45
17	30
18	28
19	40
20	28

Mode = =MODE.SNGL()

Function Arguments

MODE.SNGL

Number1 = array

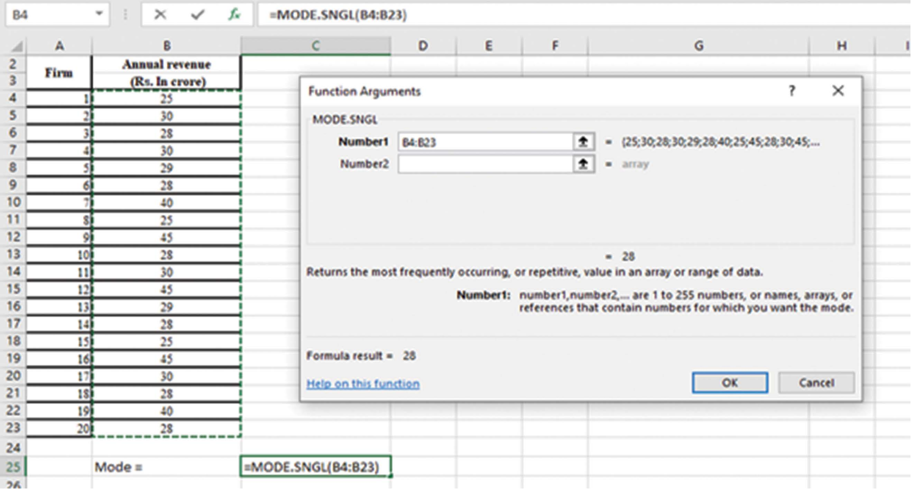
Number2 = array

Returns the most frequently occurring, or repetitive, value in an array or range of data.

Number1: number1,number2,... are 1 to 255 numbers, or names, arrays, or references that contain numbers for which you want the mode.

Formula result =

Figure 5.14 Screenshot after clicking sequence of buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ MODE.SNGL



=MODE.SNGL(B4:B23)

Firm	Annual revenue (Rs. In crore)
1	25
2	30
3	28
4	30
5	29
6	28
7	40
8	25
9	45
10	28
11	30
12	45
13	29
14	28
15	25
16	45
17	30
18	28
19	40
20	28

Mode = =MODE.SNGL(B4:B23)

Function Arguments

MODE.SNGL

Number1 B4:B23 = {25;30;28;30;29;28;40;25;45;28;30;45;...

Number2 = array

= 28

Returns the most frequently occurring, or repetitive, value in an array or range of data.

Number1: number1,number2,... are 1 to 255 numbers, or names, arrays, or references that contain numbers for which you want the mode.

Formula result = 28

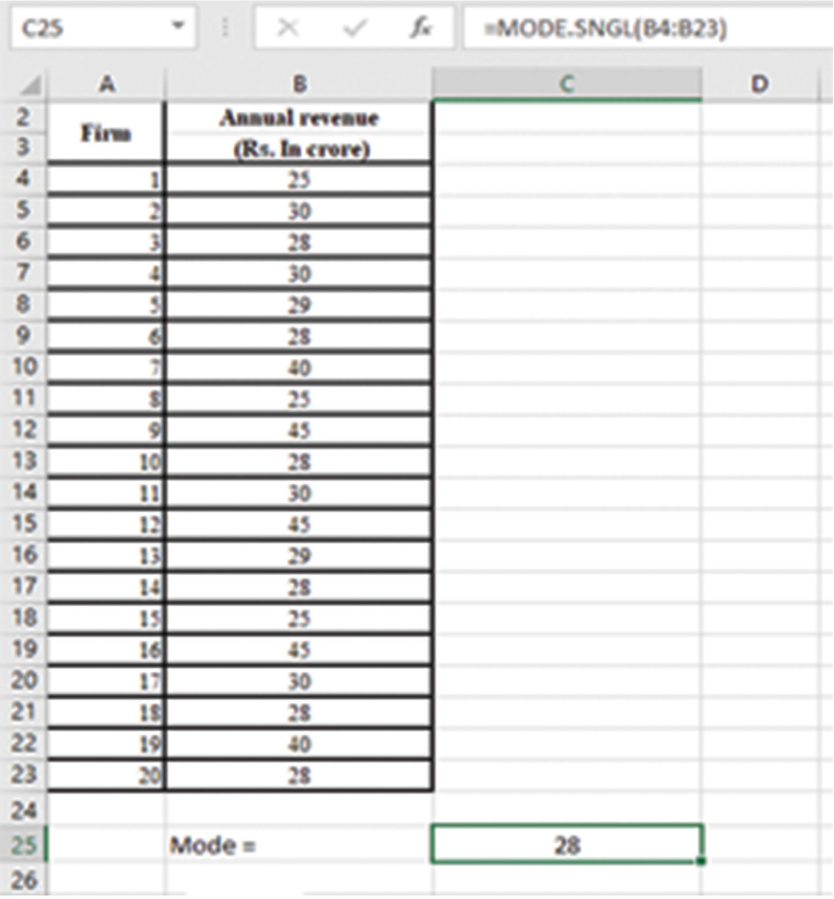
Figure 5.15 Screenshot after copying range of cells (B4:B23) in the cell against Number 1 in drop-down menu of Figure 5.14

- The screenshot after clicking the OK button in the dropdown menu of Figure 5.15 to get the mode of the given data is shown in Figure 5.16.

The mode of the given data is ₹ 28 crores, which can be seen in cell C25.

All these operations can be combined in the following formula in cell C25 to compute the mode of the given set of data.

Formula : = MODE.SNGL(B4 : B23)



=MODE.SNGL(B4:B23)

Firm	Annual revenue (Rs. In crore)
1	25
2	30
3	28
4	30
5	29
6	28
7	40
8	25
9	45
10	28
11	30
12	45
13	29
14	28
15	25
16	45
17	30
18	28
19	40
20	28

Mode = 28

Figure 5.16 Screenshot after clicking the OK button in the dropdown menu of Figure 5.15

Example 5.4

The dividend payout percentages of a company in the past 10 years are given in Table 5.10. Find the mode of the dividend payout percentages of the company.

Solution

The data for Example 5.4 are shown in Table 5.11.

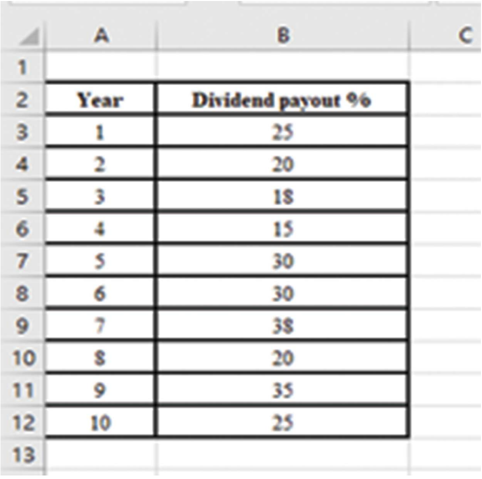
The screenshot after inputting the dividend payout percentages of the company in an Excel sheet is shown in Figure 5.17. Click cells C14, C15, C16, and 17 (vertical array to hold modes). The screenshot after clicking the sequence of buttons Formulas ⟹ More Functions ⟹ Statistical ⟹ MODE.MULT is shown in Figure 5.18. The screenshot after copying the range of cells (B3:B12) in the box against Number 1 in the dropdown menu of Figure 5.18 is shown in Figure 5.19. Simultaneously press the Shift key, Ctrl key, and Enter key, as in Figure 5.19, to show the modes, as shown in Figure 5.20 in the vertical array C14, C15, C16, and C17. Since this problem has only three modes, cell C17 shows #N/A.

Table 5.10 Dividend Payout Percentages

Year	*Dividend Payout %*
1	25
2	20
3	18
4	15
5	30
6	30
7	38
8	20
9	35
10	25

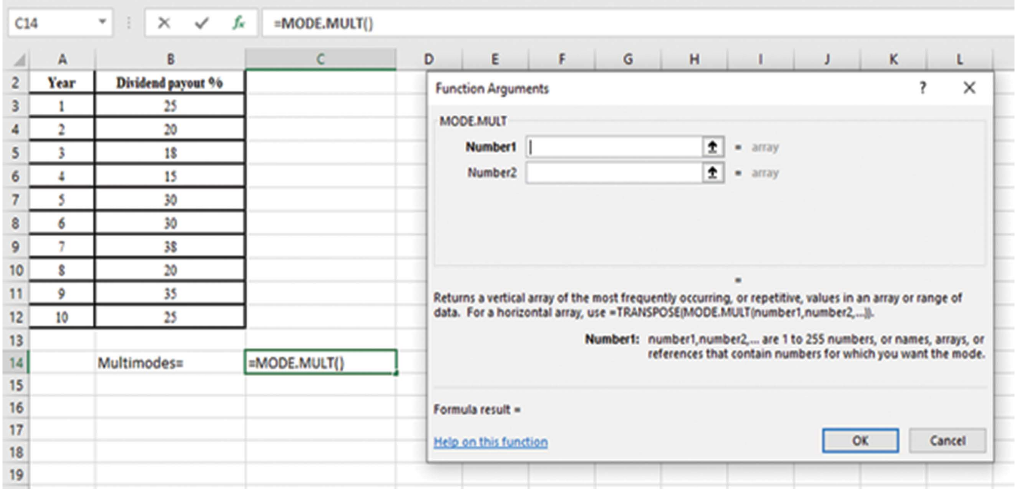
Table 5.11 Data for Example 5.4

Year	*Dividend Payout %*
1	25
2	20
3	18
4	15
5	30
6	30
7	38
8	20
9	35
10	25



	A	B	C
1			
2	**Year**	**Dividend payout %**	
3	1	25	
4	2	20	
5	3	18	
6	4	15	
7	5	30	
8	6	30	
9	7	38	
10	8	20	
11	9	35	
12	10	25	
13			

Figure 5.17 Screenshot after inputting dividend payout % of company in an Excel sheet



C14 =MODE.MULT()

	A	B	C
2	Year	Dividend payout %	
3	1	25	
4	2	20	
5	3	18	
6	4	15	
7	5	30	
8	6	30	
9	7	38	
10	8	20	
11	9	35	
12	10	25	
13			
14		Multimodes=	=MODE.MULT()

Function Arguments

MODE.MULT

Number1 = array

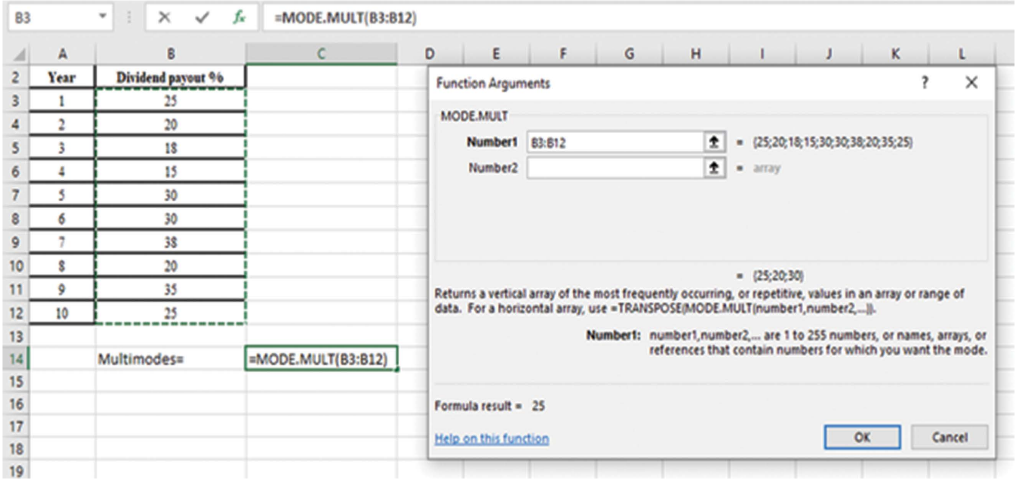
Number2 = array

Returns a vertical array of the most frequently occurring, or repetitive, values in an array or range of data. For a horizontal array, use =TRANSPOSE(MODE.MULT(number1,number2,...)).

Number1: number1,number2,... are 1 to 255 numbers, or names, arrays, or references that contain numbers for which you want the mode.

Formula result =

Figure 5.18 Screenshot after clicking sequence of buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ MODE.MULT



B3 =MODE.MULT(B3:B12)

	A	B	C
2	Year	Dividend payout %	
3	1	25	
4	2	20	
5	3	18	
6	4	15	
7	5	30	
8	6	30	
9	7	38	
10	8	20	
11	9	35	
12	10	25	
13			
14		Multimodes=	=MODE.MULT(B3:B12)

Function Arguments

MODE.MULT

Number1 B3:B12 = {25;20;18;15;30;30;38;20;35;25}

Number2 = array

= {25;20;30}

Returns a vertical array of the most frequently occurring, or repetitive, values in an array or range of data. For a horizontal array, use =TRANSPOSE(MODE.MULT(number1,number2,...)).

Number1: number1,number2,... are 1 to 255 numbers, or names, arrays, or references that contain numbers for which you want the mode.

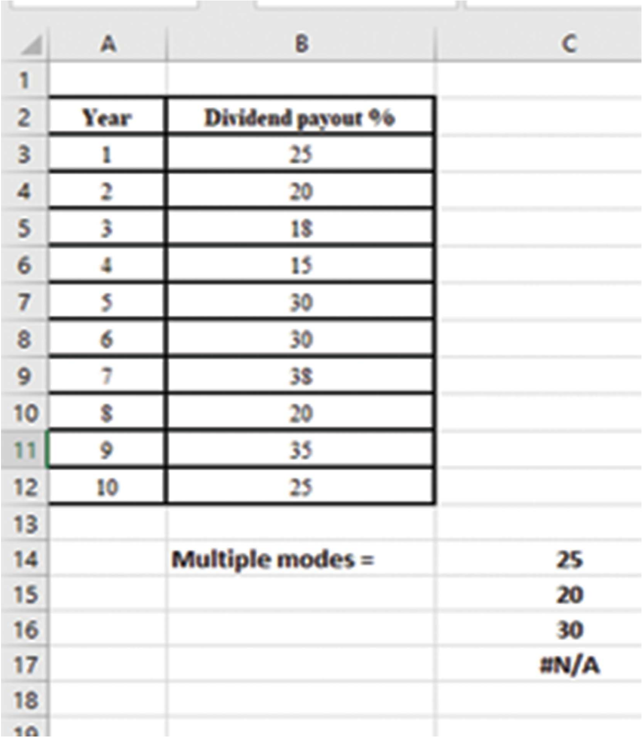
Formula result = 25

Figure 5.19 Screenshot after copying range of cells (B4:B12) in the box against Number 1 in the dropdown menu of Figure 5.18

The modes of the given data are 25%, 20%, and 30%, which can be seen in cells C14, C15, and C16, respectively, in Figure 5.20.

5.3.2 Mode of Grouped Data With Frequencies Using Excel Sheets

If the data are in the form of interval data along with frequencies, as shown in Table 5.12, then an Excel sheet can be designed by the investigators themselves.



	A	B	C
1			
2	Year	Dividend payout %	
3	1	25	
4	2	20	
5	3	18	
6	4	15	
7	5	30	
8	6	30	
9	7	38	
10	8	20	
11	9	35	
12	10	25	
13			
14		Multiple modes =	25
15			20
16			30
17			#N/A
18			

Figure 5.20 Screenshot after clicking the OK button in the dropdown menu of Figure 5.19

Table 5.12 Interval Data With Frequencies

Data Item Interval	*Frequency (f_i)*
Less than X_1	f_1
X_2–X_3	f_2
X_3–X_4	f_3
–	–
–	–
X_i–X_{i+1}	f_i
–	–
X_{n-1}–X_n	f_{n-1}
More than X_n	f_n

The mode of the data items is the value of the data item which corresponds to the maximum frequency. The formula for the mode of the grouped data is as follows.

$$\text{Mode} = X_k + \left(\frac{f_p}{f_p + f_s}\right) \times C$$

where

k is the interval (class) corresponding to maximum frequency

X_k is the lower limit of the data item of the k^{th} interval, which is the modal class

f_p is the absolute difference between the frequency of the modal class and that of the immediately preceding class

f_s is the absolute difference between the frequency of the modal class and that of the immediately succeeding class

C is the width of the class interval

Example 5.5

The distribution of the number of months holding shares of a company by its shareholders is presented in Table 5.13.

Find the mode of the number of months holding shares by the shareholders.

Solution

The data for Example 5.5 are shown in Table 5.14.

The formula for the mode of the grouped data is as follows.

$$\text{Mode} = X_k + \left(\frac{f_p}{f_p + f_s}\right) \times C$$

where

k is the interval (class) corresponding to maximum frequency

X_k is the lower limit of the data item of the k^{th} interval, which is the modal class

f_p is the absolute difference between the frequency of the modal class and that of the immediately preceding class

f_s is the absolute difference between the frequency of the modal class and that of the immediately succeeding class

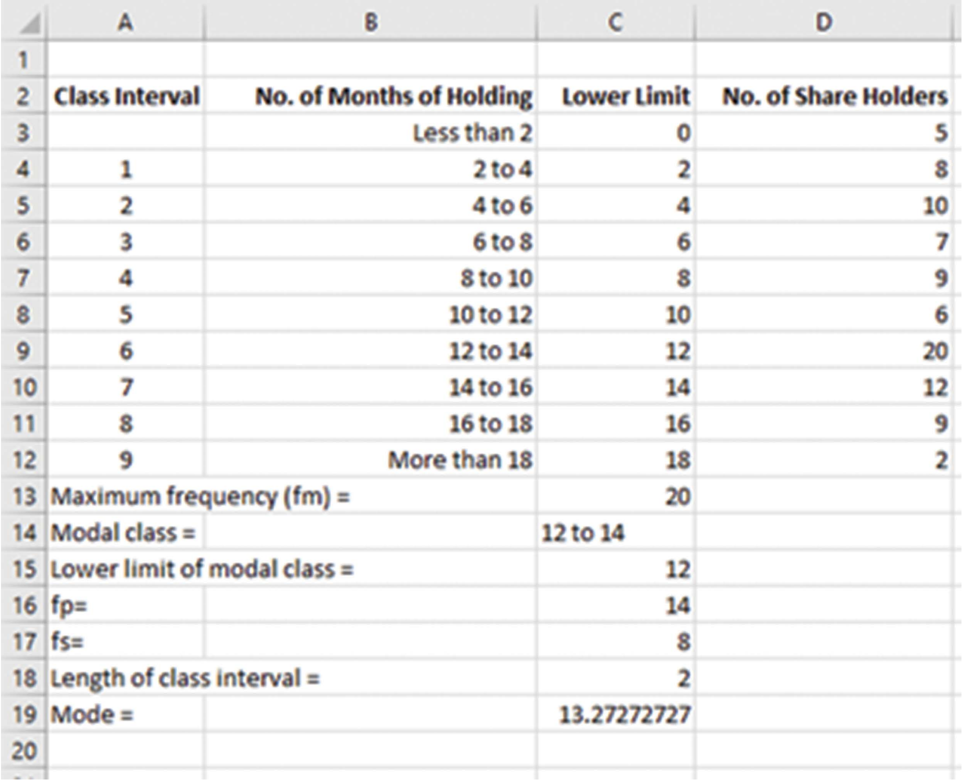
C is the width of the class interval

Table 5.13 Distribution of Number of Months Holding Shares

No. of Months Holding	*No. of Shares*
Less than 2	5
2 to 4	8
4 to 6	10
6 to 8	7
8 to 10	9
10 to 12	6
12 to 14	20
14 to 16	12
16 to 18	9
More than 18	2

Table 5.14 Data for Example 5.5

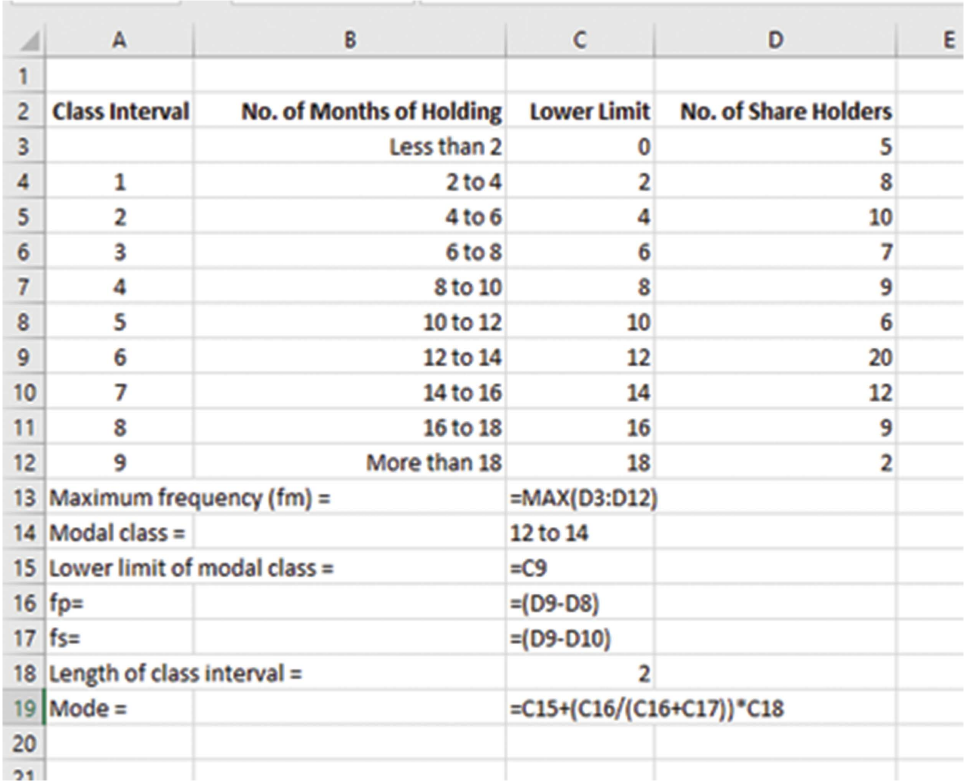
No. of Months Holding	*No. of Shares*
Less than 2	5
2 to 4	8
4 to 6	10
6 to 8	7
8 to 10	9
10 to 12	6
12 to 14	20
14 to 16	12
16 to 18	9
More than 18	2



	A	B	C	D
1				
2	**Class Interval**	**No. of Months of Holding**	**Lower Limit**	**No. of Share Holders**
3		Less than 2	0	5
4	1	2 to 4	2	8
5	2	4 to 6	4	10
6	3	6 to 8	6	7
7	4	8 to 10	8	9
8	5	10 to 12	10	6
9	6	12 to 14	12	20
10	7	14 to 16	14	12
11	8	16 to 18	16	9
12	9	More than 18	18	2
13	Maximum frequency (fm) =		20	
14	Modal class =		12 to 14	
15	Lower limit of modal class =		12	
16	fp=		14	
17	fs=		8	
18	Length of class interval =		2	
19	Mode =		13.27272727	
20				

Figure 5.21 Screenshot of working for finding the mode of Example 5.5

The data for Example 5.5 along with the working to obtain the mode of the grouped data are shown in Figure 5.21, and the corresponding formulas are shown in Figure 5.22. The mode of the number of months holding shares of the company is 13.27272727.



	A	B	C	D	E
1					
2	Class Interval	No. of Months of Holding	Lower Limit	No. of Share Holders	
3		Less than 2	0	5	
4	1	2 to 4	2	8	
5	2	4 to 6	4	10	
6	3	6 to 8	6	7	
7	4	8 to 10	8	9	
8	5	10 to 12	10	6	
9	6	12 to 14	12	20	
10	7	14 to 16	14	12	
11	8	16 to 18	16	9	
12	9	More than 18	18	2	
13	Maximum frequency (fm) =		=MAX(D3:D12)		
14	Modal class =		12 to 14		
15	Lower limit of modal class =		=C9		
16	fp=		=(D9-D8)		
17	fs=		=(D9-D10)		
18	Length of class interval =		2		
19	Mode =		=C15+(C16/(C16+C17))*C18		
20					

Figure 5.22 Screenshot of formulas for working to find the mode of Example 5.5

5.4 Percentile

A number in a range of values is said to be at the percentile if a specific proportion of observations fall below it. A percentile, on the other hand, is a value of a random variable for a specific percentage of the overall frequency of a particular set of data. This indicates that a significant portion of the observations fall below the percentile value [4]. Consider the scores received on a professional course entrance exam. With 1,50,000 individuals showing up for the test, let the maximum score on the entrance examination be 800. Thus, more than one candidate will undoubtedly receive the same mark, which results in a frequency (number of candidates) of 1 or more for each mark.

The steps to obtain the mark percentile for a given fraction (percentage) of the total frequency are as follows.

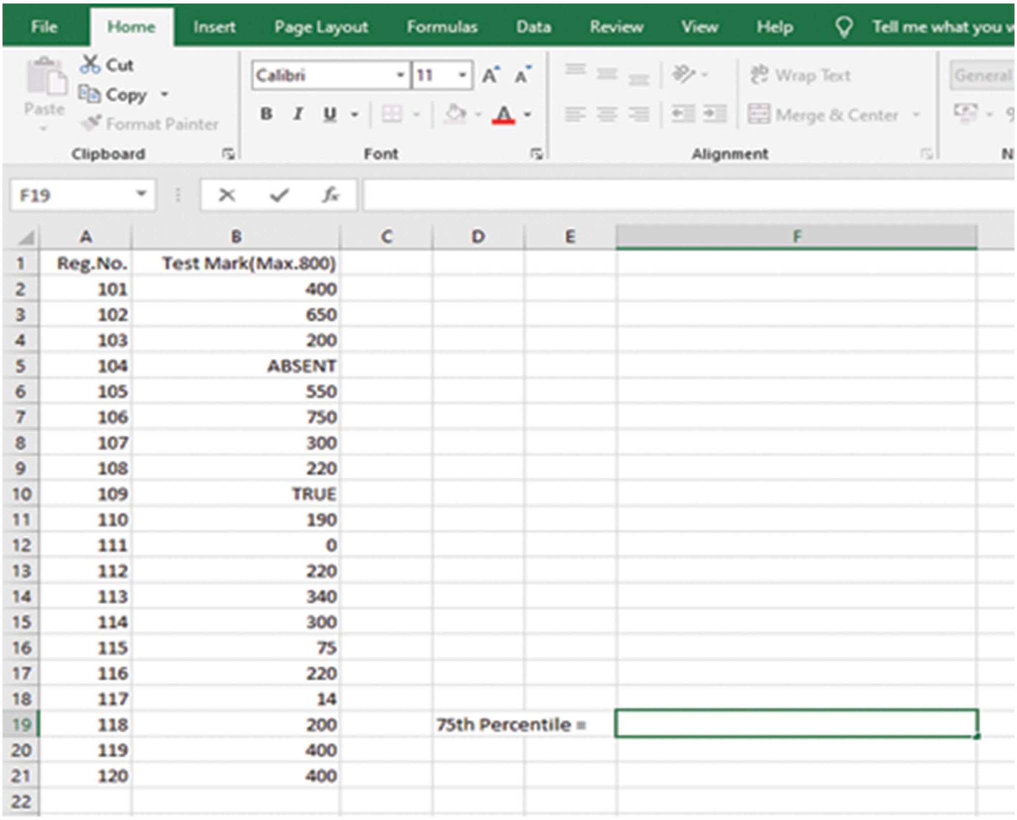
1. Input a fraction *K* in the range 0 to 1 or a percentage in the range 0 to 100 of the total frequency of the test mark.
2. Sort the data containing the mark and the frequency in ascending order.
3. Obtain the cumulative frequency of the test mark.

4. Multiply the total frequency by the percentage *K* given in Step 1 and let it be *X*.
5. Read the mark corresponding to the cumulative frequency *X* as the *K*th percentile. This means that *K* percentage of observations will be below the *K*th percentile mark.

Excel has two types of functions for percentile, PERCENTILE.EXC and PERCENTILE.INC.

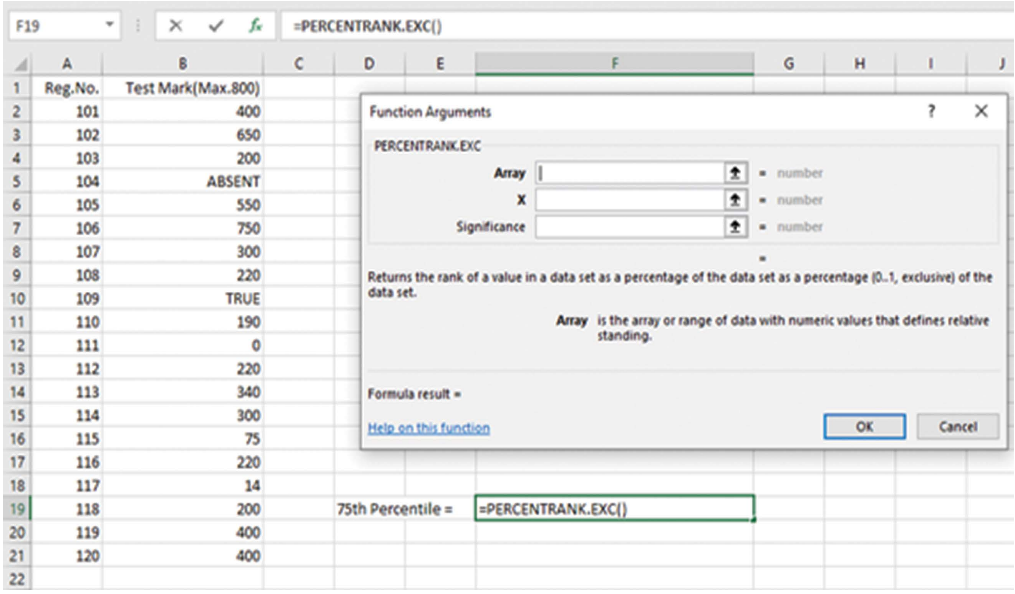
5.4.1 PERCENTILE.EXC Function

The PERCENTILE.EXC function determines percentile value by excluding 0 and 1. Logical TRUE will have a value of 1, and logical FALSE will have a value of 0. Any other string will have a value of 0. Consider the data in the screenshot shown in Figure 5.23. This contains numerical number, TRUE, and ABSENT. Here, TRUE means that the candidate has attended the test and left in the middle of the test. The screenshot after clicking the sequence of buttons Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ PERCENTILE.EXC is shown in Figure 2.24. After filling the box against Array with cells B2:B21 and the box against K with 0.75 in the dropdown menu of Figure 5.24, the



	A	B	C	D	E	F
1	Reg.No.	Test Mark(Max.800)				
2	101	400				
3	102	650				
4	103	200				
5	104	ABSENT				
6	105	550				
7	106	750				
8	107	300				
9	108	220				
10	109	TRUE				
11	110	190				
12	111	0				
13	112	220				
14	113	340				
15	114	300				
16	115	75				
17	116	220				
18	117	14				
19	118	200		75th Percentile =		
20	119	400				
21	120	400				
22						

Figure 5.23 Screenshot of data to demonstrate percentile



F19 =PERCENTRANK.EXC()

	A	B
1	Reg.No.	Test Mark(Max.800)
2	101	400
3	102	650
4	103	200
5	104	ABSENT
6	105	550
7	106	750
8	107	300
9	108	220
10	109	TRUE
11	110	190
12	111	0
13	112	220
14	113	340
15	114	300
16	115	75
17	116	220
18	117	14
19	118	200
20	119	400
21	120	400

75th Percentile = =PERCENTRANK.EXC()

Function Arguments

PERCENTRANK.EXC

Array = number

X = number

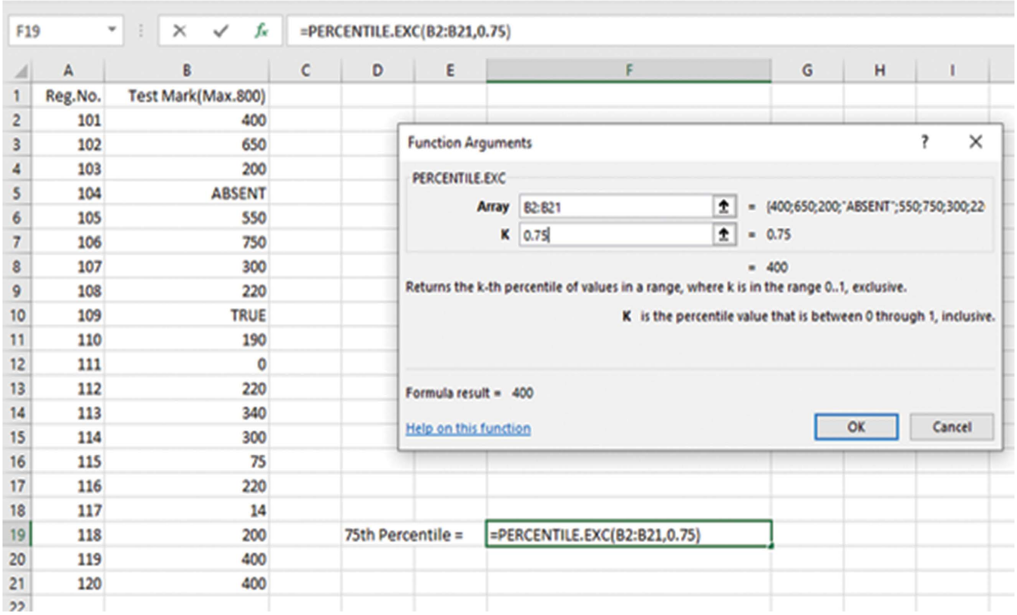
Significance = number

Returns the rank of a value in a data set as a percentage of the data set as a percentage (0..1, exclusive) of the data set.

Array is the array or range of data with numeric values that defines relative standing.

Formula result =

Figure 5.24 Screenshot after clicking sequence of buttons, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ PERCENTILE.EXC



F19 =PERCENTILE.EXC(B2:B21,0.75)

	A	B
1	Reg.No.	Test Mark(Max.800)
2	101	400
3	102	650
4	103	200
5	104	ABSENT
6	105	550
7	106	750
8	107	300
9	108	220
10	109	TRUE
11	110	190
12	111	0
13	112	220
14	113	340
15	114	300
16	115	75
17	116	220
18	117	14
19	118	200
20	119	400
21	120	400

75th Percentile = =PERCENTILE.EXC(B2:B21,0.75)

Function Arguments

PERCENTILE.EXC

Array B2:B21 = {400;650;200;"ABSENT";550;750;300;22

K 0.75 = 0.75

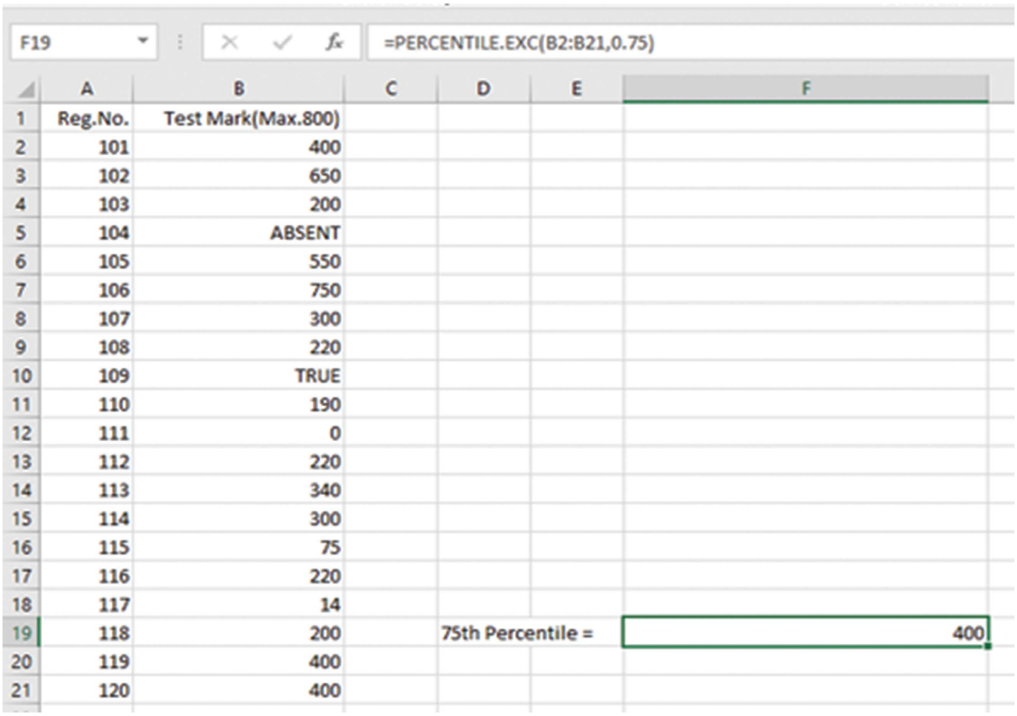
= 400

Returns the k-th percentile of values in a range, where k is in the range 0..1, exclusive.

K is the percentile value that is between 0 through 1, inclusive.

Formula result = 400

Figure 5.25 Screenshot after filling data in the dropdown menu of Figure 5.24



F19 =PERCENTILE.EXC(B2:B21,0.75)

	A	B	C	D	E	F
1	Reg.No.	Test Mark(Max.800)				
2	101	400				
3	102	650				
4	103	200				
5	104	ABSENT				
6	105	550				
7	106	750				
8	107	300				
9	108	220				
10	109	TRUE				
11	110	190				
12	111	0				
13	112	220				
14	113	340				
15	114	300				
16	115	75				
17	116	220				
18	117	14				
19	118	200		75th Percentile =		400
20	119	400				
21	120	400				

Figure 5.26 Screenshot after clicking the OK button in the dropdown menu of Figure 5.25

display will be as shown in Figure 5.25. Clicking the OK button in the dropdown menu of Figure 5.25 gives the result in Figure 5.26. The 75th percentile mark is 400.

The formula for PERCENTILE.EXE is as follows.

= PERCENTILE.EXC(range of cells containing data, Value of K)

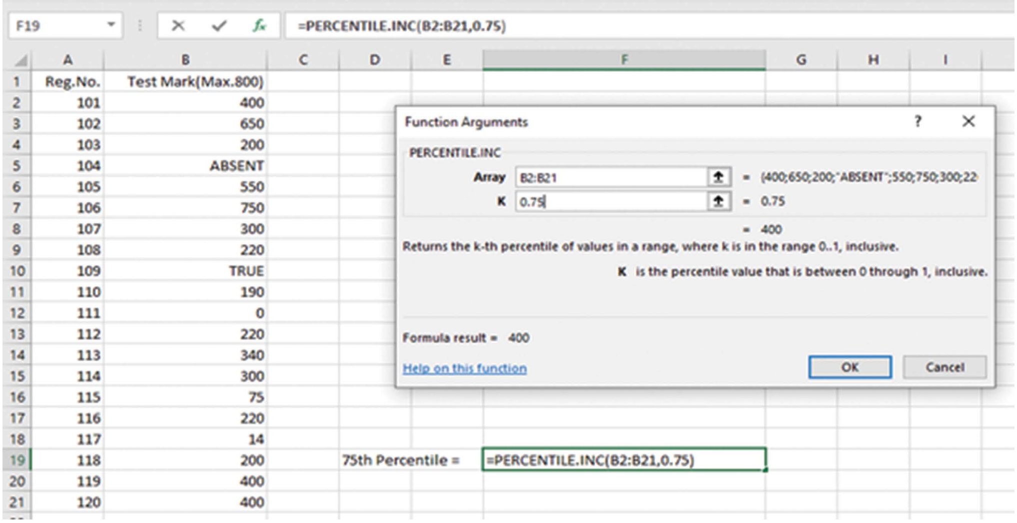
If K = 0.75, then the 75th percentile of the data given in Figure 5.23 can be obtained using the following formula.

= PERCENTILE.EXC(B2 : B21, 0.75)

5.4.2 PERCENTILE.INC Function

The PERCENTILE.INC function finds the percentile value of the given data for a given K. Consider the data which is already given in Figure 5.23 for the marks of the candidates in the entrance examination of a professional course.

The sequence of button clicks Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ PERCENTILE.INC is shown in Figure 5.27. Filling the box against Array with cells B2:B21, filling the box against K with 0.75 in the dropdown menu of Figure 5.27, and clicking the OK button in its dropdown menu gives the result in Figure 5.28. The 75th percentile mark is 400.



F19 =PERCENTILE.INC(B2:B21,0.75)

	A	B	C	D	E	F
1	Reg.No.	Test Mark(Max.800)				
2	101	400				
3	102	650				
4	103	200				
5	104	ABSENT				
6	105	550				
7	106	750				
8	107	300				
9	108	220				
10	109	TRUE				
11	110	190				
12	111	0				
13	112	220				
14	113	340				
15	114	300				
16	115	75				
17	116	220				
18	117	14				
19	118	200		75th Percentile =		=PERCENTILE.INC(B2:B21,0.75)
20	119	400				
21	120	400				

Function Arguments

PERCENTILE.INC

Array B2:B21 = {400;650;200;"ABSENT";550;750;300;22

K 0.75 = 0.75

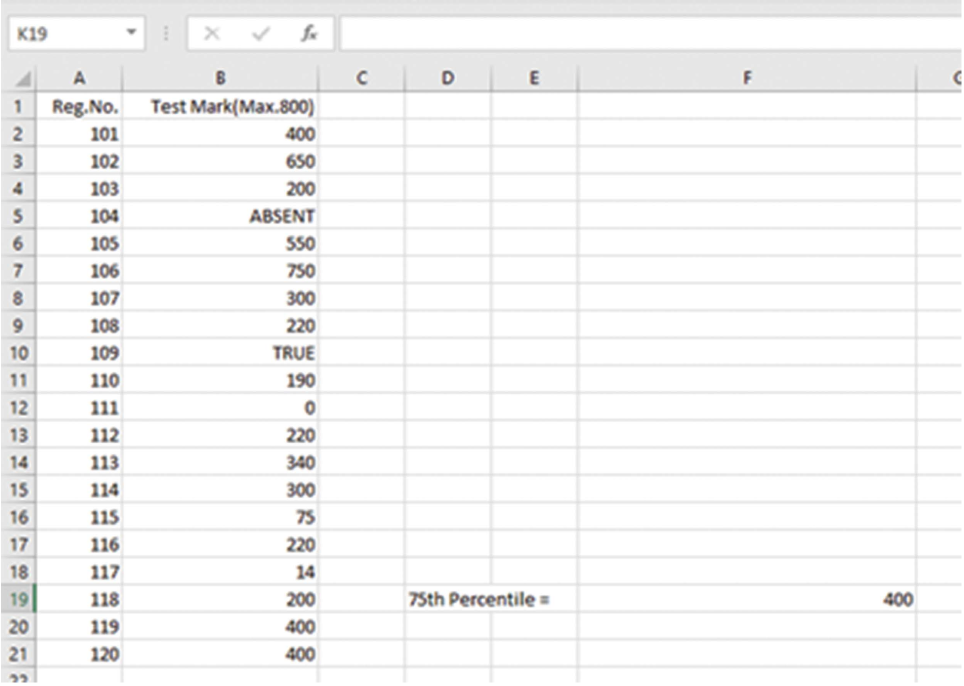
= 400

Returns the k-th percentile of values in a range, where k is in the range 0..1, inclusive.

K is the percentile value that is between 0 through 1, inclusive.

Formula result = 400

Figure 5.27 Screenshot for the sequence of button clicks, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ PERCENTILE.INC



K19

	A	B	C	D	E	F
1	Reg.No.	Test Mark(Max.800)				
2	101	400				
3	102	650				
4	103	200				
5	104	ABSENT				
6	105	550				
7	106	750				
8	107	300				
9	108	220				
10	109	TRUE				
11	110	190				
12	111	0				
13	112	220				
14	113	340				
15	114	300				
16	115	75				
17	116	220				
18	117	14				
19	118	200		75th Percentile =		400
20	119	400				
21	120	400				

Figure 5.28 Screenshot after filling the data in the dropdown menu of Figure 5.27 and clicking the OK button

The formula for PERCENTILE.INC is as follows.

= PERCENTILE.INC(Range of cells containing the data, Value of K)

If K = 0.75, then 75th percentile of the data given in Figure 5.23 can be obtained using the following formula.

= PERCENTILE.INC(B2 : B21, 0.75)

5.5 Quartile

Quartile is the value of a random variable with respect to a specified percentile out of five different percentiles, that is, minimum value (0th percentile), 25th percentile, 50th percentile, 75th percentile, and maximum value (100th percentile) of the total frequency. The values of the random variable with respect to the minimum value (0th percentile), 25th percentile, 50th percentile, 75th percentile, and maximum value (100th percentile) of the total frequency are called zeroth quartile, first quartile, second quartile, third quartile, and fourth quartile, respectively. The second quartile is called the median of the given set of data.

In Excel, the QUARTILE function is classified into two types, QUARTILE.EXC and QUARTILE.INC [5]. The QUARTILE.EXC function excludes the zeroth quartile and fourth quartile of the given data for computing the quartiles, whereas QUARTILE.INC includes the zeroth quartile and fourth quartile along with first quartile, second quartile, and third quartile of the given data.

5.5.1 QUARTILE.EXC Function

This section demonstrates the determination of the first quartile, second quartile, and third quartile of a given set of data.

Consider Example 5.1 and its screenshot shown in Figure 5.29.

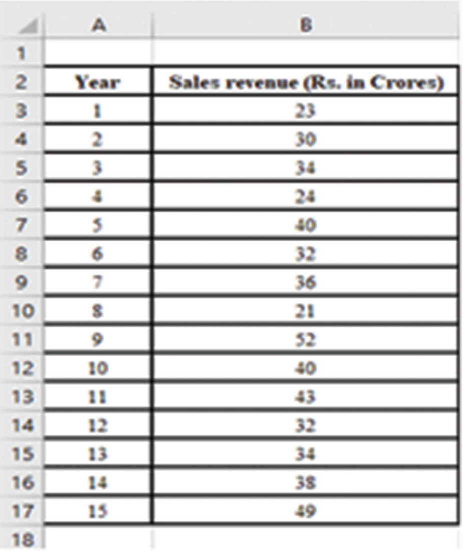
The steps of finding the value of a specific quartile, say, the 2nd quartile of the QUARTILE.EXC function for the given ungrouped data shown in Figure 5.29, are presented in the following.

1. Click the buttons in the sequence Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹QUARTILE.EXC to show the display in Figure 5.30.
2. Then fill the box against Array with cells B3:B17 and the box against Quart with 2 for the 2nd quartile in the dropdown menu of Figure 5.30 as shown in Figure 5.31.
3. Click the OK button in the dropdown menu of Figure 5.31 to show the results in Figure 5.32.

All these can be combined using the following formula.

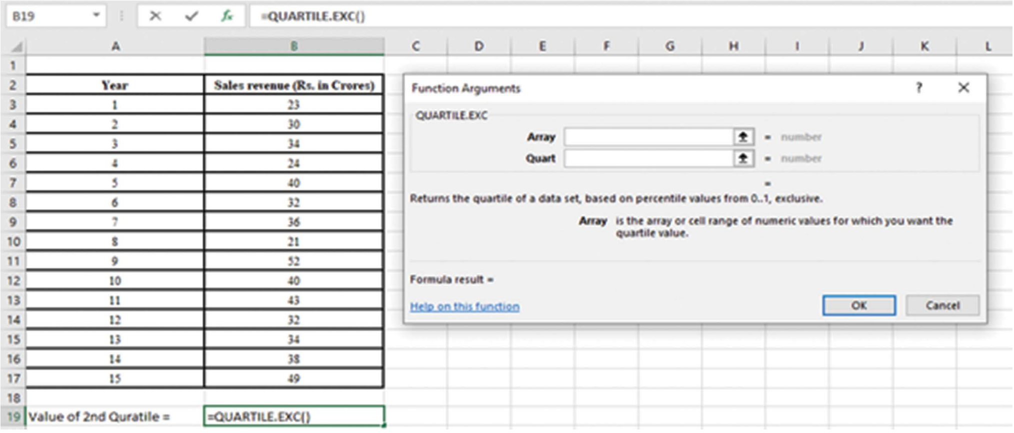
= QUARTILE.EXC(B3 : B17, 2)

The result of the 2nd quartile is 34, which is same as the result given in Figure 5.7 for the median of Example 5.1.



Year	Sales revenue (Rs. in Crores)
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49

Figure 5.29 Screenshot after inputting data for Example 5.1 to demonstrate QUARTILE



B19 =QUARTILE.EXC()

Year	Sales revenue (Rs. in Crores)
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49

Value of 2nd Quratile = =QUARTILE.EXC()

Function Arguments

QUARTILE.EXC

Array = number

Quart = number

Returns the quartile of a data set, based on percentile values from 0..1, exclusive.

Array is the array or cell range of numeric values for which you want the quartile value.

Formula result =

Figure 5.30 Screenshot for clicking the sequence of buttons, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ QUARTILE.EXC

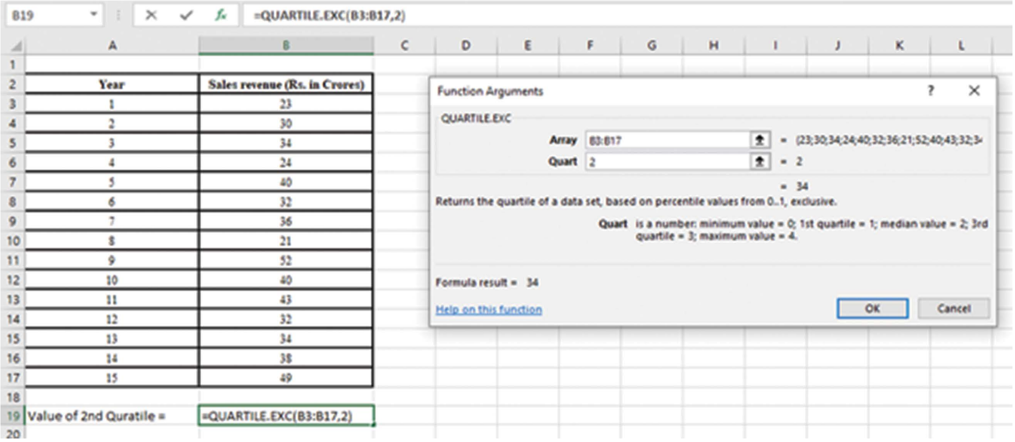
The formula for QUARTILE.EXC is as follows.

= QUARTILE.EXC(Range of cells containing data, Quartile number out of 1, 2, and 3)

The values of the first quartile and third quartile can be obtained using the following formulas.

Formula for 1^{st} quartile : = QUARTILE.EXC($B3 : B17,1$), which gives the value of 30

Formula for 3^{rd} quartile : = QUARTILE.EXC($B3 : B17,3$), which gives the value of 40



B19 =QUARTILE.EXC(B3:B17,2)

Year	Sales revenue (Rs. in Crores)
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49

Value of 2nd Quratile = =QUARTILE.EXC(B3:B17,2)

Function Arguments

QUARTILE.EXC

Array B3:B17 = {23;30;34;24;40;32;36;21;52;40;43;32;3

Quart 2 = 2

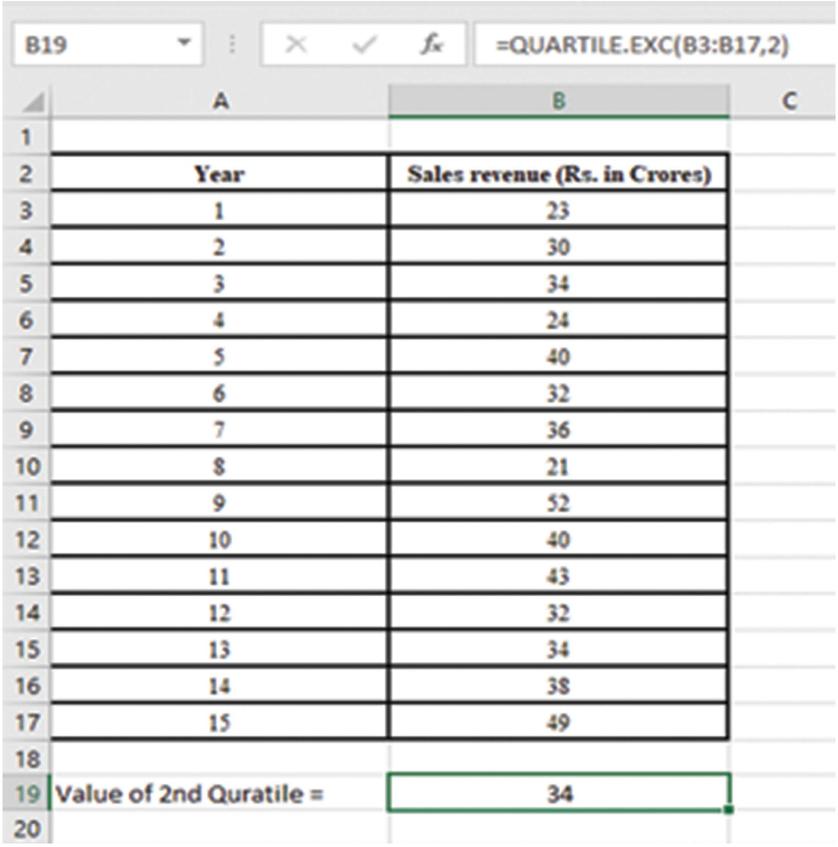
= 34

Returns the quartile of a data set, based on percentile values from 0..1, exclusive.

Quart is a number: minimum value = 0; 1st quartile = 1; median value = 2; 3rd quartile = 3; maximum value = 4.

Formula result = 34

Figure 5.31 Screenshot after filling data in the dropdown menu of Figure 5.30



B19 =QUARTILE.EXC(B3:B17,2)

Year	Sales revenue (Rs. in Crores)
1	23
2	30
3	34
4	24
5	40
6	32
7	36
8	21
9	52
10	40
11	43
12	32
13	34
14	38
15	49

Value of 2nd Quratile = 34

Figure 5.32 Screenshot after clicking the OK button in the dropdown menu of Figure 5.31

5.5.2 QUARTILE.INC Function

The QUARTILE.INC function gives the values of the zeroth quartile, first quartile, second quartile, third quartile, and fourth quartile of the given data shown in Figure 5.29.

The formula for QUARTILE.INC is as follows.

$$= \text{QUARTILE.INC}(\text{Range of cells containing data, Quartile number out of } 0, 1, 2, 3, 4)$$

The values of the zeroth quartile, first quartile, second quartile, third quartile, and fourth quartile using QUARTILE.INC can be obtained using the following formulas.

$$\text{Formula for } 0^{\text{th}} \text{ quartile} := \text{QUARTILE.INC}(\text{B3} : \text{B17}, 0), \text{ which gives the value as } 21$$

$$\text{Formula for } 1^{\text{st}} \text{ quartile} := \text{QUARTILE.INC}(\text{B3} : \text{B17}, 1), \text{ which gives the value as } 31$$

$$\text{Formula for 2nd quartile} := \text{QUARTILE.INC}(\text{B3} : \text{B17}, 1), \text{ which gives the value as } 34$$

$$\text{Formula for } 3^{\text{rd}} \text{ quartile} := \text{QUARTILE.INC}(\text{B3} : \text{B17}, 3), \text{ which gives the value as } 40$$

$$\text{Formula for } 4^{\text{th}} \text{ quartile} := \text{QUARTILE.INC}(\text{B3} : \text{B17}, 4), \text{ which gives the value as } 52$$

Summary

- The median function determines the middlemost observation of a given set of data
- The formula to obtain the median of ungrouped data is:

$$\text{Median} = X_p, \text{ if } N \text{ is odd}$$

$$= \frac{\left(X_{p_1} + X_{p_2}\right)}{2}, \text{ if } N \text{ is even}$$

where

N is the total number of observations of the data

X_i is the i^{th} observation of the given data, where i = 1, 2, 3, . . . , N

p is the middlemost point of N observations.

$$p = \frac{(N-1)}{2} + 1, \text{ if } N \text{ is odd}$$

$$p_1 = \frac{(N)}{2} \text{and}$$

$$p_2 = p_1 + 1, \text{ if } N \text{ is even}$$

- The formula for the median of ungrouped data in Excel is:

  $$= \text{MEDIAN}(\text{Range of cells containing data})$$

- The sequence of mouse clicks to find the median of ungrouped data is: HOME ⟹ FORMULAS ⟹ MORE FUNCTIONS ⟹ STATISTICAL ⟹ MEDIAN.
- The formula for the median of the grouped data is:

  $$Median = X_k + \left(\frac{\frac{N}{2} - F_k}{f_k} \right) \times C$$

  where
  N is the total frequency
  k is the interval (class) corresponding to 50% of the total frequency
  X_k is the lower limit of the data item of the k^{th} interval (median class)
  f_k is the frequency of the k^{th} interval (median class)
  F_k is the cumulative frequency of the previous interval with respect to the median class
  C is the length of the class interval

- The mode of a given set of data (observations) is the item of that data which has the maximum frequency.
- Mode is a kind of measure of central tendency.
- The formula for single mode is: = MODE.SNGL(Range of cells containing data).
- The formula of multi-mode is: MODE.MULT(Range of cells containing data).
- The formula for the mode of the grouped data is as follows.

  $$Mode = X_k + \left(\frac{f_p}{f_p + f_s} \right) \times C$$

  where
  k is the interval (class) corresponding to maximum frequency
  X_k is the lower limit of the data item of the k^{th} interval, which is the modal class
  f_p is the absolute difference between the frequency of the modal class and that of the immediately preceding class
  f_s is the absolute difference between the frequency of the modal class and that of the immediately succeeding class
  C is the width of the class interval

- Percentile is a value in the given range of values such that a given percentage of observations fall below that value.
- Quartile is the value of a random variable with respect to a specified percentage out of five different percentages, that is, minimum value (0th percentile), 25th percentile, 50th percentile, 75th percentile, and maximum value (100th percentile) of the total frequency.
- The values of the random variable with respect to minimum value (0th percentile), 25th percentile, 50th percentile, 75th percentile, and maximum value (100th percentile) of the total frequency are called the zeroth quartile, first quartile, second quartile, third quartile, and fourth quartile, respectively.
- The second quartile is called the median of the given set of data.

Keywords

The Median function determines the middlemost observation of a given set of data.

Ungrouped data do not have frequencies.

Grouped data have frequencies.

An open-ended interval means that the lower limit of the first interval and the upper limit of the last interval are absent.

The mode of a given set of data (observations) is the item of that data which has the maximum frequency.

Mode is a kind of measure of central tendency.

Percentile is a value in the given range of values such that a given percentage of observations fall below that value.

Quartile is the value of a random variable with respect to a specified percentile out of five different percentages, that is, minimum value, 25th percentile, 50th percentile, 75th percentile, and maximum value of the total frequency.

Review Questions

1. Define the median and distinguish it from the mean.
2. Give the mathematical formula for the median of ungrouped data and explain it.
3. Give the syntax of the MEDIAN function in Excel.
4. Illustrate the process of finding the median of ungrouped data using the menu-driven method in Excel.
5. The following table displays the ages of the employees at a corporation. Find the age of the employees using the MEDIAN function in Excel.

Employee code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Age in year	55	45	28	26	35	49	40	58	43	26	23	39	30	35	55

6. Give the formula for the median of group data and explain its components.
7. The following table provides an overview of the industrial estate businesses' yearly sales in crores of rupees. Utilise Excel to determine the company's annual median revenues.

Annual sales in crores of rupees	*No. of companies*
Less than 4	17
4–6	22
6–8	26
8–10	20
10–12	15
12–14	
More than 14	10

8. Define the mode and distinguish it from the mean and median.
9. Illustrate the sequence of clicks of Excel buttons to obtain the MODE of ungrouped data of your choice.
10. List the types of MODE functions in Excel and give the syntax of each of them for ungrouped data.

11. The EPS data of a company are shown in the following table for the past ten quarters. Find the mode of the EPS data using MODE.MULT in Excel.

Quarter	1	2	3	4	5	6	7	8	9	10
EPS in ₹	35	20	12	15	35	32	28	20	35	25

12. Give the mathematical formula to find the mode of a given set of grouped data and explain its components.
13. Consider the test score of the employees who underwent a training program, as shown in the following table. Find the mode of the test score of the employees using Excel.

Test Score	*Frequency*
Less than 10	5
10 to 20	8
20 to 30	10
30 to 40	7
40 to 50	9
50 to 60	6
60 to 70	20
70 to 80	12
80 to 90	9
More than 90	2

14. The following table displays the distribution of annual income in lakhs of rupees among employees of a specific cadre in an industrial park. Using Excel, determine the employees' median annual income.

Annual Income in Lakhs of Rupees	*Number of Employees*
Less than 2.0	100
2.0 to 2.5	150
2.5–3.0	175
3.0–3.5	190
3.5–4.0	120
More than 4.0	50

15. Define quartile and explain its types.
16. Illustrate the application of the PERCENTILE.EXC function using an example.
17. The marks of students in an examination out of 1000 are shown in the following table.

Reg. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Mark	980	880	900	890	500	490	478	970	690	345	560	345	765	345	879	567	345

Find the 80th percentile of the mark using Excel.

18. Illustrate the application of the PERCENTILE.INC function using an example.
19. Distinguish between the PERCENTILE.EXC function and the PERCENTILE.INC function.
20. Illustrate the application of the QUARTILE.EXC function using an example.
21. Illustrate the application of the QUARTILE.INC function using an example.
22. Distinguish between the QUARTILE.EXC function and the QUARTILE.INC function.

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. https://Exceljet.net/Excel-functions/Excel-median-function [June 25, 2020].
3. https://Exceljet.net/Excel-functions/Excel-mode-function [June 25, 2020].
4. https://Exceljet.net/Excel-functions/Excel-percentile-function [June 30, 2020].
5. https://Exceljet.net/Excel-functions/Excel-quartile-function [June 29, 2020].

6 Measures of Variation

Learning Objectives

A complete reading of this chapter will enable readers to

- Understand the role of a spread if the means of two samples are the same.
- Analyse the range and coefficient of range of a given set of data.
- Compute the quartile deviation of ungrouped data.
- Distinguish the methods to compute the quartile deviation of grouped data.
- Understand the concept of average deviation with illustrations.
- Analyse the data using standard deviation.
- Understand the method of computing the standard deviation for ungrouped data using the STDEV.P function and STDEV.S function.
- Distinguish the method of computing the standard deviation for grouped data from that for ungrouped data.
- Understand the STDEVA function, which finds the standard deviation of the observations in a sample of a population including logical values and text.
- Understand the STDEVPA function, which finds the standard deviation of the observations of the entire population including logical values and text from the STDEVA function.
- Analyse data using the coefficient of variation.

6.1 Introduction

In practice, researchers will be interested in comparing the mean values of several samples. Even when sample observations differ from one another, the samples occasionally have the same mean. Consider a scenario where paddy is grown on 10 separate plots using two common fertilisers. Following harvest, the yields (standard bags weighing 75 kg) from the 10 plots where each type of fertiliser was applied were gathered. Let them be as Table 6.1 depicts.

It is possible to confirm that the average yield for each category of fertilisers listed in Table 6.1 is 10 standard bags. Therefore, it will be challenging for the researcher to choose the fertiliser that produces the highest average yield. As a result, it is possible to distinguish the fertilisers in terms of mean yield using the spread (variation) of the observations.

DOI: 10.4324/9781032671703-6

Table 6.1 Yields (Standard Bags of 75 kg) of Paddy for Different Types of Fertiliser

Plot No.	*Type of Fertiliser*	
	X	*Y*
1	10	12
2	12	11
3	11	9
4	9	10
5	10	12
6	11	9
7	7	12
8	8	8
9	10	9
10	12	8

The different measures of variation are as listed [1].

1. Range
2. Quartile deviation
3. Average deviation
4. Standard deviation
5. Coefficient of variation

6.2 Range Using Excel Sheets

Range is a simple measure of variation, the difference between the highest value of a set of observations and the lowest value of that set of observations. The formula for the range is as follows.

$$R = H - L$$

where
H is the highest value of a set of observations
L is the lowest value of a set of observations
R is the range of a set of observations

Further, another measure called coefficient of range based on the range is the ratio between the range and the sum of the highest value of a set of observations and the lowest value of that set of observations. Its formula is as follows.

$$Coeffcient\, of\ range = \frac{H - L}{H + L}$$

where
H is the highest value of a set of observations
L is the lowest value of a set of observations

Example 6.1

The weights of the employees working in a company are summarised in Table 6.2. Find its range and coefficient of range using Excel.

Table 6.2 Weights of Employees

Employee Code	*Wight (Kg.)*
1	58
2	52
3	70
4	52
5	60
6	55
7	79
8	68
9	64
10	80
11	85
12	82
13	74
14	70

Solution

The input of Example 6.1 in an Excel sheet is shown in Figure 6.1.

The formula for the range is as follows.

$$R = H - L$$

where
H is the highest value of a set of observations
L is the lowest value of a set of observations
R is the range of a set of observations

The formula for the coefficient of range is as follows.

$$Coeffcient\,of\,range = \frac{H - L}{H + L}$$

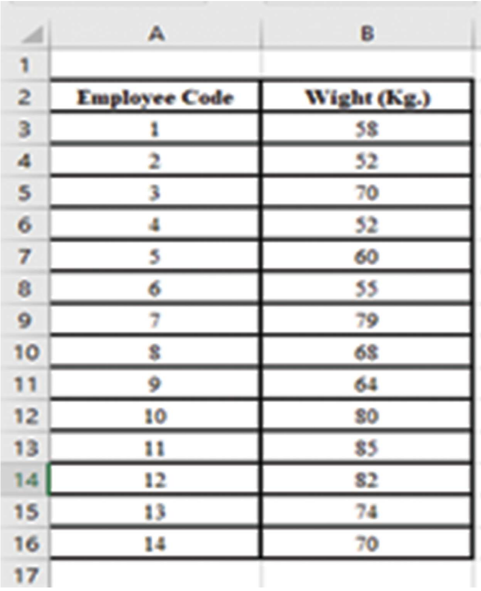
where
H is the highest value of a set of observations
L is the lowest value of a set of observations

The working of the range and the coefficient of range are shown in Figure 6.2. The formulas for the working of the range and the coefficient of range are shown in Figure 6.3. The results are listed in the following.

Range (R) = 33
Coefficient of range = 0.240876

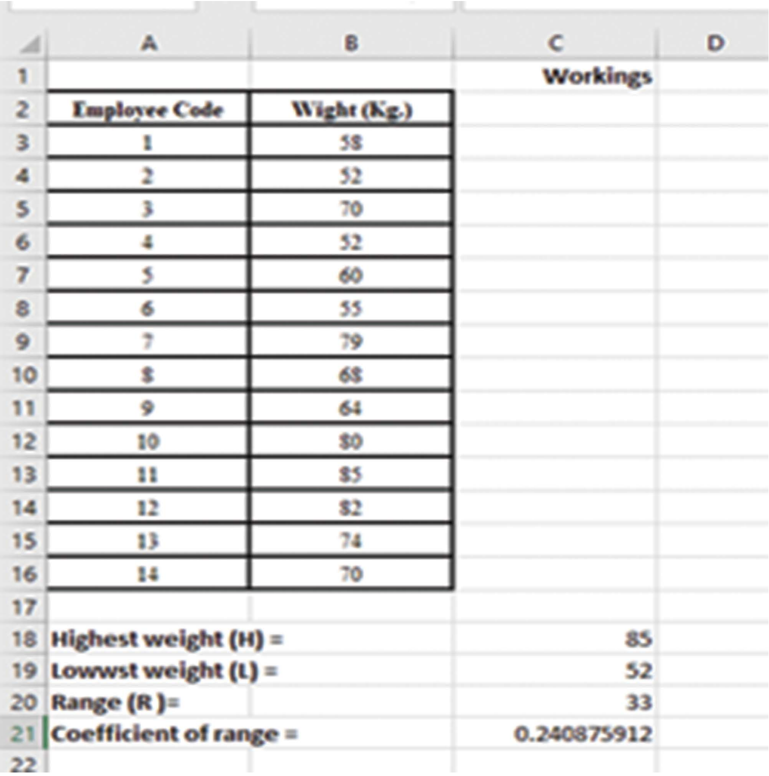
6.3 Quartile Deviation

A quartile, often known as the first, second, and third quartiles, is a deviation of a random variable based on the cumulative frequencies of a collection of observations that is divided into these three groups. The random variable's value in relation to one-fourth of the total frequency is known as the first quartile (Q1). The random variable's value



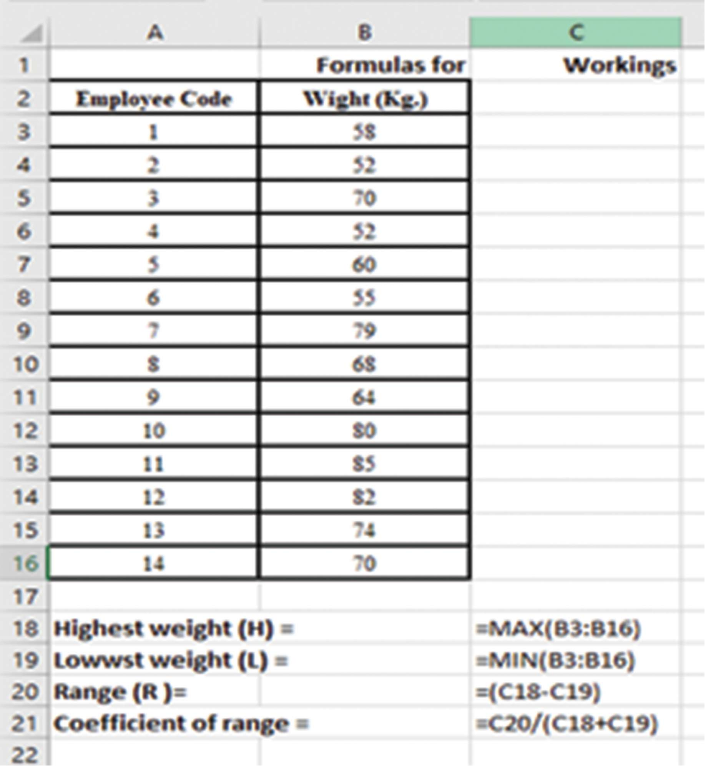
	A	B
1		
2	Employee Code	Wight (Kg.)
3	1	58
4	2	52
5	3	70
6	4	52
7	5	60
8	6	55
9	7	79
10	8	68
11	9	64
12	10	80
13	11	85
14	12	82
15	13	74
16	14	70
17		

Figure 6.1 Screenshot of input of Example 6.1



	A	B	C	D
1			Workings	
2	Employee Code	Wight (Kg.)		
3	1	58		
4	2	52		
5	3	70		
6	4	52		
7	5	60		
8	6	55		
9	7	79		
10	8	68		
11	9	64		
12	10	80		
13	11	85		
14	12	82		
15	13	74		
16	14	70		
17				
18	Highest weight (H) =		85	
19	Lowwst weight (L) =		52	
20	Range (R)=		33	
21	Coefficient of range =		0.240875912	
22				

Figure 6.2 Screenshot of working of range and coefficient of range for Example 6.1



	A	B	C
1		Formulas for	Workings
2	Employee Code	Wight (Kg.)	
3	1	58	
4	2	52	
5	3	70	
6	4	52	
7	5	60	
8	6	55	
9	7	79	
10	8	68	
11	9	64	
12	10	80	
13	11	85	
14	12	82	
15	13	74	
16	14	70	
17			
18	Highest weight (H) =		=MAX(B3:B16)
19	Lowwst weight (L) =		=MIN(B3:B16)
20	Range (R)=		=(C18-C19)
21	Coefficient of range =		=C20/(C18+C19)
22			

Figure 6.3 Screenshot of formulas of working of range and coefficient of range for Example 6.1

in relation to half of the overall frequency is represented by the second quartile (Q2). The random variable's value in relation to three-fourths of the total frequency is what is referred to as the third quartile, or Q3.

Quartile deviation (*QD*) is half of the difference between the third quartile and the first quartile, which is given by the following formula.

$$QD = \frac{Q_3 - Q_1}{2}$$

where
Q_1 is the first quartile
Q_3 is the third quartile
QD is the quartile deviation

For frequency-based data, the formulas for Q_1 and Q_3 are presented in the following.

$$First\, quartile\left(Q_1\right) = L_1 + \left(\frac{\frac{N}{4} - F}{f_1}\right) \times C$$

where
Q_1 is the first quartile
f_1 is the frequency of the first quartile class
F is the cumulative frequency of the immediate previous class interval with respect to the first quartile class
N is the sum of the frequencies of all the class intervals
C is the width of the class interval
L_1 is the lower limit of the value of the variable of the first quartile class

$$Third\,quartile(Q_3) = L_3 + \left(\frac{\frac{3N}{4} - F}{f_3}\right) \times C$$

where
Q_3 is the third quartile
f_3 is the frequency of the third quartile class
F is the cumulative frequency of the immediate previous class interval with respect to the third quartile class
N is the sum of the frequencies of all the class intervals
C is the width of the class interval
L_3 is the lower limit of the value of the variable of the third quartile class

6.3.1 Quartile Deviation of Ungrouped Data Using Quartile Function

This section presents the quartile deviation for ungrouped data. The formula for the quartile deviation is as follows.

$$QD = \frac{Q_3 - Q_1}{2}$$

where
Q_1 is the first quartile
Q_3 is the third quartile
QD is the quartile deviation

In Excel, the functions for quartiles are as listed.

QUARTILE.EXC
QUARTILE.INC

The QUARTILE.EXC function excludes the zeroth quartile, and the QUARTILE.INC function includes the zeroth quartile. Normally, QUERTILE.INC is used.

Example 6.2

The data for the height in centimetres of the employees of a small-scale company are shown in Table 6.3. Find the quartile deviation of this data.

Table 6.3 Data on Height of Employees

Employee Code	*Height (cm)*
1	158
2	170
3	180
4	152
5	160
6	155
7	179
8	168
9	164
10	180
11	172
12	170
13	155
14	152

Solution

The input of the data of this example is shown in Figure 6.4. The display after clicking the sequence of buttons Formulas ⟹> More Functions ⟹ Statistical ⟹ QUARTILE.INC is shown in Figure 6.5. The display after entering the range of cells containing data on height of employees in the box against Array and "1" in the box against Quart in the dropdown menu of the Figure 6.5 is shown in Figure 6.6. Quart is a number. When it is 0, it means 0th quartile, 1 means first quartile, 2 means median (second quartile), 3 means the third quartile, and 4 means the fourth quartile (full value). Clicking the OK button in the dropdown menu of Figure 6.6 gives the result of the first quartile in cell B18 of Figure 6.7. The computation of Q3 is done using the following formula entered in cell B19.

Formula : = QUARTILE.INC(Range of cells containing data on height of employees,

that is B3 : B16, Quartile number, that is 3)

The final formula for Q_3 is : QUARTILE.INC(B3 : B16, 3)

The formula for quartile deviation (QD) is entered in cell B20. The screenshot of the result of Q_3 and QD along with Q_1 is shown in Figure 6.8. The formulas of Q_1, Q_3, and QD are shown in Figure 6.9.

6.3.2 Quartile Deviation of Grouped Data With Frequency Using Excel Sheets

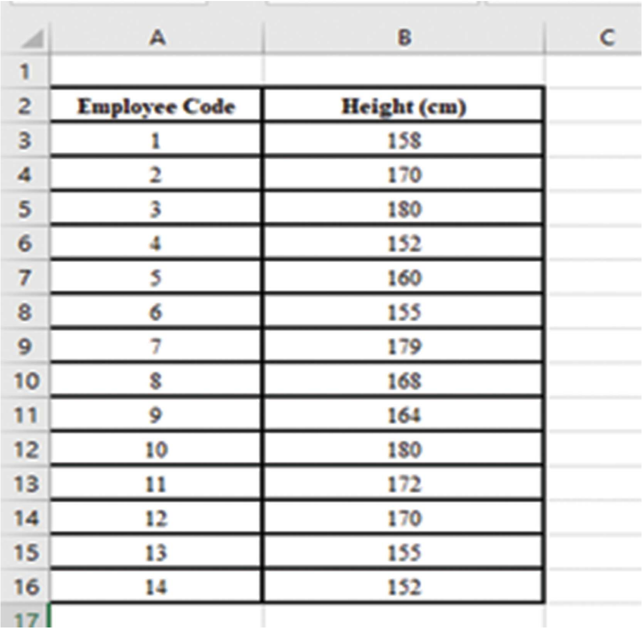
In reality, there may be grouped data with frequencies. The formula for the quartile deviation of grouped data with frequencies is as follows.

$$QD = \frac{Q_3 - Q_1}{2}$$

where

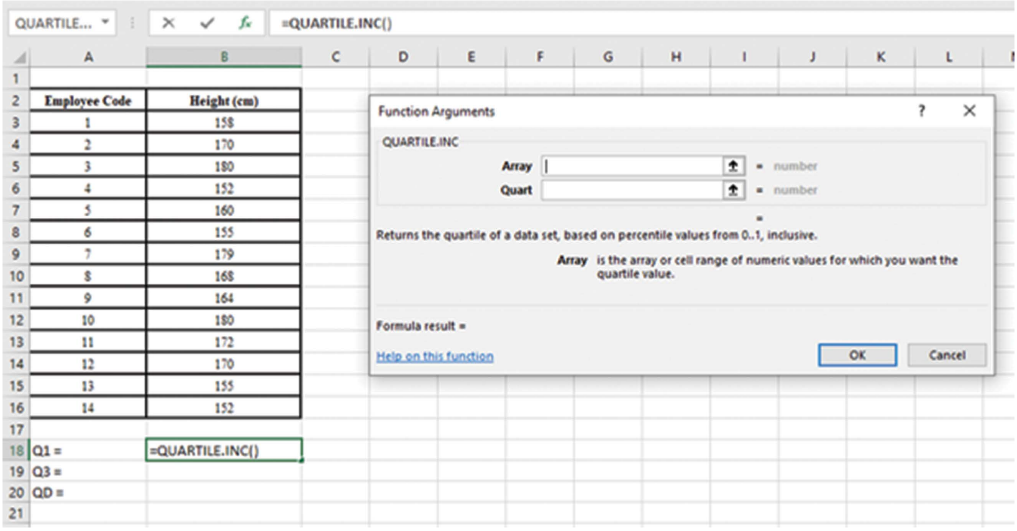
Q_1 is the first quartile

Q_3 is the third quartile



Employee Code	Height (cm)
1	158
2	170
3	180
4	152
5	160
6	155
7	179
8	168
9	164
10	180
11	172
12	170
13	155
14	152

Figure 6.4 Screenshot of input for Example 6.2



=QUARTILE.INC()

Employee Code	Height (cm)
1	158
2	170
3	180
4	152
5	160
6	155
7	179
8	168
9	164
10	180
11	172
12	170
13	155
14	152
Q1 =	=QUARTILE.INC()
Q3 =	
QD =	

Function Arguments

QUARTILE.INC

Array | = number

Quart = number

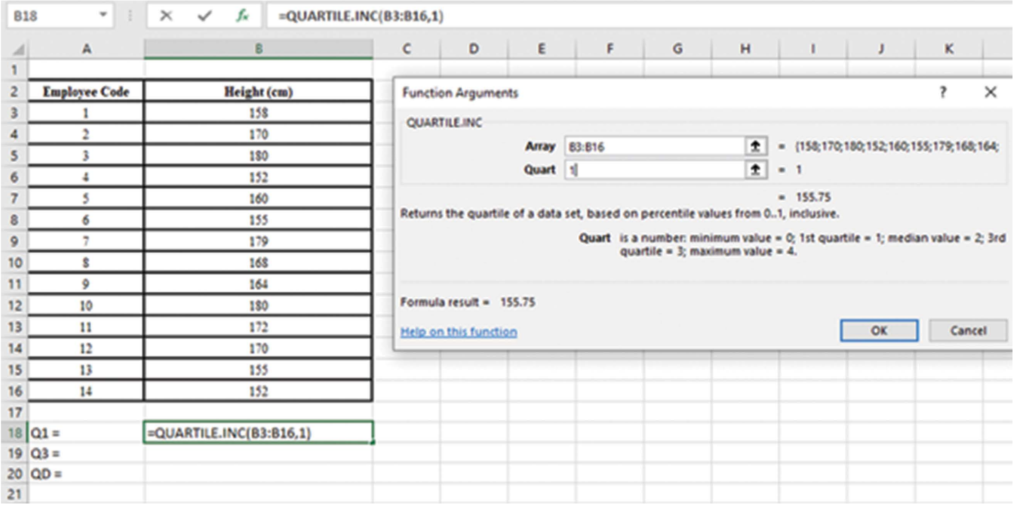
Returns the quartile of a data set, based on percentile values from 0..1, inclusive.

Array is the array or cell range of numeric values for which you want the quartile value.

Formula result =

Help on this function

Figure 6.5 Screenshot after clicking sequence of buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ QUARTILE.INC



B18 =QUARTILE.INC(B3:B16,1)

Employee Code	Height (cm)
1	158
2	170
3	180
4	152
5	160
6	155
7	179
8	168
9	164
10	180
11	172
12	170
13	155
14	152

Q1 = =QUARTILE.INC(B3:B16,1)

Q3 =

QD =

Function Arguments

QUARTILE.INC

Array B3:B16 = {158;170;180;152;160;155;179;168;164;

Quart 1 = 1

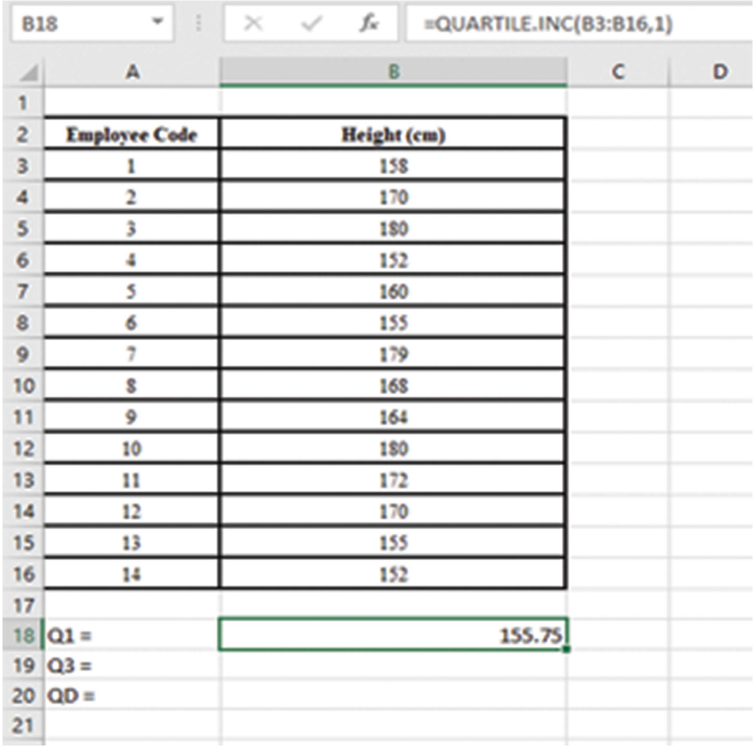
= 155.75

Returns the quartile of a data set, based on percentile values from 0..1, inclusive.

Quart is a number: minimum value = 0; 1st quartile = 1; median value = 2; 3rd quartile = 3; maximum value = 4.

Formula result = 155.75

Figure 6.6 Screenshot after entering the range of cells containing data on height and value for Quart as 1



B18 =QUARTILE.INC(B3:B16,1)

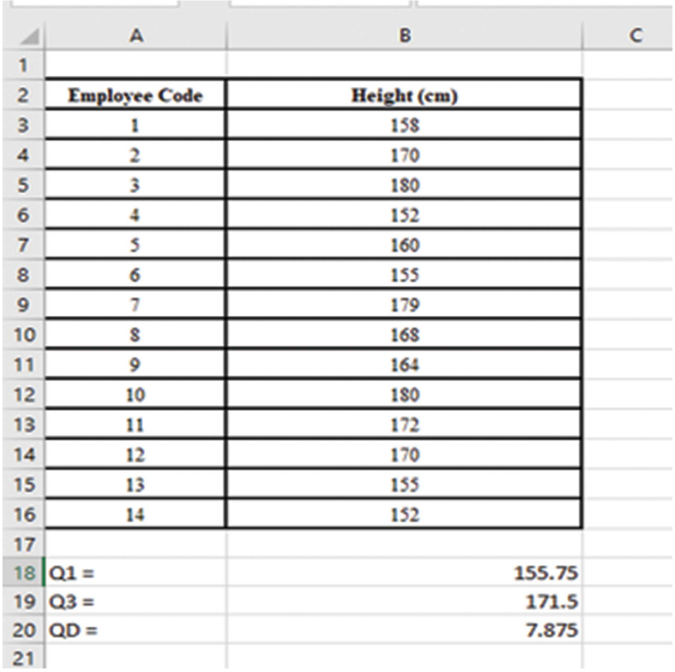
Employee Code	Height (cm)
1	158
2	170
3	180
4	152
5	160
6	155
7	179
8	168
9	164
10	180
11	172
12	170
13	155
14	152

Q1 = 155.75

Q3 =

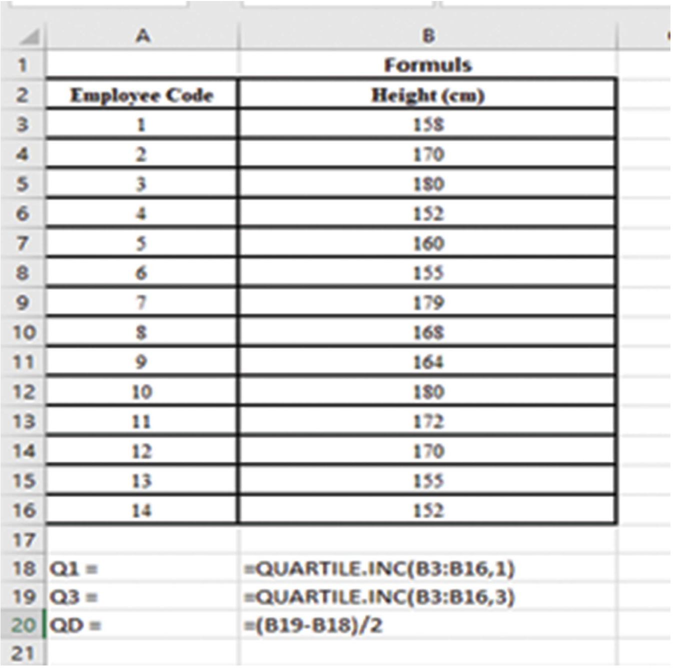
QD =

Figure 6.7 Screenshot of display after clicking the OK button in the dropdown menu of Figure 6.6



	A	B	C
1			
2	Employee Code	Height (cm)	
3	1	158	
4	2	170	
5	3	180	
6	4	152	
7	5	160	
8	6	155	
9	7	179	
10	8	168	
11	9	164	
12	10	180	
13	11	172	
14	12	170	
15	13	155	
16	14	152	
17			
18	Q1 =	155.75	
19	Q3 =	171.5	
20	QD =	7.875	
21			

Figure 6.8 Screenshot of working of Q_3 and QD of ungrouped data for Example 6.2



	A	B
1		Formuls
2	Employee Code	Height (cm)
3	1	158
4	2	170
5	3	180
6	4	152
7	5	160
8	6	155
9	7	179
10	8	168
11	9	164
12	10	180
13	11	172
14	12	170
15	13	155
16	14	152
17		
18	Q1 =	=QUARTILE.INC(B3:B16,1)
19	Q3 =	=QUARTILE.INC(B3:B16,3)
20	QD =	=(B19-B18)/2
21		

Figure 6.9 Screenshot of formulas of working of quartile deviation of ungrouped data for Example 6.2

QD is the quartile deviation of the grouped data with frequency
For frequency-based data, the formulas for Q_1 and Q_3 are presented in the following.

$$First\,quartile\left(Q_1\right) = L_1 + \left(\frac{\frac{N}{4} - F}{f_1}\right) \times C$$

where
Q_1 is the first quartile
f_1 is the frequency of the first quartile class
F is the cumulative frequency of the immediate previous class interval with respect to the first quartile class
N is the sum of the frequencies of all the class intervals
C is the width of the class interval
L_1 is the lower limit of the value of the variable of the first quartile class

$$Third\,quartile\left(Q_3\right) = L_3 + \left(\frac{\frac{3N}{4} - F}{f_3}\right) \times C$$

where
Q_3 is the third quartile
f_3 is the frequency of the third quartile class
F is the cumulative frequency of the immediate previous class interval with respect to the third quartile class
N is the sum of the frequencies of all the class intervals
C is the width of the class interval
L_3 is the lower limit of the value of the variable of the third quartile class

Example 6.3

The distribution of loan amounts (in lakhs of rupees) sanctioned to industries by a bank is shown as in Table 6.4. Find the quartile deviation of the loan amount using Excel.

Table 6.4 Distribution of Loan Amounts

Loan Amount (₹ in Lakhs)	*No. of Firms*
Below 4	20
4–8	16
8–12	20
12–16	28
16–20	40
20–24	25
24–28	23
28–32	28
32–36	6
More than 36	2

Solution

The data for Example 6.3 are shown with some modifications in Table 6.5.

For frequency-based data, the formulas for Q_1 and Q_3 are presented in the following.

$$\textit{First quartile}\left(Q_1\right) = L_1 + \left(\frac{\frac{N}{4} - F}{f_1}\right) \times C$$

$$\textit{Third quartile}\left(Q_3\right) = L_3 + \left(\frac{\frac{3N}{4} - F}{f_3}\right) \times C$$

The input of Example 6.3 is shown in Figure 6.10. The working of *Q*1, *Q*3, and *QD* are shown in Figure 6.11. The formulas for the working of *Q*1, *Q*3, and *QD* are shown in Figure 6.12. The quartile deviation of the loan amount is ₹ 7.0086956 lakhs.

6.4 Average Deviation Using Excel Sheets

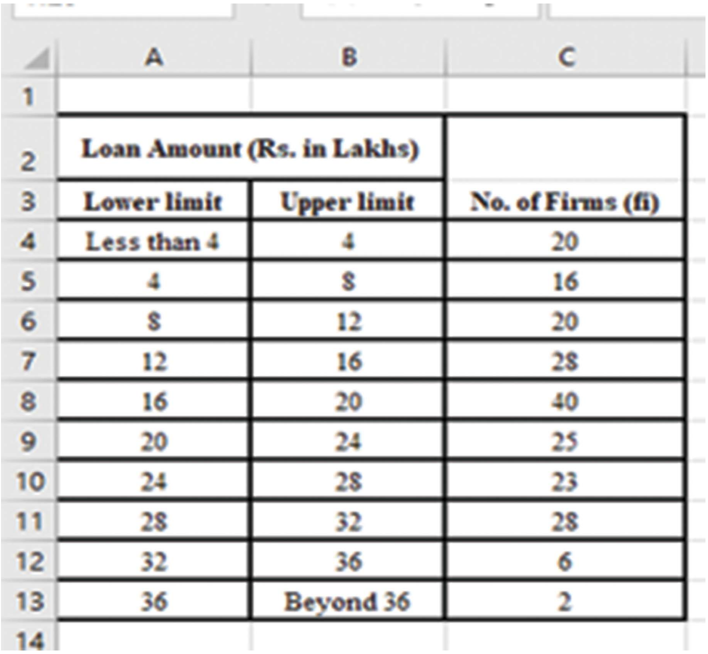
Average deviation is the mean of the absolute deviations of the observations from the mean of those observations.

The formula for the average deviation (*AD*) of ungrouped data is presented in the following.

$$AD = \frac{\sum_{i=1}^{n} \left|X_i - \bar{X}\right|}{N}$$

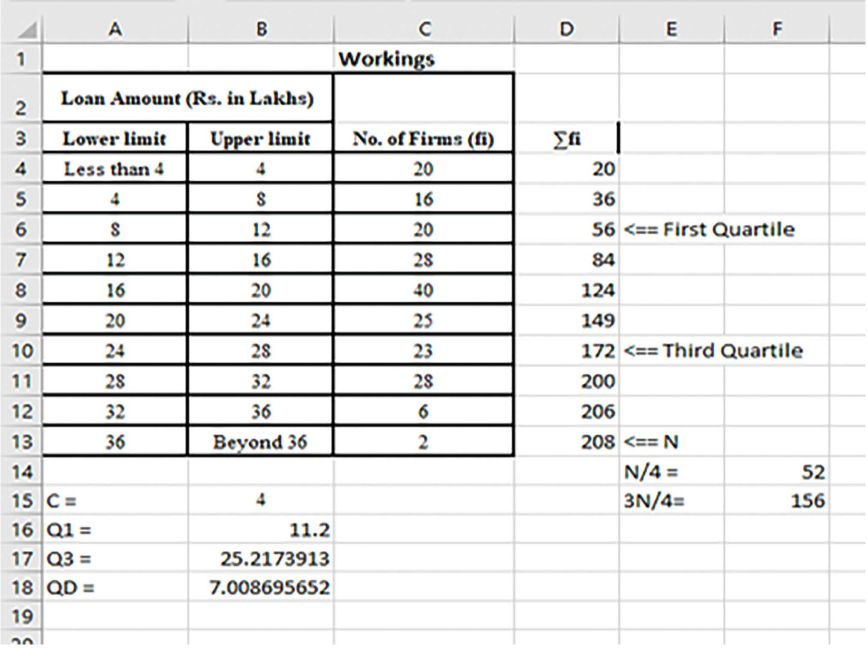
Table 6.5 Data for Example 6.3

Loan Amount (₹ in Lakhs)		*No. of Firms*
Lower Limit	*Upper Limit*	
Less than 4	4	20
4	8	16
8	12	20
12	16	28
16	20	40
20	24	25
24	28	23
28	32	28
32	36	6
36	Beyond 36	2



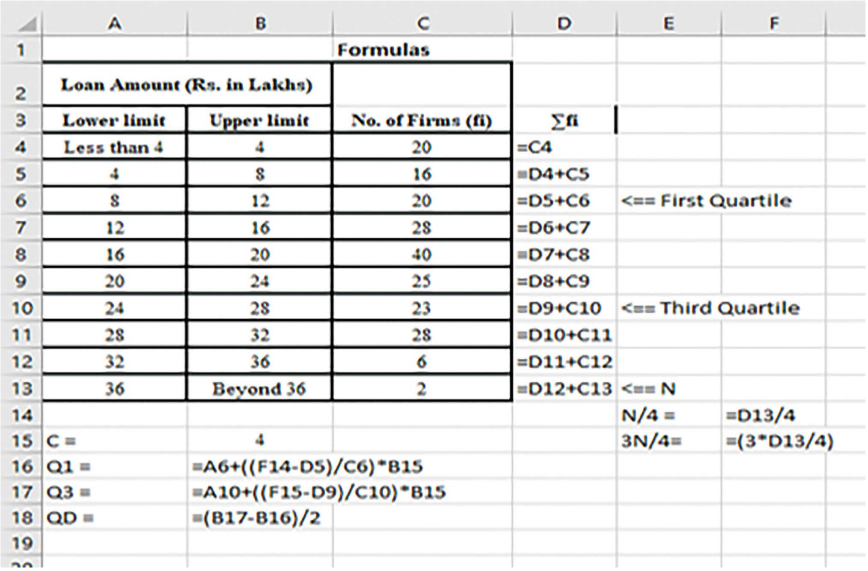
	A	B	C
1			
2	Loan Amount (Rs. in Lakhs)		
3	Lower limit	Upper limit	No. of Firms (fi)
4	Less than 4	4	20
5	4	8	16
6	8	12	20
7	12	16	28
8	16	20	40
9	20	24	25
10	24	28	23
11	28	32	28
12	32	36	6
13	36	Beyond 36	2
14			

Figure 6.10 Screenshot of input for Example 6.3



	A	B	C	D	E	F
1			Workings			
2	Loan Amount (Rs. in Lakhs)					
3	Lower limit	Upper limit	No. of Firms (fi)	∑fi		
4	Less than 4	4	20	20		
5	4	8	16	36		
6	8	12	20	56	<== First Quartile	
7	12	16	28	84		
8	16	20	40	124		
9	20	24	25	149		
10	24	28	23	172	<== Third Quartile	
11	28	32	28	200		
12	32	36	6	206		
13	36	Beyond 36	2	208	<== N	
14					N/4 =	52
15	C =	4			3N/4=	156
16	Q1 =	11.2				
17	Q3 =	25.2173913				
18	QD =	7.008695652				
19						

Figure 6.11 Screenshot of working of Q1, Q3, and QD for Example 6.3



	A	B	C	D	E	F
1			Formulas			
2	Loan Amount (Rs. in Lakhs)					
3	Lower limit	Upper limit	No. of Firms (fi)	∑fi		
4	Less than 4	4	20	=C4		
5	4	8	16	=D4+C5		
6	8	12	20	=D5+C6	<== First Quartile	
7	12	16	28	=D6+C7		
8	16	20	40	=D7+C8		
9	20	24	25	=D8+C9		
10	24	28	23	=D9+C10	<== Third Quartile	
11	28	32	28	=D10+C11		
12	32	36	6	=D11+C12		
13	36	Beyond 36	2	=D12+C13	<== N	
14					N/4 =	=D13/4
15	C =	4			3N/4=	=(3*D13/4)
16	Q1 =	=A6+((F14-D5)/C6)*B15				
17	Q3 =	=A10+((F15-D9)/C10)*B15				
18	QD =	=(B17-B16)/2				
19						

Figure 6.12 Screenshot of formulas for working of Q1, Q3, and QD for Example 6.3

or

$$AD = \frac{\sum_{i=1}^{n} \left| X_i - Median \right|}{N}$$

where
n is the number of observations
X_i is the i^{th} observation, i = 1, 2, 3, . . ., n
$\bar{X}$ is the mean of the observations
AD is the average deviation
N is the sum of the frequencies of all class intervals

The formulas for the average deviation of grouped data with frequencies are presented in the following.

$$AD = \frac{\sum_{i=1}^{n} f_i \times \left| X_i - \bar{X} \right|}{N}$$

or

$$AD = \frac{\sum_{i=1}^{n} f_i \times \left| X_i - Median \right|}{N}$$

where
n is the number of class intervals
X_i is the mid-point of the i^{th} class interval, i = 1, 2, 3, . . ., n
f_i is the frequency of the i^{th} class interval, i = 1, 2, 3, . . . , n

N is the sum of the frequencies of all class intervals
$\bar{X}$ is the mean of the observations
AD is the average deviation
The coefficient of average deviation based on the mean as well as the median is as follows.

$$Coefficient\,of\ AD = \frac{AD}{Mean}\,or\,\frac{AD}{Median}$$

Example 6.4

The distribution of the number of shares applied to the recently concluded equity issue of an automobile company is shown in Table 6.6. Compute the average deviation as well as coefficient of average deviation based on the mean for the number of shares applied.

Solution

The data for this example with some modifications in the format are shown in Table 6.7.

The formula for the average deviation of grouped data with frequencies is presented in the following.

$$AD = \frac{\sum_{i=1}^{n} f_i \times |X_i - \bar{X}|}{N}$$

Table 6.6 Distribution of Number of Shares Applied

Number of Shares Applied	*Number of Applications*
50–100	1,500
100–150	11,500
150–200	11,800
200–250	11,600
250–300	11,000
300–350	1,850
350–400	1,800
400–450	1,600
450–500	1,400

Table 6.7 Data for Example 6.4

Number of Shares Applied		*Midpoint*	*Number of Applications*
Lower Limit	*Upper Limit*		
50	100	75	1,500
100	150	125	11,500
150	200	175	11,800
200	250	225	11,600
250	300	275	11,000
300	350	325	1,850
350	400	375	1,800
400	450	425	1,600
450	500	475	1,400

where
n is the number of class intervals
X_i is the mid-point of the i^{th} class interval, i = 1, 2, 3, . . ., n
f_i is the frequency of the i^{th} class interval, i = 1, 2, 3, . . . , n
N is the sum of the frequencies of all class intervals
$\bar{X}$ is the mean of the observations
AD is the average deviation
The coefficient of average deviation based on the mean is as follows.

$$Coefficient\,of\;AD = \frac{AD}{Mean}$$

The input for Example 6.4 in an Excel sheet is shown in Figure 6.13. The working of average deviation (AD) and the coefficient of AD are shown in Figure 6.14. The formulas for the working of average deviation and the coefficient of AD are shown in Figure 6.15.

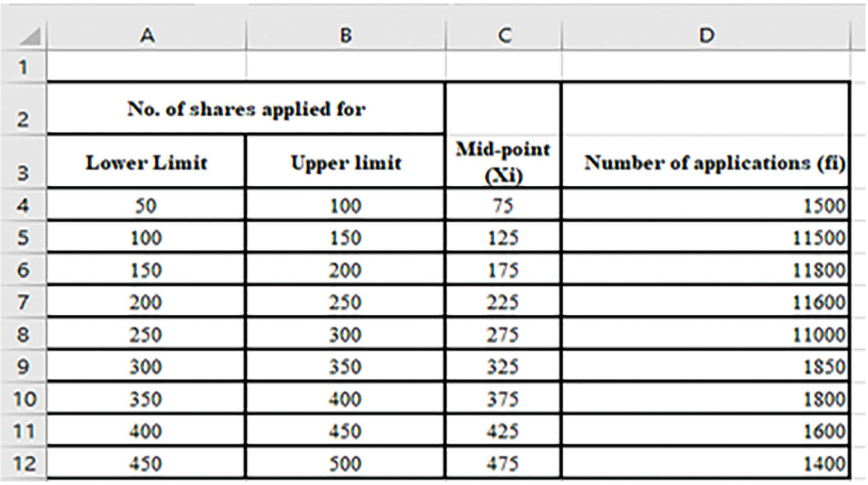
From Figure 6.14, the results are as follows.

Average deviation (AD) = 67.7867908 shares
Coefficient of AD = 0.30863439

6.5 Standard Deviation

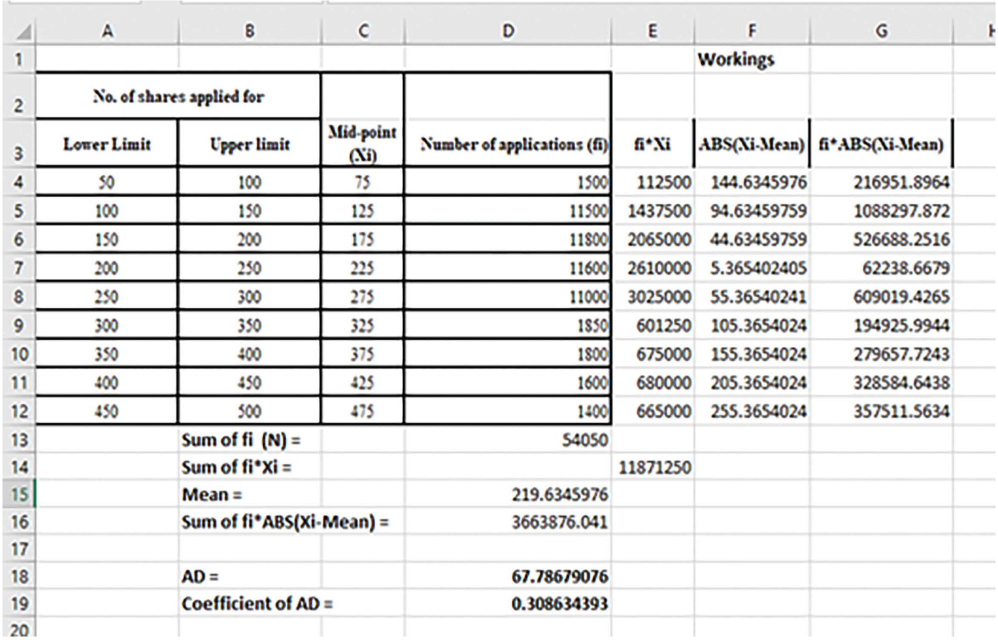
The average of the squares of departures from the mean of a set of data from individual observations is known as the variance, which is a measure of variation [1, 2]. The square root of the variance is the standard deviation. It shows how the data are distributed around their mean.

In this section, the standard deviation of ungrouped data and that for grouped data are presented using Excel.



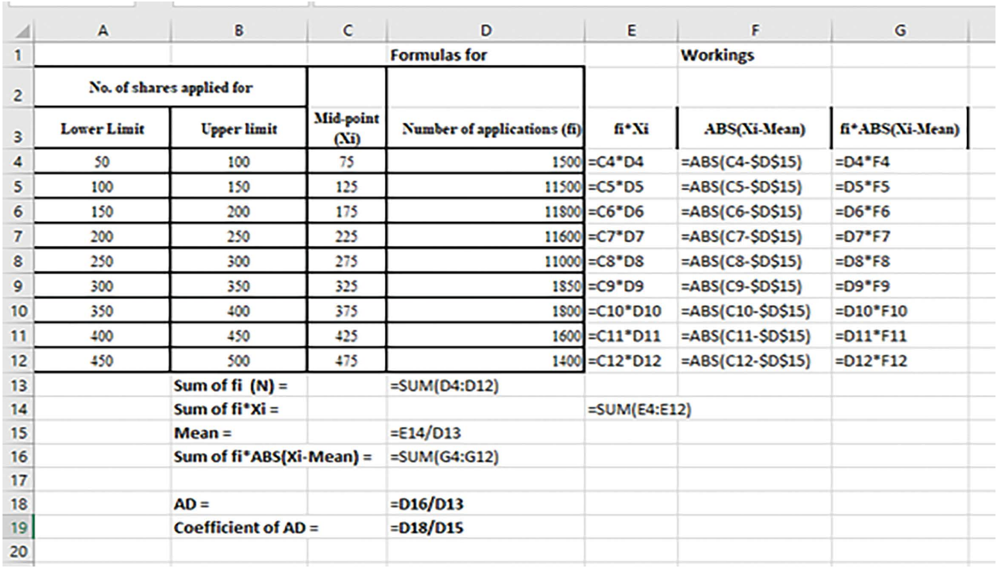
	A	B	C	D
1				
2	No. of shares applied for		Mid-point (Xi)	Number of applications (fi)
3	Lower Limit	Upper limit		
4	50	100	75	1500
5	100	150	125	11500
6	150	200	175	11800
7	200	250	225	11600
8	250	300	275	11000
9	300	350	325	1850
10	350	400	375	1800
11	400	450	425	1600
12	450	500	475	1400

Figure 6.13 Screenshot of input for Example 6.4



	A	B	C	D	E	F	G
1						Workings	
2	No. of shares applied for						
3	Lower Limit	Upper limit	Mid-point (Xi)	Number of applications (fi)	fi*Xi	ABS(Xi-Mean)	fi*ABS(Xi-Mean)
4	50	100	75	1500	112500	144.6345976	216951.8964
5	100	150	125	11500	1437500	94.63459759	1088297.872
6	150	200	175	11800	2065000	44.63459759	526688.2516
7	200	250	225	11600	2610000	5.365402405	62238.6679
8	250	300	275	11000	3025000	55.36540241	609019.4265
9	300	350	325	1850	601250	105.3654024	194925.9944
10	350	400	375	1800	675000	155.3654024	279657.7243
11	400	450	425	1600	680000	205.3654024	328584.6438
12	450	500	475	1400	665000	255.3654024	357511.5634
13		Sum of fi (N) =		54050			
14		Sum of fi*Xi =			11871250		
15		Mean =		219.6345976			
16		Sum of fi*ABS(Xi-Mean) =		3663876.041			
17							
18		AD =		**67.78679076**			
19		Coefficient of AD =		**0.308634393**			
20							

Figure 6.14 Screenshot of working of AD and coefficient of AD for Example 6.4



	A	B	C	D	E	F	G
1				Formulas for		Workings	
2	No. of shares applied for						
3	Lower Limit	Upper limit	Mid-point (Xi)	Number of applications (fi)	fi*Xi	ABS(Xi-Mean)	fi*ABS(Xi-Mean)
4	50	100	75	1500	=C4*D4	=ABS(C4-$D$15)	=D4*F4
5	100	150	125	11500	=C5*D5	=ABS(C5-$D$15)	=D5*F5
6	150	200	175	11800	=C6*D6	=ABS(C6-$D$15)	=D6*F6
7	200	250	225	11600	=C7*D7	=ABS(C7-$D$15)	=D7*F7
8	250	300	275	11000	=C8*D8	=ABS(C8-$D$15)	=D8*F8
9	300	350	325	1850	=C9*D9	=ABS(C9-$D$15)	=D9*F9
10	350	400	375	1800	=C10*D10	=ABS(C10-$D$15)	=D10*F10
11	400	450	425	1600	=C11*D11	=ABS(C11-$D$15)	=D11*F11
12	450	500	475	1400	=C12*D12	=ABS(C12-$D$15)	=D12*F12
13		Sum of fi (N) =		=SUM(D4:D12)			
14		Sum of fi*Xi =			=SUM(E4:E12)		
15		Mean =		=E14/D13			
16		Sum of fi*ABS(Xi-Mean) =		=SUM(G4:G12)			
17							
18		AD =		**=D16/D13**			
19		Coefficient of AD =		**=D18/D15**			
20							

Figure 6.15 Screenshot of formulas for working of AD and coefficient of AD for Example 6.4

6.5.1 Standard Deviation of Ungrouped Data Using STDEV Function

The standard deviation (σ) of the ungrouped data is the square root of the variance (σ^2) of the observations of a set of observations pertaining to a variable X_i, where i varies from 1 to n and n is the number of observations. The formula for the variance of the ungrouped data is as follows.

$$Variance\left(\sigma^2\right) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$$

where
n is the number of observations of a sample/population
X_i is the i^{th} observation of the variable of interest for i = 1, 2, 3, . . . , n
$\bar{X}$ is the mean of the observations
σ^2 is the variance of the observations

This formula gives the variance of the sample (S^2). In the formula, if the denominator is changed to n, then it becomes the variance of the population (σ^2).

Therefore, the standard deviation (σ) of the ungrouped data of the population is given by the following formula.

$$Standard\,deviation\left(\sigma\right) = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n}}$$

In Excel, the standard deviation of a set of observations of a sample and that of a set of observations of a population can be obtained using the following functions, respectively [3, 4].

=STDEV.S(Range of cells containing data of sample excluding logical and text)
=STDEV.P(Range of cells containing data of population excluding logical and text)

The first function uses the following formula to find the standard deviation of a sample.

$$Standard\,deviation\left(S\right) = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}}$$

The second function uses the following formula to find the standard deviation of a population.

$$Standard\,deviation\left(\sigma\right) = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n}}$$

Example 6.5

The yield quantities in terms of 75-kg bags of paddy in 10 different plots for application of each of the two types of fertiliser, Type *X* and Type *Y*, are summarised in Table 6.8.

Table 6.8 Yields (Standard Bags of 75 kg) of Paddy for Different Type of Fertilisers

Plot No.	*Type of Fertiliser*	
	X	*Y*
1	10	12
2	10	11
3	10	9
4	10	10
5	10	12
6	10	9
7	10	12
8	10	8
9	10	9
10	10	8

1. Identify the best fertiliser based on the average yield per plot.
2. If the average yield per plot is the same for both fertiliser types, then find the best fertiliser based on the least standard deviation.

Solution

The data for Example 6.5 are shown in Table 6.9.

1. Determination of Mean Yield for Each Fertiliser Type
   The input of the data for Example 6.5 in an Excel sheet is shown in Figure 6.16. The computations of the mean yield quantities for different fertiliser types are shown in Figure 6.17, and their formulas are shown in Figure 6.18. The mean yield for fertiliser type *X* is 10 bags, and that for the fertiliser type *Y* is also 10 bags. So, the investigator cannot distinguish between the fertiliser types to find the best type.
2. Determination of Standard Deviation of Yield for Each Fertiliser Type

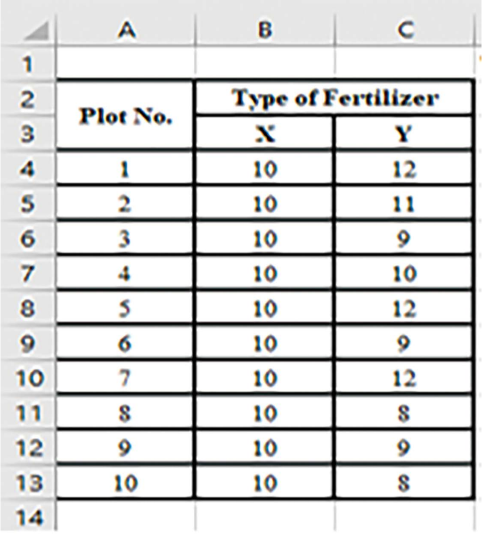
Clicking the sequence of button Formulas $\Longrightarrow$ More Functions $\Longrightarrow$ Statistical $\Longrightarrow$ STDEV.S gives the display shown in Figure 6.19. After copying the range of cells from B4 to B13, which contain data of Fertiliser type *X* in the box against Number 1 of the dropdown menu of Figure 6.19, the display against the box, "Number 1" is shown in Figure 6.20. Then clicking the OK button in the dropdown menu of Figure 6.20 gives the result of the standard deviation of yield for fertiliser type *X*, as shown in cell F15 of Figure 6.21. The standard deviation of yield of paddy for fertiliser type *Y* is computed using the following formula, as entered in cell F18 in Figure 6.22.

$$= \text{STDEV.S}(\text{C4} : \text{C13})$$

Pressing the Enter key gives the standard deviation of yield of paddy for fertiliser type *Y*, as shown in cell F18 of Figure 6.23.

Table 6.9 Data for Example 6.5

Plot No.	*Type of Fertiliser*	
	X	*Y*
1	10	12
2	10	11
3	10	9
4	10	10
5	10	12
6	10	9
7	10	12
8	10	8
9	10	9
10	10	8



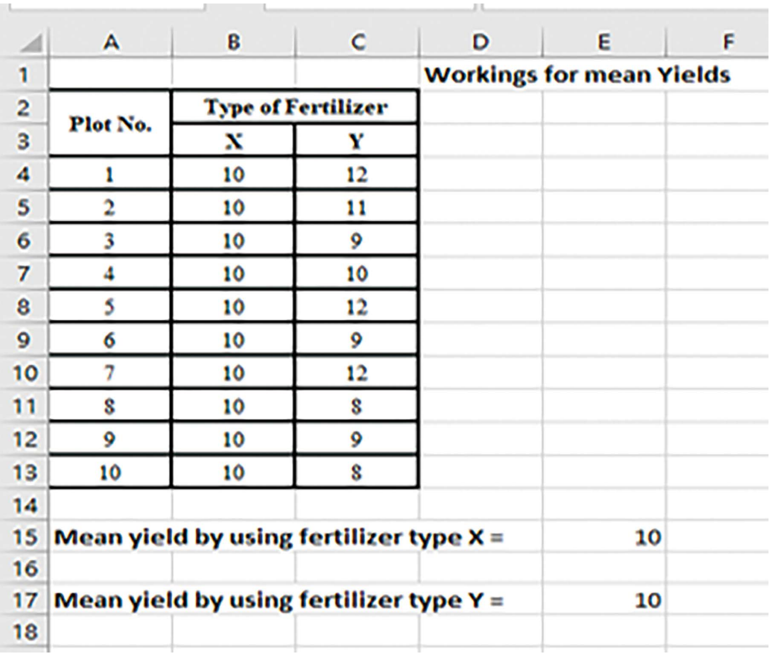
	A	B	C
1			
2	**Plot No.**	**Type of Fertilizer**	
3		**X**	**Y**
4	1	10	12
5	2	10	11
6	3	10	9
7	4	10	10
8	5	10	12
9	6	10	9
10	7	10	12
11	8	10	8
12	9	10	9
13	10	10	8
14			

Figure 6.16 Screenshot of input for Example 6.5 in Excel sheet

The standard deviation of yield for fertiliser type *X* is 0 bags, and that for the fertiliser type *Y* is 1.632993162 bags. So, fertiliser type *X* is the best in terms of least standard deviation.

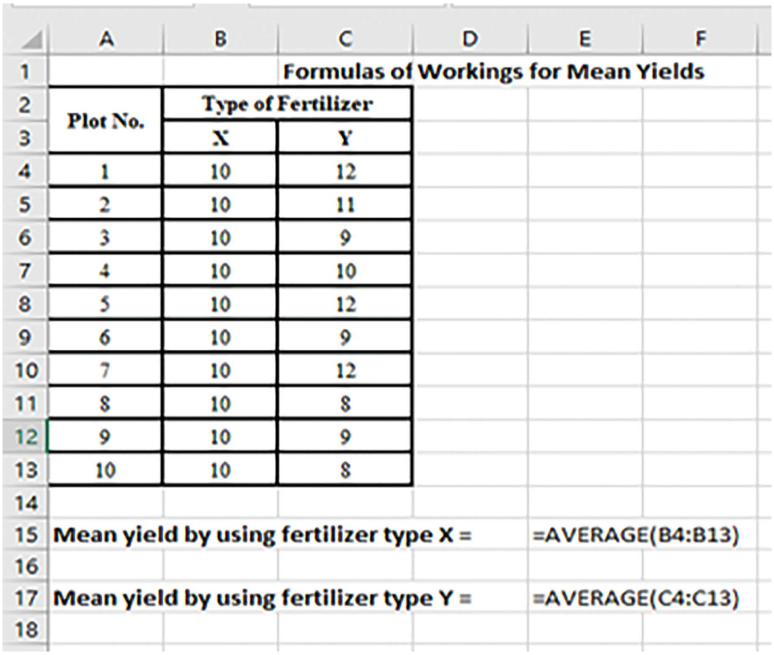
6.5.2 Standard Deviation of Ungrouped Data With Frequencies

In some cases, the values of an important variable exist continuously in discrete form, with frequencies ranging from a lower value to a higher value. The everyday need for bread packets with frequency is the finest illustration. Table 6.10 displays some sample data for this illustration.



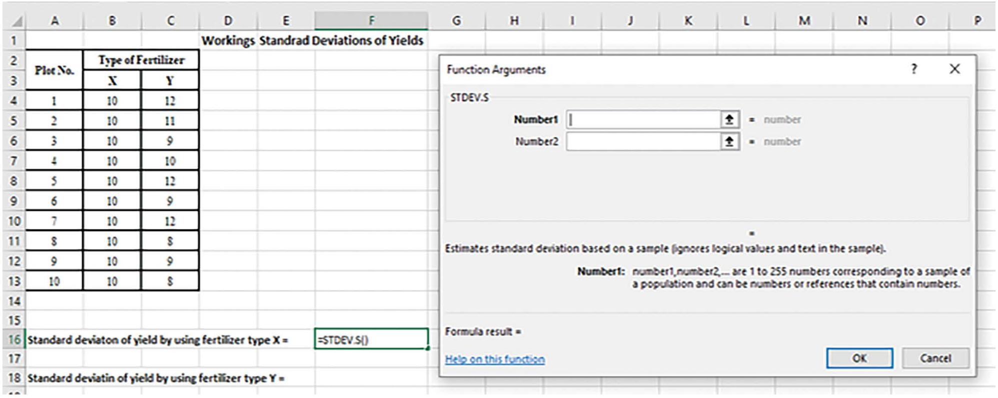
	A	B	C	D	E	F
1				Workings for mean Yields		
2	Plot No.	Type of Fertilizer				
3		X	Y			
4	1	10	12			
5	2	10	11			
6	3	10	9			
7	4	10	10			
8	5	10	12			
9	6	10	9			
10	7	10	12			
11	8	10	8			
12	9	10	9			
13	10	10	8			
14						
15	Mean yield by using fertilizer type X =				10	
16						
17	Mean yield by using fertilizer type Y =				10	
18						

Figure 6.17 Screenshot of computation of mean yields per plot of fertiliser types



	A	B	C	D	E	F
1			Formulas of Workings for Mean Yields			
2	Plot No.	Type of Fertilizer				
3		X	Y			
4	1	10	12			
5	2	10	11			
6	3	10	9			
7	4	10	10			
8	5	10	12			
9	6	10	9			
10	7	10	12			
11	8	10	8			
12	9	10	9			
13	10	10	8			
14						
15	Mean yield by using fertilizer type X =				=AVERAGE(B4:B13)	
16						
17	Mean yield by using fertilizer type Y =				=AVERAGE(C4:C13)	
18						

Figure 6.18 Screenshot of formulas for computation of mean yields per plot of fertiliser types



Workings Standrad Deviations of Yields

Plot No.	Type of Fertilizer	
	X	Y
1	10	12
2	10	11
3	10	9
4	10	10
5	10	12
6	10	9
7	10	12
8	10	8
9	10	9
10	10	8

Standard deviaton of yield by using fertilizer type X = =STDEV.S()

Standard deviatin of yield by using fertilizer type Y =

Function Arguments

STDEV.S

Number1 = number

Number2 = number

Estimates standard deviation based on a sample (ignores logical values and text in the sample).

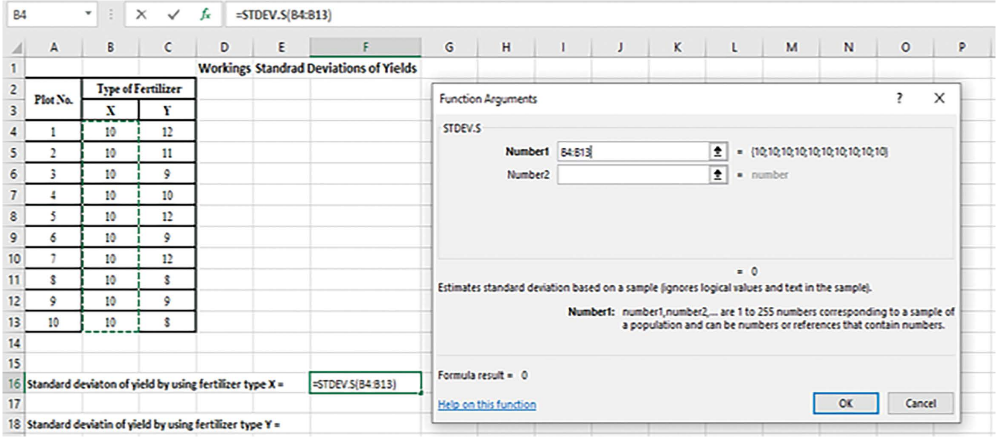
Number1: number1,number2,... are 1 to 255 numbers corresponding to a sample of a population and can be numbers or references that contain numbers.

Formula result =

Help on this function

OK Cancel

Figure 6.19 Screenshot of display of sequence of button clicks, Formulas ⟹ More Functions ⟹ Statistical ⟹ STDEV.S



B4 =STDEV.S(B4:B13)

Workings Standrad Deviations of Yields

Plot No.	Type of Fertilizer	
	X	Y
1	10	12
2	10	11
3	10	9
4	10	10
5	10	12
6	10	9
7	10	12
8	10	8
9	10	9
10	10	8

Standard deviaton of yield by using fertilizer type X = =STDEV.S(B4:B13)

Standard deviatin of yield by using fertilizer type Y =

Function Arguments

STDEV.S

Number1 B4:B13 = {10;10;10;10;10;10;10;10;10;10}

Number2 = number

= 0

Estimates standard deviation based on a sample (ignores logical values and text in the sample).

Number1: number1,number2,... are 1 to 255 numbers corresponding to a sample of a population and can be numbers or references that contain numbers.

Formula result = 0

Help on this function

OK Cancel

Figure 6.20 Screenshot after entering range of cells from B4 to B13 in the cell against Number 1 in the dropdown menu of Figure 6.19

The formula for the standard deviation of demand for this type of data is shown in the following.

$$Standard\,deviation(\sigma) = \sqrt{\frac{\sum_{i=1}^{n} f_i \times \left(X_i - \bar{X}\right)^2}{N}}$$

where

n is the number of demand values

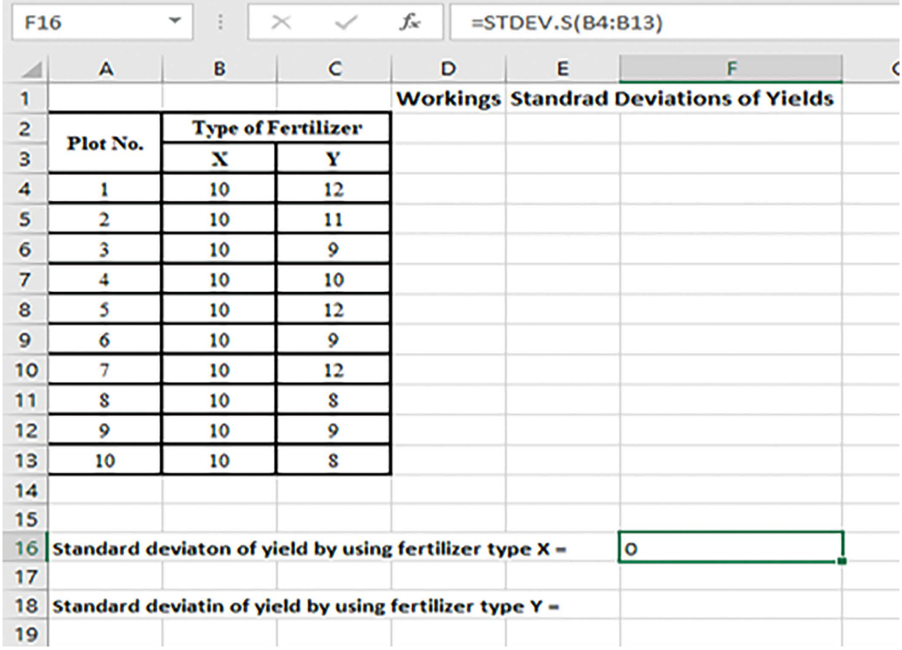
f_i is the frequency of the i^{th} demand value for i = 1, 2, 3, . . . , n

N is the sum of the frequencies

X_i is the i^{th} demand value, where i = 1, 2,3, . . . , n

$\bar{X}$ is the mean demand

σ is the standard deviation of the demand



F16 =STDEV.S(B4:B13)

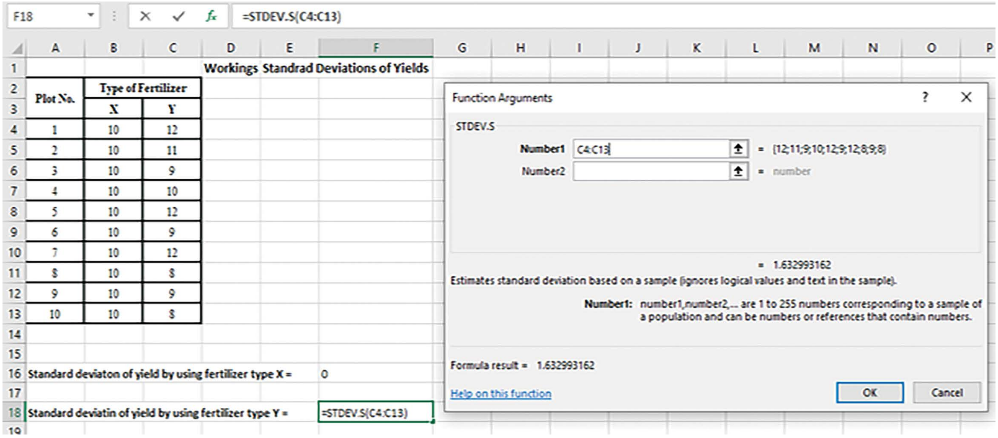
Workings Standrad Deviations of Yields

Plot No.	Type of Fertilizer	
	X	Y
1	10	12
2	10	11
3	10	9
4	10	10
5	10	12
6	10	9
7	10	12
8	10	8
9	10	9
10	10	8

Standard deviaton of yield by using fertilizer type X = 0

Standard deviatin of yield by using fertilizer type Y =

Figure 6.21 Screenshot after clicking the OK button in the dropdown menu of Figure 6.20



F18 =STDEV.S(C4:C13)

Workings Standrad Deviations of Yields

Plot No.	Type of Fertilizer	
	X	Y
1	10	12
2	10	11
3	10	9
4	10	10
5	10	12
6	10	9
7	10	12
8	10	8
9	10	9
10	10	8

Standard deviaton of yield by using fertilizer type X = 0

Standard deviatin of yield by using fertilizer type Y = =STDEV.S(C4:C13)

Function Arguments

STDEV.S

Number1 C4:C13 = {12;11;9;10;12;9;12;8;9;8}

Number2 = number

= 1.632993162

Estimates standard deviation based on a sample (ignores logical values and text in the sample).

Number1: number1,number2,... are 1 to 255 numbers corresponding to a sample of a population and can be numbers or references that contain numbers.

Formula result = 1.632993162

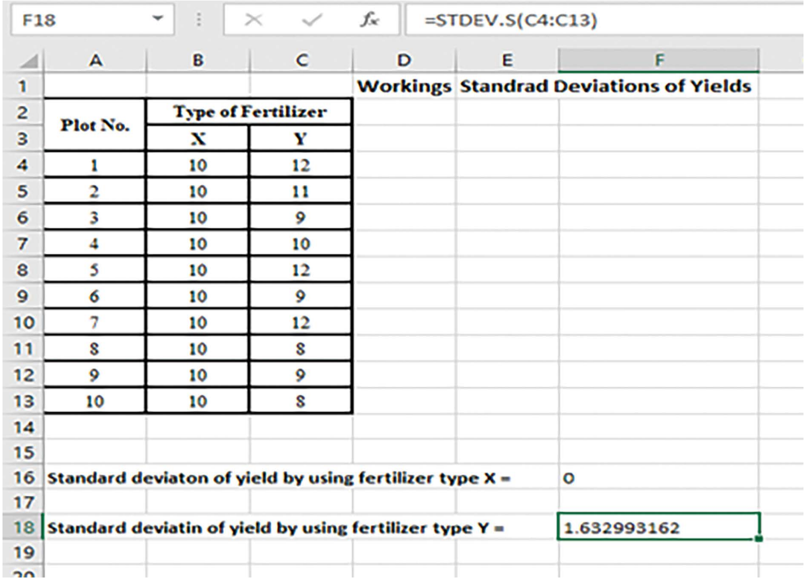
Figure 6.22 Screenshot of formula for STDEV in cell F18 for fertiliser type *Y*

Example 6.6

The distribution of the daily share price of a blue-chip company is shown in Table 6.11. Find the standard deviation of the daily share price of the company.

Solution

The data for this problem are shown in Table 6.12.



F18 =STDEV.S(C4:C13)

Workings Standrad Deviations of Yields

Plot No.	Type of Fertilizer	
	X	Y
1	10	12
2	10	11
3	10	9
4	10	10
5	10	12
6	10	9
7	10	12
8	10	8
9	10	9
10	10	8

Standard deviaton of yield by using fertilizer type X = 0

Standard deviatin of yield by using fertilizer type Y = 1.632993162

Figure 6.23 Screenshot of results of standard deviation of yield of paddy in cell F18 for fertiliser type *Y*

Table 6.10 Daily Demand for Bread Packets With Frequencies

Demand (No. of Packets)	*Number of Days of Occurrence of Demand*
21	20
22	22
23	26
24	30
25	35
26	25
27	20
28	15

The formula for the standard deviation of demand for this type of data is shown in the following.

$$Standard\,deviation(\sigma) = \sqrt{\frac{\sum_{i=1}^{n} f_i \times \left(X_i - \bar{X}\right)^2}{N}}$$

where

n is the number of demand values

f_i is the frequency of the i^{th} demand value for i = 1, 2, 3, . . . , n

N is the sum of the frequencies

X_i is the i^{th} demand value, where i = 1, 2,3, . . . , n

$\bar{X}$ is the mean demand

Table 6.11 Distribution of Daily Share Price of Blue Chip Company

Daily Share Price (₹)	*Frequency of Quoting of Share Price*
220	20
221	22
222	26
223	30
224	35
225	25
226	20
227	16
228	15
229	10
230	19
231	18
232	23
233	12
234	10
235	7

Table 6.12 Data for Example 6.6

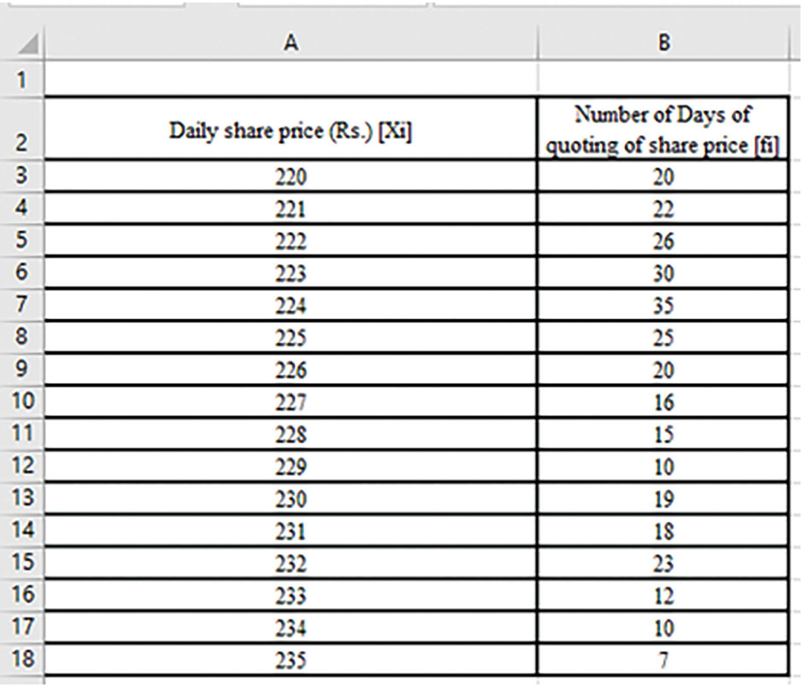
Daily Share Price (₹)	*Frequency of Quoting of Share Price*
220	20
221	22
222	26
223	30
224	35
225	25
226	20
227	16
228	15
229	10
230	19
231	18
232	23
233	12
234	10
235	7

σ is the standard deviation of the demand

The input of the data for this problem in an Excel sheet is shown in Figure 6.24. The working of the standard deviation for the daily share price is shown in Figure 6.25. The formulas for working of the standard deviation for the daily share price are shown in Figure 6.26. The standard deviation of the daily share price is ₹ 4.28.

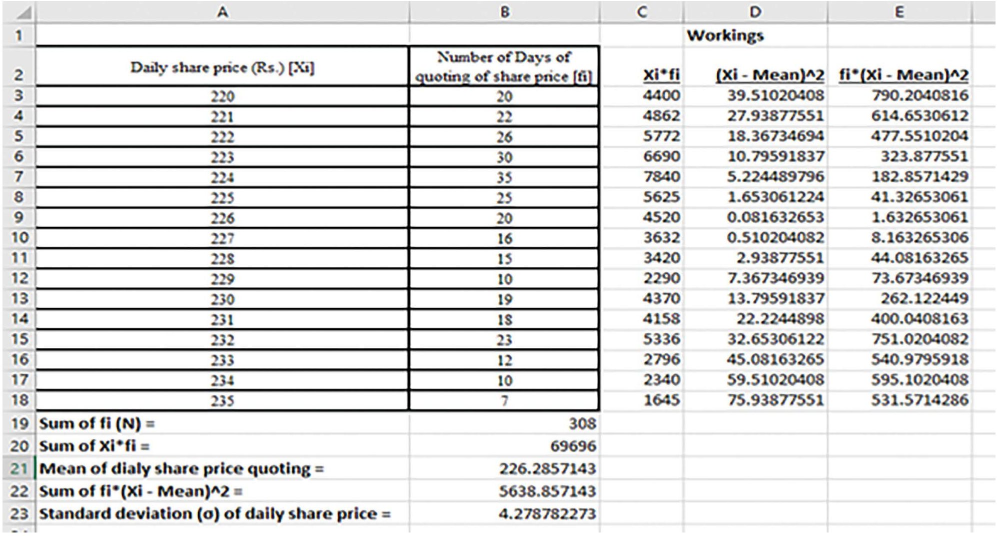
6.5.3 Standard Deviation of Grouped Data With Frequencies

The data for a reality pertaining to a variable of interest may be in the form of a frequency table. This means that the number of occurrences of each possible interval (class)



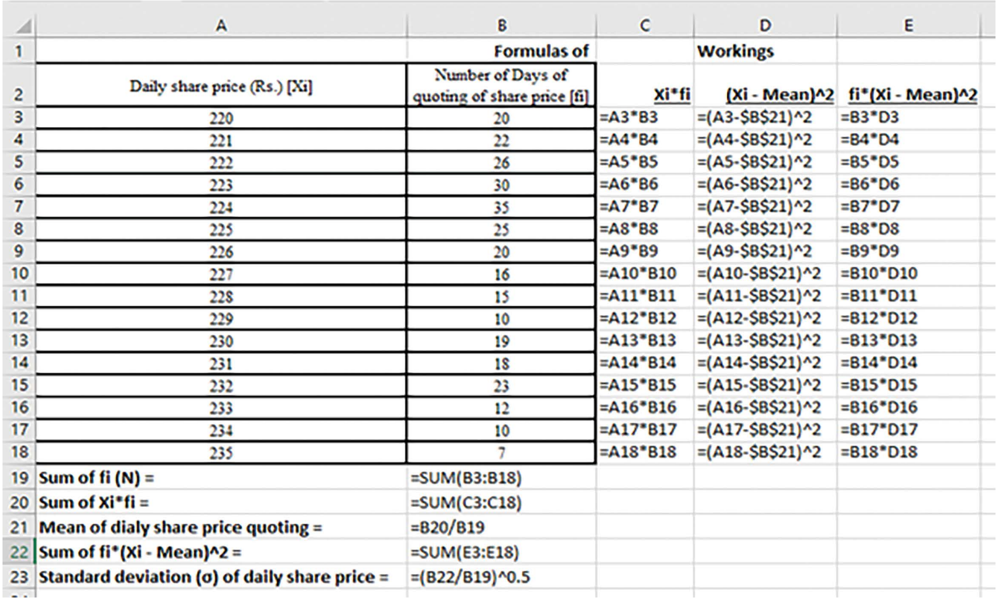
	A	B
1		
2	Daily share price (Rs.) [Xi]	Number of Days of quoting of share price [fi]
3	220	20
4	221	22
5	222	26
6	223	30
7	224	35
8	225	25
9	226	20
10	227	16
11	228	15
12	229	10
13	230	19
14	231	18
15	232	23
16	233	12
17	234	10
18	235	7

Figure 6.24 Screenshot of input of Example 6.6



	A	B	C	D	E
1				Workings	
2	Daily share price (Rs.) [Xi]	Number of Days of quoting of share price [fi]	Xi*fi	(Xi - Mean)^2	fi*(Xi - Mean)^2
3	220	20	4400	39.51020408	790.2040816
4	221	22	4862	27.93877551	614.6530612
5	222	26	5772	18.36734694	477.5510204
6	223	30	6690	10.79591837	323.877551
7	224	35	7840	5.224489796	182.8571429
8	225	25	5625	1.653061224	41.32653061
9	226	20	4520	0.081632653	1.632653061
10	227	16	3632	0.510204082	8.163265306
11	228	15	3420	2.93877551	44.08163265
12	229	10	2290	7.367346939	73.67346939
13	230	19	4370	13.79591837	262.122449
14	231	18	4158	22.2244898	400.0408163
15	232	23	5336	32.65306122	751.0204082
16	233	12	2796	45.08163265	540.9795918
17	234	10	2340	59.51020408	595.1020408
18	235	7	1645	75.93877551	531.5714286
19	Sum of fi (N) =	308			
20	Sum of Xi*fi =	69696			
21	Mean of dialy share price quoting =	226.2857143			
22	Sum of fi*(Xi - Mean)^2 =	5638.857143			
23	Standard deviation (σ) of daily share price =	4.278782273			

Figure 6.25 Screenshot of working of Example 6.6



	A	B	C	D	E
1		Formulas of		Workings	
2	Daily share price (Rs.) [Xi]	Number of Days of quoting of share price [fi]	Xi*fi	(Xi - Mean)^2	fi*(Xi - Mean)^2
3	220	20	=A3*B3	=(A3-$B$21)^2	=B3*D3
4	221	22	=A4*B4	=(A4-$B$21)^2	=B4*D4
5	222	26	=A5*B5	=(A5-$B$21)^2	=B5*D5
6	223	30	=A6*B6	=(A6-$B$21)^2	=B6*D6
7	224	35	=A7*B7	=(A7-$B$21)^2	=B7*D7
8	225	25	=A8*B8	=(A8-$B$21)^2	=B8*D8
9	226	20	=A9*B9	=(A9-$B$21)^2	=B9*D9
10	227	16	=A10*B10	=(A10-$B$21)^2	=B10*D10
11	228	15	=A11*B11	=(A11-$B$21)^2	=B11*D11
12	229	10	=A12*B12	=(A12-$B$21)^2	=B12*D12
13	230	19	=A13*B13	=(A13-$B$21)^2	=B13*D13
14	231	18	=A14*B14	=(A14-$B$21)^2	=B14*D14
15	232	23	=A15*B15	=(A15-$B$21)^2	=B15*D15
16	233	12	=A16*B16	=(A16-$B$21)^2	=B16*D16
17	234	10	=A17*B17	=(A17-$B$21)^2	=B17*D17
18	235	7	=A18*B18	=(A18-$B$21)^2	=B18*D18
19	Sum of fi (N) =	=SUM(B3:B18)			
20	Sum of Xi*fi =	=SUM(C3:C18)			
21	Mean of dialy share price quoting =	=B20/B19			
22	Sum of fi*(Xi - Mean)^2 =	=SUM(E3:E18)			
23	Standard deviation (σ) of daily share price =	=(B22/B19)^0.5			

Figure 6.26 Screenshot of formulas for working of Example 6.6

Table 6.13 EPS Values and Their Frequencies of Companies in an Industry

EPS (₹)	*No. of Companies*
10–15	60
15–20	62
20–25	70
25–30	68
30–35	60
35–40	50
40–45	20
45–50	10

of that variable will be given in that table. Consider the EPS (earnings per share) of 400 companies with equal capital in an industry with frequencies as shown in Table 6.13.

Now, the objective is to find the standard deviation (σ) of EPS for this grouped data. The formula to find the standard deviation of the grouped data is presented in the following.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i \times X_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i \times X_i}{N}\right)^2} = \sqrt{\frac{\sum_{i=1}^{n} f_i \times X_i^2}{N} - \bar{X}^2}$$

where

n is the number of class intervals

X_i is mid-point of the i^{th} class interval for i = 1, 2, 3, . . ., *n*

f_i is the frequency of the i^{th} class interval, i = 1, 2, 3, . . . , *n*

N is the sum of the frequencies of the class intervals
$\bar{X}$ is the mean of the variable of interest
σ is the standard deviation of grouped data with frequencies

Example 6.7

The distribution of the price of equity shares of a company in a stock exchange is as shown in Table 6.14. Find the standard deviation of the share price of the company.

Solution

The data for Example 6.7 with some modifications in presentation are shown in Table 6.15.

The formula to find the standard deviation of the grouped data is presented as follows.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i \times X_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i \times X_i}{N}\right)^2} = \sqrt{\frac{\sum_{i=1}^{n} f_i \times X_i^2}{N} - \bar{X}^2}$$

where
n is the number of class intervals
X_i is the mid-point of the i^{th} class interval for i = 1, 2, 3, . . ., *n*
f_i is the frequency of the i^{th} class interval, i = 1, 2, 3, . . . , *n*
N is the sum of the frequencies of the class intervals
$\bar{X}$ is the mean of the variable of interest
σ is the standard deviation of grouped data with frequencies

Table 6.14 Distribution of Share Price

Share Price (₹)	*No. of Sessions*
50–60	50
60–70	90
70–80	96
80–90	80
90–100	40

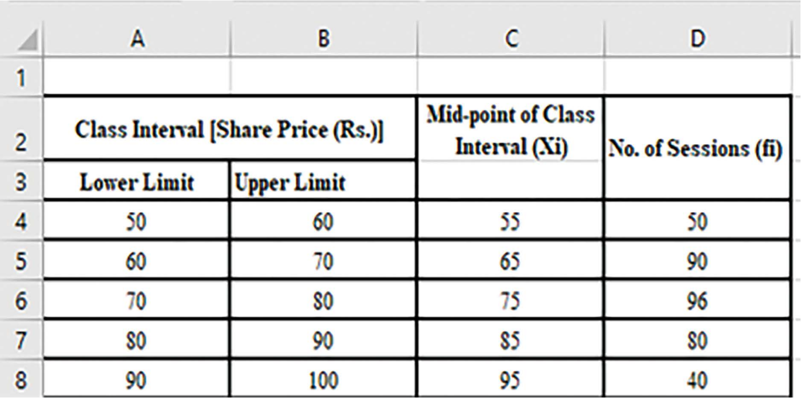
Table 6.15 Data for Example 6.7

Class Interval [Share Price (₹)]		*No. of Sessions*
Lower Limit	*Upper Limit*	
50	60	50
60	70	90
70	80	96
80	90	80
90	100	40

The input of the data for Example 6.7 is shown in Figure 6.27. The working to compute the standard deviation of the grouped data with frequencies for Example 6.7 is shown in Figure 6.28. The formulas for the working to compute the standard deviation of the grouped data with frequencies of Example 6.7 are shown in Figure 6.29. The standard deviation of the share price is ₹ 12.17.

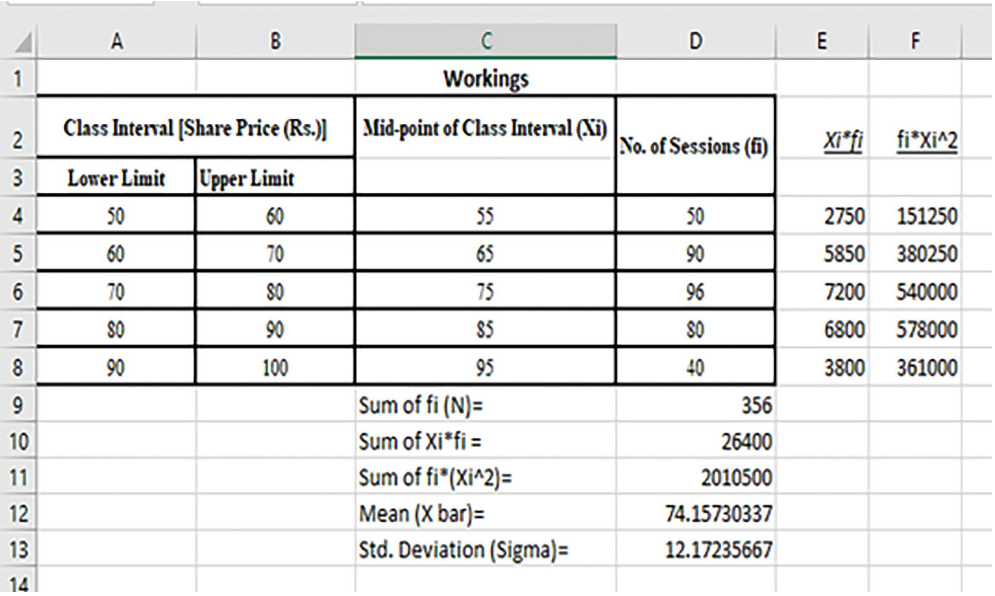
6.5.4 STDEVA/STDEVPA Functions

With logical values and text included in the sample population, the STDEVA function calculates the standard deviation of the data. The value of the text and the logical value



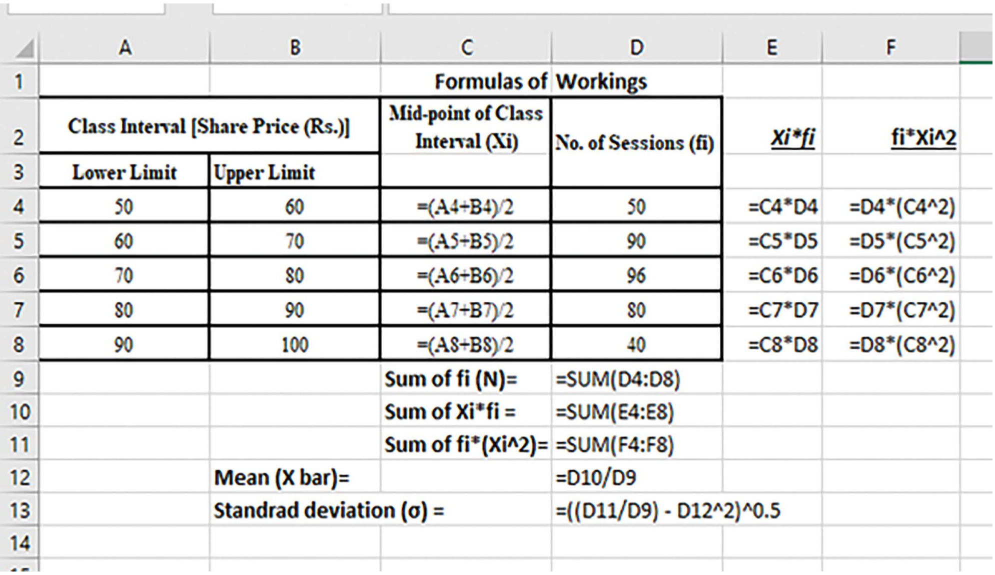
	A	B	C	D
1				
2	Class Interval [Share Price (Rs.)]		Mid-point of Class Interval (Xi)	No. of Sessions (fi)
3	Lower Limit	Upper Limit		
4	50	60	55	50
5	60	70	65	90
6	70	80	75	96
7	80	90	85	80
8	90	100	95	40

Figure 6.27 Screenshot of input of Example 6.7



	A	B	C	D	E	F
1			Workings			
2	Class Interval [Share Price (Rs.)]		Mid-point of Class Interval (Xi)	No. of Sessions (fi)	Xi*fi	fi*Xi^2
3	Lower Limit	Upper Limit				
4	50	60	55	50	2750	151250
5	60	70	65	90	5850	380250
6	70	80	75	96	7200	540000
7	80	90	85	80	6800	578000
8	90	100	95	40	3800	361000
9			Sum of fi (N)=	356		
10			Sum of Xi*fi =	26400		
11			Sum of fi*(Xi^2)=	2010500		
12			Mean (X bar)=	74.15730337		
13			Std. Deviation (Sigma)=	12.17235667		
14						

Figure 6.28 Screenshot of working of standard deviation of Example 6.7



	A	B	C	D	E	F
1			Formulas of	Workings		
2	Class Interval [Share Price (Rs.)]		Mid-point of Class Interval (Xi)	No. of Sessions (fi)	*Xi*fi*	fi*Xi^2
3	Lower Limit	Upper Limit				
4	50	60	=(A4+B4)/2	50	=C4*D4	=D4*(C4^2)
5	60	70	=(A5+B5)/2	90	=C5*D5	=D5*(C5^2)
6	70	80	=(A6+B6)/2	96	=C6*D6	=D6*(C6^2)
7	80	90	=(A7+B7)/2	80	=C7*D7	=D7*(C7^2)
8	90	100	=(A8+B8)/2	40	=C8*D8	=D8*(C8^2)
9			Sum of fi (N)=	=SUM(D4:D8)		
10			Sum of Xi*fi =	=SUM(E4:E8)		
11			Sum of fi*(Xi^2)=	=SUM(F4:F8)		
12		Mean (X bar)=		=D10/D9		
13		Standrad deviation (σ) =		=((D11/D9) - D12^2)^0.5		
14						

Figure 6.29 Screenshot of formulas of working of standard deviation of Example 6.7

with a FALSE instance is 0, and the value of the logical value with a TRUE instance is 1. Up to 255 values, or a sample of a population, will be represented by the values in the arrays Value 1, Value 2, and so on. These values can be values, names, arrays, or references that contain values. In this instance, *n* – 1 will serve as the denominator in the standard deviation formula. The standard deviation of the sample is what is referred to as an unbiased estimator of standard deviation.

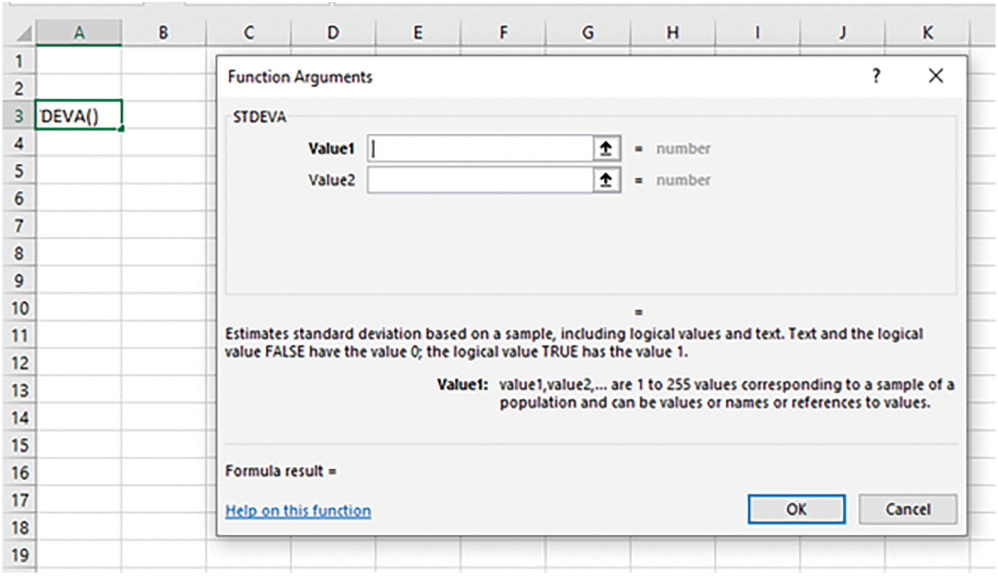
The STDEVPA function calculates the standard deviation of all observations made on the population as a whole, including text and logical values. The value of a text or logical value with a FALSE instance is 0, and the value of a logical value with a TRUE instance is 1. The values 1, 2, . . . will increase to 255 values, corresponding to a population, and they can be values, names, arrays, or references to values. *In this instance, n will serve as the denominator in the standard deviation formula. A biased estimator is one with such a standard deviation. This will be covered later.*

Pressing the sequence of buttons Formulas ⟹ More Functions ⟹ Statistical ⟹ STDEVA gives a display as in Figure 6.30. Entering the range of cells which contains numbers, text, and logical values in the box against Value 1 of the dropdown menu of Figure 6.30 and clicking the OK button will give the result for the standard deviation. The same can be achieved using the following formula.

Formula for STDEVA : = STDEVA (range of cells containing numbers, text and
logical values of sample in a population)

Example 6.8

Find the standard deviation of the data shown in Table 6.16 using the STDEVA function. By treating the data as a population, find its standard deviation using its formula.



Function Arguments

STDEVA

Value1 = number

Value2 = number

=

Estimates standard deviation based on a sample, including logical values and text. Text and the logical value FALSE have the value 0; the logical value TRUE has the value 1.

Value1: value1,value2,... are 1 to 255 values corresponding to a sample of a population and can be values or names or references to values.

Formula result =

Help on this function

Figure 6.30 Screenshot after clicking sequence of buttons, Formulas ⇨ More Functions ⇨ Statistical ⇨ STDEVA

Table 6.16 Data With Numbers and Logical Text

S. No.	*Data*
1	34
2	TRUE
3	43
4	56
5	FALSE
6	45
7	76
8	36

Solution

The data for Example 6.8 are shown in Table 6.17.

The input of this example is shown in Figure 6.31. The usage of the following formula gives the standard deviation of the given data in cell B12 as shown in Figure 6.32, which is 25.78448.

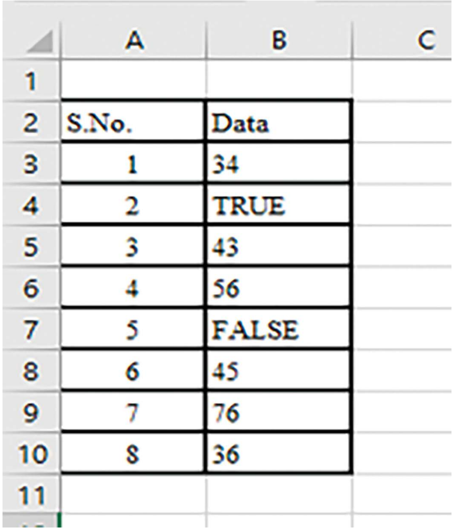
$$\text{Formula} := \text{STDEVA}(\text{B3} : \text{B10})$$

6.5.5 STDEVP or STDEV.P Function

By eliminating text and logical values from the set of observations, the STDEVP or STDEV.P function calculates the standard deviation for the population. The snapshot

Table 6.17 Data for Example 6.8

S. No.	*Data*
1	34
2	TRUE
3	43
4	56
5	FALSE
6	45
7	76
8	36



	A	B	C
1			
2	S.No.	Data	
3	1	34	
4	2	TRUE	
5	3	43	
6	4	56	
7	5	FALSE	
8	6	45	
9	7	76	
10	8	36	
11			

Figure 6.31 Screenshot of input of Example 6.8

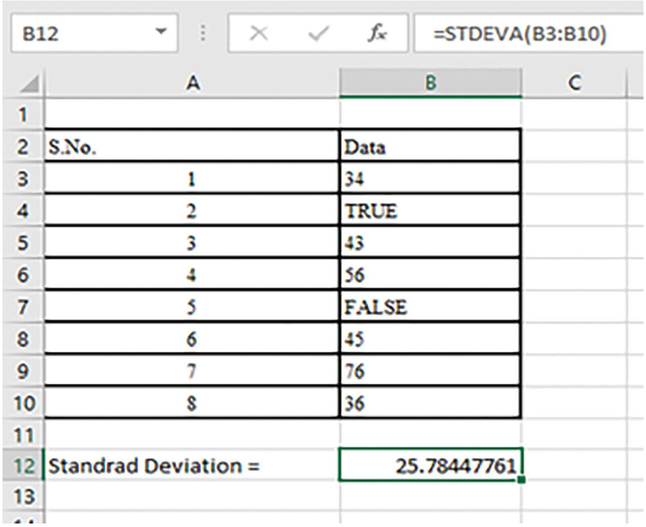
that results from pressing Formulas ⟹ More Functions ⟹ Statistical ⟹ STDEV.P in that order is similar to that in Figure 6.33. The standard deviation will be determined by entering the range of cells that contain numbers, text, and logical values in the box next to Value 1 in the dropdown menu of Figure 6.33 and clicking the OK button. The following formula can be used to do the same thing.

Formula for STDEV.P : = STDEV.P (Range of cells containing numbers, text and logical values)

or

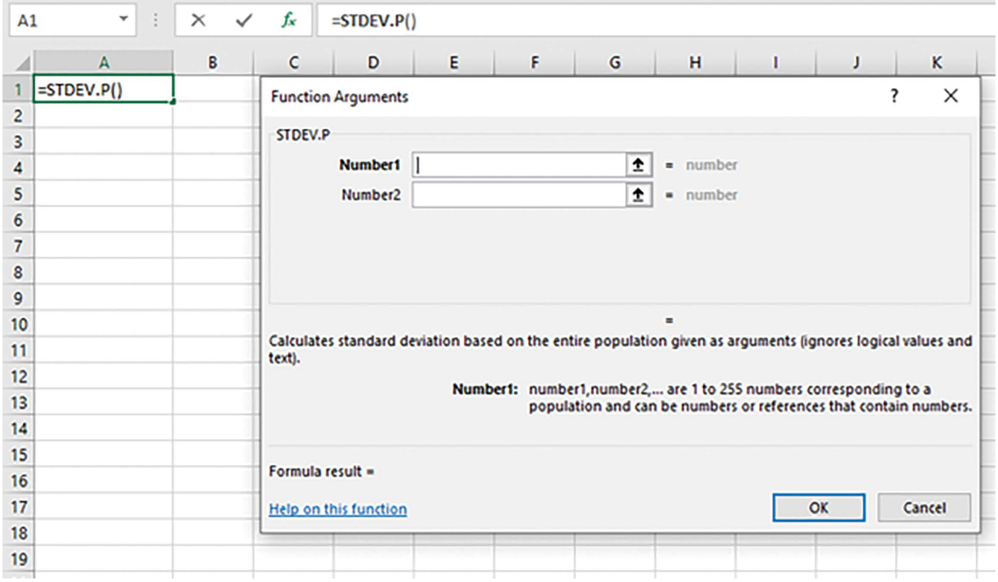
= STDEVP(Range of cells containing numbers, text and logical values)

[This can be given as formula in a cell; there is no function in the dropdown menue of Excel for this function]


B12 =STDEVA(B3:B10)

	A	B
2	S.No.	Data
3	1	34
4	2	TRUE
5	3	43
6	4	56
7	5	FALSE
8	6	45
9	7	76
10	8	36
12	Standrad Deviation =	25.78447761

Figure 6.32 Screenshot of result of STDEVA function applied to Example 6.8


A1 =STDEV.P()

Function Arguments

STDEV.P

Number1 = number

Number2 = number

Calculates standard deviation based on the entire population given as arguments (ignores logical values and text).

Number1: number1,number2,... are 1 to 255 numbers corresponding to a population and can be numbers or references that contain numbers.

Formula result =

Figure 6.33 Screenshot of clicks of sequence of buttons, Formula ⟹ More functions ⟹ Statistical ⟹ STDEV.P

Figure 6.33 is a screenshot of clicks of the sequence of buttons Formulas ⟹ More Functions ⟹ Statistical ⟹ STDEV.P.

Example 6.9

Find the standard deviation of the data shown in Table 6.18 using the STDEV.P function.

Table 6.18 Data With Numbers and Logical Text

S. No.	*Data*
1	44
2	TRUE
3	43
4	TRUE
5	FALSE
6	45
7	56
8	26

Solution

The data for Example 6.9 are shown in Table 6.19.

The input for Example 6.9 is shown in Figure 6.34. The usage of the following formula gives the standard deviation of the given data in cell B12 as shown in Figure 6.35, which is 9.620811.

$$\text{Formula} := \text{STDEVP}\left(\text{B3}:\text{B10}\right) \text{ or } = \text{STDEV.P(B3:B10)}$$

One can note the same result in cell C12 and cell C13 of Figure 6.35 for using the two formulas in the same order.

6.5.6 STDEVPA Function

The STDEVPA function finds the standard deviation of the observations in the entire population, including logical values and text. Text and a logical value with a FALSE instance will have the value 0, and a logical value with a True instance will have the value 1. Value 1, Value 2, . . . will extend up to 255 values corresponding to a population, and they can be values, names, arrays, or references that contain values. *In this case, the formula for the standard deviation will have n as the denominator. This is known as biased estimator of standard deviation.*

The screenshot after pressing the sequence of buttons Formulas ⟹ More Functions ⟹ Statistical ⟹ STDEVPA is shown in Figure 6.36. Entering the range of cells, which contains numbers, text and logical values in the cell, against Value 1 of the dropdown menu of Figure 6.36 and clicking the OK button will give the result for the standard deviation. The same can be achieved using the following formula.

$$\text{Formula for STDEVPA} := \text{STDEVPA (Range of cells containing numbers, text and logical values of population)}$$

Example 6.10

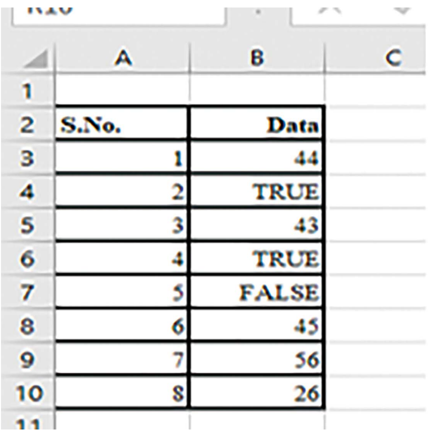
Find the standard deviation of the data shown in Table 6.20 using the STDEVPA function.

Solution

The data for Example 6.10 are shown in Table 6.21.

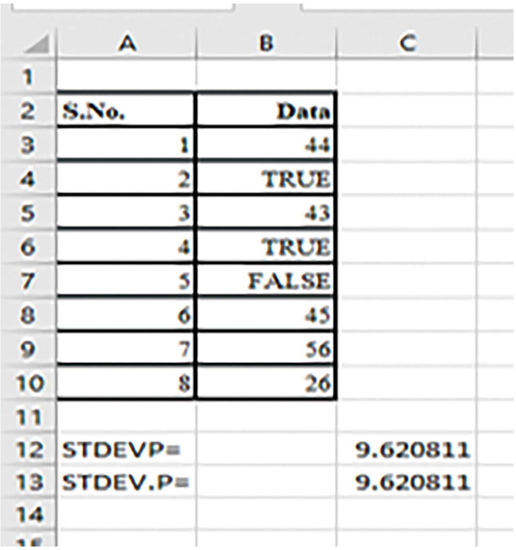
Table 6.19 Data for Example 6.9

S. No.	*Data*
1	44
2	TRUE
3	43
4	TRUE
5	FALSE
6	45
7	56
8	26



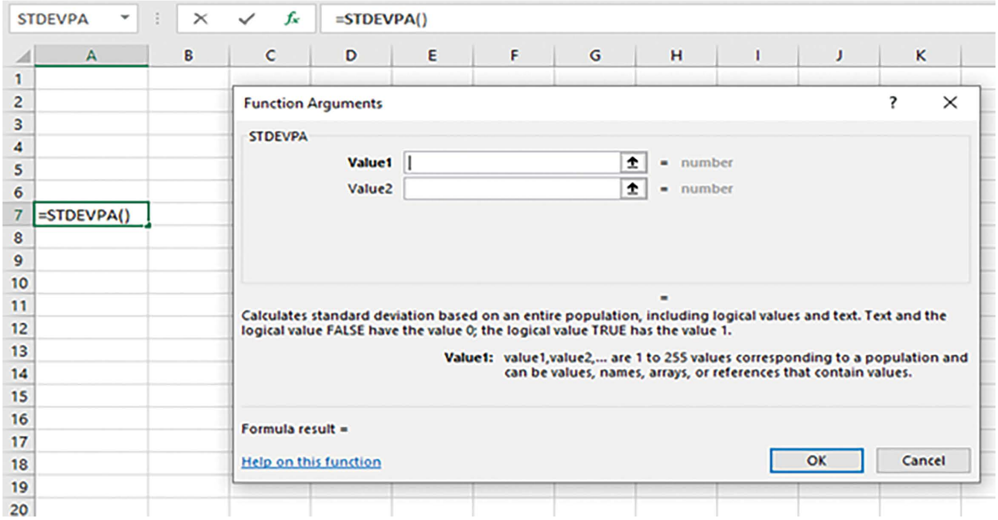
	A	B	C
1			
2	S.No.	Data	
3	1	44	
4	2	TRUE	
5	3	43	
6	4	TRUE	
7	5	FALSE	
8	6	45	
9	7	56	
10	8	26	

Figure 6.34 Screenshot of input of Example 6.9



	A	B	C
1			
2	S.No.	Data	
3	1	44	
4	2	TRUE	
5	3	43	
6	4	TRUE	
7	5	FALSE	
8	6	45	
9	7	56	
10	8	26	
11			
12	STDEVP=		9.620811
13	STDEV.P=		9.620811
14			

Figure 6.35 Screenshot of result of STDEV.P function applied to Example 6.9



Calculates standard deviation based on an entire population, including logical values and text. Text and the logical value FALSE have the value 0; the logical value TRUE has the value 1.

Value1: value1,value2,... are 1 to 255 values corresponding to a population and can be values, names, arrays, or references that contain values.

Formula result =

Figure 6.36 Screenshot of clicking the sequence of buttons, Formulas $\Longrightarrow$ More Functions $\Longrightarrow$ Statistical $\Longrightarrow$ STDEVPA

Table 6.20 Data With Numbers and Logical Text

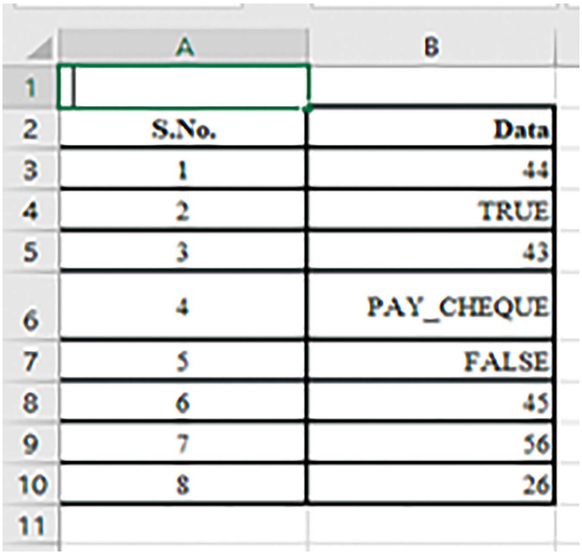
S. No.	*Data*
1	44
2	TRUE
3	43
4	PAY_CHEQUE
5	FALSE
6	45
7	56
8	26

Table 6.21 Data for Example 6.10

S. No.	*Data*
1	44
2	TRUE
3	43
4	PAY_CHEQUE
5	FALSE
6	45
7	56
8	26

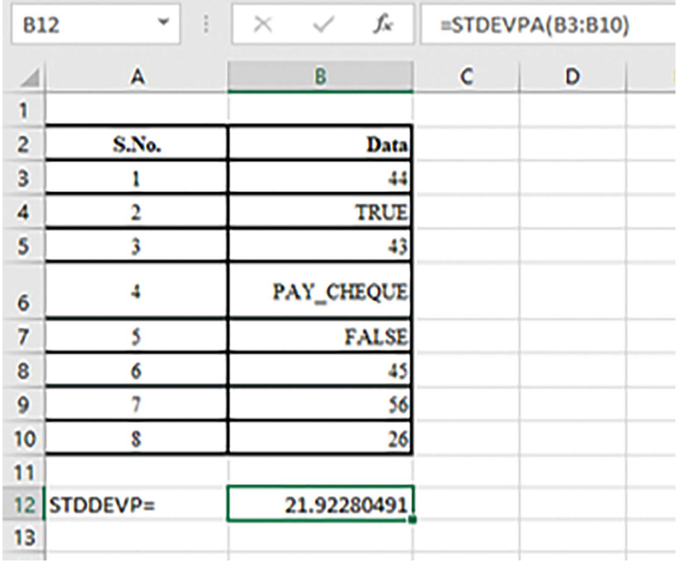
The input of this example is shown in Figure 6.37. The usage of the following formula gives the standard deviation of the given data in cell B12, as shown in Figure 6.38, which is 21.92280491.

$$\text{Formula}: \ = \text{STDEVPA}(\text{B3}:\text{B10})$$



	A	B
1		
2	S.No.	Data
3	1	44
4	2	TRUE
5	3	43
6	4	PAY_CHEQUE
7	5	FALSE
8	6	45
9	7	56
10	8	26
11		

Figure 6.37 Screenshot of input of Example 6.10



B12 =STDEVPA(B3:B10)

	A	B	C	D
1				
2	S.No.	Data		
3	1	44		
4	2	TRUE		
5	3	43		
6	4	PAY_CHEQUE		
7	5	FALSE		
8	6	45		
9	7	56		
10	8	26		
11				
12	STDDEVP=	21.92280491		
13				

Figure 6.38 Screenshot of result of STDEVPA function applied to Example 6.10

6.6 Variance of Ungrouped Data

The variance (σ^2) of ungrouped data with observations pertaining to a variable X_i, where I varies from 1 to n and n is the number of observations, is given by the following formula

$$Variance\left(\sigma^2\right) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{(n-1)\, or\, n}$$

where
n is the number of observations of a sample/population
X_i is the i^{th} observation of the variable of interest for i = 1, 2, 3, . . . , n
$\bar{X}$ is the mean of the observations
σ^2 is the variance of the observations

In Excel, the variance of a set of observations can be obtained using the functions VAR.S, VAR.P, VARA, and VARPA.

The first and the third functions use the following formula to find the standard deviation of a sample.

$$Standard\,deviation(S) = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}}$$

The second and the fourth functions use the following formula to find the standard deviation of a population.

$$Standard\,deviation(\sigma) = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n}}$$

6.6.1 VAR.S Function

With the exception of logical and text data, the VAR.S function calculates the variance of a sample of data. Take a look at the screenshot in Figure 6.39, which displays a company's annual sales for the previous 10 years in crores of rupees.

Clicking the sequence of buttons Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ VAR.S gives a screenshot as in Figure 6.40 [5]. The display after filling the box against Number1 with cells B2:B11 in the dropdown menu of Figure 6.40 is as in Figure 6.41. Clicking the OK button in the dropdown menu of Figure 6.41 gives the result as shown in Figure 6.42. The variance is 10843.07143.

6.6.2 VAR.P Function

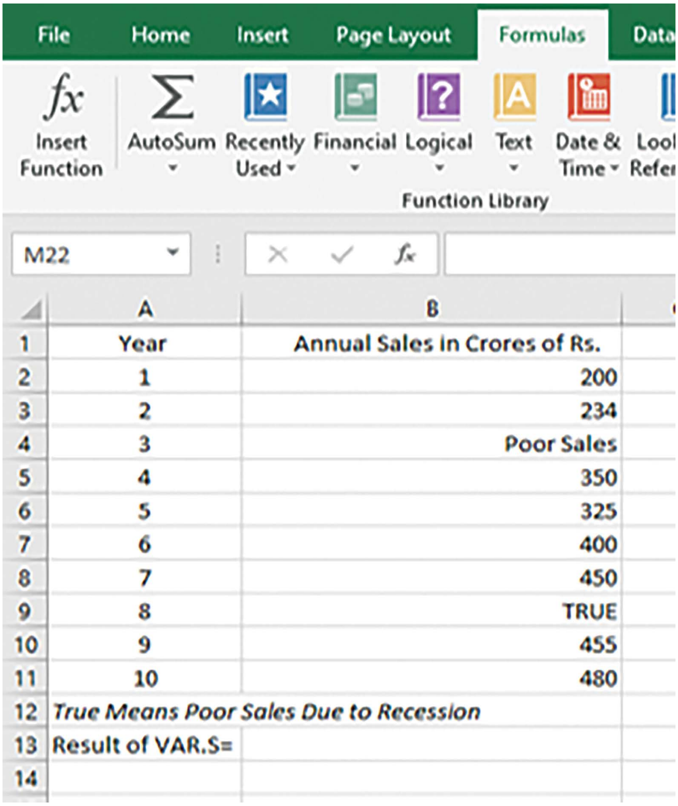
The VAR.P function, which ignores logical and text data, calculates the variance of a population of data. The following is the formula for this function.

$$= \text{VAR.P}(\text{Range of cells containing data of population excluding logical and text})$$

Consider again the screenshot shown in Figure 6.39, which contains the annual sales in crores of rupees of a company for the past 10 years.

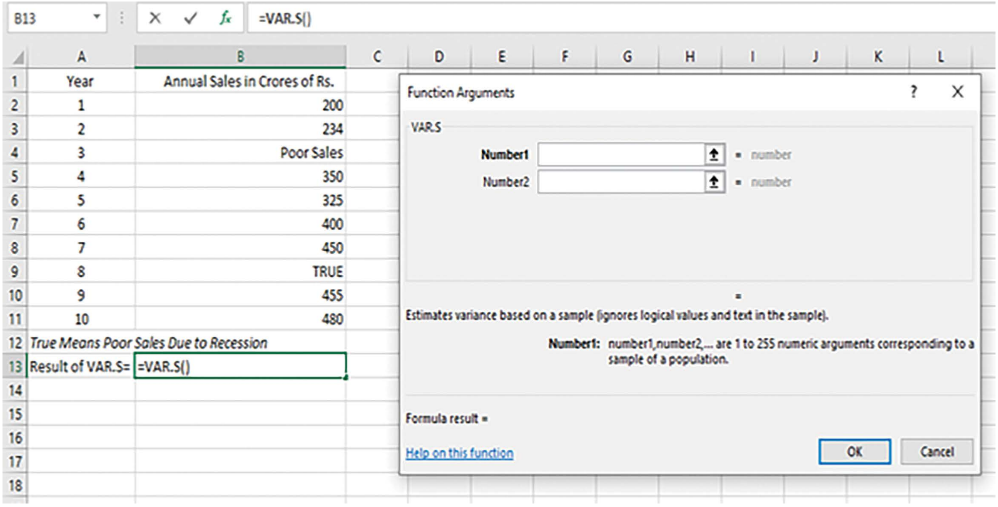
The formula applied to the data in the screenshot shown in Figure 6.39 at cell B13 is as follows, which gives the result as in Figure 6.43 [4].

$$= \text{VAR.P}(\text{B2 : B11})$$



	A	B
1	Year	Annual Sales in Crores of Rs.
2	1	200
3	2	234
4	3	Poor Sales
5	4	350
6	5	325
7	6	400
8	7	450
9	8	TRUE
10	9	455
11	10	480
12	*True Means Poor Sales Due to Recession*	
13	Result of VAR.S=	
14		

Figure 6.39 Screenshot of data of annual sales of company



=VAR.S()

	A	B
1	Year	Annual Sales in Crores of Rs.
2	1	200
3	2	234
4	3	Poor Sales
5	4	350
6	5	325
7	6	400
8	7	450
9	8	TRUE
10	9	455
11	10	480
12	*True Means Poor Sales Due to Recession*	
13	Result of VAR.S=	=VAR.S()
14		
15		
16		
17		
18		

Function Arguments

VAR.S

Number1 = number

Number2 = number

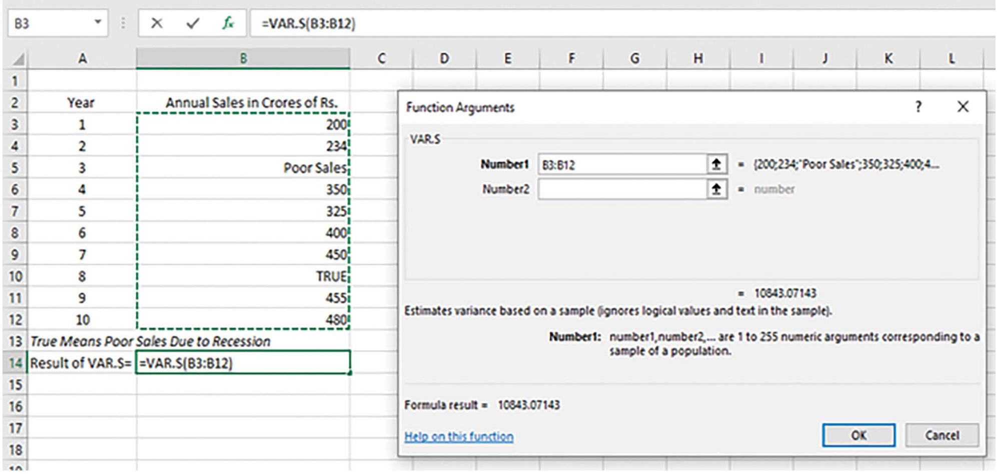
=

Estimates variance based on a sample (ignores logical values and text in the sample).

Number1: number1,number2,... are 1 to 255 numeric arguments corresponding to a sample of a population.

Formula result =

Figure 6.40 Screenshot after clicking sequence of buttons, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ VAR.S



	A	B
1		
2	Year	Annual Sales in Crores of Rs.
3	1	200
4	2	234
5	3	Poor Sales
6	4	350
7	5	325
8	6	400
9	7	450
10	8	TRUE
11	9	455
12	10	480
13	True Means Poor Sales Due to Recession	
14	Result of VAR.S=	=VAR.S(B3:B12)

Number1 B3:B12 = {200;234;"Poor Sales";350;325;400;4...

Number2 = number

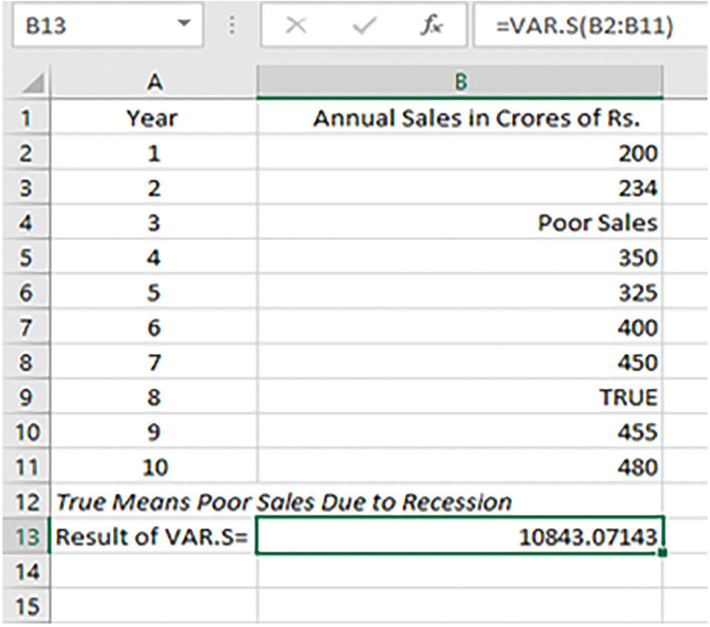
= 10843.07143

Estimates variance based on a sample (ignores logical values and text in the sample).

Number1: number1,number2,... are 1 to 255 numeric arguments corresponding to a sample of a population.

Formula result = 10843.07143

Figure 6.41 Screenshot after filling data in the dropdown menu of Figure 6.40



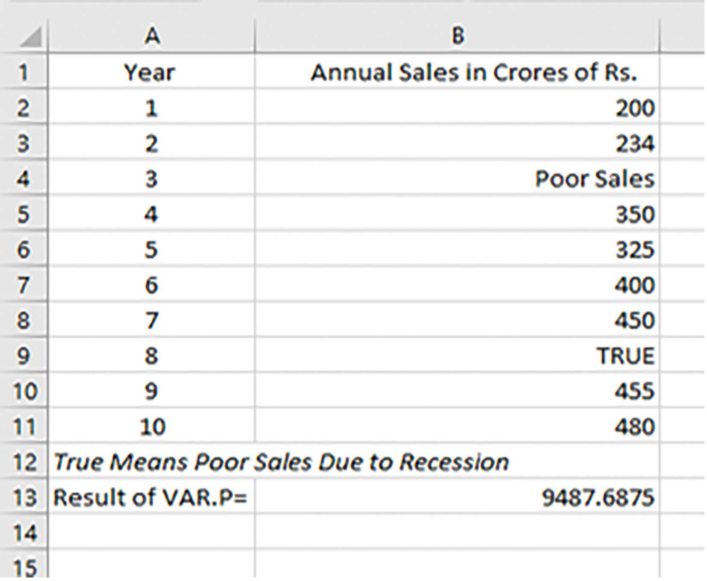
	A	B
1	Year	Annual Sales in Crores of Rs.
2	1	200
3	2	234
4	3	Poor Sales
5	4	350
6	5	325
7	6	400
8	7	450
9	8	TRUE
10	9	455
11	10	480
12	True Means Poor Sales Due to Recession	
13	Result of VAR.S=	10843.07143
14		
15		

Figure 6.42 Screenshot after clicking the OK button in the dropdown menu of Figure 6.41

6.6.3 VARA Function

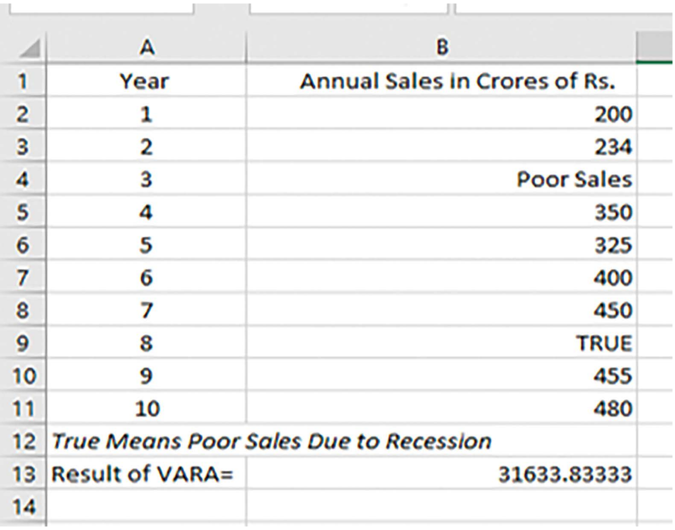
With logical and text data included, the VARA function calculates the variance of the *sample data*. Logical TRUE will make an assumption of 1, logical FALSE will make an assumption of 0, and text will make an assumption of 0. The following is the formula for this function.

$$= \text{VARA}(\text{Range of cells containing data of sample including logical and text})$$



	A	B
1	Year	Annual Sales in Crores of Rs.
2	1	200
3	2	234
4	3	Poor Sales
5	4	350
6	5	325
7	6	400
8	7	450
9	8	TRUE
10	9	455
11	10	480
12	*True Means Poor Sales Due to Recession*	
13	Result of VAR.P=	9487.6875
14		
15		

Figure 6.43 Screenshot for the result of the formula: =VAR.P(B2:B11)



	A	B
1	Year	Annual Sales in Crores of Rs.
2	1	200
3	2	234
4	3	Poor Sales
5	4	350
6	5	325
7	6	400
8	7	450
9	8	TRUE
10	9	455
11	10	480
12	*True Means Poor Sales Due to Recession*	
13	Result of VARA=	31633.83333
14		

Figure 6.44 Screenshot for the result of the formula: =VARA(B2:B11)

Consider again the screenshot shown in Figure 6.39, which contains the annual sales in crores of rupees of a company for the past 10 years along with logical and text data.

The formula applied to the data in the screenshot shown in Figure 6.39 at cell B13 is as follows, which gives the result as in Figure 6.44.

$$= \text{VARA}(\text{B2} : \text{B11})$$

6.6.4 VARPA Function

With logical and text data included, the VARPA function calculates the variance of the *population*. Logical TRUE will make an assumption of 1, logical FALSE will make an assumption of 0, and text will make an assumption of 0. The following is the formula for this function.

$$= \text{VARPA}(\text{Range of cells containing data of population including logical and text})$$

Consider again the screenshot shown in Figure 6.39, which contains the annual sales in crores of rupees of a company for the past 10 years along with logical and text data.

The formula applied to the data in the screenshot shown in Figure 6.39 at cell B13 is as follows, which gives the result as in Figure 6.45.

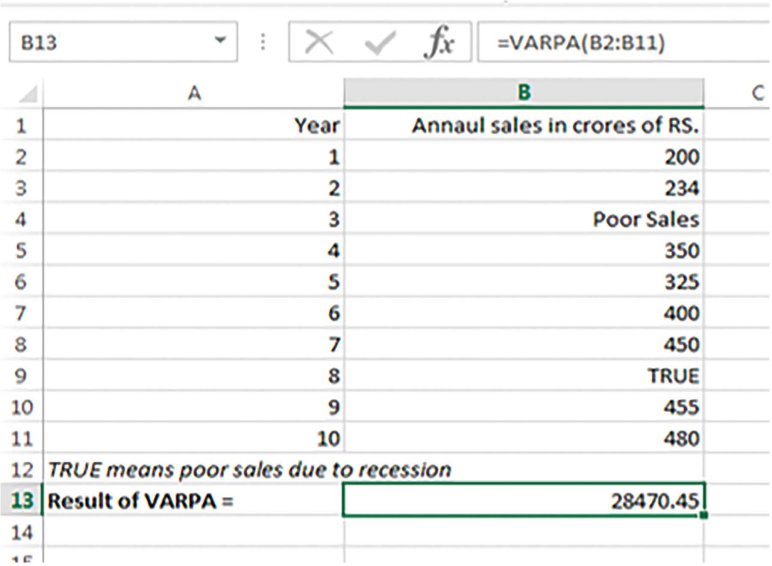
$$= \text{VARPA}(\text{B2} : \text{B11})$$

6.7 Coefficient of Variation Using Excel Sheets

Coefficient of variation aims to check the consistency of the observations of a variable of interest. Sometimes the mean of two different sets of observations, say annual sales of companies in industrial estate *X* and annual sales in industrial estate *Y*, may be the same. But there will be variations among the observations of each sample. One can use the coefficient of variation as a measure to find which industrial estate has more consistent annual sales.

The formula for the coefficient of variation is presented as follows.

$$\textit{Coefficient of variation}(CV) = \sigma / \bar{X}$$



B13

=VARPA(B2:B11)

	A	B
1	Year	Annaul sales in crores of RS.
2	1	200
3	2	234
4	3	Poor Sales
5	4	350
6	5	325
7	6	400
8	7	450
9	8	TRUE
10	9	455
11	10	480
12	*TRUE means poor sales due to recession*	
13	**Result of VARPA =**	28470.45
14		

Figure 6.45 Screenshot for the result of the formula: =VARPA(B2:B11)

where
σ is the standard deviation of the observations
$\bar{X}$ is the mean of the observations

This explains the extent of variation per unit value of the random variable. If this measure is used to check the consistency of the observations of k different samples, then the sample which has the lowest coefficient of variation is said to have more consistency in its observations.

Example 6.11

Consider two different salespeople of a company whose annual sales figures in lakhs of rupees for the past 5 years are as shown in Table 6.22. Identify the salesperson who is more consistent in terms of annual sales made.

Solution

The data for Example 6.11 are shown in Table 6.23.

The formula for the coefficient of variation is presented as follows.

$$Coefficient\ of\ variation(CV) = \sigma / \bar{X}$$

where
σ is the standard deviation of the observations
$\bar{X}$ is the mean of the observations

Table 6.22 Annual Sales of Salesperson A and Salesperson B

Year	*Salesperson A (Lakhs of Rupees)*	*Salesperson B (Lakhs of Rupees)*
1	85	100
2	90	120
3	95	50
4	150	60
5	200	125

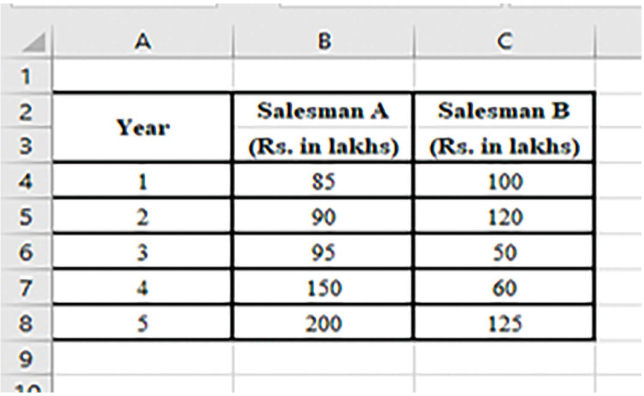
Table 6.23 Data for Example 6.11

Year	*Salesperson A (Lakhs of Rupees)*	*Salesperson B (Lakhs of Rupees)*
1	85	100
2	90	120
3	95	50
4	150	60
5	200	125

The input of Example 6.11 is shown in Figure 6.46. The working of the coefficient of variation of salesperson A and that of salesperson B are shown in Figure 6.47. The formulas for the working of the coefficient of variation of salesperson A and that of salesperson B are shown in Figure 6.48.

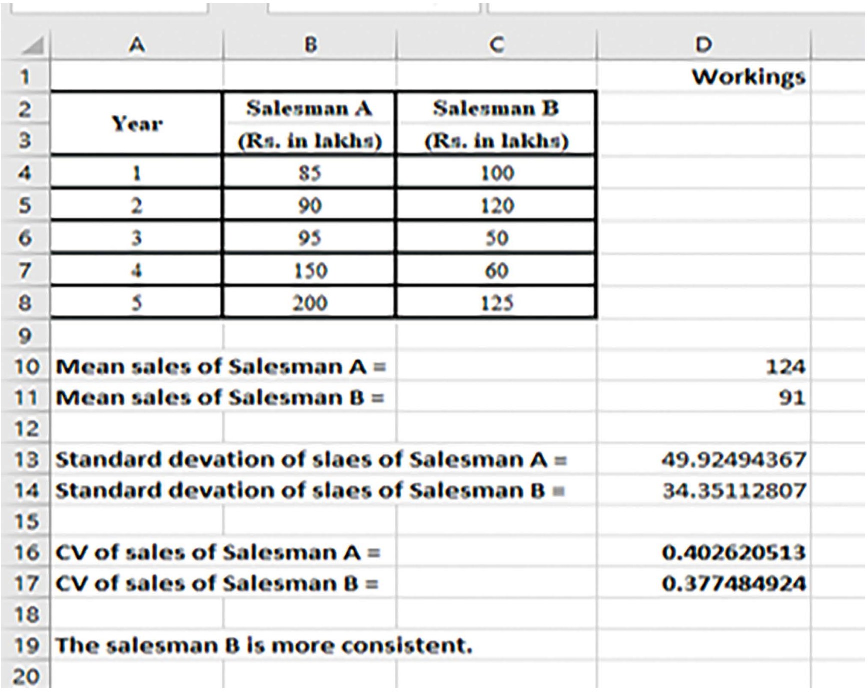
From Figure 6.47, the coefficient of variations of the salespeople are as follows.

CV of Salesperson A = 0.402620513
CV of Salesperson B = 0.377484924



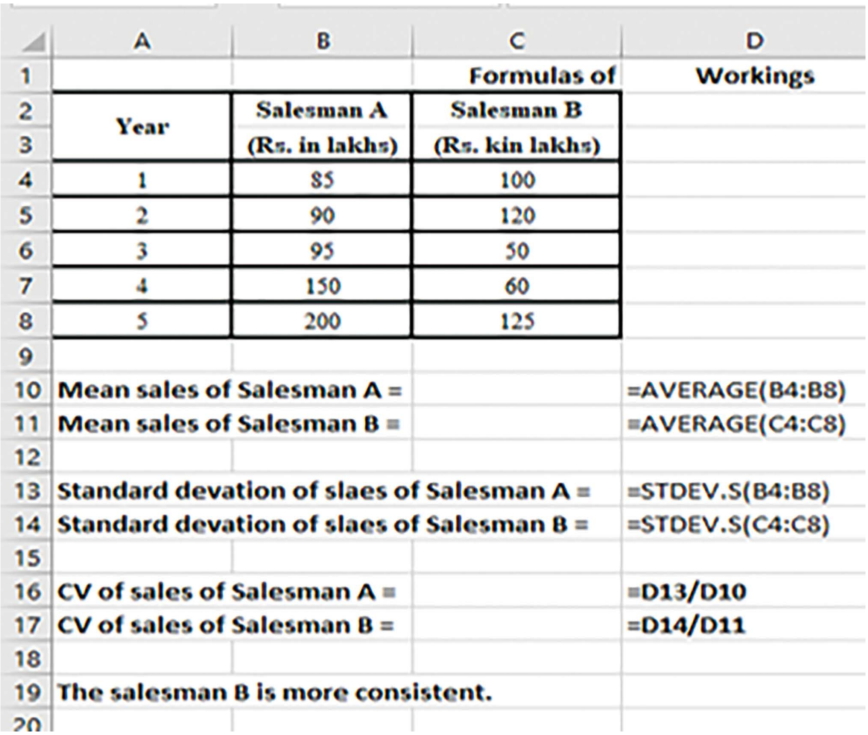
	A	B	C
1			
2	Year	Salesman A	Salesman B
3		(Rs. in lakhs)	(Rs. in lakhs)
4	1	85	100
5	2	90	120
6	3	95	50
7	4	150	60
8	5	200	125
9			

Figure 6.46 Screenshot of input of Example 6.11 in Excel sheet



	A	B	C	D
1				Workings
2	Year	Salesman A	Salesman B	
3		(Rs. in lakhs)	(Rs. in lakhs)	
4	1	85	100	
5	2	90	120	
6	3	95	50	
7	4	150	60	
8	5	200	125	
9				
10	Mean sales of Salesman A =			124
11	Mean sales of Salesman B =			91
12				
13	Standard devation of slaes of Salesman A =			49.92494367
14	Standard devation of slaes of Salesman B =			34.35112807
15				
16	CV of sales of Salesman A =			0.402620513
17	CV of sales of Salesman B =			0.377484924
18				
19	The salesman B is more consistent.			
20				

Figure 6.47 Screenshot of working of CVs of salespeople



	A	B	C	D
1			Formulas of	Workings
2	Year	Salesman A	Salesman B	
3		(Rs. in lakhs)	(Rs. kin lakhs)	
4	1	85	100	
5	2	90	120	
6	3	95	50	
7	4	150	60	
8	5	200	125	
9				
10	Mean sales of Salesman A =			=AVERAGE(B4:B8)
11	Mean sales of Salesman B =			=AVERAGE(C4:C8)
12				
13	Standard devation of slaes of Salesman A =			=STDEV.S(B4:B8)
14	Standard devation of slaes of Salesman B =			=STDEV.S(C4:C8)
15				
16	CV of sales of Salesman A =			=D13/D10
17	CV of sales of Salesman B =			=D14/D11
18				
19	The salesman B is more consistent.			
20				

Figure 6.48 Screenshot of formulas for the working of CVs of annual sales of salespeople

Since the coefficient of variation of salesperson B is less when compared to that of salesperson A, it is concluded that salesperson B has more consistent sales.

Summary

- Spread distinguishes samples with the same mean.
- Range is a simple measure of variation, which is the difference between the highest value of a set of observations and the lowest value of that set of observations.
- The formula for the range is as follows.

  $R = H - L$
  where
  H is the highest value of a set of observations
  L is the lowest value of a set of observations
  R is the range of a set of observations

- The coefficient of range based on the range is the ratio between the range and the sum of the highest value of a set of observations and the lowest value of that set of observations.
- The STDEVA function finds the standard deviation of the observations in a sample of a population including logical values and text.
- The formula for STDEVA is: =STDEVA(Range of cells containing numbers, text and logical values of sample in a population).

- The STDEVP or STDEV.P function finds the standard deviation of a set of observations, excluding text and logical values in it.
- The formula for STDEV.P is: =STDEV.P(Range of cells containing numbers, text and logical values).
- The STDEVPA function finds the standard deviation of the observations in the entire population including logical values and text. *In this case, the formula for the standard deviation will have n as the denominator. This is known as a biased estimator of standard deviation.*
- The formula for STDEVPA is: =STDEVPA(Range of cells containing numbers, text and logical values of population).
- The coefficient of variation aims to check the consistency of the observations of a variable of interest.
- VAR.S finds the variance of a sample by excluding logical and text.
- VAR.P finds the variance of a population by excluding logical and text.
- VARA finds the variance of a sample by including logical and text.
- VARPA finds the variance of a population by including logical and text.
- The formula for the coefficient of variation is presented as follows.

$$Coefficient\,of\,variation(CV) = \frac{\sigma}{\bar{X}}$$

where
σ is the standard deviation of the observations
$\bar{X}$ is the mean of the observations

Keywords

Spread distinguishes samples with the same mean.

Range is a simple measure of variation, which is the difference between the highest value of a set of observations and the lowest value of that set of observations.

Coefficient of range based on the range is the ratio between the range and the sum of the highest value of a set of observations and the lowest value of that set of observations.

Quartile is a deviation of a random variable based on the cumulative frequencies of a set of observations, which is classified into three categories, first quartile, second quartile, and third quartile.

Quartile deviation (QD) is half of the difference between the third quartile and the first quartile.

Average deviation is the mean of the absolute deviations of the observations from the mean of those observations.

Variance is a measure of variation, which is the average of the squares of deviations of the mean of a set of observations from individual observations of that set of observations.

Standard deviation is the square root of the variance. It represents the spread of the data around the mean of that data.

The STDEVA function finds the standard deviation of the observations in a sample of a population including logical values and text.

The STDEVPA function finds the standard deviation of the observations in a population including logical values and text.

The STDEVS or STDEV.S function finds the standard deviation of a set of observations of a sample, excluding text and logical values in it.

The STDEVP or STDEV.P function finds the standard deviation of a set of observations of a population, excluding text and logical values in it.

The biased estimator of standard deviation contains *n* as the denominator.

The coefficient of variation aims to check the consistency of the observations of a variable of interest.

Review Questions

1. Discuss the importance of spread over mean of a given set of data using an example.
2. Define range as well as coefficient of range.
3. The heights of the employees working in the foundry section of a company are summarised in the following table. Find its range and coefficient of range using Excel.

Employee Code	*Height in cm*
1	158
2	152
3	170
4	152
5	160
6	155
7	179
8	168
9	164
10	170
11	175
12	162
13	172
14	165

4. a. What is quartile deviation? Give its mathematical formula.

   b. Give the mathematical formula for each of the following.
   – First quartile
   – Second quartile
   – Third quartile

5. List different QUARTILE functions in Excel and give the syntax of each of them.
6. The data for the monthly Advertising Expenditure of the last 12 months in lakhs of rupees of a company are shown in the following table.

   Find the quartile deviation of these data using QUARTILE.INC function in Excel.

Month	*Monthly Advertising Expenditure (Lakhs of Rupees)*
1	50
2	68
3	75
4	52

Month	*Monthly Advertising Expenditure (Lakhs of Rupees)*
5	60
6	59
7	79
8	68
9	67
10	82
11	72
12	75

7. The distribution of a deposit amount (in lakhs of rupees) in a bank is shown in the following table. Find the quartile deviation of the deposit amount using Excel.

Deposit Amount (₹ in Lakhs)	*No. of Firms*
Below 2	20
2–4	30
4–8	24
8–12	20
12–16	12
16–20	10
More than 20	1

8. The following table displays how a product's daily manufacturing units are distributed inside a heavy engineering organisation. Using Excel, calculate the average deviation and coefficient of average deviation based on the mean for the product's daily manufacturing units.

Daily Production Units	*Number of Days*
90–95	35
95–100	40
100–105	75
105–110	50
110–115	75
115–120	50
120–125	25
125–130	15

9. The following table displays the student pass percentage for instruction using a blackboard and ICT for eight distinct batches.

   a. Using Excel, determine the ideal teaching strategy based on the mean pass percentage.
   b. If both teaching strategies have the same mean pass percentage, use Excel to determine which strategy has the lowest standard deviation.

Batch No	*Teaching Through*	
	Blackboard	*ICT*
1	95	98
2	98	97

Batch No	*Teaching Through*	
	Blackboard	*ICT*
3	90	85
4	88	92
5	92	88
6	87	90
7	95	85
8	85	95

10. The distributions of the daily sales of cake in kgs at a bakery are shown in the following table.

    Find the standard deviation of the sales of the cake in the bakery using Excel.

Daily sales of cake (kgs)	*Number of days of sales*
120	10
121	12
122	16
123	20
124	25
125	15
126	10
127	16
128	14
129	10
130	18
131	15
132	20
133	10
134	8
135	5

11. Find the Excel STDEVPA function using an example.
12. Explain the purpose of the STDEVP or STDEV.P function in Excel and give their syntax.
13. Explain the purpose of the STDEVPA function in Excel and give its syntax.
14. Distinguish between the VAR.S function and VAR.P function
15. Distinguish between the VARA function and VARPA function.
16. Define coefficient of variation and give its mathematical formula.
17. The yield amounts for each of the two types of fertiliser, Type X and Type Y, in 10 separate plots are summarised in the following table in terms of 75 kg bags of paddy. Find the best fertiliser using the coefficient of variation of the yield with regard to each type of fertiliser.

Plot No.	*Type of Fertiliser*	
	X	*Y*
1	11	12
2	10	11
3	8	9
4	12	10
5	10	12
6	13	9
7	9	12
8	11	8
9	8	9
10	10	8

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. www.Excel-easy.com/examples/standard-deviation.html [June 25, 2020].
3. https://support.microsoft.com/en-us/office/varp-function-26a541c4-ecee-464d-a731-bd4c575b1a6b [June 30, 2020].
4. https://Exceljet.net/Excel-functions/Excel-stdev.s-function [July 3, 2020].
5. https://support.microsoft.com/en-us/office/var-s-function-913633de-136b-449d-813e-65a00b2b990b [June 30, 2020].
6. https://Exceljet.net/Excel-functions/Excel-stdeva-function [July 3, 2020].

7 Measures of Skewness

Learning Objectives

The topics of this chapter will enable the readers to

- Understand skewness and its importance.
- Analyse ungrouped data based on Karl Pearson's coefficient of skewness.
- Understand the steps to analyse grouped data based on Pearson's coefficient of skewness.
- Compute Bowley's coefficient of skewness for analysing ungrouped data.
- Understand the procedure to analyse grouped data using Bowley's coefficient of skewness.
- Analyse kurtosis for different types of data.

7.1 Introduction

The mean and standard deviation of two separate samples may really be the same. Therefore, it won't be possible to tell the samples are different from one another. In this case, one can make use of certain distributional shape information. The following three groups are used to categorise distribution shapes.

- Symmetrical distribution
- Positively skewed distribution
- Negatively skewed distribution

The left half and the right half of the distribution will be equally symmetrical in the case of the symmetrical distribution. On the right side of the distribution, the favourably skewed distribution will have a slender tail component. On the left side of the distribution, the negatively skewed distribution will have a part with a thinner tail.

The coefficient of skewness is used to quantify them. It is also known as a characterisation of a distribution's degree of asymmetry around its mean. The coefficient of skewness has a range of –1 to +1. The distribution is symmetrical if the coefficient of skewness is zero. If it is positive, the distribution's right side will contain a section with a thinner tail. In the event that it is negative, the distribution's left side will have a section with a thinner tail.

DOI: 10.4324/9781032671703-7

When the coefficient of skewness is positive, the relationship among the mean, median, and mode is as follows.

Mean > Median > Mode

When the coefficient of skewness is negative, the relationship among the mean, median, and mode is as follows.

Mean < Median < Mode

The different types of coefficient of skewness are as follows [1].

- Pearson's coefficient of skewness
- Bowley's coefficient of skewness

7.2 Pearson's Coefficient of Skewness Using Excel Sheets

Assume that the mean, median, and mode of a given set of observations are known. Then the formula to compute Pearson's coefficient of skewness (*CS*) is as follows.

$$CS = \frac{(Mean - Mode)}{\sigma}$$

where
CS is Pearson's coefficient of skewness
σ is the standard deviation of the given set of observations
Mean is the arithmetic mean of the given set of observations
Mode is the value of variable with respect to the maximum frequency

An approximate formula for Pearson's coefficient of skewness based on the relation mode is equal to (3Median – 2Mean), as follows.

$$CS = \frac{3 \times (Mean - Median)}{\sigma}$$

where
CS is Pearson's coefficient of skewness
σ is the standard deviation of the given set of observations
Mean is the arithmetic mean of the given set of observations
Median is the value of variable with respect to the half of the total frequency

Example 7.1

Table 7.1 displays a company's annual sales in crores of rupees for the previous 12 years.

1. Using Excel, calculate the annual sales' Pearson's coefficient of skewness and remark on the distribution's shape.
b. Repeat part 1 for the population parameter of the standard deviation.

Table 7.1 Data on Annual Sales

Year	*Sales in Crores of ₹*
1	10
2	12
3	14
4	12
5	16
6	10
7	14
8	10
9	15
10	12
11	15
12	16

Table 7.2 Data for Example 7.1

Year	*Sales in Crores of ₹*
1	10
2	12
3	14
4	12
5	16
6	10
7	14
8	10
9	15
10	12
11	15
12	16

Solution

The data for Example 7.1 are shown in Table 7.2.

The formula to compute Pearson's coefficient of skewness (*CS*) is as follows.

$$CS = \frac{3\times(Mean - Median)}{\sigma}$$

where
CS is Pearson's coefficient of skewness
σ is the standard deviation of the given set of observations
Mean is the arithmetic mean of the given set of observations
Median is the value of variable with respect to the half of the total frequency

1. Determination of Pearson's Coefficient of Skewness Using Sample Parameter of σ

   If the denominator of the formula for the standard deviation is $n - 1$, the standard deviation is called the sample parameter. Instead, if the denominator is n, then the standard deviation is the population parameter.

This section gives the coefficient of skewness using the sample parameter of the standard deviation for the given problem.

The input for Example 7.1 is shown in Figure 7.1. The screenshot for the button clicks Home ⟹ Formulas ⟹ More Functions ⟹ Statistical is shown in Figure 7.2. Clicking the SKEW function in the dropdown menu of Figure 7.2 gives the screenshot in Figure 7.3. Filling the range B3:B14 in the box against Number 1 in the dropdown menu of Figure 7.3 and clicking the OK button gives the screenshot in Figure 7.4. Clicking the OK button in the dropdown menu of Figure 7.4 gives the screenshot in Figure 7.5, which contains the value of Pearson's coefficient of skewness. The value of this coefficient is –0.108122691, which means that the distribution of the given data is almost symmetrical.

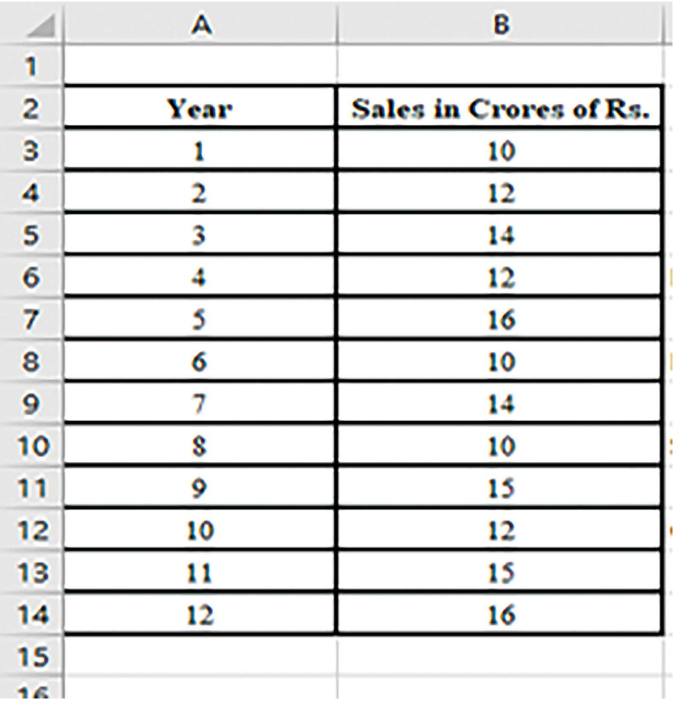
The syntax for the skewness is: = SKEW(Range of cells containing data)

The formula for the given problem to find its coefficient skewness assuming a sample parameter for the standard deviation is: $= \text{SKEW}(\text{B3} : \text{B14})$

2. Determination of Pearson's Coefficient of Skewness Assuming Population Parameter for σ

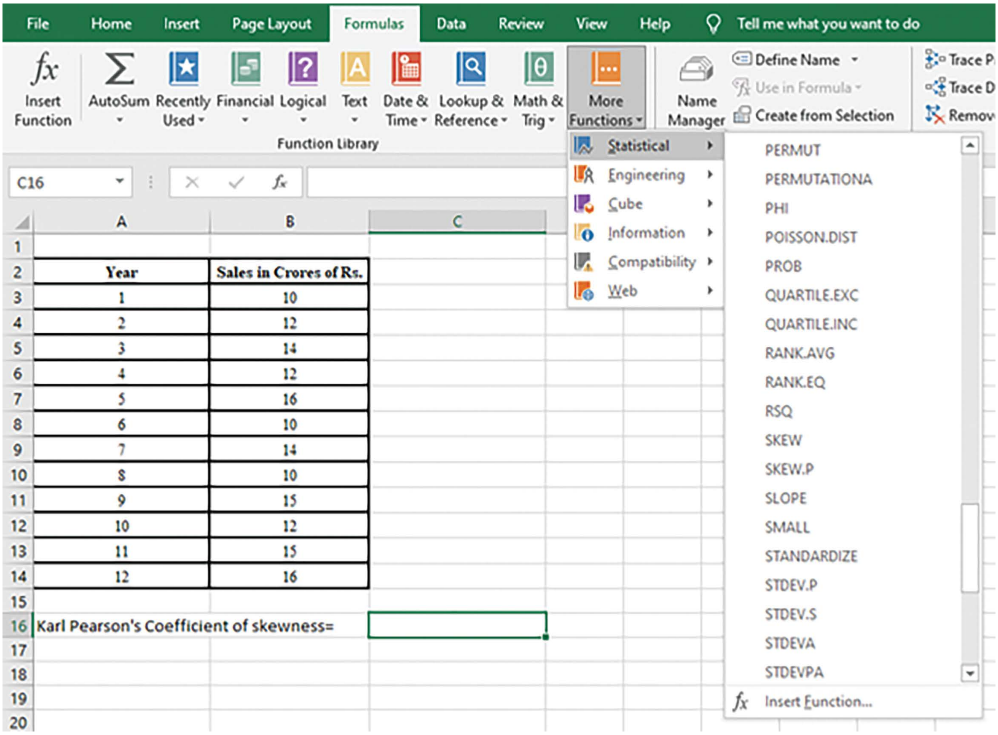
   Since the menu-driven approach for the determination of Pearson's coefficient of skewness with sample parameter for the standard deviation has been demonstrated in part 1 of this question, the final screenshot for Pearson's coefficient of skewness with the population parameter for the standard deviation is shown in Figure 7.6.

   The same may be obtained using the formula: =SKEW.P(Range of cells containing data)



	A	B
1		
2	Year	Sales in Crores of Rs.
3	1	10
4	2	12
5	3	14
6	4	12
7	5	16
8	6	10
9	7	14
10	8	10
11	9	15
12	10	12
13	11	15
14	12	16
15		

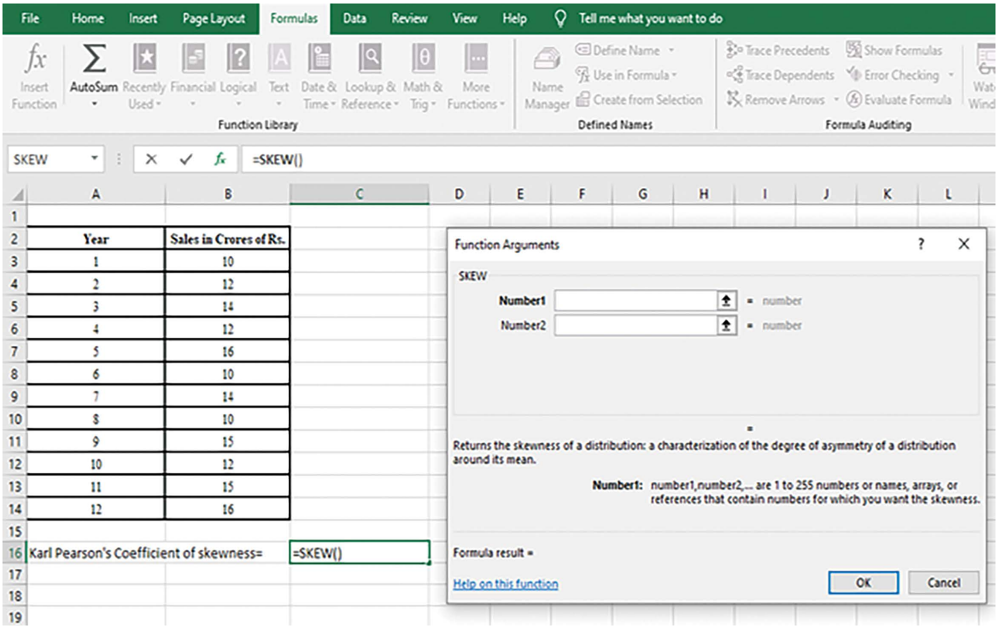
Figure 7.1 Screenshot of input of Example 7.1 in Excel sheet



Year	Sales in Crores of Rs.
1	10
2	12
3	14
4	12
5	16
6	10
7	14
8	10
9	15
10	12
11	15
12	16

Karl Pearson's Coefficient of skewness=

Figure 7.2 Screenshot after clicking buttons, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical



=SKEW()

Year	Sales in Crores of Rs.
1	10
2	12
3	14
4	12
5	16
6	10
7	14
8	10
9	15
10	12
11	15
12	16

Karl Pearson's Coefficient of skewness= =SKEW()

SKEW

Number1 = number

Number2 = number

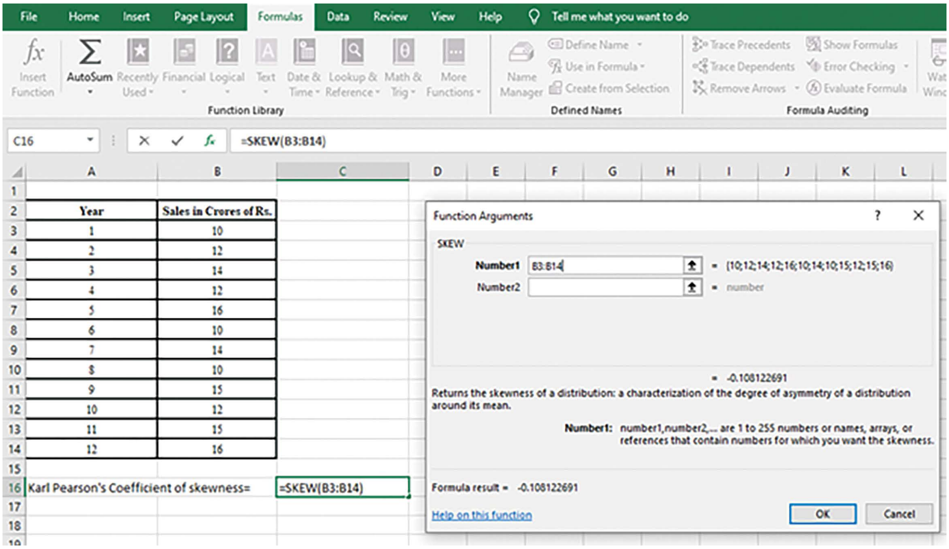
=

Returns the skewness of a distribution: a characterization of the degree of asymmetry of a distribution around its mean.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the skewness.

Formula result =

Figure 7.3 Screenshot after clicking SKEW() function in the dropdown menu of Figure 7.2



=SKEW(B3:B14)

	A	B	C
1			
2	Year	Sales in Crores of Rs.	
3	1	10	
4	2	12	
5	3	14	
6	4	12	
7	5	16	
8	6	10	
9	7	14	
10	8	10	
11	9	15	
12	10	12	
13	11	15	
14	12	16	
15			
16	Karl Pearson's Coefficient of skewness=		=SKEW(B3:B14)

Function Arguments

SKEW

Number1 B3:B14 = {10;12;14;12;16;10;14;10;15;12;15;16}

Number2 = number

= -0.108122691

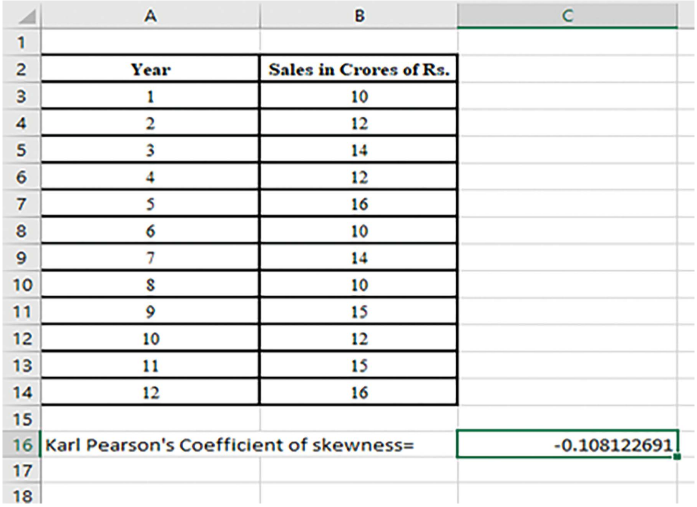
Returns the skewness of a distribution: a characterization of the degree of asymmetry of a distribution around its mean.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the skewness.

Formula result = -0.108122691

Help on this function

Figure 7.4 Screenshot after filling the range of data (B3:B14) in the dropdown menu of Figure 7.3



	A	B	C
1			
2	Year	Sales in Crores of Rs.	
3	1	10	
4	2	12	
5	3	14	
6	4	12	
7	5	16	
8	6	10	
9	7	14	
10	8	10	
11	9	15	
12	10	12	
13	11	15	
14	12	16	
15			
16	Karl Pearson's Coefficient of skewness=		-0.108122691
17			
18			

Figure 7.5 Screenshot after clicking the OK button in the dropdown menu of Figure 7.4

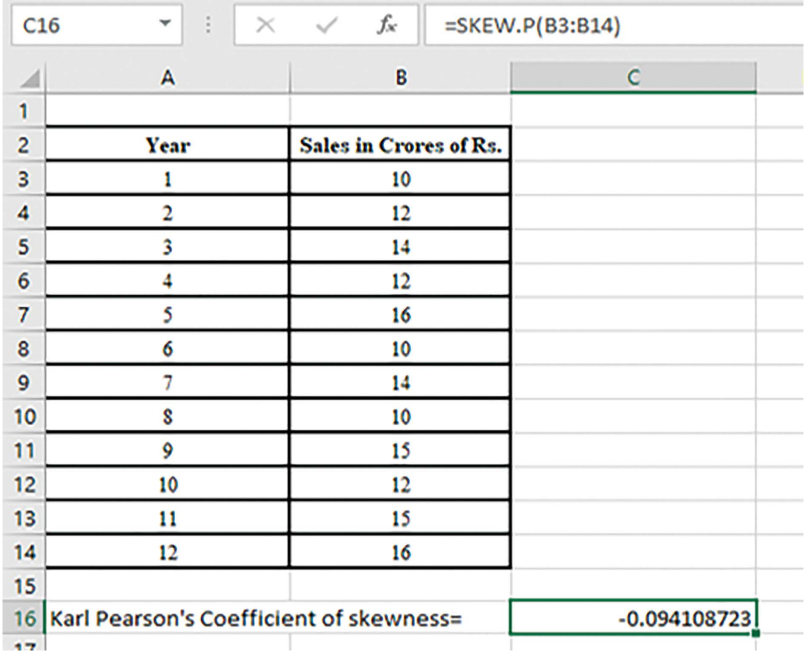
For the given data, this formula is as follows [2, 3, 4].

=SKEW.P(B3:B14)

From Figure 7.6, one can see the value of Pearson's coefficient of skewness is –0.094108723.

Example 7.2

Table 7.3 displays information on employee ages at a company. Find the coefficient of skewness of that data and comment on the shape of the distribution.



=SKEW.P(B3:B14)

	A	B	C
1			
2	Year	Sales in Crores of Rs.	
3	1	10	
4	2	12	
5	3	14	
6	4	12	
7	5	16	
8	6	10	
9	7	14	
10	8	10	
11	9	15	
12	10	12	
13	11	15	
14	12	16	
15			
16	Karl Pearson's Coefficient of skewness=		-0.094108723

Figure 7.6 Screenshot of final result for Karl Pearson's coefficient of skewness with population parameter for σ

Table 7.3 Data on Age of Employees

Age in years	*No. of Employees*
22–26	30
26–30	50
30–34	70
34–38	80
38–42	70
42–46	60
46–50	40
50–54	30

Solution

The data for Example 7.2 are shown in Table 7.4.

The formula to compute Pearson's coefficient of skewness (CS) is as follows.

$$CS = \frac{(Mean - Mode)}{\sigma}$$

where
CS is Pearson's coefficient of skewness
σ is the standard deviation of the given set of observations

Table 7.4 Data for Example 7.2

Age in years	*No. of Employees*
22–26	30
26–30	50
30–34	70
34–38	80
38–42	70
42–46	60
46–50	40
50–54	30

Mean is the arithmetic mean of the given set of observations
Mode is the maximum frequency of the given set of observations

The formulas for mean, mode, and standard deviation, which are used in the formula for the coefficient of skewness, are as follows.

$$\text{Mean} = \frac{\sum_{i=1}^{n}(f_i X_i)}{\sum_{i=1}^{n} f_i}$$

where
n is the number of class intervals
X_i is the mid-point of the class interval i, i = 1, 2, 3, . . ., n
f_i is the frequency of the class interval i, i = 1, 2, 3, . . ., n

$$\text{Mode} = L + \left(\frac{f_1}{f_1 + f_2}\right) \times C$$

where
L is the lower limit of the modal class
f_1 is the absolute difference between the cumulative frequency of the modal class and that of its preceding class
f_2 is the absolute difference between the cumulative frequency of the modal class and that of its succeeding class
C is the width of the class interval

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i \times X_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i \times X_i}{N}\right)^2}$$

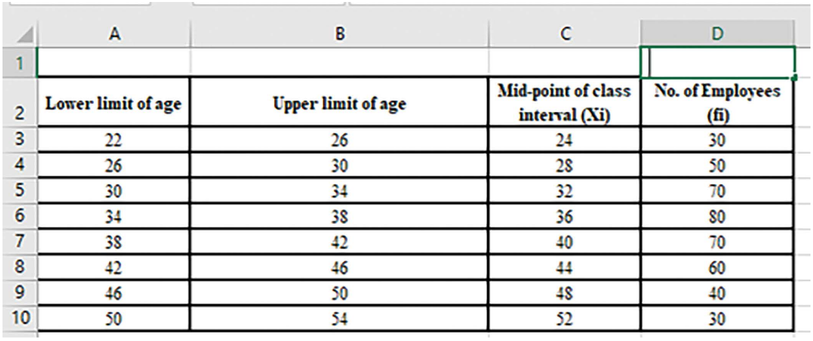
where
n is the number of class intervals
X_i is the mid-point of the class interval i, i = 1, 2, 3, . . . , n
f_i is the frequency of the class interval i, i = 1, 2, 3, . . . , n
N is the sum of the frequencies of all class intervals
σ is the standard deviation

The input of Example 7.2 is shown in Figure 7.7. The working of Pearson's coefficient of skewness is shown in Figure 7.8. The formulas for the working of Pearson's coefficient of skewness are shown in Figure 7.9. From Figure 7.8, it is clear that the coefficient of skewness value is 0.18635032. Since, the coefficient of skewness is positive, the distribution of the data of the age of the employees is positively skewed, which means that the distribution has a thinner-tailed portion on its right tail.

7.3 Bowley's Coefficient of Skewness Using Excel Sheets

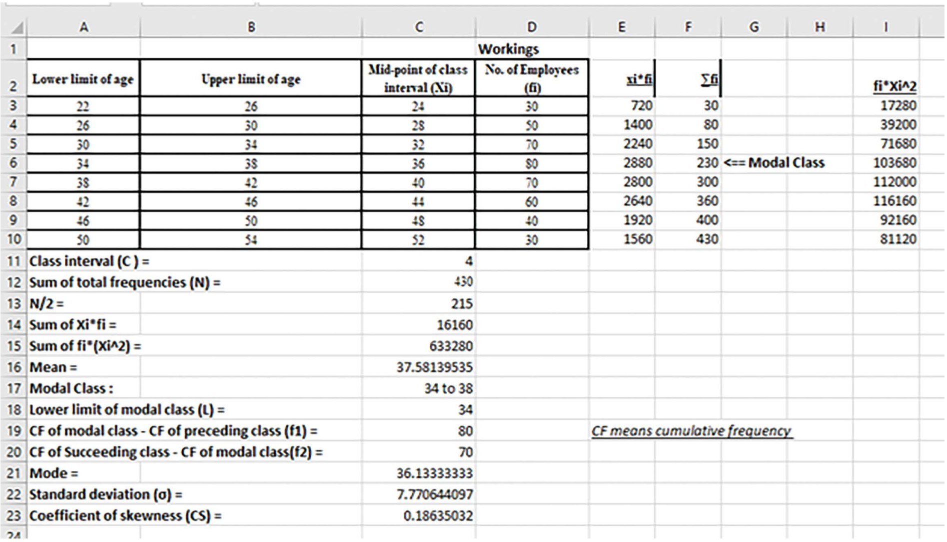
Bowley's coefficient of skewness (*CS*) is computed for grouped data with open-ended class intervals. The formula for this coefficient of skewness is as follows.

$$CS = \frac{Q_3 + Q_1 - 2 \times Median}{Q_3 - Q_1}$$



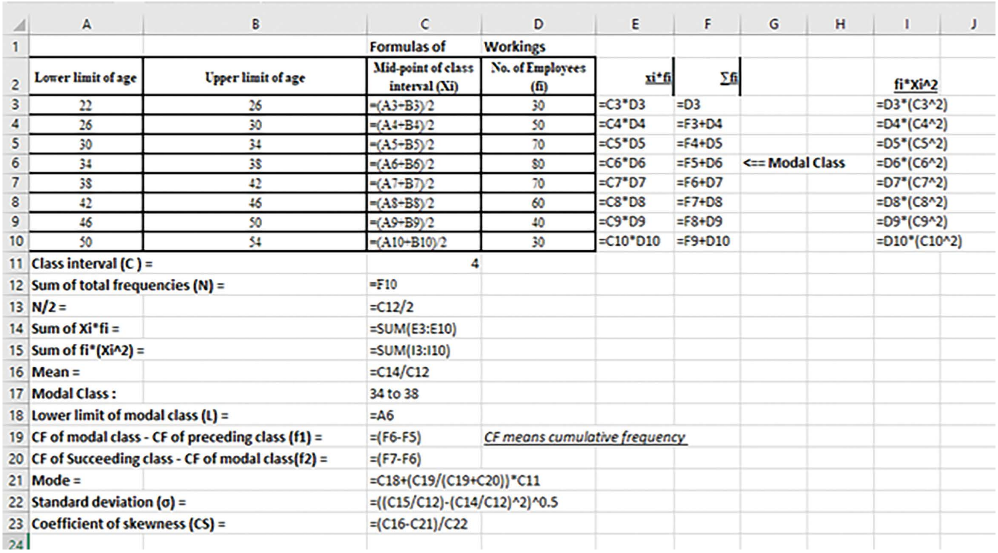
	A	B	C	D
1				
2	Lower limit of age	Upper limit of age	Mid-point of class interval (Xi)	No. of Employees (fi)
3	22	26	24	30
4	26	30	28	50
5	30	34	32	70
6	34	38	36	80
7	38	42	40	70
8	42	46	44	60
9	46	50	48	40
10	50	54	52	30

Figure 7.7 Screenshot of data for Example 7.2



	A	B	C	D	E	F	G	H	I
1				Workings					
2	Lower limit of age	Upper limit of age	Mid-point of class interval (Xi)	No. of Employees (fi)	xi*fi	Σfi			fi*Xi^2
3	22	26	24	30	720	30			17280
4	26	30	28	50	1400	80			39200
5	30	34	32	70	2240	150			71680
6	34	38	36	80	2880	230	<== Modal Class		103680
7	38	42	40	70	2800	300			112000
8	42	46	44	60	2640	360			116160
9	46	50	48	40	1920	400			92160
10	50	54	52	30	1560	430			81120
11	Class interval (C) =		4						
12	Sum of total frequencies (N) =		430						
13	N/2 =		215						
14	Sum of Xi*fi =		16160						
15	Sum of fi*(Xi^2) =		633280						
16	Mean =		37.58139535						
17	Modal Class :		34 to 38						
18	Lower limit of modal class (L) =		34						
19	CF of modal class - CF of preceding class (f1) =		80		CF means cumulative frequency				
20	CF of Succeeding class - CF of modal class(f2) =		70						
21	Mode =		36.13333333						
22	Standard deviation (σ) =		7.770644097						
23	Coefficient of skewness (CS) =		0.18635032						
24									

Figure 7.8 Screenshot of working of Karl Pearson's coefficient of skewness of Example 7.2



	A	B	C	D	E	F	G	H	I	J
1			Formulas of	Workings						
2	Lower limit of age	Upper limit of age	Mid-point of class interval (Xi)	No. of Employees (fi)	xi*fi	Σfi			fi*Xi^2	
3	22	26	=(A3+B3)/2	30	=C3*D3	=D3			=D3*(C3^2)	
4	26	30	=(A4+B4)/2	50	=C4*D4	=F3+D4			=D4*(C4^2)	
5	30	34	=(A5+B5)/2	70	=C5*D5	=F4+D5			=D5*(C5^2)	
6	34	38	=(A6+B6)/2	80	=C6*D6	=F5+D6	<== Modal Class		=D6*(C6^2)	
7	38	42	=(A7+B7)/2	70	=C7*D7	=F6+D7			=D7*(C7^2)	
8	42	46	=(A8+B8)/2	60	=C8*D8	=F7+D8			=D8*(C8^2)	
9	46	50	=(A9+B9)/2	40	=C9*D9	=F8+D9			=D9*(C9^2)	
10	50	54	=(A10+B10)/2	30	=C10*D10	=F9+D10			=D10*(C10^2)	
11	Class interval (C) =		4							
12	Sum of total frequencies (N) =		=F10							
13	N/2 =		=C12/2							
14	Sum of Xi*fi =		=SUM(E3:E10)							
15	Sum of fi*(Xi^2) =		=SUM(I3:I10)							
16	Mean =		=C14/C12							
17	Modal Class :		34 to 38							
18	Lower limit of modal class (L) =		=A6							
19	CF of modal class - CF of preceding class (f1) =		=(F6-F5)	*CF means cumulative frequency*						
20	CF of Succeeding class - CF of modal class(f2) =		=(F7-F6)							
21	Mode =		=C18+(C19/(C19+C20))*C11							
22	Standard deviation (σ) =		=((C15/C12)-(C14/C12)^2)^0.5							
23	Coefficient of skewness (CS) =		=(C16-C21)/C22							
24										

Figure 7.9 Screenshot of formulas for working of Karl Pearson's coefficient of skewness of Example 7.2

where
Q_1 is the first quartile of the distribution
Q_3 is the third quartile of the distribution
CS is the coefficient of skewness

For frequency-based data, the formulas for Q_1, Q_3, and Median are as follows.

$$\text{First quartile}(Q_1) = L_1 + \left(\frac{\frac{N}{4} - F}{f_1} \right) \times C$$

where
Q_1 is the first quartile
f_1 is the frequency of the first quartile class
F is the cumulative frequency of the immediate previous class interval with respect to the first quartile class
N is the sum of the frequencies of all the class intervals
C is the width of the class interval
L_1 is the lower limit of the value of the variable of the first quartile class

$$\text{Third quartile}(Q_3) = L_3 + \left(\frac{\frac{3N}{4} - F}{f_3} \right) \times C$$

where

Q_3 is the third quartile

f_3 is the frequency of the third quartile class

F is the cumulative frequency of the immediate previous class interval with respect to the third quartile class

N is the sum of the frequencies of all the class intervals

C is the width of the class interval

L_3 is the lower limit of the value of the variable of the third quartile class

$$\text{Median}\,(Q_2) = L_2 + \left(\frac{\frac{N}{2} - F}{f_2}\right) \times C$$

where

Q_2 is the second quartile

f_2 is the frequency of the second quartile class with respect to the cumulative frequency of $N/2$

F is the cumulative frequency of the immediate previous class interval with respect to the second quartile class

N is the sum of the frequencies of all the class intervals

C is the width of the class interval

L_2 is the lower limit of the value of the variable of the second quartile class

Example 7.3

The distribution of a loan amount in lakhs of rupees sanctioned to industries by a bank is shown in Table 7.5.

Find the coefficient skewness of this distribution and comment on the nature of the distribution.

Table 7.5 Data on Distribution of Loan Amounts

Loan Amount (Lakhs of ₹)	*No. of Firms*
Below 4	24
4–8	20
8–12	24
12–16	32
16–20	44
20–24	29
24–28	27
28–32	32
32–36	12
More than 36	8

Solution

The data for Example 7.3 are shown in Table 7.6. Since the data given in Table 7.6 have open-ended class intervals, Bowley's coefficient of skewness (*CS*) is to be found to determine the nature of the distribution of the data.

The formula for Bowley's coefficient of skewness (*CS*) is as follows.

$$CS = \frac{Q_3 + Q_1 - 2 \times Median}{Q_3 - Q_1}$$

where
Q_1 is the first quartile of the distribution
Q_3 is the third quartile of the distribution
CS is the coefficient of skewness

For frequency-based data, the formulas for Q_1, Q_3, and Median are as follows.

$$\text{First quartile}(Q_1) = L_1 + \left(\frac{\frac{N}{4} - F}{f_1}\right) \times C$$

where
Q_1 is the first quartile
f_1 is the frequency of the first quartile class with respect to the cumulative frequency of *N*/4
F is the cumulative frequency of the immediate previous class interval with respect to the first quartile class
N is the sum of the frequencies of all the class intervals
C is the width of the class interval

Table 7.6 Data for Example 7.3

Loan Amount (Lakhs of ₹)		*No. of Firms*
Lower Limit	*Upper Limit*	
–	4	24
4	8	20
8	12	24
12	16	32
16	20	44
20	24	29
24	28	27
28	32	32
32	36	12
36	–	8

L_1 is the lower limit of the value of the variable of the first quartile class

$$\text{Third quartile}(Q_3) = L_3 + \left(\frac{\frac{3N}{4} - F}{f_3}\right) \times C$$

where
Q_3 is the third quartile
f_3 is the frequency of the third quartile class with respect to the cumulative frequency of $3N/4$
F is the cumulative frequency of the immediate previous class interval with respect to the third quartile class
N is the sum of the frequencies of all the class intervals
C is the width of the class interval
L_3 is the lower limit of the value of the variable of the third quartile class

$$\text{Median }(Q_2) = L_2 + \left(\frac{\frac{N}{2} - F}{f_2}\right) \times C$$

where
Q_2 is the second quartile
f_2 is the frequency of the second quartile class with respect to the cumulative frequency of $N/2$
F is the cumulative frequency of the immediate previous class interval with respect to the second quartile class
N is the sum of the frequencies of all the class intervals
C is the width of the class interval
L_2 is the lower limit of the value of the variable of the second quartile class

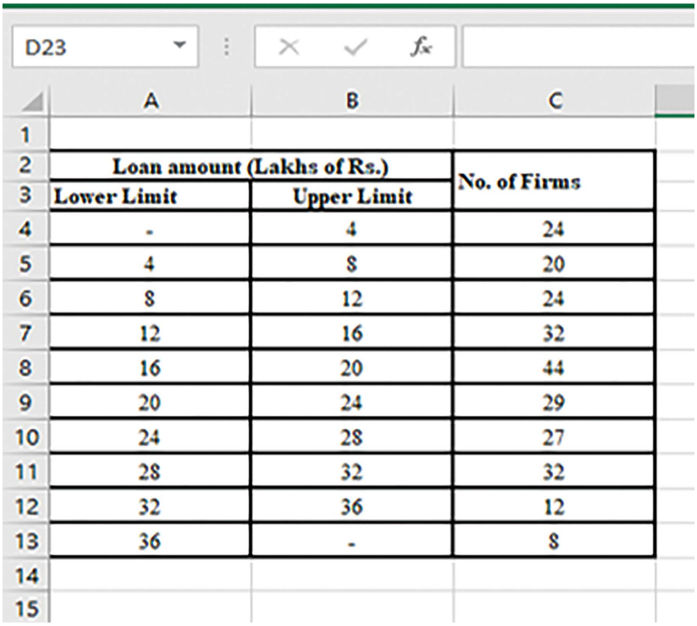
The input of Example 7.3 is shown in Figure 7.10. The working of Bowley's coefficient of skewness is shown in Figure 7.11. The formulas for the working of Bowley's coefficient of skewness are shown in Figure 7.12. From Figure 7.11, it is clear that the coefficient skewness value is 0.05326099, which is almost 0.

Since the coefficient of skewness is almost zero, the distribution of data of loan sanctioned is symmetrical about its mean.

7.4 Kurtosis

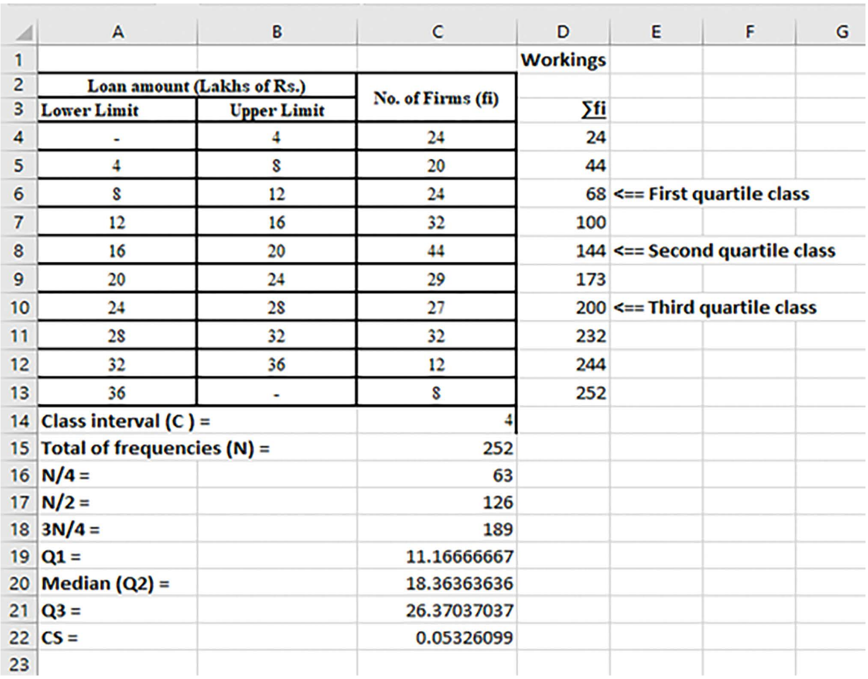
In addition to skewness, kurtosis is a measurement used to examine the properties of distributions. The kurtosis of a distribution provides information about the heaviness in terms of peaked-ness or flatness of the distribution at the tails, whereas the skewness of a distribution provides information about the symmetry of that distribution.

$$\text{Kurtosis} = \left[\frac{n \times (n+1)}{(n-1)\times(n-2)\times(n-3)} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S}\right)^4\right] - \frac{3\times(n-1)^2}{(n-2)\times(n-3)}$$



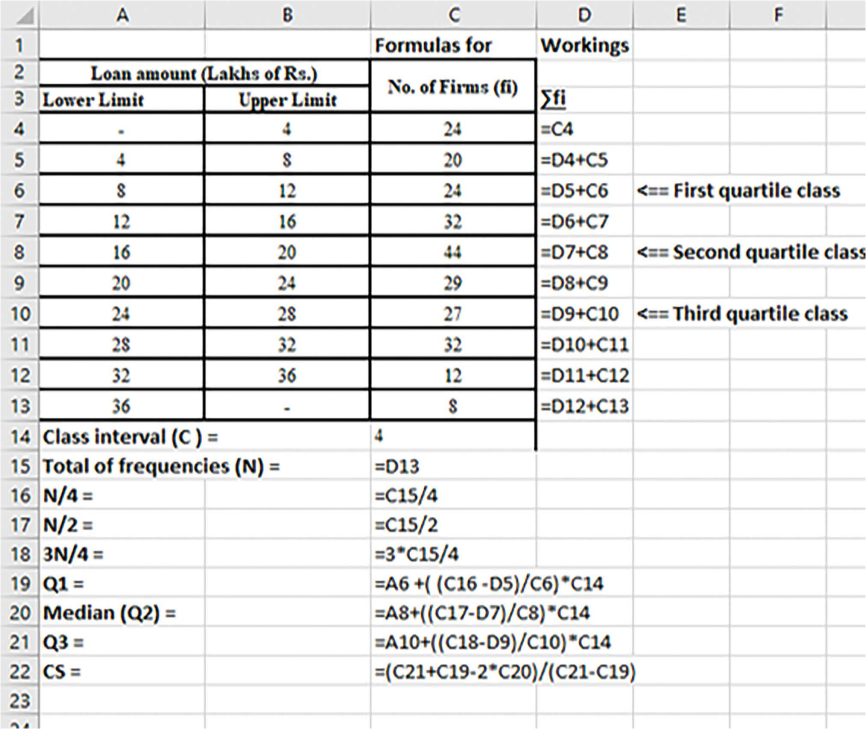
Loan amount (Lakhs of Rs.)		No. of Firms
Lower Limit	Upper Limit	
-	4	24
4	8	20
8	12	24
12	16	32
16	20	44
20	24	29
24	28	27
28	32	32
32	36	12
36	-	8

Figure 7.10 Screenshot of input of Example 7.3



Loan amount (Lakhs of Rs.)		No. of Firms (fi)	Workings	
Lower Limit	Upper Limit		∑fi	
-	4	24	24	
4	8	20	44	
8	12	24	68	<== First quartile class
12	16	32	100	
16	20	44	144	<== Second quartile class
20	24	29	173	
24	28	27	200	<== Third quartile class
28	32	32	232	
32	36	12	244	
36	-	8	252	
Class interval (C) =		4		
Total of frequencies (N) =		252		
N/4 =		63		
N/2 =		126		
3N/4 =		189		
Q1 =		11.16666667		
Median (Q2) =		18.36363636		
Q3 =		26.37037037		
CS =		0.05326099		

Figure 7.11 Screenshot of working of Bowley's coefficient of skewness for Example 7.3



	A	B	C	D	E	F
1			Formulas for	Workings		
2	Loan amount (Lakhs of Rs.)		No. of Firms (fi)			
3	Lower Limit	Upper Limit		∑fi		
4	-	4	24	=C4		
5	4	8	20	=D4+C5		
6	8	12	24	=D5+C6	<== First quartile class	
7	12	16	32	=D6+C7		
8	16	20	44	=D7+C8	<== Second quartile class	
9	20	24	29	=D8+C9		
10	24	28	27	=D9+C10	<== Third quartile class	
11	28	32	32	=D10+C11		
12	32	36	12	=D11+C12		
13	36	-	8	=D12+C13		
14	Class interval (C) =		4			
15	Total of frequencies (N) =		=D13			
16	N/4 =		=C15/4			
17	N/2 =		=C15/2			
18	3N/4 =		=3*C15/4			
19	Q1 =		=A6 +((C16 -D5)/C6)*C14			
20	Median (Q2) =		=A8+((C17-D7)/C8)*C14			
21	Q3 =		=A10+((C18-D9)/C10)*C14			
22	CS =		=(C21+C19-2*C20)/(C21-C19)			
23						

Figure 7.12 Screenshot of formulas of working of Bowley's coefficient of skewness of Example 7.3

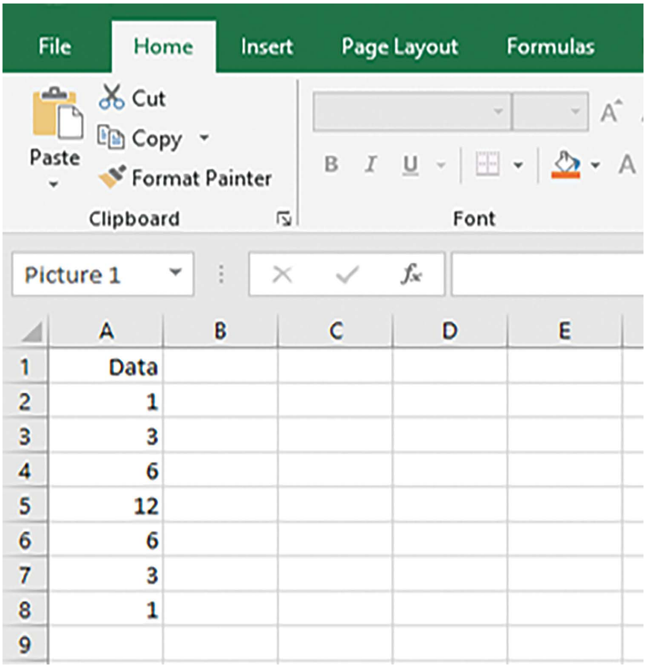
where
n is the number of observations
Xi is the i^{th} observation, i = 1, 2, 3, . . ., n
$\bar{X}$ is the mean of the observations
S is the standard deviation of the observations

The determination of the kurtosis of a given set of data in Excel using the KURT function is demonstrated as follows.

Consider a sequence of data, that is, 1 3, 6, 12, 6, 3, 1. A screenshot of this data is shown in Figure 7.13. Clicking the sequence of buttons Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ KURT gives the screenshot in Figure 7.14. The screenshot after filling the data range A2: A8 in the box against Number 1 in the dropdown menu of Figure 7.14 is shown in Figure 7.15. Clicking the OK button in Figure 7.15 gives the value of the kurtosis of the given data as 1.678364234. This is a positive kurtosis.

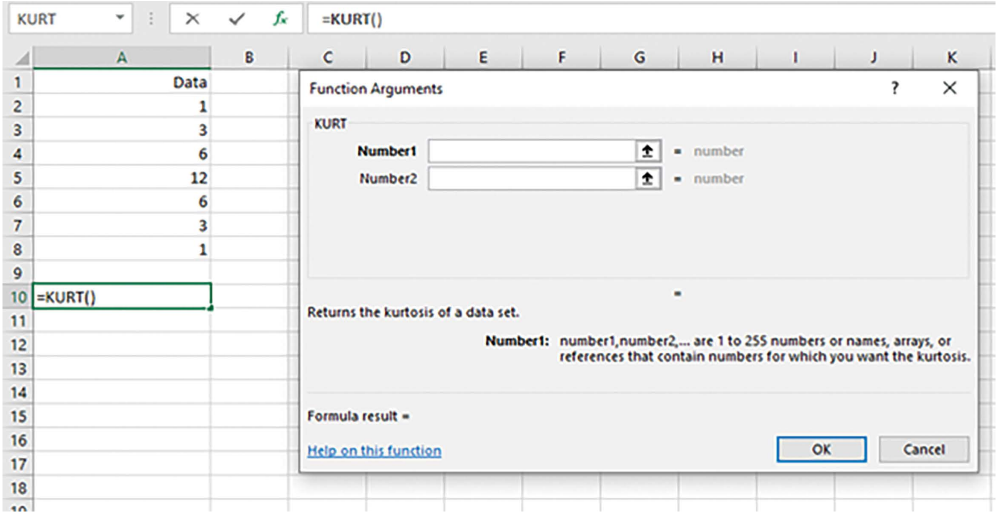
The same may be obtained by using the following Excel formula [5].

$$= \text{KURT}(\text{Range of cells containing data})$$



	A
1	Data
2	1
3	3
4	6
5	12
6	6
7	3
8	1
9	

Figure 7.13 Screenshot of sample data in Excel



=KURT()

	A
1	Data
2	1
3	3
4	6
5	12
6	6
7	3
8	1
9	
10	=KURT()

KURT

Number1 = number

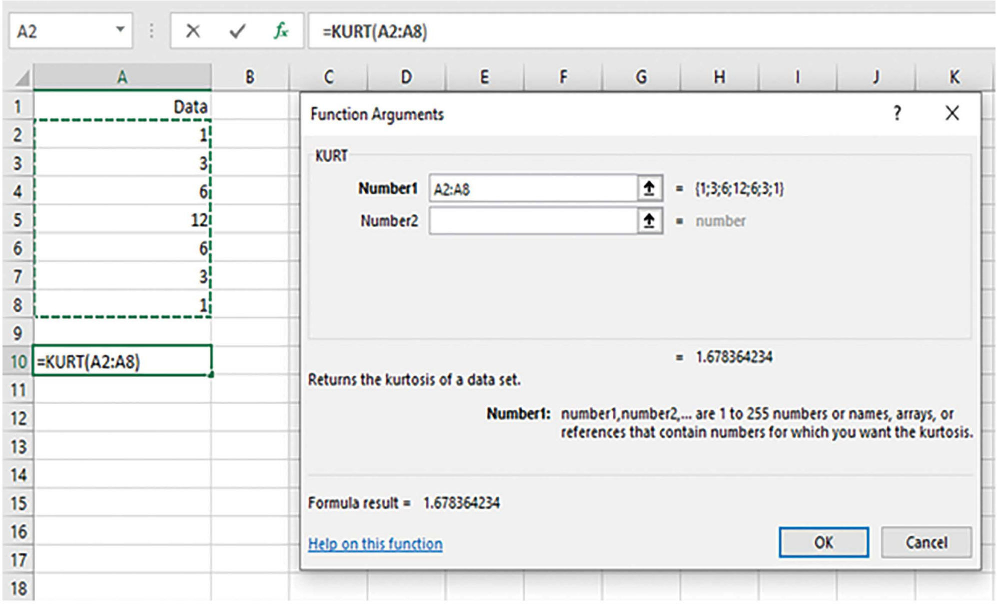
Number2 = number

Returns the kurtosis of a data set.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the kurtosis.

Formula result =

Figure 7.14 Screenshot for the sequence of button clicks, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ KURT



A2 =KURT(A2:A8)

	A
1	Data
2	1
3	3
4	6
5	12
6	6
7	3
8	1
9	
10	=KURT(A2:A8)

Function Arguments

KURT

Number1 A2:A8 = {1;3;6;12;6;3;1}

Number2 = number

= 1.678364234

Returns the kurtosis of a data set.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the kurtosis.

Formula result = 1.678364234

Figure 7.15 Screenshot after filling the data range in the box against Number 1 in the dropdown menu of Figure 7.14

There is another measure called excess kurtosis. The formula for the excess kurtosis is as follows.

Excess kurtosis of a distribution = Kurtosis of the distribution – 3

where
The value 3 in the formula is the kurtosis of the normal distribution.
Kurtosis is classified into the following types.

- Mesokurtic
- Leptokurtic
- Platykurtic

7.4.1 Mesokurtic Kurtosis

The type of kurtosis is determined by the value of the excess kurtosis. A distribution is regarded as being mesokurtic if the excess kurtosis value is zero or near zero. The normal distribution is a mesokurtic distribution, which has the excess kurtosis of 0 and its picture is shown in Figure 7.16. The spread of the mesokurtic distribution is moderate.

Example 7.4

Consider 15 sales regions in a rural area. The yearly sales revenues of these regions for a product are shown in Table 7.7. Find the kurtosis and excess kurtosis and comment on the distribution of the given data on the yearly sales revenue.

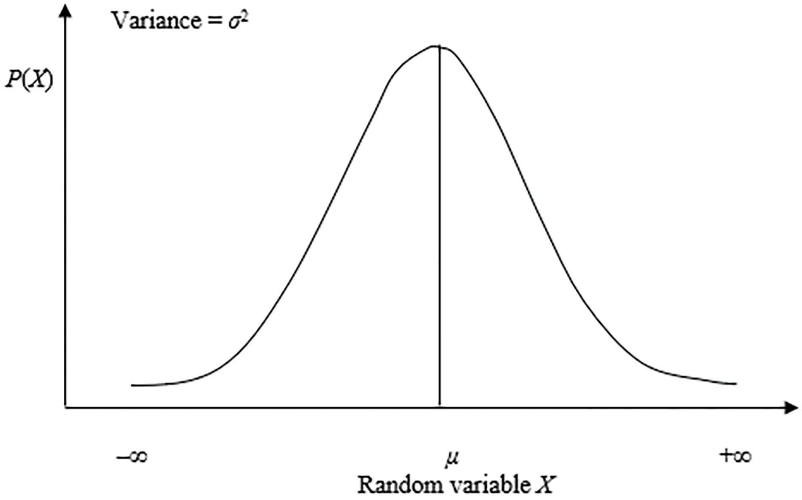


Figure 7.16 Mesokurtic distribution (normal distribution)

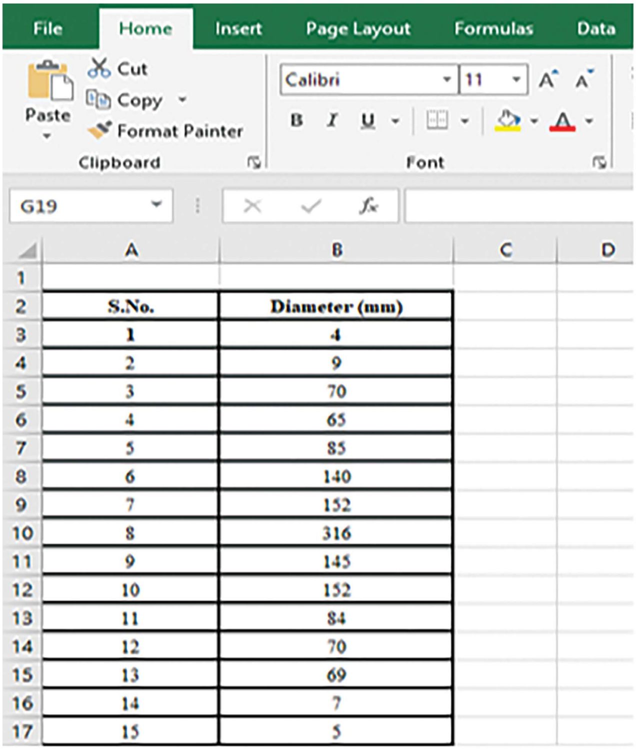
Table 7.7 Yearly Sales Revenue of Regions

Sales Region	*Revenue (Lakhs of Rupees)*
1	4
2	5
3	70
4	65
5	85
6	140
7	152
8	316
9	145
10	152
11	84
12	70
13	69
14	6
15	5

Solution

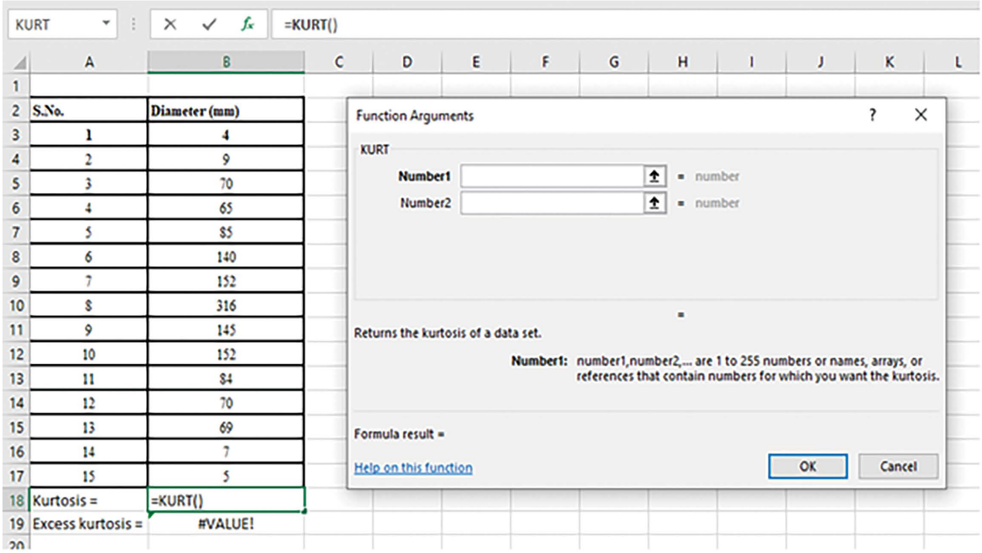
A screenshot of the given data on the yearly sales revenues of 15 sales regions in Excel is shown in Figure 7.17. Clicking the sequence of buttons Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ KURT in Figure 7.17 gives the screenshot in Figure 7.18. A screenshot after filling the data in the box against Number 1 in the dropdown menu of Figure 7.18 is shown in Figure 7.19. Clicking the OK button in the dropdown menu of Figure 7.19 is shown in Figure 7.20.

From Figure 7.20, it can be seen that the value of the kurtosis is 3.001480065 and that of the excess kurtosis is 0.001480065. Since the value of the excess kurtosis is almost 0, the given data follows mesokurtic distribution, which is a normal distribution.



	A	B
1		
2	S.No.	Diameter (mm)
3	1	4
4	2	9
5	3	70
6	4	65
7	5	85
8	6	140
9	7	152
10	8	316
11	9	145
12	10	152
13	11	84
14	12	70
15	13	69
16	14	7
17	15	5

Figure 7.17 Screenshot of data for Example 7.4



	A	B
1		
2	S.No.	Diameter (mm)
3	1	4
4	2	9
5	3	70
6	4	65
7	5	85
8	6	140
9	7	152
10	8	316
11	9	145
12	10	152
13	11	84
14	12	70
15	13	69
16	14	7
17	15	5
18	Kurtosis =	=KURT()
19	Excess kurtosis =	#VALUE!

Function Arguments

KURT

Number1 = number

Number2 = number

=

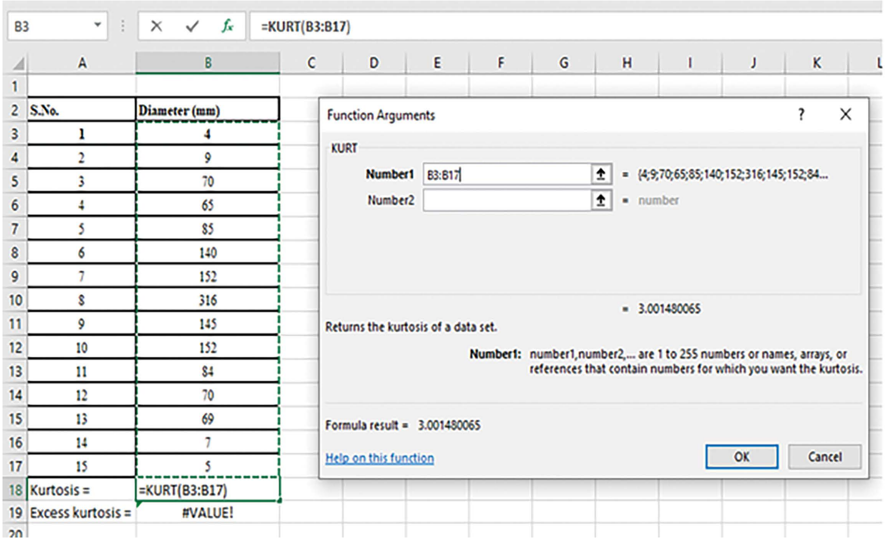
Returns the kurtosis of a data set.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the kurtosis.

Formula result =

Help on this function

Figure 7.18 Screenshot for the sequence of button clicks, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ KURT



B3 =KURT(B3:B17)

S.No.	Diameter (mm)
1	4
2	9
3	70
4	65
5	85
6	140
7	152
8	316
9	145
10	152
11	84
12	70
13	69
14	7
15	5
Kurtosis =	=KURT(B3:B17)
Excess kurtosis =	#VALUE!

Function Arguments

KURT

Number1 B3:B17 = {4;9;70;65;85;140;152;316;145;152;84...

Number2 = number

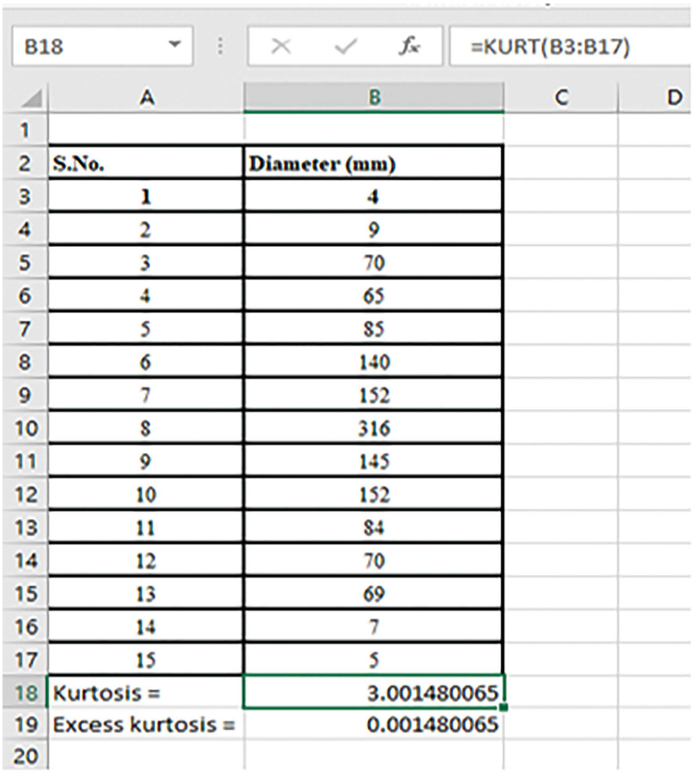
= 3.001480065

Returns the kurtosis of a data set.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the kurtosis.

Formula result = 3.001480065

Figure 7.19 Screenshot after filling the data range in the box against Number 1 in the dropdown menu of Figure 7.18



B18 =KURT(B3:B17)

S.No.	Diameter (mm)
1	4
2	9
3	70
4	65
5	85
6	140
7	152
8	316
9	145
10	152
11	84
12	70
13	69
14	7
15	5
Kurtosis =	3.001480065
Excess kurtosis =	0.001480065

Figure 7.20 Screenshot after clicking the OK button in the dropdown menu of Figure 7.19

This may be obtained using the following formulas.

Excel formula for kurtosis in cell B18:

$= \mathrm{KURT}(\mathrm{B3} : \mathrm{B17})$

Excel formula for excess kurtosis in cell B19 :

$= \mathrm{B18} - 3$

7.4.2 Leptokurtic Kurtosis

Leptokurtic is a degree of heaviness of a distribution when the excess kurtosis is positive. This means that the heaviness of the frequency will be closer to the mean of the distribution, as shown in Figure 7.21. The leptokurtic distribution has less spread.

Example 7.5

The monthly maximum share prices of a particular script during last the 16 months are summarised in Table 7.8. Find the kurtosis and excess kurtosis of this data and comment on its distribution.

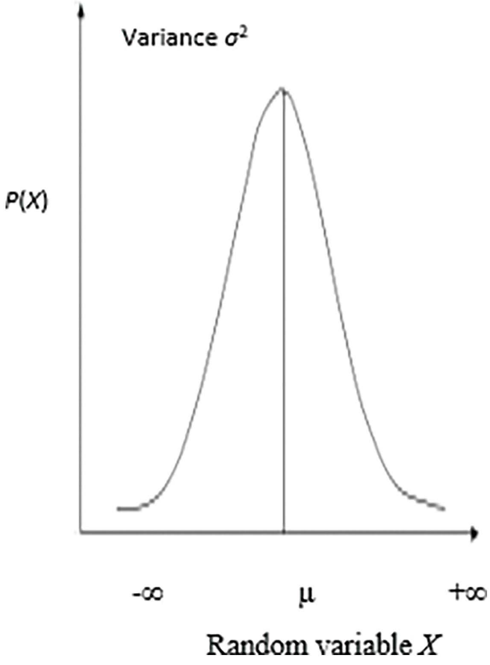


Figure 7.21 Leptokurtic distribution

Table 7.8 Monthly Maximum Share Prices

Month	*Monthly Maximum Share Price (₹)*
1	6
2	200
3	75
4	100
5	40
6	50
7	50
8	60
9	75
10	60
11	50
12	60
13	60
14	51
15	40
16	100
17	120
18	6

Solution

The screenshot of the given data for Example 7.5 on the monthly maximum share prices of a script is shown in Figure 7.22.

The generalised formula for the kurtosis is as follows.

=KURT(Range of cells containing data)

Enter the formula for kurtosis in cell B21 and the formula for the excess kurtosis in cell B22 in the screenshot shown in Figure 7.23, which are as follows. This will give the results as in Figure 7.23.

Formula for kurtosis: $= \text{KURT}(\text{B3} : \text{B20})$
Formula for excess kurtosis: $= \text{B21} - 3$

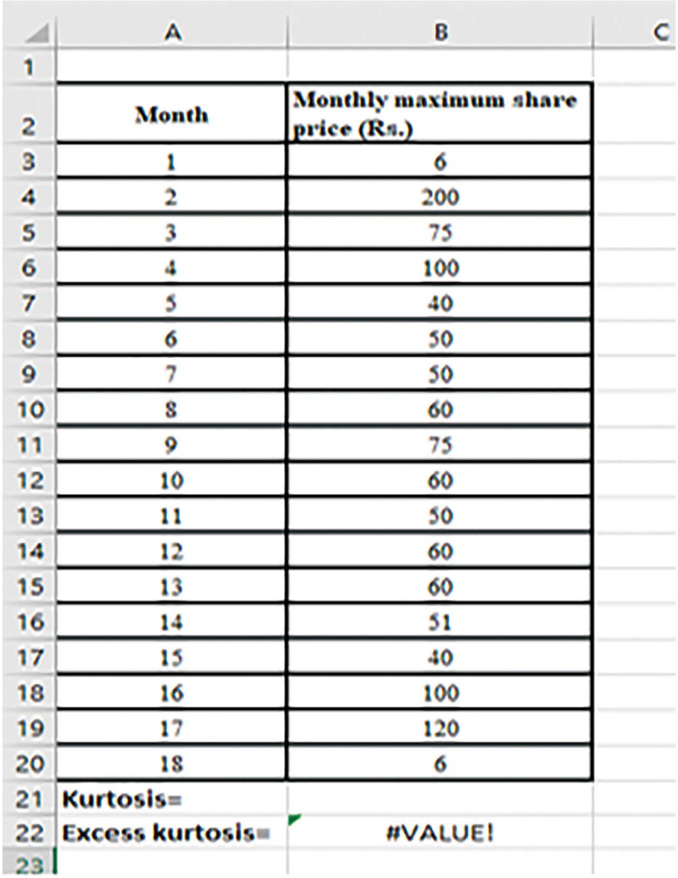
From Figure 7.23, the results are given in the following.

Kurtosis of the given data = 4.243367046
Excess kurtosis of the given data = 1.243367046

Since the excess kurtosis is positive, the distribution of the given data is leptokurtic, which means that the heaviness of the frequencies is around the mean of the distribution.

7.4.3 Platykurtic Kurtosis

A distribution's platykurtic measure can be used to determine how heavy it is in terms of frequency. In terms of the heaviness of a distribution's frequency toward its tails, a



	A	B	C
1			
2	Month	Monthly maximum share price (Rs.)	
3	1	6	
4	2	200	
5	3	75	
6	4	100	
7	5	40	
8	6	50	
9	7	50	
10	8	60	
11	9	75	
12	10	60	
13	11	50	
14	12	60	
15	13	60	
16	14	51	
17	15	40	
18	16	100	
19	17	120	
20	18	6	
21	Kurtosis=		
22	Excess kurtosis=	#VALUE!	
23			

Figure 7.22 Screenshot of data for Example 7.5

distribution is said to be platykurtic if the excess kurtosis is negative, as seen in Figure 7.24. Platykurtic dispersion has a wider range.

Example 7.6

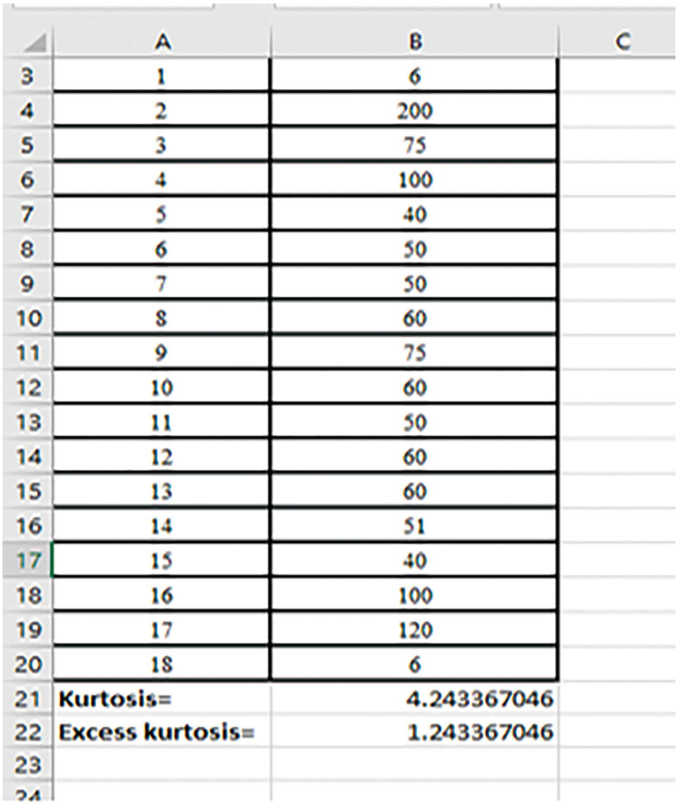
The quality manager of a manufacturing firm producing bearings wants to study the kurtosis of the diameter of the bearings produced in the firm. The data on the diameters in mm of the bearings in a sample are shown in Table 7.9. Find the kurtosis and excess kurtosis and comment on the distribution of the data.

Solution

The screenshot of the data for Example 7.6 is shown in Figure 7.25.

The generalised formula for the kurtosis is as follows.

=KURT(Range of cells containing data)



	A	B	C
3	1	6	
4	2	200	
5	3	75	
6	4	100	
7	5	40	
8	6	50	
9	7	50	
10	8	60	
11	9	75	
12	10	60	
13	11	50	
14	12	60	
15	13	60	
16	14	51	
17	15	40	
18	16	100	
19	17	120	
20	18	6	
21	**Kurtosis=**	4.243367046	
22	**Excess kurtosis=**	1.243367046	
23			
24			

Figure 7.23 Screenshot of the result of Example 7.5

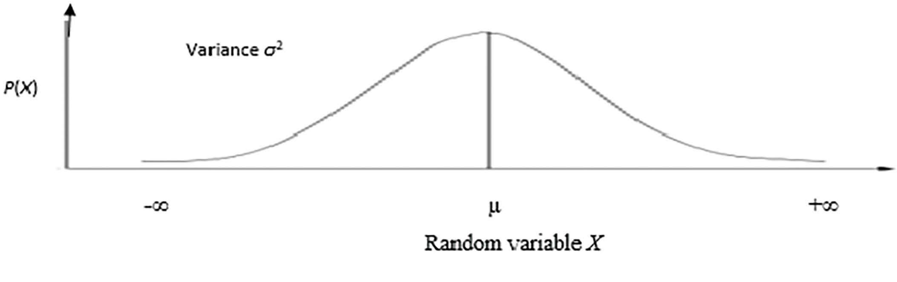


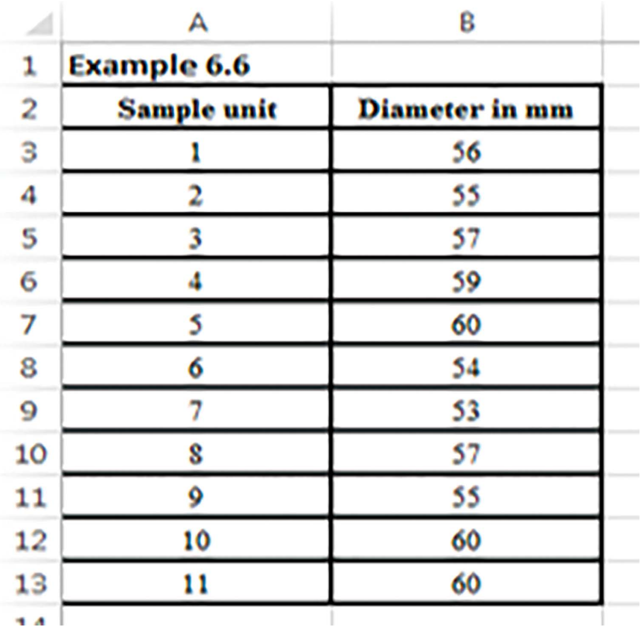
Figure 7.24 Platykurtic distribution

Enter the formula for kurtosis in cell B15 and the formula for the excess kurtosis in cell B16 in the screenshot shown in Figure 7.26, which are as follows. This will give the results as in Figure 7.26.

Formula for kurtosis: = KURT(B3 : B13)
Formula for excess kurtosis: = B15 − 3

Table 7.9 Diameters of Bearings in Millimetres

Sample Unit	*Diameter in mm*
1	56
2	55
3	57
4	59
5	60
6	54
7	53
8	57
9	55
10	60
11	60



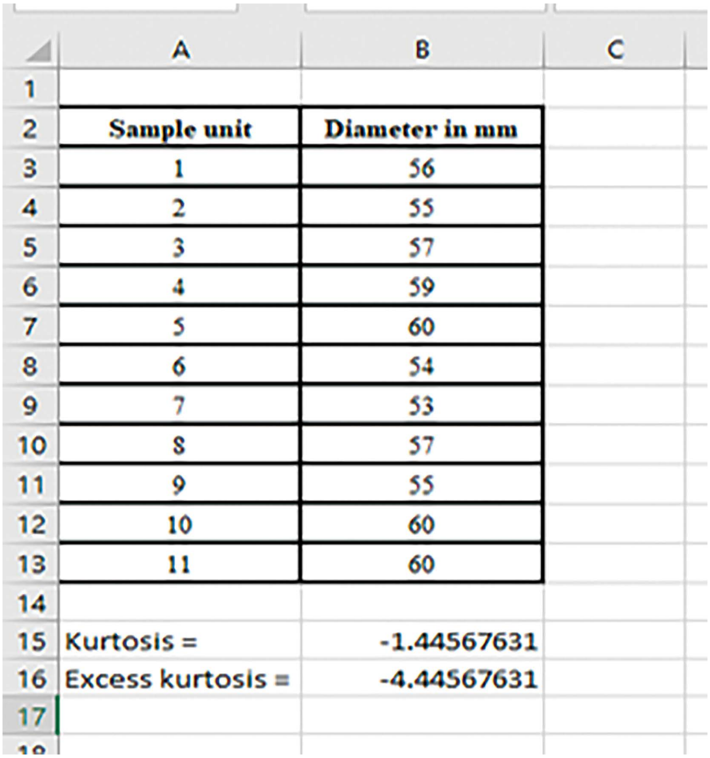
	A	B
1	**Example 6.6**	
2	**Sample unit**	**Diameter in mm**
3	**1**	**56**
4	**2**	**55**
5	**3**	**57**
6	**4**	**59**
7	**5**	**60**
8	**6**	**54**
9	**7**	**53**
10	**8**	**57**
11	**9**	**55**
12	**10**	**60**
13	**11**	**60**

Figure 7.25 Screenshot of Example 7.6

From Figure 7.26, the results are given in the following.

Kurtosis of the given data = –1.44567631
Excess kurtosis of the given data = –4.44567631

Since the excess kurtosis is negative, the distribution of the given data is platykurtic, which means that the heaviness of the frequencies is around the tails of the distribution. Alternatively, this distribution has more flatness.



Sample unit	Diameter in mm
1	56
2	55
3	57
4	59
5	60
6	54
7	53
8	57
9	55
10	60
11	60
Kurtosis =	-1.44567631
Excess kurtosis =	-4.44567631

Figure 7.26 Screenshot of result of Example 7.6

Summary

- One can use information on the shape of the distributions, which is called skewness.
- For a symmetrical distribution, the left half of the distribution and the right half of the distribution will be symmetrical.
- A positively skewed distribution will have a thinner-tailed portion on the right side of the distribution.
- A negatively skewed distribution will have a thinner-tailed portion on the left side of the distribution.
- The coefficient of skewness is called a characterisation of the degree of asymmetry of a distribution around its mean.
- The range of the coefficient of skewness is from –1 to + 1. If the coefficient of skewness is zero, then the distribution is symmetrical. If it is positive, then the distribution will have a thinner-tailed portion on the right tail of the distribution. If it is negative, then the distribution will have a thinner-tailed portion on the left tail of the distribution.
- When the coefficient of skewness is positive, the relationship among the mean, median, and mode is as follows.

  Mean > Median > Mode

- When the coefficient of skewness is negative, the relationship among the mean, median, and mode is as follows.

  Mean < Median < Mode

- The formula to compute Pearson's coefficient of skewness (*CS*) is as follows.

  $$CS = \frac{(Mean - Mode)}{\sigma}$$

  where
  CS is Pearson's coefficient of skewness
  σ is the standard deviation of the given set of observations
  Mean is the arithmetic mean of the given set of observations
  Mode is the maximum frequency of the given set of observations

- Bowley's coefficient of skewness (*CS*) is computed for grouped data with open-ended class intervals.

  The formula for this coefficient of skewness is as follows.

  $$CS = \frac{Q_3 + Q_1 - 2 \times Median}{Q_3 - Q_1}$$

  where
  Q_1 is the first quartile of the distribution
  Q_3 is the third quartile of the distribution
  CS is the coefficient of skewness

- Kurtosis of a distribution gives information about the heaviness in terms of peakedness or flatness of the distribution at tails.
- Excess kurtosis is equal to kurtosis of the distribution minus 3.
- A normal distribution has zero excess kurtosis.
- If the value of the excess kurtosis of a distribution is zero or closer to zero, then the distribution is said to be mesokurtic.
- Leptokurtic is a degree of heaviness of a distribution when the excess kurtosis is positive.
- Platykurtic is a measure of a distribution to study the heaviness of a distribution in terms of the presence of frequency.

Keywords

- A symmetrical distribution has symmetricity of the left half of the distribution and right half of the distribution.
- A positively skewed distribution contains a thinner-tailed portion on the right tail of the distribution.
- A negatively skewed distribution will have a thinner-tailed portion on the left tail of the distribution.
- Coefficient of skewness is called a characterisation of the degree of asymmetry of a distribution around its mean.
- Coefficient of skewness is zero for symmetrical distributions.
- When the coefficient of skewness is positive, then the distribution will have a thinner-tailed portion on the right tail of the distribution.
- When the coefficient of skewness is negative, then the distribution will have a thinner-tailed portion on the left tail of the distribution.
- Bowley's coefficient of skewness (*CS*) is computed for grouped data with open-ended class intervals.

- The kurtosis of a distribution gives information about the heaviness in terms of peakedness or flatness of the distribution at tails.
- Excess kurtosis is equal to kurtosis of the distribution minus 3.
- A normal distribution has zero excess kurtosis.
- Mesokurtic is degree of heaviness of a distribution when the excess kurtosis is zero.
- Leptokurtic is a degree of heaviness of a distribution when the excess kurtosis is positive.
- Platykurtic is a measure of a distribution to study the heaviness of a distribution in terms of the presence of frequency.

Review Questions

1. a. Define skewness.
   b. Define co-efficient of skewness.
   c. How do you classify a distribution based on coefficient of skewness?
2. Give the mathematical formula for Pearson's coefficient of skewness and explain the variables in it.
3. The following table displays a company's gross profit over 12 quarters in crores of rupees. Using Excel, calculate the quarterly gross profit's Pearson's coefficient of skewness and remark on the distribution's shape.

Quarter	*Quarterly Gross Profit (Crores of Rupees)*
1	80
2	120
3	140
4	130
5	160
6	100
7	140
8	100
9	150
10	120
11	150
12	160

4. The following table displays information about the respondents' annual salaries in lakhs of rupees. Utilising Excel, calculate the data's Pearson's coefficient of skewness and make a comment about the shape of the distribution.

Annual Salary (Lakhs of Rupees)	*No. of Respondents*
4–8	55
8–12	65
12–16	50
16–20	30
20–24	20
24–28	15
28–32	10

5. The following table displays the annual compensation of the CEOs of automobile firms in lakhs of rupees. Find the distribution's Bowley's coefficient skewness and describe the nature of the distribution.

Annual Salary (Lakhs of Rupees)	*No. of CEOs*
Below 40	20
40–80	16
80–120	20
120–160	28
160–200	40
200–240	25
240–280	21
More than 280	4

6. What is kurtosis? Explain its types.
7. What is excess kurtosis? Discuss its usefulness in classifying kurtosis.
8. Illustrate mesokurtic kurtosis using a suitable example in Excel.
9. Illustrate leptokurtic kurtosis with an example in Excel.
10. Illustrate platykurtic kurtosis with a suitable example in Excel.

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. www.statisticshowto.com/find-pearsons-coefficient-skewness-Excel/ [June 25, 2020].
3. https://support.microsoft.com/en-us/office/skew-p-function-76530a5c-99b9-48a1-8392-26632d542fcb [July 4, 2020].
4. www.tutorialspoint.com/advanced_Excel_functions/advanced_Excel_statistical_skewp_function.htm [July 4, 2020].
5. www.simplypsychology.org/kurtosis.html [July 7, 2020].

8 Probability Distributions

Learning Objectives

After reading this chapter, you will be able to

- Analyse binomial distribution and its applications using Excel.
- Understand the process where the Poisson distribution can be used and applied in practice using Excel.
- Study the areas of applications of exponential distribution and its theoretical aspects for implementation using Excel.
- Understand the scope of normal distribution and standard normal distribution and their theories for implementation in practice using Excel.
- Analyse the uniform distribution and its applications in practice using Excel.
- Understand the scope of the *t* distribution and its theory for practical implementation using Excel.
- Determine the confidence interval for large samples as well as small samples using Excel commands.

8.1 Introduction

Most real-world occurrences in society or in business are probabilistic in nature. The ability to make decisions requires some kind of historical data. The presentation of this form uses probability distributions. There will always be a mean and a variance for any probability distribution. The following probability distributions are presented in this chapter along with examples that use Excel to estimate confidence intervals.

1. Binomial distribution
2. Poisson distribution
3. Exponential distribution
4. Normal distribution
5. Uniform distribution
6. t distribution
7. Estimation of confidence interval

8.2 Binomial Distribution BINOM.DIST Function

Binomial distribution comes under discrete probability distribution. It is based on the *Bernoulli process*. In the experiment of tossing an unbiased coin, there are two outcomes,

DOI: 10.4324/9781032671703-8

heads or tails. The occurrence of heads will prevent the occurrence of tails, and vice versa. The experiment has only two events, which are the occurrences of heads or tails. These events are mutually exclusive and collectively exhaustive. The probability of occurrence of heads or tails is 0.5 [1].

Assume that two people are playing a game during this experiment. If the outcome of the trial is assumed to be the success of one person in this situation, it will be the failure of the other person. p represents the likelihood of a trial's success. Therefore, $1 - p$, which is taken as equal to q, is the chance of failure in a trial. Since the likelihood of one person succeeding is equal to 0.5, the probability of another person failing is equal to 0.5 as well.

Bernoulli Process: If n trials are conducted repeatedly in an experiment, one could be curious to know the likelihood that a person will achieve X successes. The Bernoulli process is the name given to such an experiment.

The assumptions of the Bernoulli process in which n repeated trials are performed are listed here.

1. The experiment has only two exclusive and collectively exhaustive events.
2. In all the trials, each of the two events has the same probability of occurrence.
3. The observations of the events are independent of one another in all n trials.

The formula for the binomial distribution is as follows.

$$P(X \text{ Successes in } n \text{ trials given } p) = n_{C_X} \; p^X q^{n-X}$$

where
$X = 0, 1, 2, 3, \ldots, n$
p is the probability of success
q is the probability of failure

A generalised representation of this distribution in terms of the binomial random variable X and two parameters, that is, number of total trials and the probability of success (p), is as follows.

$$P(X,n,p) = n_{C_X} \; p^X q^{n-X}$$

where $X = 0, 1, 2, 3, \ldots, n$, and it is a random variable
p is the probability of success
q is the probability of failure, which is $1 - p$

The way of reading $P(X, n, p)$ is the probability of having X successes in n trials given that the probability of success in a trial is p.

The cumulative distribution function of the binomial distribution is as follows.

$$F(X,n,p) = \sum_{X=0}^{n} P(X,n,p) = \sum_{X=0}^{n} n_{C_X} \; p^X q^{n-X}$$

In Excel, there are three types of functions for binomial distribution, as listed.

BINOM.DIST
BINOM.DIST.RANGE
BINOM.INV

8.2.1 BINOM.DIST Function

This section presents the use of the BINOM.DIST function [2], which finds the binomial probability of the binomial mass function as well as the probability of a binomial cumulative probability distribution.

The Excel template to compute binomial probability and binomial cumulative probability is shown in Figure 8.1. The sequence of clicks to get the display shown in Figure 8.1 is as follows.

HOME ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ BINOM.DIST

The data items which are to be fed into the boxes in the dropdown menu of Figure 8.1 are explained as follows.

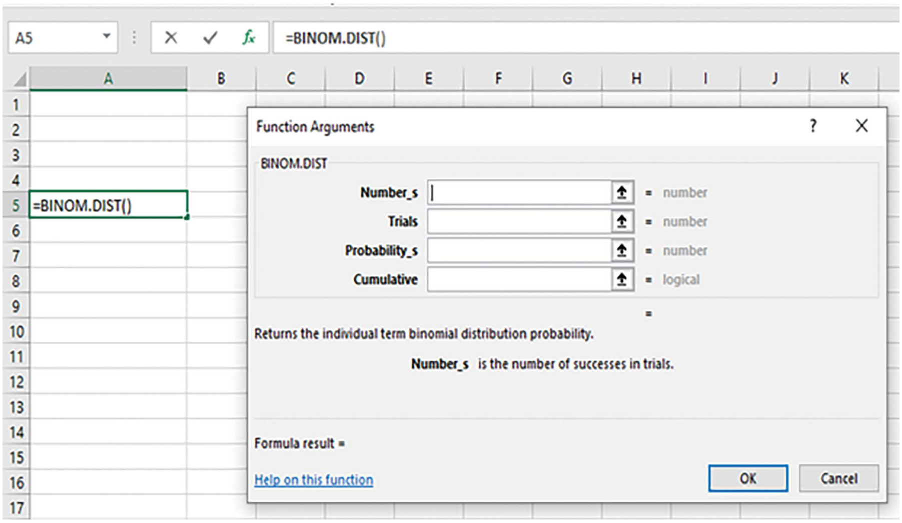
- Number_s is the number of successes in trials
- Trials is the number of independent trials
- Probability_s is the probability of success in each trial
- Cumulative is a logical value; for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE

The Excel formula to compute the cumulative distribution function is as follows.

= BINOM.DIST(Number_s, Trials, Probability_s, TRUE)

The Excel formula to compute the probability mass function (individual term of binomial distribution) is as follows.

= BINOM.DIST(Number_s, Trials, Probability_s, FALSE)



A5

=BINOM.DIST()

Function Arguments

BINOM.DIST

Number_s = number

Trials = number

Probability_s = number

Cumulative = logical

Returns the individual term binomial distribution probability.

Number_s is the number of successes in trials.

Formula result =

Figure 8.1 Screenshot for sequence of clicks, Home ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ BINOM.DIST

Example 8.1

If a coin is tossed 15 times, what is the probability of having:

1. Zero heads?
2. Five heads?
3. At most three heads?
4. At least five heads?

Solution

The probability of having heads while tossing a coin (p) = 0.5.

For each of the subsections of this problem, the result can be obtained by entering the respective formula in a convenient cell in an Excel sheet.

1. The probability of having zero heads out of 15 trials is computed using the following Excel formula, which gives a value of 0.0000305176.

   = BINOM.DIST(0,15,0.5,FALSE)

2. The probability of having five heads out of 15 trials is computed using the following Excel formula, which gives a value of 0.091644287.

   = BINOM.DIST(5,15,0.5,FALSE)

   3. The probability of having at most three heads = $\sum_{X=0}^{3} P(X,15,0.5)$

   It is computed using the following Excel formula, which gives a value of 0.017578125.

   = BINOM.DIST(3,15,0.5,TRUE))

   4. The probability of having at least five heads out of15 trials = $1-\sum_{X=0}^{4} P(X,15,0.5)$

   It is computed using the following formula in Excel, which gives a value of .940734863.

   = 1 – BINOM.DIST(4,15,0.5,TRUE))

Example 8.2

The head of a consultancy firm calculated that the likelihood of finishing each project on time is 0.7 based on prior experience. Twelve of these projects will be carried out by the business in the upcoming year. Find the following.

1. Probability of completing no project in time.
2. Probability of completing four projects in time.
3. Probability of completing at most two projects in time.
4. Probability of completing at least three projects in time.

Solution

The estimated probability of completing each project in time (p) = 0.7.

Number of projects to be executed in an year (n) = 12.

For each of the subsections of this problem, the result can be obtained by entering the respective formula in a convenient cell of an Excel sheet.

1. Probability(Completing no project in time) = $P(0,12, 0.7)$

   The Excel formula to compute this probability is shown as follows, which gives the value 5.31441 × 10^{-07}.

   = BINOM.DIST(0, 12, 0.7, FALSE)

2. Probability(Completing four projects in time) = $P(4,12, 0.7)$

   The Excel formula to compute this probability is shown as follows, which gives the value 0.007797716.

   = BINOM.DIST(4, 12, 0.7, FALSE)

3. $\text{Probability}\left(\text{Completing at most two projects in time}\right) = \sum_{X=0}^{2} P\left(X,12,0.7\right)$

   The Excel formula to compute this probability is shown as follows, which gives the value 0.000206376.

   = BINOM.DIST(2, 12, 0.7, TRUE)

4. $\text{Probability}\left(\text{Completing at least three projects in time}\right) = 1 - \sum_{X=0}^{2} P\left(X,12,0.7\right)$

   The Excel formula to compute this probability is shown as follows, which gives the value 0.000206376.

   $= 1 - \text{BINOM.DIST}(2, 12, 0.7, \text{TRUE}) = 0.000206376$

8.2.2 BINOM.DIST.RANGE Function

The BINOM.DIST.RANGE function gives the probability that the number of successful trials lies in a given range of trials (s1 and s2). The screenshot for the sequence of button clicks Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ BINOM.DIST.RANGE is shown in Figure 8.2.

The dropdown menu of Figure 8.2 contains the following data items.

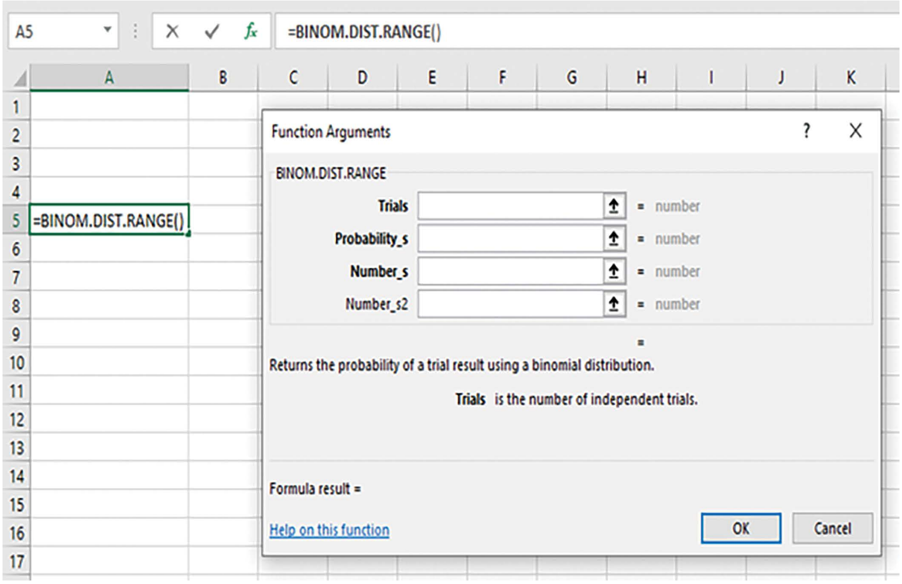
1. Trials, which is the total number of independent trials carried out in an experiment
2. Probability_s is the probability of occurrence of the successful event (*p*)
3. Number_s is the number of successes in trials (lower limit)
4. Number_s2 is the number of successes in trials (upper limit)

The formula for this functions as follows.

=BINOM.DIST.RANGE(Trials, Probability_s, Number_s, Number_s2)

Example 8.3

A leading R&D company carries out numerous research projects. In the upcoming three years, it intends to complete 15 R&D initiatives. The R&D initiatives' success



=BINOM.DIST.RANGE()

Function Arguments

BINOM.DIST.RANGE

Trials = number

Probability_s = number

Number_s = number

Number_s2 = number

=

Returns the probability of a trial result using a binomial distribution.

Trials is the number of independent trials.

Formula result =

Help on this function

Figure 8.2 Screenshot for sequence of button clicks, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ BINOM.DIST.RANGE

probability is 0.65. Find the likelihood that there will be between six and nine successful R&D projects.

Solution

The number of R&D projects to be carried out in the next three years (Trials) = 15
The probability of occurrence of successful R&D project (Probability_s) = 0.65
Lower limit for the number of successful R&D projects (Number_s) = 6
Upper limit for the number of successful projects (Number_s2) = 9
The formula for the probability that the number of successful R&D projects is between six and nine is as follows, which gives a value of 0.423275.

$$= \text{BINOM.DIST.RANGE}(15, 0.65, 6, 9)$$

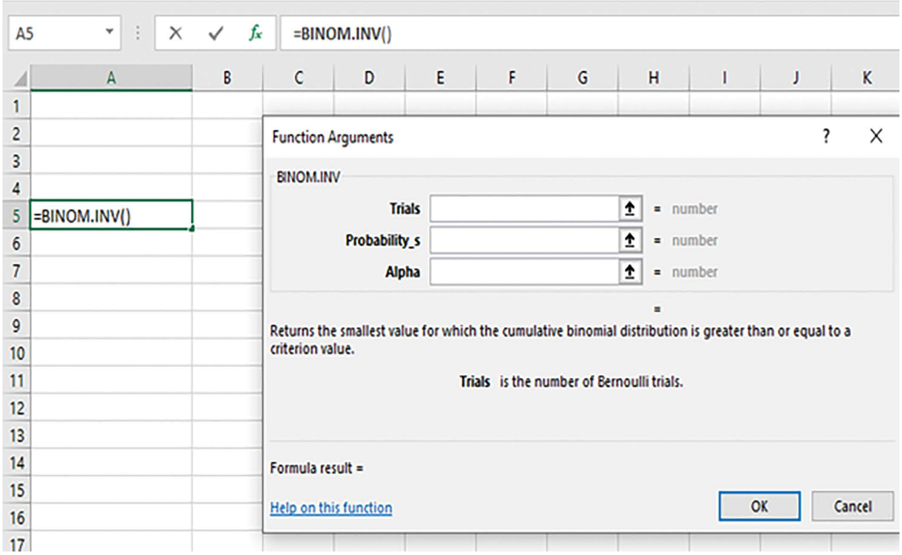
8.2.3 BINOM.INV Function

The BINOM.INV function [3] finds the value of the binomial random variable for the criterion that the binomial cumulative probability is more than a specified value.

The screenshot for the sequence of button clicks, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ BINOM.INV is shown in Figure 8.3.

The data items in the dropdown menu of Figure 8.3 are explained as follows.

- Trials is the total number of trials of the experiment
- Probability_s is the probability of success of the trials
- Alpha is a criterion value between 0 and 1, and the smallest value returned by the cumulative binomial distribution is greater than or equal to this criterion value



=BINOM.INV()

Returns the smallest value for which the cumulative binomial distribution is greater than or equal to a criterion value.

Trials is the number of Bernoulli trials.

Figure 8.3 Screenshot for button clicks, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ BINOM.INV

The formula of this function is as follows.

=BINOM.INV(Trials, Probability_s, Alpha)

Example 8.4

A company is working on a project to explore for oil and gas. It intends to drill in 20 distinct locations. The probability of this drilling operation producing oil or gas is 0.7. In this drilling exercise, determine the number of locations where oil or gas should be present so that the cumulative probability is 0.8.

Solution

The number of drilling places for oil and gas exploration (Trials) = 20

The probability of occurrence of oil/gas in drilling place (Probability_s) = 0.7

Criterion probability (Alpha) = 0.8

The formula to find the value of the binomial distribution such that the cumulative probability is more than 0.8 is given as follows, which gives a value of 16.

$$= \text{BINOM.INV}(20, 0.7, 0.8)$$

The number of places where the company should get oil/gas such that the cumulative probability is more than 0.8 is 16.

8.3 Poisson Distribution Using Poisson.Dist Function

The arrival rate of consumers at a service station, for instance, can be captured using the discrete Poisson distribution [4]. Consider the counter where customers check into purchase tickets for their travel at a transport corporation. The number of clients arriving at the booking counter per hour, or the arrival rate, is typically assumed to follow a Poisson distribution. The mean inter-arrival time of customers at the booking counter can be used to calculate the data on the arrival rate. The investigator should collect data on consecutive arrivals of customers. The difference between these two arrival times gives the inter-arrival time.

The arrival rate is represented by λ, and the inter-arrival time is represented by $1/\lambda$.

Let X be the random variable representing the specific rate of occurrence of an event. λ is the mean occurrence rate of the event.

Consider the arrival of airplanes in an airport. Assume that a study is conducted for 150 one-hour intervals. The number of arrivals of airplanes per hour and the corresponding frequencies of occurrences are as shown in Table 8.1.

The Poisson probability distribution is given by the following formula.

$$P(X) = \frac{\lambda^X e^{-\lambda}}{X!}$$

where
λ is the arrival rate of a specific occurrence of an event
X is the random variable representing the occurrence of specific event

A sample Poisson distribution is shown in Figure 8.4. This distribution is skewed towards the right.

The formulas for the mean and variance of the Poisson distribution are one and the same, which is:

$$Mean(\mu) = Variance(\sigma^2) = \lambda$$

Table 8.1 Arrival Rate of Airplanes and Their Frequencies

Arrival Rate of Airplanes (X_i) in One-Hour Intervals	2	3	4	5	6	7	8	9
Frequency (f_i)	18	25	30	35	20	10	7	5

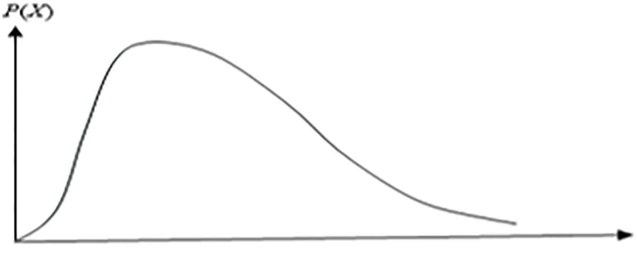
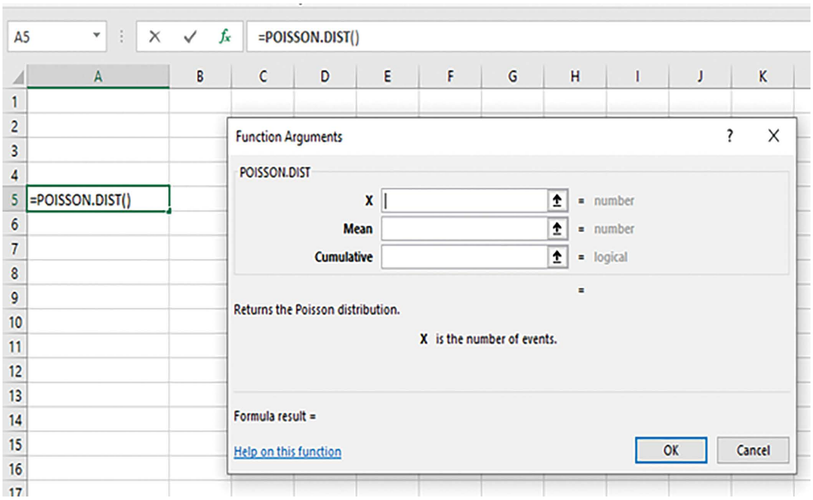


Figure 8.4 Screenshot of confidence interval after clicking the OK button in the dropdown menu of Figure 8.23



=POISSON.DIST()

Returns the Poisson distribution.

X is the number of events.

Figure 8.5 Screenshot after clicking sequence of buttons, Home ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ POISSON.DIST

The Excel template to compute the Poisson probability as well as the Poisson cumulative probability is shown in Figure 8.5. The sequence of mouse clicks to get the display shown in Figure 8.5 is as follows.

HOME ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ POISSON.DIST

The data items which are to be fed into the data boxes in the dropdown menu of Figure 8.5 are explained as follows.

- *X* is the number of events
- Mean is the expected numeric value of the Poisson distribution, a positive number
- Cumulative is a logical value; for the cumulative Poisson probability, use TRUE; for the Poisson probability mass function, use FALSE

The Excel formula to compute the cumulative Poisson distribution function is as follows.

$$= \text{POISSON.DIST}\left(\text{X,Mean, TRUE}\right)$$

The Excel formula to compute the Poisson probability mass function is as follows.

$$= \text{POISSON.DIST}(\left(\text{X, Mean, FALSE}\right)$$

Example 8.5

The arrival rate of vehicles at a petrol station follows Poisson distribution with a mean arrival rate of 10 per 15-minute interval. Find the

1. probability of no customers arriving in a 15-minute interval.
2. probability of exactly three customers arriving in a 15-minute interval.
3. probability of at most three customers arriving in a 15-minute interval.
4. probability of at least four customers arriving in a 15-minute interval.

Solution

The arrival rate of vehicles in 15 minutes interval at the petrol station follows a Poisson distribution.

The arrival rate of vehicles, λ in a 15-minute interval = 10 vehicles.

For each of the subsections of this problem, the result can be obtained by entering the respective formula in a convenient cell of an Excel sheet.

$$1.\ \text{Probability}\left(\text{No customer arrives in 15-minute interval}\right) = P\left(X,\lambda\right) = P\left(X = 0,\ \lambda = 10\right) = \frac{10^0 e^{-10}}{0!}$$

The Excel command to obtain the value of P(X = 0, 10) is as follows, which gives the probability as shown in the same line.

$$= \text{POISSON.DIST}\left(0, 10, \text{FALSE}\right) = 0.0000453999$$

$$2.\ \text{Probability}\left(\text{Exactly three customers arrive in 15-minute interval}\right) = P\left(X,\lambda\right) = P\left(X = 3,\ \lambda = 10\right)$$

The Excel command to obtain the value of $P(X$ = 3, 10) is as follows, which gives the probability as shown in the same line.

$$= \text{POISSON.DIST}\left(3, 10, \text{FALSE}\right) = 0.007566655$$

$$3.\ \text{Probability}\left(\text{At most three customers arrive in 15-minute interval}\right) = P\left(X \le 3,\lambda\right) = P\left(X \le 3,\ \lambda = 10\right)$$

The Excel command to obtain the value of $P(X$ ≤3, 10) is as follows, which gives the probability as shown in the same line.

$$= POISSON.DIST\left(3, 10, TRUE\right) = 0.01033605$$

$$d)\ \text{Probability}\left(\text{At least three customers arrive in 15-minute interval}\right) = 1 - P\left(X \le 2,\lambda\right)$$

$$= 1 - P\left(X \le 2,\ \lambda = 10\right)$$

The Excel command to obtain the value of 1 – $P(X$ ≤2, 10) is as follows, which gives the probability as shown in the same line.

$$= 1 - POISSON.DIST\left(2, 10, TRUE\right) = 0.997230604$$

8.4 Exponential Distribution Using Expon.Dist Function

In queuing theory, the service time spent on clients is represented by a continuous probability distribution called the exponential probability distribution. Actually, this is known as a negative exponential distribution.

The service rate is μ; hence the service time is1/μ.

Let the service time be τ, which is 1/μ.

A sketch of the exponential distribution is shown in Figure 8.6.

The formula of the exponential distribution in terms of service time is as follows.

$$P\left(X\right) = \frac{1}{\tau}e^{-X\tau},\ where\ X > 0\ and\ \tau > 0$$

The formula of the exponential distribution in terms of service rate is as follows.

$$P\left(X\right) = \mu e^{-\mu X},\ where\ X > 0\ and\ \mu > 0$$

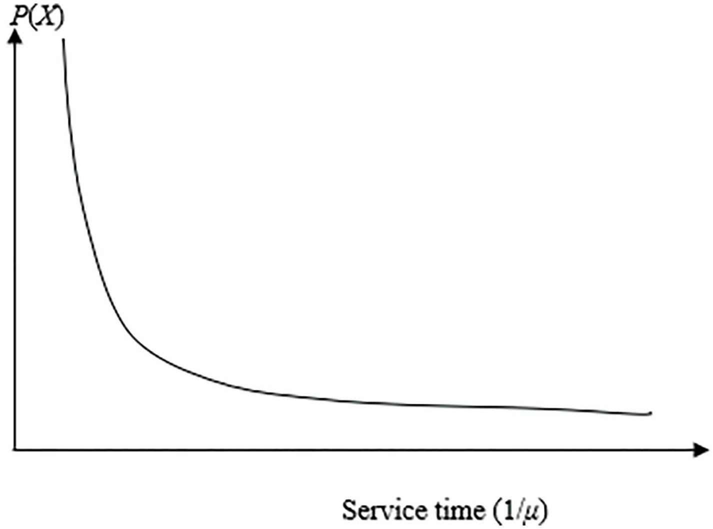


Figure 8.6 Sketch of exponential distribution

The cumulative density function of the exponential distribution is given as follows.

In terms of service time

$$F(X) = 1 - \mu e^{-X/\tau}, \textit{ where } X > 0, \tau > 0$$

In terms of service rate

$$F(X) = 1 - e^{-\mu X}, \textit{ where } X > 0, \mu > 0$$

The Excel template to compute the exponential probability and exponential cumulative probability is shown in Figure 8.7. The sequence of mouse clicks to get the display as shown in Figure 8.7 is as follows [5].

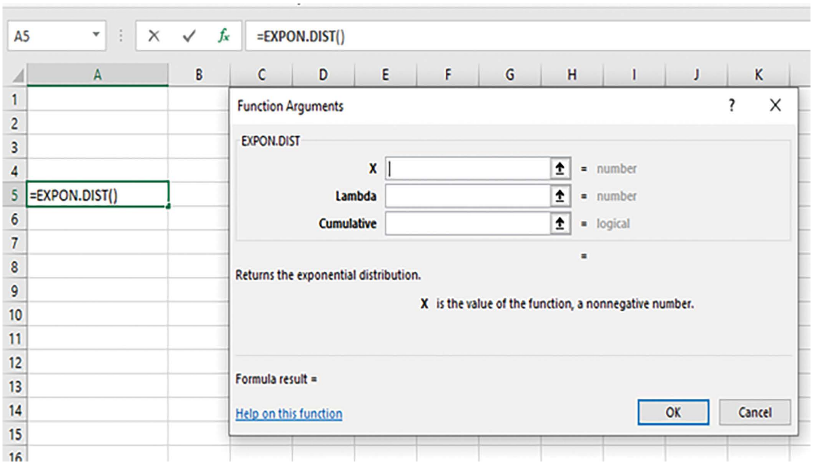
HOME ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ EXPON.DIST

The data items which are to be fed into the data boxes in the dropdown menu of Figure 8.7 are explained as follows.

- *X* is the value of the function, a non-negative number, which may be treated as a given service time
- Lambda is the parameter value, a positive number, which may treated as the mean service rate, µ (this is equal to $1/\tau$, where τ is the service time)
- Cumulative is a logical value for the function to return; the cumulative distribution function = TRUE; the probability density function = FALSE

The Excel formula to compute the cumulative Exponential distribution function is as follows.

$$= \text{EXPON.DIST}(\text{X}, \text{Lambda}, \text{TRUE})$$



=EXPON.DIST()

=EXPON.DIST()

X = number

Lambda = number

Cumulative = logical

Returns the exponential distribution.

X is the value of the function, a nonnegative number.

Formula result =

Figure 8.7 Screenshot after clicking sequence of buttons, Home $\Longrightarrow$ Formulas $\Longrightarrow$ More functions $\Longrightarrow$ Statistical $\Longrightarrow$ EXPON.DIST

The Excel formula to compute the Exponential probability density function is as follows.

$$= \text{EXPON.DIST}(\text{X}, \text{Lambda}, \text{FALSE})$$

Example 8.6

In a mainframe computer centre, execution times of programs follow exponential distribution. The average execution time ($1/\mu$) is 0.75 minute. Find the

1. Probability that the execution time is equal to 2 minute.
2. Probability that the execution time is less than 1.5 minute.
3. Probability that the execution time is more than 1.25 minute.
4. Probability that the execution time is between 1 and 1.75 minutes

Solution

The execution times of programs in a mainframe computer centre follow an exponential distribution.

The average execution time of programs (τ) = 0.75 minute.

$$\mu = \frac{1}{\tau} = \frac{1}{0.75}$$

For each of the subsections of this problem, the result can be obtained by entering the respective formula in a convenient cell of an Excel sheet.

$$1.P(Execution\ time = 2) = P\left(X = 2,\ \mu = \frac{1}{0.75}\right) = \frac{e^{-\frac{2}{0.75}}}{0.75}$$

The Excel formula to compute the cumulative Exponential distribution function is as follows.

$$= \text{EXPON.DIST}\left(2, \frac{1}{0.75}, \text{FALSE}\right)$$

The formula returns the probability of 0.092644602.

$$2.P\left(\textit{Execution time is less than } 1.5 \textit{ minutes} = F(X \le 1.5, \mu = \frac{1}{0.75}\right) = 1 - e^{-\frac{1.5}{0.75}}$$

The Excel formula to compute the cumulative Exponential distribution function is as follows.

$$= \text{EXPON.DIST}\left(1.5, \frac{1}{0.75}, \text{TRUE}\right)$$

The formula returns the probability of 0.864664717.

$$3.P(\textit{Execution time is more than } 1.25 \textit{ minutes}) = F\left(X \ge 1.25, \mu = \frac{1}{0.75}\right)$$

$$= 1 - F\left(X \le 1.25, \mu = \frac{1}{0.75}\right)$$

$$= 1 - \left(1 - e^{-\frac{X}{\tau}}\right) = e^{-\frac{1.25}{0.75}}$$

The Excel formula to compute the probability is as follows.

$$= \left(1 - \text{EXPON.DIST}\left(1.25, \frac{1}{0.75}, \text{TRUE}\right)\right)$$

The formula returns the probability of 0.188875603.

$$4.\ \text{P}(\text{Execution time is between1 and1.75 minutes}) = F\left(X \le 1.75,\ \mu = \frac{1}{0.75}\right) - F\left(X \le 1,\ \mu = \frac{1}{0.75}\right)$$

The Excel formula to compute the probability is as follows.

$$= \text{EXPON.DIST}\left(1.75, \frac{1}{0.75}, \text{TRUE}\right) - \text{EXPON.DIST}\left(1, \frac{1}{0.75}, \text{TRUE}\right)$$

The formula returns the probability of 0.16662517.

8.5 Normal Distribution Using NORM.DIST, NORM.INV, NORM.S.DIST, and NORM.S.INV Functions

Normal distribution is a continuous probability distribution. Most of real-life situations can be captured in the form of a normal distribution.

Some of the applications where data follow a normal distribution are as follows.

- Income of employees in organisations
- The marks of the students in a subject
- The internal diameter of bearings produced in a company
- Height of employees in an organisation
- Weight of employees in an organisation
- Sales turnover of companies in an industrial estate

A sketch of the normal distribution is shown in Figure 8.8. It is a symmetrical distribution, and the value of the random variable (*X*) varies from – ∞ to + ∞. The mean and variance of the normal distribution are μ and σ^2, respectively.

The formula for the normal probability density distribution is as follows.

$$P(X) = \frac{1}{\sigma \times \sqrt{2\pi}} \times e^{\left(-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right)}, \textit{ where} -\infty < X < +\infty$$

where
μ is the mean of the normal distribution
σ is the standard deviation of the normal distribution
X is the normal random variable

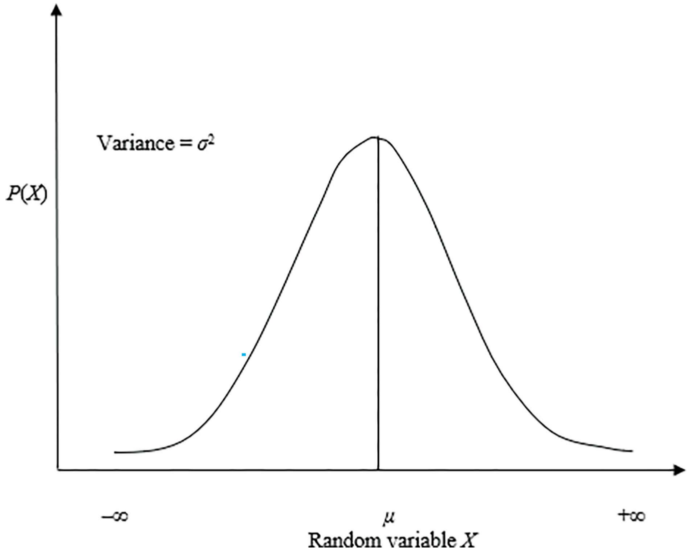


Figure 8.8 Normal distribution

It is very difficult to provide tables for different combinations of μ and σ^2, because the number of such combinations is infinite. Hence, another type of normal probability density function with a mean of 0 and variance of 1 is designed using the central limit theorem by replacing the normal random variable X with a standard normal variable Z as follows.

$$Z = \frac{X-\mu}{\sigma}$$

The formula for the standard normal probability density distribution, which has the mean and the variance as 0 and 1, respectively, is as follows.

$$P(Z) = \frac{1}{\sqrt{2\pi}} \times e^{\left(-\frac{Z^2}{2}\right)}, where - \infty < Z < +\infty$$

where
Z is the standard normal random variable

The formula for the cumulative normal density distribution of the normal distribution is as follows.

$$F(X) = \int_{-\infty}^{+X} \frac{1}{\sigma \times \sqrt{2\pi}} \times e^{\left(-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right)} dX, \ where - \infty < X < +X$$

where
μ is the mean of the normal distribution
σ is the standard deviation of the normal distribution
X is the normal random variable
$F(X)$ is the cumulative normal density function

Excel Commands for Normal Distribution

The Excel screenshot to compute the normal probability and normal cumulative probability is shown in Figure 8.9. The sequence of clicks to get the display shown in Figure 8.9 is as follows.

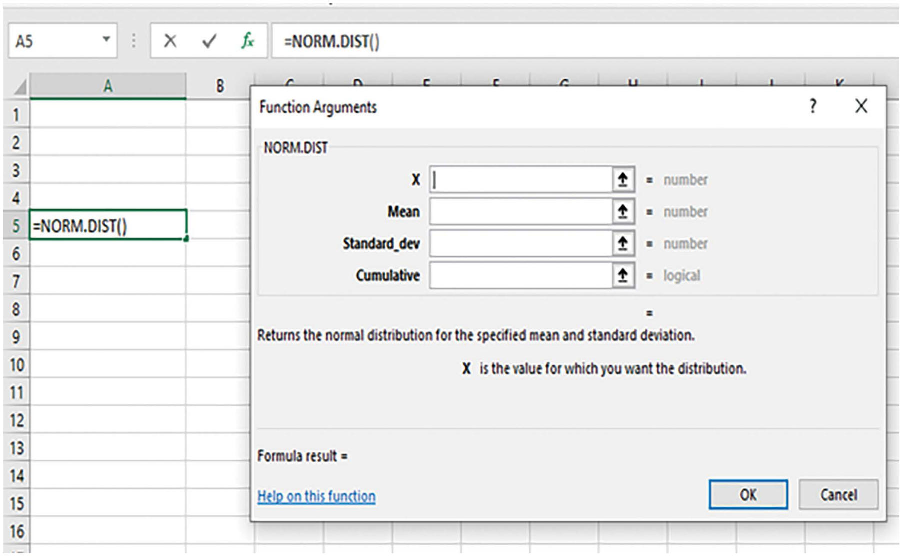
HOME ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ NORM.DIST

The data items which are to be fed into the data boxes in the dropdown menu of Figure 8.9 are explained as follows.

- X is the value for which the probability/cumulative probability of the distribution is required
- Mean is the arithmetic mean of the distribution
- Standard_dev is the standard deviation of the distribution, a positive number
- Cumulative is a logical value for the function to return; for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE

The Excel formula to compute the cumulative normal distribution function is as follows.

$$= \text{NORM.DIST}(\text{X, Mean, Standrad\_dev, TRUE})$$



=NORM.DIST()

Returns the normal distribution for the specified mean and standard deviation.

X is the value for which you want the distribution.

Figure 8.9 Template to compute exponential value

The Excel formula to compute the normal probability density function is as follows.

= NORM.DIST(X, Mean, Standrad_dev, FALSE)

Excel Commands to Get Value of Random Variable for a Given Cumulative Probability

An Excel screenshot to compute the value of a normal random variable for a given cumulative probability is shown in Figure 8.10. The sequence of clicks to get the display shown in Figure 8.10 is as follows.

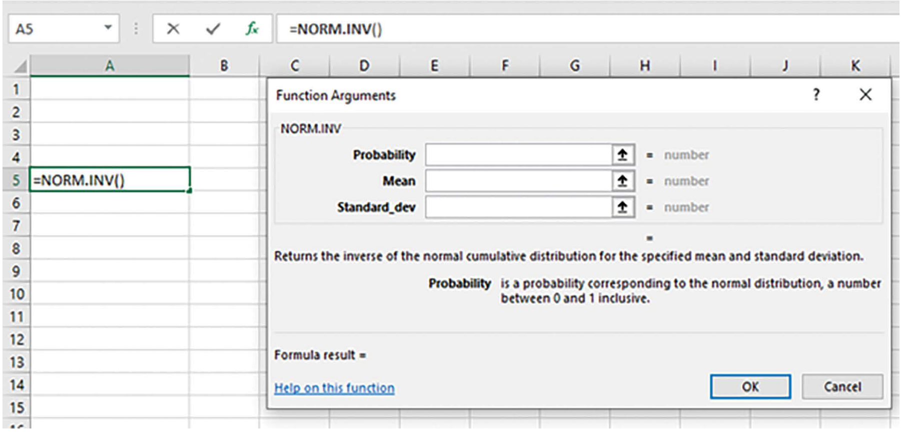
HOME $\Longrightarrow$ Formulas $\Longrightarrow$ More functions $\Longrightarrow$ Statistical $\Longrightarrow$ NORM.INV

The data items which are to be fed into the cells of the dropdown menu of Figure 8.10 are explained as follows.

- Probability is a cumulative probability corresponding to the normal distribution in the range from 0 to 1
- Mean is the arithmetic mean of the distribution
- Standard_dev is the standard deviation of the distribution, a positive number

The Excel formula to compute the value of the normal random variable is shown as follows.

= NORM.INV(Probability, Mean, Standard_dev)



Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.

Probability is a probability corresponding to the normal distribution, a number between 0 and 1 inclusive.

Figure 8.10 Excel screenshot to obtain the value of normal random variable for a given cumulative probability

Excel Commands for Standard Normal Distribution

An Excel screenshot to compute the standard normal probability and standard normal cumulative probability is shown in Figure 8.11. The sequence of clicks to get the display shown in Figure 8.11 is as follows [6].

HOME ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ NORM.S.DIST

The data items which are to be fed in to the cell of the dropdown menu of Figure 8.11 are explained as follows.

- Z is the value for which the cumulative probability from – ∞ to Z is to be found.
- Cumulative is a logical value for the function to return. For the cumulative distribution function, it is TRUE; for the probability density function, it is FALSE. *The second option of using FALSE has no relevance in practice.*

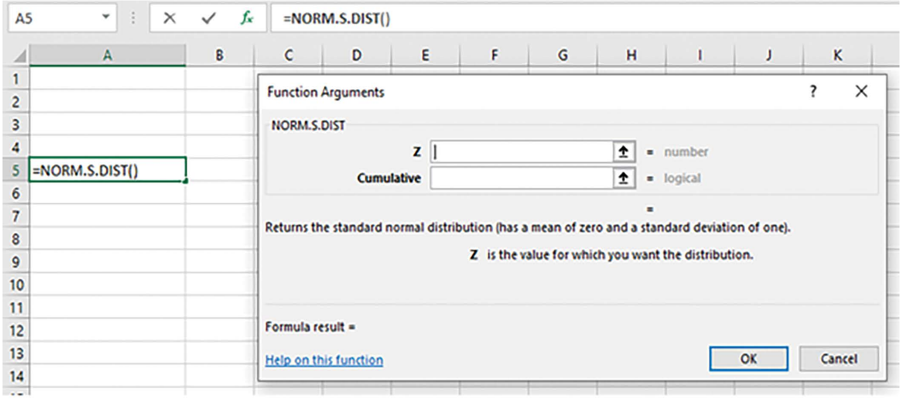
The Excel formula to compute the standard normal cumulative distribution function is as follows.

$$= \text{NORM.S.DIST}(\text{Z}, \text{TRUE})$$

Excel Commands to Get Value of Standard Normal Variable for a Given Cumulative Probability

An Excel screenshot to compute the standard normal value for a given cumulative probability is shown in Figure 8.12. The sequence of clicks to get the display shown in Figure 8.12 is as follows.

HOME ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ NORM.S.INV



=NORM.S.DIST()

Function Arguments

NORM.S.DIST

Z = number

Cumulative = logical

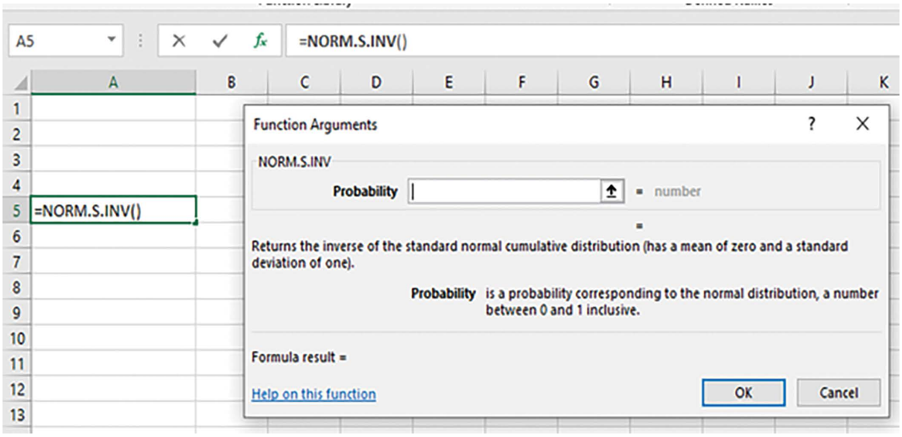
Returns the standard normal distribution (has a mean of zero and a standard deviation of one).

Z is the value for which you want the distribution.

Formula result =

Help on this function

Figure 8.11 Screenshot after issuing the sequence of clicks, Home ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ NORM.S.DIST



=NORM.S.INV()

Function Arguments

NORM.S.INV

Probability = number

Returns the inverse of the standard normal cumulative distribution (has a mean of zero and a standard deviation of one).

Probability is a probability corresponding to the normal distribution, a number between 0 and 1 inclusive.

Formula result =

Help on this function

Figure 8.12 Excel screenshot for mouse click sequence, Home ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ NORM.S.INV

The data item which is to be fed in to the cell of the dropdown menu of Figure 8.12 is explained as follows.

- Probability is a cumulative probability corresponding to the normal distribution in the range from 0 to 1.

The Excel formula to compute the standard normal statistic (Z) for a given cumulative probability is as follows.

= NORM.S.INV(Probability)

Example 8.7

The monthly income of the respondents in a survey with a sample of 300 people follows a normal distribution, with a mean and standard deviation of ₹ 15,000 and ₹ 3,000, respectively. Answer the following.

1. What is the probability that the monthly income is less than ₹ 12,000? Also, find the number of respondents who have income less than ₹ 12,000.
2. What is the probability that the monthly income is more than ₹ 16,000? Also, find the number of respondents who have income more than ₹ 16,000.
3. What is the probability that the monthly income is between ₹ 10,000 and ₹ 17.000? Also, find the number of respondents who have monthly income between ₹ 10,000 and ₹ 17,000.
4. Find the value of the monthly income (X) when the cumulative probability is 0.1587 and develop the corresponding constructs of probability.

   Construct means P(X≥ K) or P(X ≤ K), where K is the mirror replica on X axis for the X computed about mean.

5. Find the value of the monthly income (X) when the cumulative probability is 0.3694 and develop the corresponding constructs of probability.

Solution

The monthly income of the respondents follows a normal distribution with mean (μ) and standard deviation (σ) of ₹ 15,000 and ₹ 3,000, respectively.

That is, $X \sim N(\mu, \sigma^2)$
$\sim N(15{,}000, 3000^2)$

Sample size, $n = 300$

1. The Excel formula to compute P($X \leq 12000$) is shown as follows.

   $= \text{NORM.DIST}(12000, 15000, 3000, \text{TRUE})$
   The cumulative probability of the previous formula is 0.158655254.
   The number of respondents whose income is less than or equal to ₹ 12,000 $= 0.158655254 \times n$
   $= 0.158655254 \times 300 = 47.5966 \approx 48$

2. The formula to compute P($X \geq 16000$) is 1 – P($X \leq 16000$), and the Excel formula to compute the same is shown as follows.

   $= 1 - \text{NORM.DIST}(16000, 15000, 3000, \text{TRUE})$
   The cumulative probability of the previous formula is 0.36944134.
   The number of respondents whose income is more than or equal to ₹ 16,000 $= 0.36944134 \times n$
   $= 0.36944134 \times 300$
   $= 110.83 \approx 111$

3. The formula to compute $P(10000 \leq X \leq 17000)$ is $P(X \leq 17000) - P(X \leq 10000)$ and the Excel formula to compute this value is shown as follows.
   = NORM.DIST(17000, 15000, 3000, TRUE) – NORM.DIST(10000, 15000, 3000, TRUE)
   = 0.69971711
   The number of respondents whose income is between ₹ 10000 and ₹ 17000
   $= 0.6997171 \times n$
   $= 0.6997171 \times 300$
   $= 209.92 \approx 210$

4. Computation of the value of X when the cumulative probability is 0.1587.

   The Excel formula to find the value of X when the cumulative probability is 0.1587 is shown as follows.
   = NORM.INV(0.1587, 15000, 3000)
   The value of the X of the normal distribution per the previous formula is 12000.554729 ≈ 12001.
   The constructs of probability may be each of the following.
   $P(X \leq 12001)$
   $P(X \geq 17999)$

5. Find the value of X when the cumulative probability is 0.3694.

   The Excel formula to find the value of X when the cumulative probability is 0.3694 is shown as follows.
   = NORM.INV(0.3694, 15000, 3000)
   The value of the X of the normal distribution per the formula is 13999.67136 ≈ 14000.
   The constructs of probability may be each of the following.
   $P(X \leq 14000)$
   $P(X \geq 16000)$

Example 8.8

The daily number of patients treated in a survey with a sample of 200 private clinics follows a normal distribution, with a mean and standard deviation of 75 and 15, respectively. Answer the following.

1. What is the probability that the daily number of patients treated is less than 80? Also, find the number of clinics with a daily number of patients treated less than 80.
2. What is the probability that the daily number of patients treated is more than 60? Also, find the number of clinics with a daily number of patients treated more than 60.
3. What is the probability that the daily number of patients treated is less than 60? Also, find the number of clinics with a daily number of patients treated less than 60.
4. What is the probability that the daily number of patients treated is between 65 and 85? Also, find the number of clinics with a daily number of patients treated between 65 and 85.
5. Compute the value of X when the cumulative probability is 0.6696.
6. Compute the value of X when the cumulative probability is 0.158655
7. Compute X_1 and X_2 when the corresponding probabilities are 0.251429 and 0.748571.

Solution

The daily number of patients treated follows normal distribution, with a mean and standard deviation of 75 and 15, respectively.

That is, $X \sim N(\mu, \sigma^2)$
$\sim N(75, 15^2)$
Sample size, $n = 200$

1. $P(X \leq 80) = P\left(\frac{X-\mu}{\sigma} \leq \frac{80-\mu}{\sigma}\right) = P\left(Z \leq \frac{80-75}{15}\right) = P(Z \leq 0.3333333)$

   The Excel formula to compute the previous probability is shown as follows.

   $= \text{NORM.S.DIST}(0.3333333, \text{TRUE})$

   The value of the cumulative probability is 0.630558647.
   Number of clinics in this group $= 0.630558647 \times 200 = 126.111 \approx 126$

2. $P(X \geq 60) = P\left(\frac{X-\mu}{\sigma} \geq \frac{60-\mu}{\sigma}\right) = P\left(Z \geq \frac{60-75}{15}\right) = P(Z \geq -1)$

   $= 1 - P(Z \leq -1) = 1 - \left[1 - P(Z \leq 1)\right] = P(Z \leq 1)$

   The Excel formula to compute the probability for $P(Z \leq 1]$ is shown as follows.

   $= \text{NORM.S.DIST}(1, \text{TRUE})$

   The value of the probability for the previous formula is 0.841344746.

   Number of clinics in this group $= 0.841344746 \times 200 = 168.269 \approx 168$

3. $$P(X \leq 60) = P\left(\frac{X-\mu}{\sigma} \leq \frac{60-\mu}{\sigma}\right) = P\left(Z \leq \frac{60-75}{15}\right)$$
   $$= P(Z \leq -1) = 1 - P(Z \leq 1)$$

   The Excel formula to compute $1 - P[Z \leq 1]$ is shown as follows.

   $= (1 - \text{NORM.S.DIST}(1, \text{TRUE}))$

   The probability of the previous formula is 0.158655254.
   Number of clinics in this group $= 0.158655254 \times 200 = 31.73 \approx 32$

4.

$$P(65 \leq X \leq 85) = P\left(\frac{65-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{85-\mu}{\sigma}\right) = P\left(\frac{X-\mu}{\sigma} \leq \frac{85-\mu}{\sigma}\right) - P\left(\frac{X-\mu}{\sigma} \leq \frac{65-\mu}{\sigma}\right)$$

$$= P\left(Z \leq \frac{85-75}{15}\right) - P\left(Z \leq \frac{65-75}{15}\right) = P(Z \leq 0.67) - P(Z \leq -0.67)$$

$= P(Z \le 0.67) - [1 - P(Z \le 0.67)] = -1 + 2 \times P(Z \le 0.67)$

The Excel formula to compute the previous probability is shown as follows.

$= (-1 + 2 * \text{NORM.S.DIST}(0.67, \text{TRUE}))$

The computed probability of the previous formula is 0.49714221.
Number of clinics in this group $= 0.49714221 \times 200 = 99.43 \approx 99$

5. The given cumulative probability is 0.6696.

The Excel command to get the Z value for the cumulative probability 0.6696 is as follows.

$= \text{NORM.S.INV}(0.6696)$

The value of Z is 0.438808915.

$$\frac{X-\mu}{\sigma} = \frac{X-75}{15} = 0.438808915$$

$X = 0.438808915 \times 15 + 75 = 81.58213373 \approx 82 \textit{ clinics}$

6. The cumulative probability is 0.158655.

The Excel command to get Z value for the cumulative probability 0.158655 is as follows.

$= \text{NORM.S.INV}(0.158655)$

The Z value for the previous formula is –1.000001049 = –1.

$$\frac{X-\mu}{\sigma} = \frac{X-75}{15} = -1$$

$X = -1 \times 15 + 75 = 60 \textit{clinics}$

7. The cumulative probability with respect to $X1$ is 0.251429.

The Excel command to get Z value for the cumulative probability 0.251429 is as follows.

$= \text{NORM.S.INV}(0.251429)$

The Z value for the previous formula is – 0.669999671.

$$\frac{X-\mu}{\sigma} = \frac{X_1-75}{15} = -0.669999671$$

$X_1 = -0.669999671 \times 15 + 75 = 64.95 \approx 65 \textit{ clinics}$

The cumulative probability with respect to X_2 is 0.748571.

The Excel command to get the Z value for the cumulative probability 0.748571 is as follows.

$$= \text{NORM.S.INV}(0.748571)$$

The Z value for the previous formula is 0.669999671.

$$\frac{X_2 - \mu}{\sigma} = \frac{X_2 - 75}{15} = 0.669999671$$

$$X_2 = 0.669999671 \times 15 + 75 = 85.04999506 \approx 85 \textit{ clinics}$$

8.6 Uniform Distribution Using Excel Sheets

The uniform distribution is a continuous distribution that, in an interval with a lower limit and an upper limit of a and b, respectively, has an equal probability of occurring for each value of the random variable. Figure 8.13 displays a plot of the uniform distribution.

The formula for the probability density function of the uniform distribution is as follows.

$$P(X) = \frac{1}{b-a}, \qquad a \leq X \leq b$$
$$= 0, \textit{ otherwsie}$$

where

a is the lower limit of the uniform random variable
b is the upper limit of the uniform random variable
X is the uniform random variable

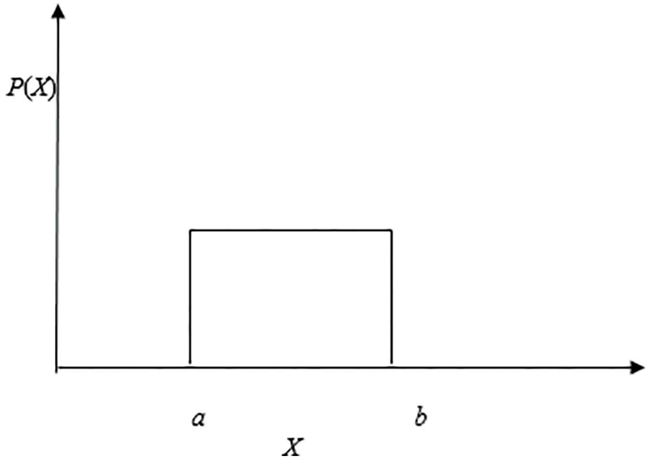


Figure 8.13 Plot of uniform distribution

$P(X)$ is the probability density function of uniform distribution
The cumulative density function of the uniform distribution is as follows.

$$F(X)=\int_a^X \frac{1}{(b-a)}dX, \textit{ where } a \le X \le b = \frac{X-a}{b-a}$$

where
a is the lower limit of the uniform random variable
b is the upper limit of the uniform random variable
X is the uniform random variable
$F(X)$ is the cumulative density function of uniform distribution

The formulas for the mean (μ) and the variance (σ^2) of the uniform distribution are given as follows.

$$\mu=\frac{a+b}{2}$$

$$\sigma^2=\frac{(b-a)^2}{12}$$

Example 8.9

The demand of a product follows uniform distribution, and its distribution is as follows.

$$P(X)=\frac{1}{5000-2000}, \textit{ where } 2000 \le X \le 5000$$

$$= 0, \textit{ otherwise}$$

1. If the maximum demand is 4,500, find the service level of the company in satisfying the demand of the product.

2. If the service level is 0.9, then find the maximum demand which can be satisfied.

Solution

The uniform distribution of the product is as follows.

$$P(X)=\frac{1}{5000-2000}, \textit{ where } 2000 \le X \le 5000$$

$$= 0, \textit{ otherwise}$$

1. If the maximum demand is 4,500, then the corresponding service level is computed as follows.

$$F(X) = \int_{a}^{4500} \frac{1}{(b-a)} dX, \textit{ where } a(2000) \le X \le b(5000)$$

$$= \frac{X-a}{b-a} = \frac{4500-2000}{5000-2000} = 0.8333 = 83.33\%$$

2. If the service level is 0.9, the maximum demand that can be satisfied is computed as follows.

$$\frac{X-a}{b-a} = \frac{X-2000}{5000-2000} = 0.9$$

$$X = 0.9 \times (5000-2000) + 2000 = 4700 \textit{ units}$$

Though there is no formula in Excel for the uniform distribution, one can use Excel sheet as a calculator to calculate the results of parts 1 and 2.

8.7 *t*-Distribution (Sampling Distribution of Mean When Normal Population Variance Is Unknown) USING T.DIST, T.DIST.2T, T.DIST.RT, T.INV, and T.INV.2T Functions

If the variance of the normal population is unknown, then the sampling distribution of the mean is called a *t*-distribution. Assume that a measurement (X) of a population follows a Gaussian distribution. Then the measurement X of a sample drawn from this population will follow a normal distribution with mean μ and variance σ^2. That is, $X \sim N(\mu, \sigma^2)$.

Let the sample mean be $\bar{X}$, which is given by the following formula.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where Xi is the i^{th} observation of the sample, i = 1, 2, 3, n
n is the sample size

By linearity, the sample mean follows a normal distribution with mean 0 and variance 1, which is as follows.

$$\bar{X} \sim \mu + \frac{\sigma}{\sqrt{n}} N(0,1)$$

Hence, $\dfrac{\bar{X}-\mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$ follows $N(0,1)$, which is known as the student *t*-distribution with (n – 1) degrees of freedom.

The shape of the *t*-distribution is shown in Figure 8.14.

Excel Formula to Find Cumulative Probability for a Given Value of Random Variable of *t* Distribution

The Excel screenshot to find the cumulative probability of the random variable of the *t*-distribution is shown in Figure 8.15. The sequence of button clicks to get the display shown in Figure 8.15 is as follows.

HOME ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ T.DIST

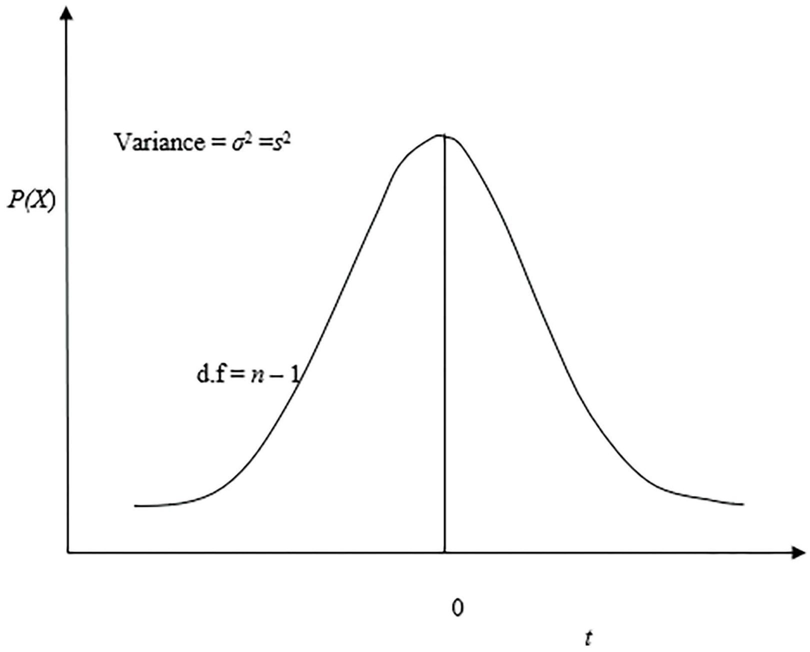
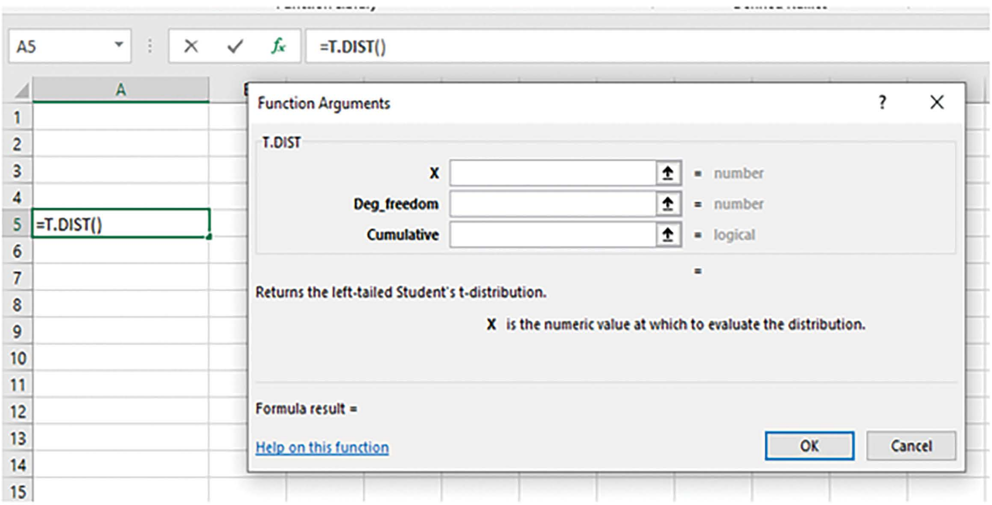


Figure 8.14 Sketch of *t*-distribution


A5

=T.DIST()

5 =T.DIST()

Function Arguments

T.DIST

X = number

Deg_freedom = number

Cumulative = logical

=

Returns the left-tailed Student's t-distribution.

X is the numeric value at which to evaluate the distribution.

Formula result =

Help on this function

OK Cancel

Figure 8.15 Screenshot of response to mouse clicks, Home ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ T.DIST

The data items which are to be fed into the data boxes in the dropdown menu of Figure 8.15 are explained as follows.

- *X* is the numeric value at which to evaluate the distribution
- Deg_freedom is an integer indicating the number of degrees of freedom that characterises the distribution
- Cumulative is a logical value, which can be TRUE/FALSE; for cumulative distribution function, use TRUE, and for the probability density function, use FALSE.

The Excel formula to compute the cumulative probability of the *t* distribution is as follows.

=T.DIST(X, Deg_freedom, TRUE)

Excel Formulas to Get the Value of a Random Variable for a Given Probability

The Excel template to find the value of a random variable for a given probability of the *t*-distribution is shown in Figure 8.16.The sequence of button clicks to get the display shown in Figure 8.16 is as follows.

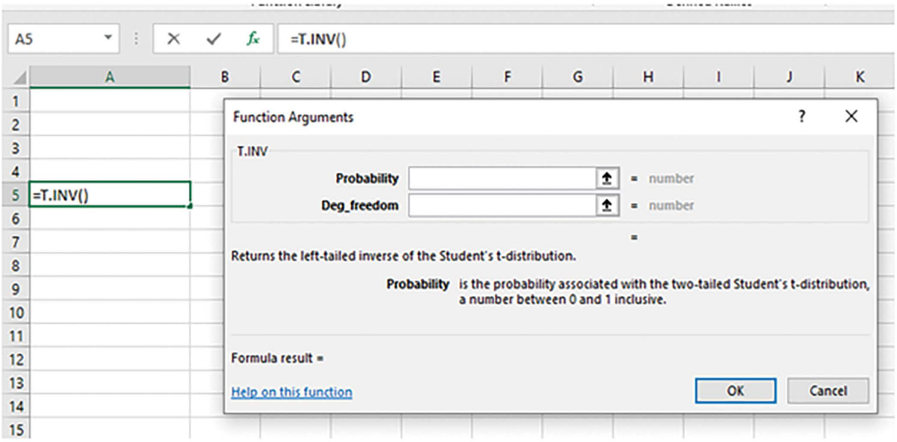
HOME ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ T.INV

The data items which are to be fed into the data boxes in Figure 8.16 are explained as follows.

- Probability is the probability associated with a two-tailed *t*-distribution, a number between 0 and 1 inclusive
- Deg_freedom is an integer indicating the number of degrees of freedom that characterises the distribution

The Excel formula to compute the value of a random variable of the *t*-distribution is as follows.

=T.INV(Probability, Deg_freedom)



A5

=T.INV()

=T.INV()

Function Arguments

T.INV

Probability = number

Deg_freedom = number

=

Returns the left-tailed inverse of the Student's t-distribution.

Probability is the probability associated with the two-tailed Student's t-distribution, a number between 0 and 1 inclusive.

Formula result =

Figure 8.16 Screenshot in response to mouse clicks, Home ⟹ Formulas ⟹ More functions ⟹ Statistical ⟹ T.INV

Example 8.10

From a normal population, 15 dealers of a particular company are chosen at random. The population's annual sales have a mean of ₹ 50 lakhs and a variance of ₹ 100 lakhs.

1. Find the probability that the mean annual sales of the sample is less than ₹ 55 lakhs
2. Find the probability that the mean annual sales of the sample is more than ₹ 55 lakhs.
3. Find the sales if the p value (significance value) is 0.04.

Solution

Size of the random sample, $n = 15$

μ = ₹ 50 lakh & $\sigma^2 = s^2$ = ₹ 100 lakh,
σ = ₹ 10 lakh

$$\frac{s}{\sqrt{n}} = ₹\frac{10}{\sqrt{15}}\ lakh$$

$$t = \frac{\bar{X}-\mu}{\left(\frac{s}{\sqrt{n}}\right)} with(15-1)d.f$$

$$1.P\left(\bar{X} \le 55\right) = P\left\{\frac{\bar{X}-\mu}{\left(\frac{s}{\sqrt{n}}\right)} \le \frac{55-50}{\left(\frac{10}{\sqrt{15}}\right)}\right\} = P(t \le 1.936492)$$

The Excel formula to find the cumulative probability when t is less than or equal to 1.936492 is as follows.

= T.DIST(1.936492, 14, TRUE)

The probability of the previous formula is 0.963371.

$$2.\ P\left(\bar{X} \ge 55\right) = P\left\{\frac{\bar{X}-\mu}{\left(\frac{s}{\sqrt{n}}\right)} \ge \frac{55-50}{\left(\frac{10}{\sqrt{15}}\right)}\right\} = P(t \ge 1.936492) = 1 - P(t \le 1.936492)$$

The Excel formula to find the cumulative probability when t is more than or equal to 1.936492 is as follows.

= 1 – T.DIST(1.936492,14,TRUE)

The probability of the previous formula is 0.036629.

3. Computation of sales if p is 0.04.
p value is 0.04.

Cumulative probability = $1 - p = 1 - 0.04 = 0.96$

The Excel formula to obtain the value of t when $(1 - p)$ is 0.96 and degrees of freedom is 14 is given as follows.

$$= \text{T.INV}(0.96, 14)$$

The t value of this formula is 1.887496.

$$t = \frac{\bar{X} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} = 1.887496$$

$$\bar{X} = 1.887496 \times \frac{s}{\sqrt{n}} + \mu = 1.887496 \times \frac{10}{\sqrt{15}} + 50 = 54.87349 \textit{ lakh}$$

The sales when the value of p is 0.04 are ₹ 54.87349 lakh.

8.8 Confidence Interval When Sample Size Is Large

In a test of a hypothesis, the investigator can use any one of the following.

1. Comparing the computed statistic with the corresponding table value of the assumed distribution or computing p for a given statistic and comparing it with the given significance level (α).
2. Checking whether the computed statistic is within an estimated confidence interval, which is the range of values of a statistic. If the computed statistic falls within this interval, accept H_0; otherwise, reject H_0.

The earlier sections were oriented towards the first option. If the second option is to be followed, the investigator should know the method of estimating the confidence interval for a given problem. In this section, the determination of the confidence interval when the sample size is large is presented.

If the sample size is more than 30, then that sample is regarded as a large sample, and the corresponding distribution for X of the population is assumed to follow normal distribution with a mean of μ and a variance of σ^2.

The sampling distribution of this reality for $\bar{X}$ follows normal distribution with a mean of μ and a variance of σ^2/n, where n is the sample size.

The confidence interval with rejection regions at both tails is shown in Figure 8.17.

$$Z_{\bar{X}_{\alpha/2}} \text{ statistic at the right tail} = \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

$$-Z_{\bar{X}_{\alpha/2}} \text{ statistic at the right tail} = -\frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

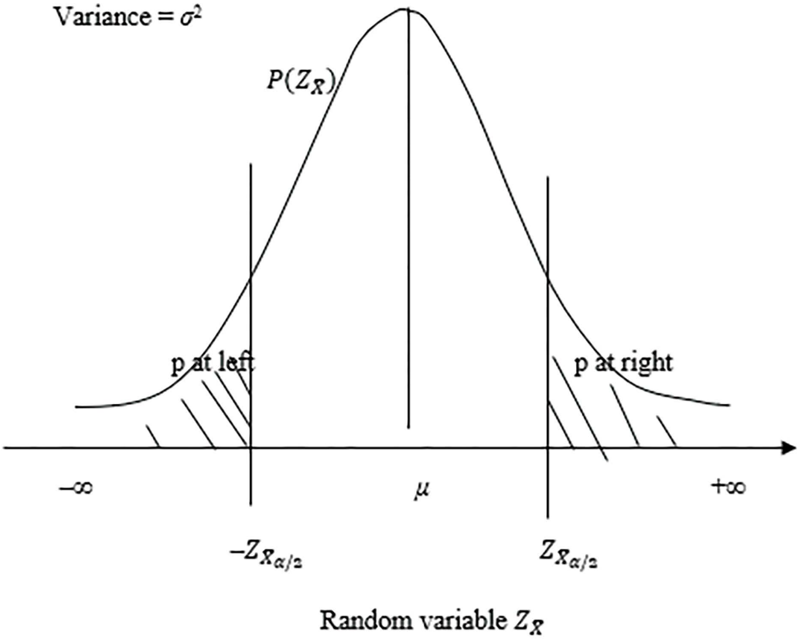


Figure 8.17 Confidence interval with rejection regions at both tails

Solving for μ in these two equations gives the confidence interval as follows, which is the confidence interval of the corresponding population.

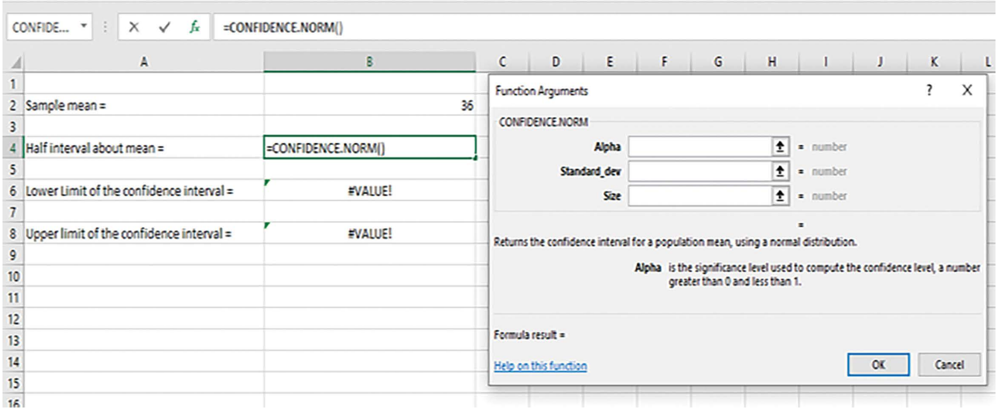
$$\mu = \bar{X} \pm Z_{\bar{X}_{\alpha/2}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Consider an example in which a company manufactures spindles in a lathe, which follows a normal distribution with a variance of 121 mm.

A random sample of 49 spindles has been taken, and the mean diameter of these spindles is 36 mm. Find the confidence interval of the diameter of a spindle that is manufactured using the lathe using a significance level of 0.05. The standard deviation of the population is 11 mm.

In Excel, the function for the confidence interval can be established. Click the sequence of buttons Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ CONFIDENCE.NORM, which gives the screenshot in Figure 8.18. Now, fill the significance level of 0.05 in the box against *Alpha*; the standard deviation of the population, which is 11, in the box against *Standard_dev*; and the sample size in the box against *Size* with 49 in the dropdown menu of Figure 8.18, as shown in Figure 8.19. Clicking the OK button in the dropdown menu of Figure 8.19 gives the screenshot in Figure 8.20 giving the half interval about the mean in cell E4 and then introducing the following formulas to compute the lower limit and upper limit in cells E6 and E8, respectively.

Formula to compute lower limit at cell E6:	= E2 + E4
Formula to compute upper limit at cell E8:	= E2 – E4



CONFIDE... =CONFIDENCE.NORM()

	A	B
1		
2	Sample mean =	36
3		
4	Half interval about mean =	=CONFIDENCE.NORM()
5		
6	Lower Limit of the confidence interval =	#VALUE!
7		
8	Upper limit of the confidence interval =	#VALUE!

Function Arguments

CONFIDENCE.NORM

Alpha = number

Standard_dev = number

Size = number

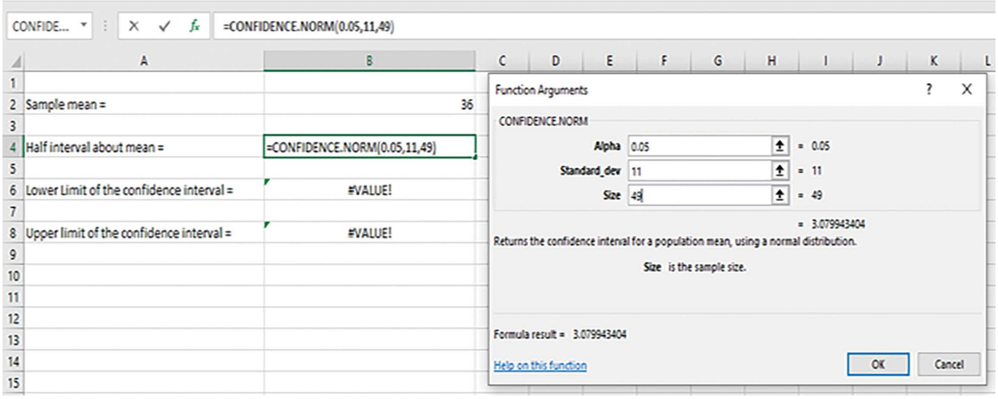
=

Returns the confidence interval for a population mean, using a normal distribution.

Alpha is the significance level used to compute the confidence level, a number greater than 0 and less than 1.

Formula result =

Figure 8.18 Screenshot for sequence of buttons, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ CONFIDENCE.NORM



CONFIDE... =CONFIDENCE.NORM(0.05,11,49)

	A	B
1		
2	Sample mean =	36
3		
4	Half interval about mean =	=CONFIDENCE.NORM(0.05,11,49)
5		
6	Lower Limit of the confidence interval =	#VALUE!
7		
8	Upper limit of the confidence interval =	#VALUE!

Function Arguments

CONFIDENCE.NORM

Alpha 0.05 = 0.05

Standard_dev 11 = 11

Size 49 = 49

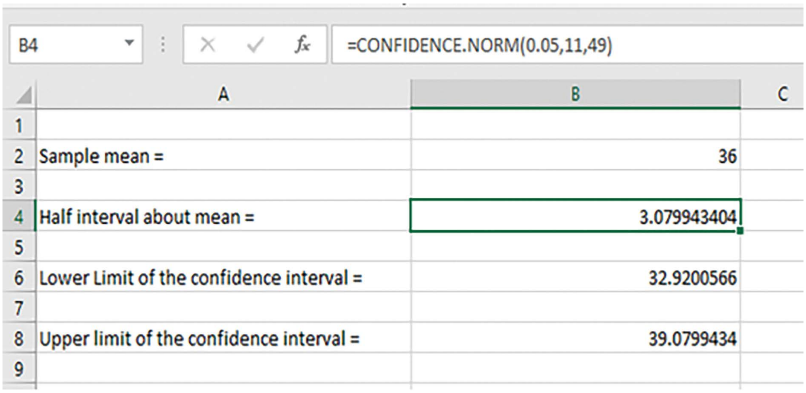
= 3.079943404

Returns the confidence interval for a population mean, using a normal distribution.

Size is the sample size.

Formula result = 3.079943404

Figure 8.19 Screenshot after filling data in the dropdown menu of Figure 8.18



B4 =CONFIDENCE.NORM(0.05,11,49)

	A	B
1		
2	Sample mean =	36
3		
4	Half interval about mean =	3.079943404
5		
6	Lower Limit of the confidence interval =	32.9200566
7		
8	Upper limit of the confidence interval =	39.0799434
9		

Figure 8.20 Screenshot of confidence interval after clicking the OK button in the dropdown menu of Figure 8.19

The lower limit and upper limit of the diameter of the spindle are 32.9200566 mm and 39.0799434 mm, respectively.

8.9 Confidence Interval When Sample Size Is Small

If the sample size is less than or equal to 30, then that sample is regarded as a small sample, and the corresponding distribution for X of the population with a mean of μ and variance of σ^2 is assumed to follow a t distribution. Here, the sample variance is used in establishing the confidence interval, since the population variance is unknown.

The sampling distribution of this reality for X follows a t distribution with a mean of $\bar{X}$ and a variance of S^2/n, where n is the sample size.

The confidence interval with rejection regions at both tails is shown in Figure 8.21.

$$t_{\alpha/2} \textit{ statistic at the right tail} = \frac{\bar{X} - \mu}{\left(\dfrac{S}{\sqrt{n}}\right)}$$

$$-t_{\alpha/2} \textit{ statistic at the right tail} = -\frac{\bar{X} - \mu}{\left(\dfrac{S}{\sqrt{n}}\right)}$$

Solving for μ in these two equations gives the confidence interval as follows, which is the confidence interval of the corresponding population.

$$\mu = \bar{X} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

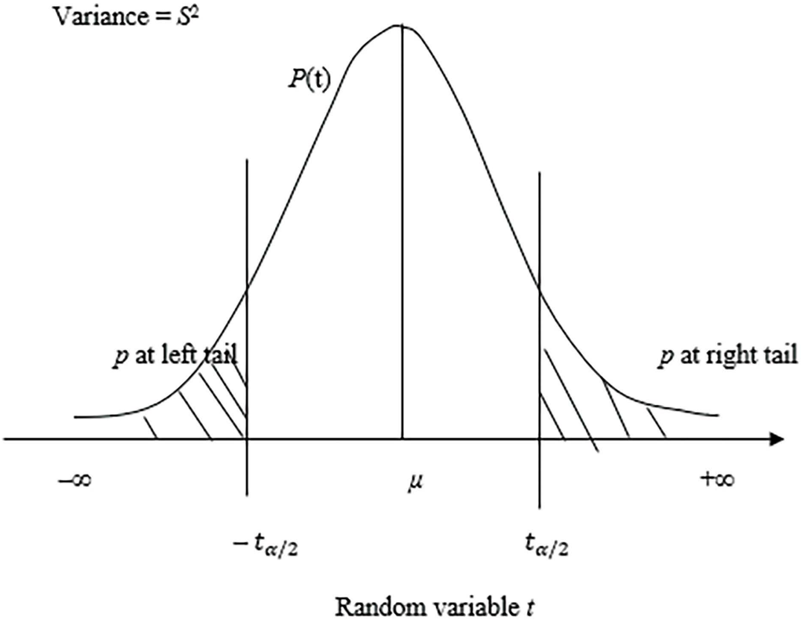


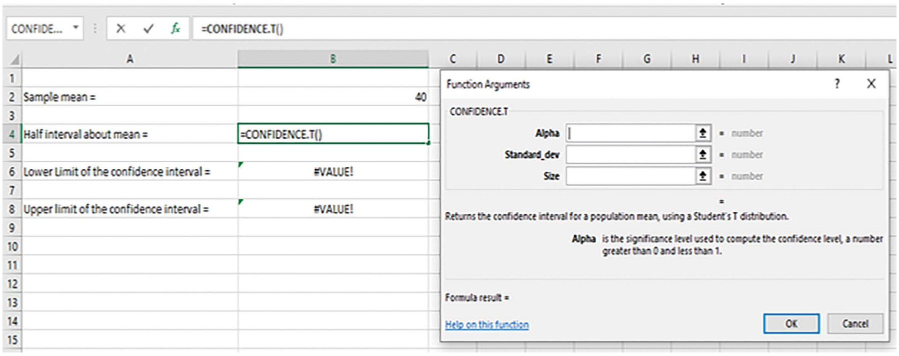
Figure 8.21 Confidence interval with rejection regions at both tails

Consider an example in which a company manufactures spindles in a lathe, which follows a normal distribution whose variance is unknown.

A random sample of 25 spindles has been taken, and the mean diameter of these spindles is 40 mm with a variance of 144 mm. Find the confidence interval of the diameter of a spindle that is manufactured using the lathe using a significance level of 0.05. The standard deviation of the sample is 12 mm.

In Excel, the function for the confidence interval can be established. Click the sequence of buttons Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ CONFIDENCE.T, which gives the screenshot in Figure 8.22. Now, fill the significance level of 0.05 in the box against Alpha; the standard deviation of the population, which is 12, in the box against Standard_dev; and the sample size in the box against size with 25 in the dropdown menu of Figure 8.22, as shown in Figure 8.23. Clicking the OK button in the dropdown menu of Figure 8.23 gives the screenshot in Figure 8.24, giving the half interval about the mean in cell E4 and then introducing the following formulas to compute the lower limit and upper limit in cells E6 and E8, respectively.

Formula to compute lower limit at cell E6:	= E2 + E4
Formula to compute upper limit at cell E8:	= E2 – E4



CONFIDE... =CONFIDENCE.T()

	A	B
1		
2	Sample mean =	40
3		
4	Half interval about mean =	=CONFIDENCE.T()
5		
6	Lower Limit of the confidence interval =	#VALUE!
7		
8	Upper limit of the confidence interval =	#VALUE!

Function Arguments

CONFIDENCE.T

Alpha = number

Standard_dev = number

Size = number

=

Returns the confidence interval for a population mean, using a Student's T distribution.

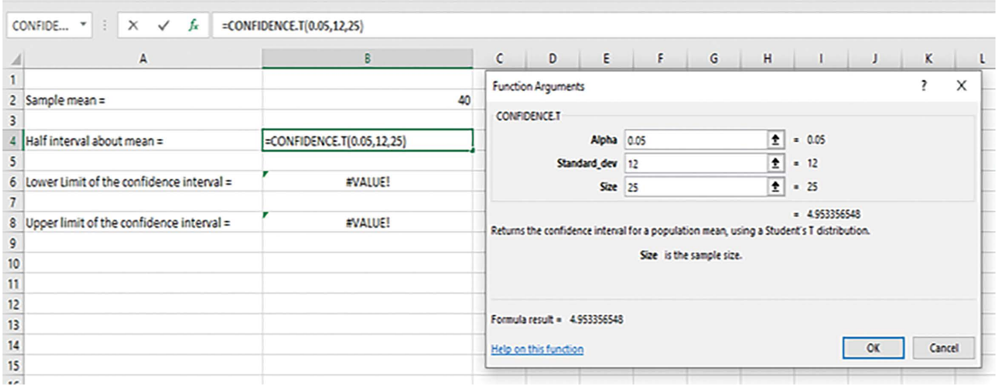
Alpha is the significance level used to compute the confidence level, a number greater than 0 and less than 1.

Formula result =

Help on this function

OK Cancel

Figure 8.22 Screenshot for sequence of buttons, Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ CONFIDENCE.T



CONFIDE... =CONFIDENCE.T(0.05,12,25)

	A	B
1		
2	Sample mean =	40
3		
4	Half interval about mean =	=CONFIDENCE.T(0.05,12,25)
5		
6	Lower Limit of the confidence interval =	#VALUE!
7		
8	Upper limit of the confidence interval =	#VALUE!

Function Arguments

CONFIDENCE.T

Alpha 0.05 = 0.05

Standard_dev 12 = 12

Size 25 = 25

= 4.953356548

Returns the confidence interval for a population mean, using a Student's T distribution.

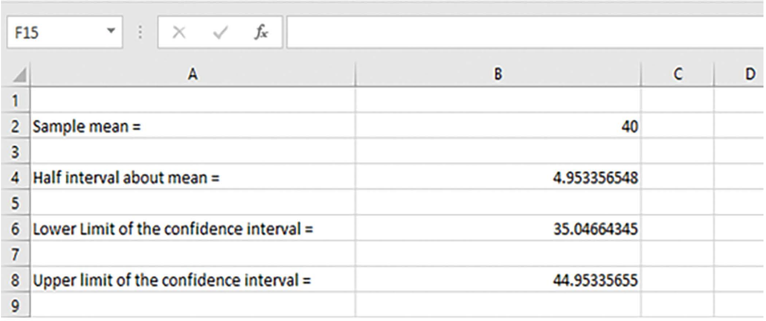
Size is the sample size.

Formula result = 4.953356548

Help on this function

OK Cancel

Figure 8.23 Screenshot after filling data in the dropdown menu of Figure 8.22



	A	B	C	D
1				
2	Sample mean =	40		
3				
4	Half interval about mean =	4.953356548		
5				
6	Lower Limit of the confidence interval =	35.04664345		
7				
8	Upper limit of the confidence interval =	44.95335655		
9				

Figure 8.24 Screenshot of confidence interval after clicking the OK button in the dropdown menu of Figure 8.23

The lower limit and upper limit of the diameter of the spindle are 35.04664345 and 44.95335655, respectively.

Summary

- In an experiment, if *n* repeated trials are performed, then one may be interested to find the probability of having *X* number of successes for the person. Such an experiment is termed the Bernoulli process.
- The *binomial distribution* comes under discrete probability distribution. It is based on the Bernoulli process.
- The *Poisson distribution* is a discrete distribution, which is used to capture, for example, the arrival rate of customers at a service station.
- The exponential probability distribution is a continuous probability distribution, which is used in queuing theory to represent the service time spent on customers in the queuing system. This is actually called a negative exponential distribution.
- A normal distribution is a symmetrical distribution, and the value of the random variable (X) varies from $-\infty$ to $+\infty$.
- The normal probability density function with a mean of 0 and variance of 1 is designed using the central limit theorem by replacing the normal random variable X with a standard normal variable Z.
- The uniform distribution is a continuous distribution, which has an equal probability of occurrence for each of the values of the random variable in an interval with a lower limit and upper limit of a and b, respectively.
- If the variance of the normal population is unknown, then the sampling distribution of the mean is called a t-distribution with sample size less than or equal to 30.
- The confidence interval is the range of values of a statistic. If the computed statistic falls within this interval, accept H_0; otherwise, reject H_0.

Keywords

- In an experiment, if *n* repeated trials are performed, then one may be interested to find the probability of having *X* number of successes by a person. Such an experiment is termed a Bernoulli process.

- The binomial distribution comes under the discrete probability distribution. It is based on the Bernoulli process.
- The Poisson distribution is a discrete distribution, which is used to capture, for example, the arrival rate of customers at a service station.
- The exponential probability distribution is a continuous probability distribution, which is used in queuing theory to represent the service time spent on customers in the queuing system. This is actually called a negative exponential distribution.
- A normal distribution is a symmetrical distribution, and the value of the random variable (X) varies from $-\infty$ to $+\infty$.
- The normal probability density function with a mean of 0 and variance of 1 is designed using the central limit theorem by replacing the normal random variable X with a standard normal variable Z.
- The uniform distribution is a continuous distribution, which has an equal probability of occurrence for each of the values of the random variable in an interval with a lower limit and an upper limit of a and b, respectively.
- If the variance of the normal population is unknown, then the sampling distribution of mean is called a t-distribution with sample size less than 30.
- The confidence interval is the range of values of a statistic.

Review Questions

1. List and explain the assumptions of the Bernoulli process.
2. Give the mathematical formula of the following and explain the variables and parameters in each.

   a. Bernoulli distribution
   b. Bernoulli cumulative distribution function

3. Give the sequence of button clicks in Excel to obtain binomial probability and explain the arguments to be filled in the last click.
4. Give the formula to obtain the binomial probability and binomial cumulative probability in Excel. Also, explain the arguments in the formula.
5. If a coin is tossed ten times, find the probability of each of the following using an Excel formula.

   a. Zero heads?
   b. Four heads?
   c. At most four heads?
   d. At least four heads?

6. Give a sketch of the Poisson distribution and also the formula for a Poisson probability distribution.
7. Give the sequence of button clicks to obtain the Poisson probability in Excel.
8. The arrival rate of flights at an airport follows a Poisson distribution with a mean arrival rate of 4 per 15-minute interval. Find the

   a. Probability of no flight arriving in a 15-minute interval.
   b. Probability of exactly two flights arriving in a 15-minute interval.
   c. Probability of at most two flights arriving in a 15-minute interval.
   d. Probability of at least two flights arriving in a 15-minute interval.

9. Give a sketch of an exponential distribution and also the formulas for the exponential probability distribution and exponential cumulative density function.
10. Give the sequence of button clicks to obtain the exponential probability as well as exponential cumulative probability in Excel.
11. In a machine centre of a fastener manufacturing company, the execution times of components follow an exponential distribution. The average execution time ($1/\mu$) is 1.25 minutes. Find the following in Excel.

    a. Probability that the execution time is equal to 3 minutes.
    b. Probability that the execution time is less than 2 minutes.
    c. Probability that the execution time is more than 2 minutes.
    d. Probability that the execution time is between 1 and 2 minutes

12. List the applications of normal distributions.
13. Give a sketch of the normal distribution and explain its characteristics.
14. Give the formula for the normal probability density distribution and its conversion into a standard normal distribution using the central limit theorem.
15. Give the Excel formula for the cumulative normal density distribution of the normal distribution and explain its variables and parameters.
16. Give the sequence of button clicks to obtain the exponential probability as well as exponential cumulative probability in Excel.
17. Give the formula to obtain the normal probability as well as the normal cumulative probability for a given value of the normal statistic in Excel.
18. Give the Excel formula to obtain the normal statistic for a given cumulative probability of the normal distribution.
19. Give the formula to obtain the standard normal probability as well as standard normal cumulative probability for a given value of the standard normal statistic in Excel.
20. Give the formula to obtain the standard normal statistic for a given cumulative probability in Excel.
21. The monthly income of companies in a survey of 250 small-scale companies follows a normal distribution, with a mean and standard deviation of ₹ 1,25,000 and ₹ 80,000, respectively. Answer the following.

    a. What is the probability that the monthly income is less than ₹ 1,00,00,000? Also, find the number of companies with income less than ₹ 1,00,00,000.
    b. What is the probability that the monthly income is more than ₹ 1,00,00,000? Also, find the number of companies with income more than ₹ 1,00,00,000.
    c. What is the probability that the monthly income is between ₹ 80, 00,000 and ₹ 1,20,00,000? Also, find the number of companies with monthly income between ₹ 80,00,000 and ₹ 1,20,00,000.
    d. Find the value of the monthly income (X) when the cumulative probability is 0.85 and develop the corresponding constructs of probability.

       Construct means P(X≥ K) or P(X ≤ K), where K is the mirror replica on the X axis for the X computed about mean.

22. Find the value of the monthly income (X) when the cumulative probability is 0.3 and develop the corresponding constructs of probability.
23. Give the sketch of the uniform distribution along with formulas for uniform probability and uniform cumulative probability.

24. The demand of a product follows a uniform distribution, and its distribution is as follows.

    $P(X) = 1/(6000–1500), 1500 \leq X \leq 6000$
    = 0, otherwise

    a. If the maximum demand is 5,200, find the service level of the company in satisfying the demand of the product.
    b. If the service level is 0.88, then find the maximum demand which can be satisfied.

25. Give a sketch of the t-distribution and its formula.
26. Give a sequence of button clicks for the following functions of the *t* distribution in Excel.

    a. One-tailed t-distribution
    b. Two-tailed t-distribution
    c. Right-tailed t-distribution
    d. One-tailed t inverse
    e. Two-tailed t inverse

27. A normal population is used to draw a random sample of 20 clinics from a city. The population's annual income has a mean of ₹ 40 lakhs and a variance of ₹ 80 lakhs, respectively.

    a. Find the probability that the mean annual income of the sample is less than ₹ 45 lakhs.
    b. Find the probability that the mean annual income of the sample is more than ₹ 45 lakhs.
    c. Find the annual income if the *p* value (significance value) is 0.35.

28. A shop floor manager wants to establish the confidence interval of bearings produced in the shop. Forty-five units of the bearing are selected as a sample. Bearing diameters have a mean of 45 mm and a standard deviation of 9 mm. Find the confidence interval of the bearing's diameter using a significance level of 0.05.

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. https://corporatefinanceinstitute.com/resources/Excel/functions/binomial-distribution-Excel/ [June 25, 2020].
3. https://support.microsoft.com/en-us/office/binom-inv-function-80a0370c-ada6-49b4-83e7-05a91ba77ac9 [July 4, 2020].
4. www.real-statistics.com/binomial-and-related-distributions/poisson-distribution/ [June 25, 2020].
5. https://support.microsoft.com/en-us/office/expon-dist-function-4c12ae24-e563-4155-bf3e-8b78b6ae140e [June 25, 2020].
6. https://support.microsoft.com/en-us/office/norm-s-dist-function-1e787282-3832-4520-a9ae-bd2a8d99ba88 [June 27, 2020].

9 Sampling Distributions of Mean and Variance

Learning Objectives

After reading this chapter, you will be able to

- Analyse the sampling distribution of the mean when the population is infinite, which is a normal distribution.
- Understand the theory of the sampling distribution of the mean when the population is finite, which is a normal distribution with a finite population multiplier in the variance.
- Study the scope of the sampling distribution of the mean when the normal population variance is unknown (*t* distribution) and apply in practice.
- Understand the chi-square distribution and its extension to the sampling distribution.
- Analyse the F distribution and its extension to the sampling distribution

9.1 Introduction

A sample is a subpopulation of a population. Population-based data collection and performing data analysis will be time consuming. As a result, in reality, analysis of a sample will aid in the drawing of conclusions about the population. The population size in this study may be taken to be both infinite and finite. For every case, sample parameters are derived using the population parameters [1]. The sampling distributions of mean, differences between means, and variance are also presented.

This chapter presents the use of Excel for each of the cases after presenting a brief introduction in each of the sections of these cases.

9.2 Sampling Distributions for Mean

This section presents the sampling distributions for means, which are as listed as follows.

- Normal distribution
- t distribution

DOI: 10.4324/9781032671703-9

9.2.1 Sampling Distribution of Mean When the Population Is Infinite Using NORM.S.DIST and NORM.S.INV Functions

Consider a sampling distribution where the population size is infinite and the normal population variance is known. The mean and variance calculations for this distribution are provided as follows.

Mean = μ, which is same as the population mean
Variance = σ^2/n, where σ^2 is the variance of the population and n is the sample size
Standard deviation = $\sigma/\sqrt{n}$, which is known as standard error

In the process of sampling, there may be some error. The standard error is a measure to represent the deviations of different sample means from the true means due to sampling error.

Let the size of a sample be n and the sample mean be $\bar{X}$. The formula for the sample mean is given by the following formula.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Then $\bar{X}$ also follows a normal distribution with mean μ and variance σ^2/n.

These are presented in notation form as follows.

$$\text{If X} \sim \text{N}\left(\mu, \sigma^2\right), \text{then } \bar{\text{X}} \sim \text{N}\left(\mu, \frac{\sigma^2}{\text{n}}\right)$$

$$\textit{The standard normal statistic for } \bar{X} \textit{ is} \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

The Excel formula to get the value of the cumulative probability for a given value of Z is as follows.

= NORM.S.DIST

The screenshot that is obtained after clicking the sequence of buttons Formulas ⟹ More Functions ⟹ Statistics ⟹ NORM.S.DIST is shown in Figure 9.1 [2].

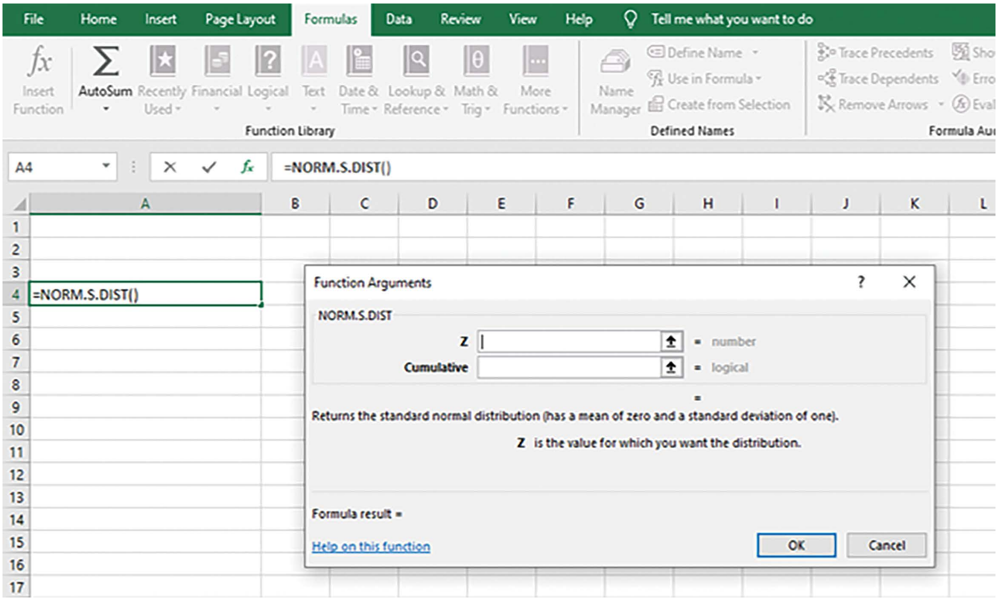
The Excel formula to get the value of value of Z for a given cumulative probability is as follows.

= NORM.S.INV

This screenshot that is obtained after clicking the sequence of buttons Formulas⟹ More Functions ⟹ Statistics ⟹ NORM.S.INV is shown in Figure 9.2 [3].

Example 9.1

An infinite normal population with a mean and variance of 200 and 750, respectively, is sampled at random with a size of 70. What is the probability that the sample mean exceeds 205?



A4 =NORM.S.DIST()

Function Arguments

NORM.S.DIST

Z = number

Cumulative = logical

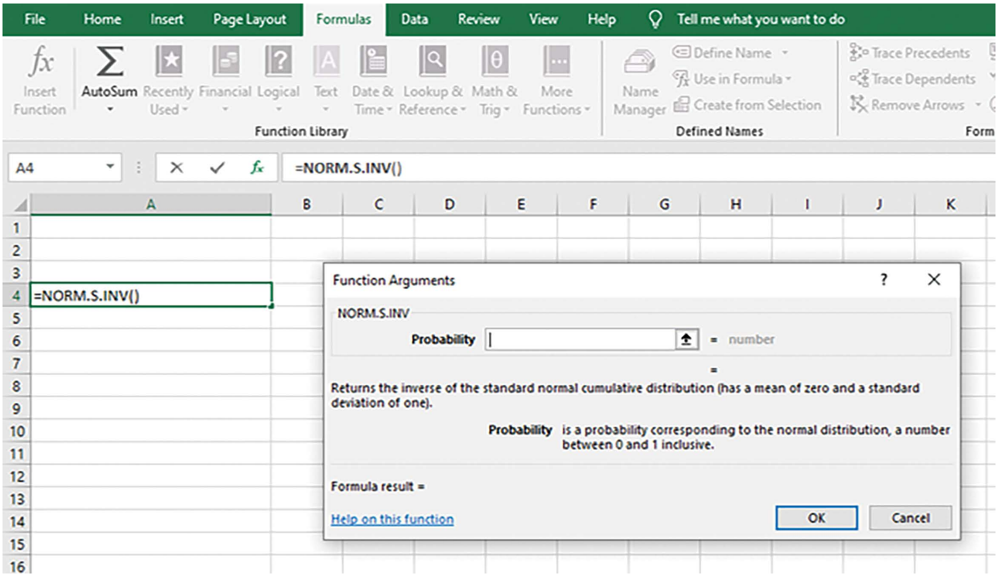
Returns the standard normal distribution (has a mean of zero and a standard deviation of one).

Z is the value for which you want the distribution.

Formula result =

Help on this function OK Cancel

Figure 9.1 Screenshot for button clicks, Formulas ⟹ More Functions ⟹ Statistics ⟹ NORM.S.DIST



A4 =NORM.S.INV()

Function Arguments

NORM.S.INV

Probability = number

Returns the inverse of the standard normal cumulative distribution (has a mean of zero and a standard deviation of one).

Probability is a probability corresponding to the normal distribution, a number between 0 and 1 inclusive.

Formula result =

Help on this function OK Cancel

Figure 9.2 Screenshot after clicking the sequence of buttons, Formulas ⟹ More Functions ⟹ Statistics ⟹ NORM.S.INV

Solution

Size of the random sample, $n = 70$

$\mu = 200$ and $\sigma^2 = 750$
$\sigma = 27.386$

$$\mathrm{X} \sim \mathrm{N}\left(\mu, \sigma^2\right) \sim \mathrm{N}(200, 750)$$

$$\bar{\mathrm{X}} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{\mathrm{n}}\right) \sim \mathrm{N}\left(200, \frac{750}{70}\right)$$

$$P\left(\bar{X} \geq 205\right) = P\left[\frac{\bar{X}-\mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \geq \frac{205-\mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}\right] = P\left(Z_{\bar{X}} \geq \frac{205-200}{\left(\frac{27.386}{\sqrt{70}}\right)}\right)$$

$$= 1 - P\left(Z_{\bar{X}} \geq \frac{205-200}{\left(\frac{27.386}{\sqrt{70}}\right)}\right)$$

Note: The simple calculation in the formula may be done using Excel.

The working to obtain the probability of the equation is shown in Figure 9.3. From Figure 9.3, it is clear that the probability that the sample mean is greater than or equal to 205 is 0.063315229, which is approximated to 0.0633. The Excel formulas of the working shown in Figure 9.3 are shown in Figure 9.4.

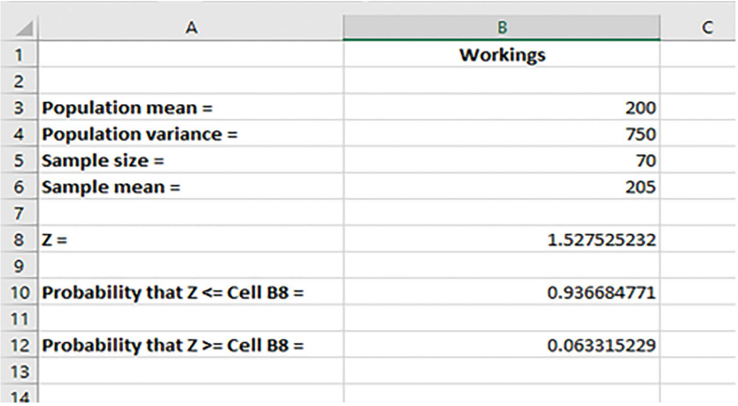
9.2.2 *Sampling Distribution of Mean When the Population Is Finite Using Excel Sheets*

The sampling distribution of the mean with an infinite population has the variance of σ^2/n. But the variance of the sampling distribution with a finite population will have a different variance, which can be obtained from the variance of the sampling distribution with an infinite population by multiplying by a finite population multiplier.

Let

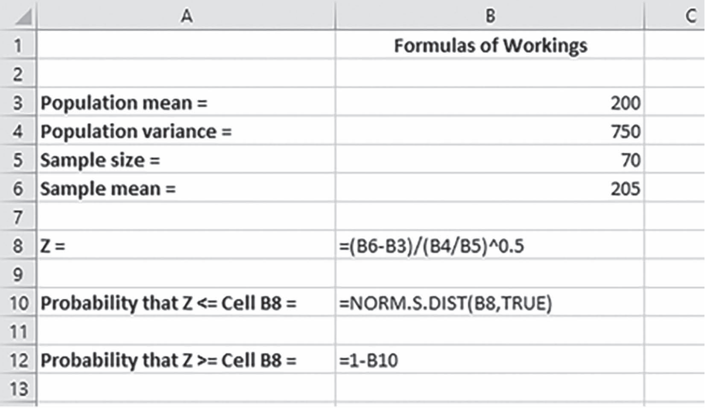
N be the size of the population

n be the size of the sample taken from the finite population



	A	B	C
1		**Workings**	
2			
3	**Population mean =**	200	
4	**Population variance =**	750	
5	**Sample size =**	70	
6	**Sample mean =**	205	
7			
8	**Z =**	1.527525232	
9			
10	**Probability that Z <= Cell B8 =**	0.936684771	
11			
12	**Probability that Z >= Cell B8 =**	0.063315229	
13			
14			

Figure 9.3 Screenshot of working of Example 9.1



	A	B	C
1		**Formulas of Workings**	
2			
3	**Population mean =**	200	
4	**Population variance =**	750	
5	**Sample size =**	70	
6	**Sample mean =**	205	
7			
8	**Z =**	=(B6-B3)/(B4/B5)^0.5	
9			
10	**Probability that Z <= Cell B8 =**	=NORM.S.DIST(B8,TRUE)	
11			
12	**Probability that Z >= Cell B8 =**	=1-B10	
13			

Figure 9.4 Screenshot of formulas for working of Example 9.1

The variance of the sampling distribution with a finite population is given by the following formula.

$$\textit{Variance} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

In this formula, $\frac{N-n}{N-1}$ is called a finite population multiplier.

Let X be a random variable following a normal distribution, with mean μ and variance σ^2. Then X is a random variable, which also follows a normal distribution with mean μ and variance $\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$.

These are stated in notation form as follows.

$$\text{If } X \sim N\left(\mu,\sigma^2\right), \text{ then } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right)$$

The standard normal statistic Z for X is $\frac{X-\mu}{\sigma}$.

The standard normal statistic $Z_{\bar{X}}$ for $\bar{X}$ is $\dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$

Example 9.2

From a finite normal population of size 1200, with a mean and variance of 85 and 64, respectively, a random sample of size 44 is taken.

1. What is the probability that the sample mean is less than 88?
2. What is the probability that the sample mean is greater than 70?
3. What is the probability that the sample mean will be between 83 and 87?

Solution

Size of the random sample, n = 44
Population size, N = 1200
Population mean, μ = 85
Population variance, σ^2 = 64
Population standard deviation, σ = 8

Sample variance = $\dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right)$.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right)$$

$$\bar{X} \sim N\left(85, \frac{64}{44}\left(\frac{1200-44}{1200-1}\right)\right)$$

The standard normal statistic $Z_{\bar{X}}$ *for* $\bar{X}$ *is* $\dfrac{\bar{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}}$

1. When $\bar{X} \le 88$,

$$P(\bar{X} \le 88) = P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \le \frac{88-85}{\frac{8}{\sqrt{44}}\sqrt{\frac{1200-44}{1200-1}}}\right)$$

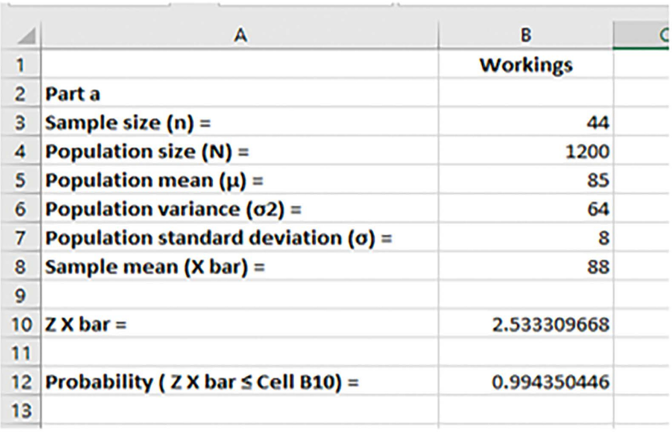
$$= P\left(Z_{\bar{X}} \le \frac{88-85}{\frac{8}{\sqrt{44}}\sqrt{\frac{1200-44}{1200-1}}}\right)$$

The working to obtain the probability of the equation is shown in Figure 9.5. From Figure 9.5, it is clear that the probability that the sample mean is less than or equal to 88 is 0.995.

The Excel formulas of the working shown in Figure 9.5 are shown in Figure 9.6.

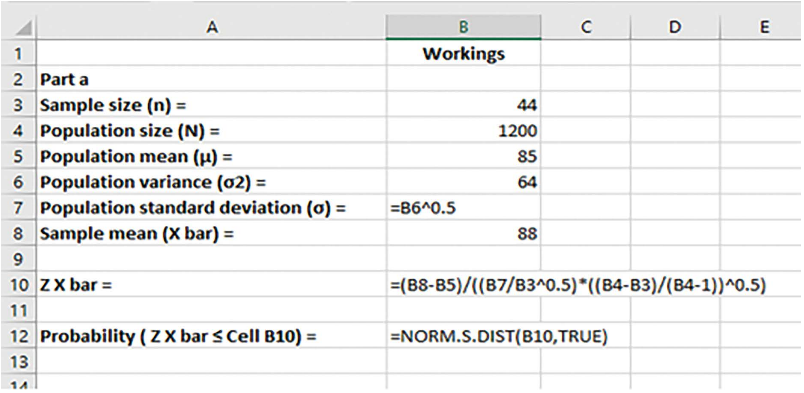
2.

$$P(\bar{X} \ge 70) = P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \ge \frac{70-85}{\frac{8}{\sqrt{44}}\sqrt{\frac{1200-44}{1200-1}}}\right)$$



	A	B
1		Workings
2	Part a	
3	Sample size (n) =	44
4	Population size (N) =	1200
5	Population mean (μ) =	85
6	Population variance (σ2) =	64
7	Population standard deviation (σ) =	8
8	Sample mean (X bar) =	88
9		
10	Z X bar =	2.533309668
11		
12	Probability (Z X bar ≤ Cell B10) =	0.994350446
13		

Figure 9.5 Screenshot of working to compute probability for condition 1 of Example 9.2

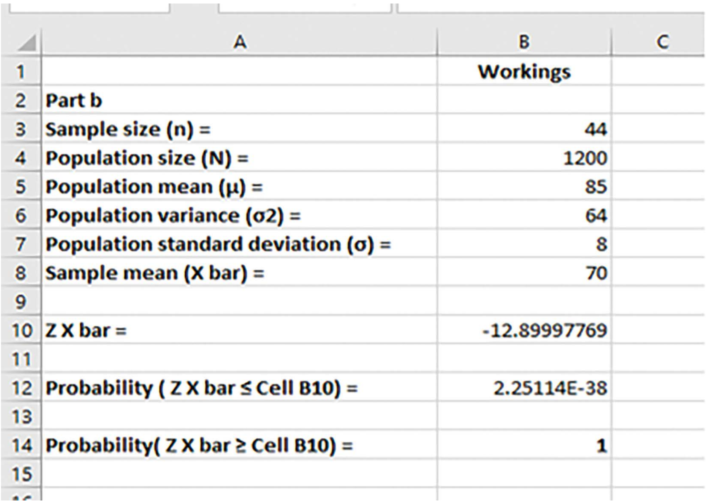


	A	B	C	D	E
1		Workings			
2	Part a				
3	Sample size (n) =	44			
4	Population size (N) =	1200			
5	Population mean (μ) =	85			
6	Population variance (σ2) =	64			
7	Population standard deviation (σ) =	=B6^0.5			
8	Sample mean (X bar) =	88			
9					
10	Z X bar =	=(B8-B5)/((B7/B3^0.5)*((B4-B3)/(B4-1))^0.5)			
11					
12	Probability (Z X bar ≤ Cell B10) =	=NORM.S.DIST(B10,TRUE)			
13					

Figure 9.6 Screenshot of formulas for working to compute probability for condition 1 of Example 9.2

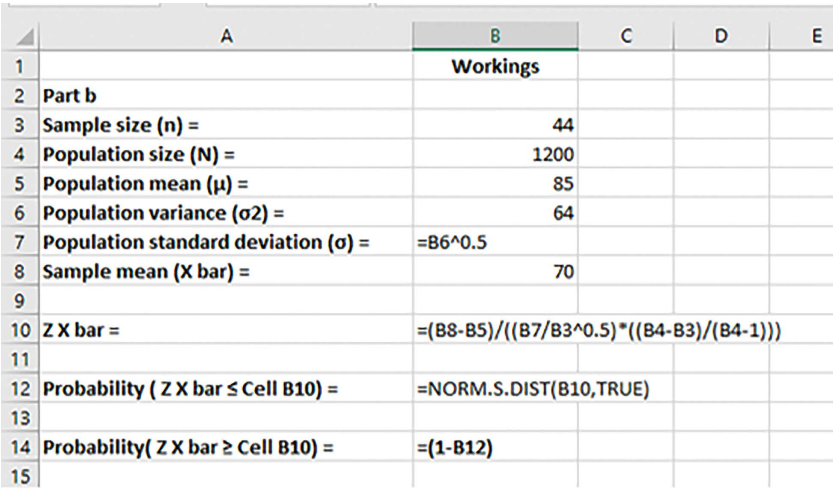
$$= \mathrm{P}\left(Z_{\bar{X}} \geq \frac{70-85}{\dfrac{8}{\sqrt{44}}\sqrt{\dfrac{1200-44}{1200-1}}} \right) = 1 - \mathrm{P}\left(Z_{\bar{X}} \leq \frac{70-85}{\dfrac{8}{\sqrt{44}}\sqrt{\dfrac{1200-44}{1200-1}}} \right)$$

The working to obtain the probability of the equation is shown in Figure 9.7. From Figure 9.7, it is clear that the probability that the sample mean is greater than or equal 70 is 1. The formulas for the working of probability are shown in Figure 9.8.



	A	B	C
1		Workings	
2	Part b		
3	Sample size (n) =	44	
4	Population size (N) =	1200	
5	Population mean (μ) =	85	
6	Population variance (σ2) =	64	
7	Population standard deviation (σ) =	8	
8	Sample mean (X bar) =	70	
9			
10	Z X bar =	-12.89997769	
11			
12	Probability (Z X bar ≤ Cell B10) =	2.25114E-38	
13			
14	Probability(Z X bar ≥ Cell B10) =	1	
15			

Figure 9.7 Screenshot for working to compute probability for condition 2 of Example 9.2



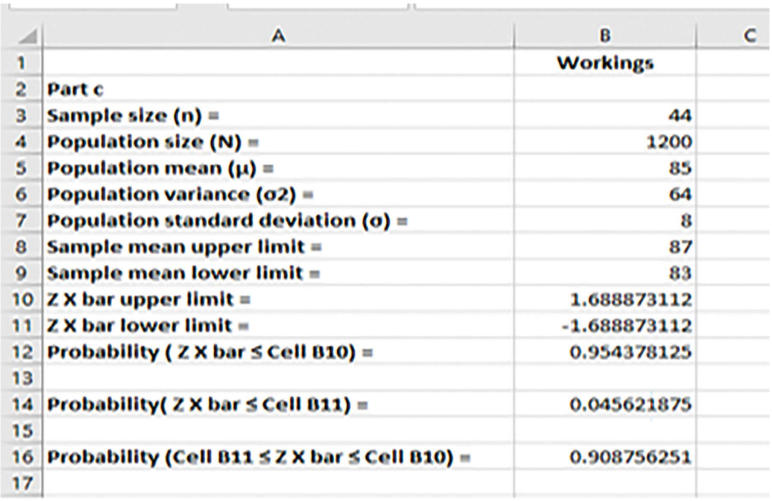
	A	B	C	D	E
1		Workings			
2	Part b				
3	Sample size (n) =	44			
4	Population size (N) =	1200			
5	Population mean (μ) =	85			
6	Population variance (σ2) =	64			
7	Population standard deviation (σ) =	=B6^0.5			
8	Sample mean (X bar) =	70			
9					
10	Z X bar =	=(B8-B5)/((B7/B3^0.5)*((B4-B3)/(B4-1)))			
11					
12	Probability (Z X bar ≤ Cell B10) =	=NORM.S.DIST(B10,TRUE)			
13					
14	Probability(Z X bar ≥ Cell B10) =	=(1-B12)			
15					

Figure 9.8 Excel sheet with formulas for the working of probability of condition 2 of Example 9.2

3.

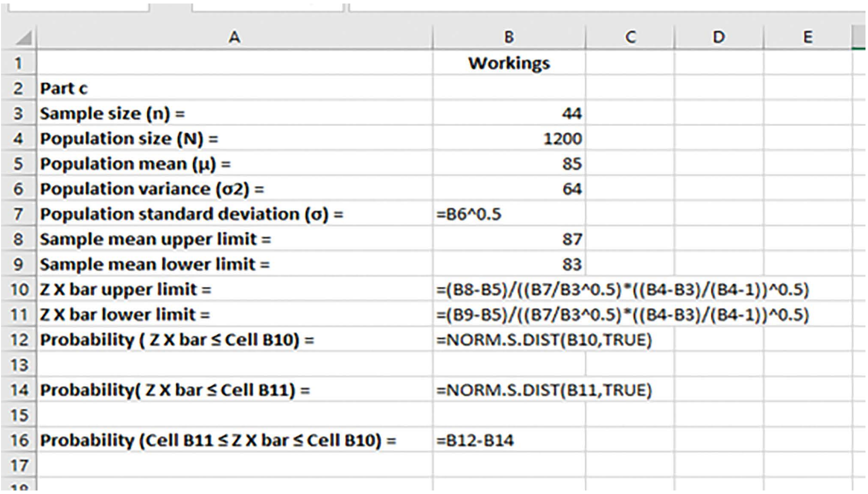
$$\mathrm{P}\left(83 \le \bar{X} \le 87\right) = \mathrm{P}\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \le \frac{87-85}{\frac{8}{\sqrt{44}}\sqrt{\frac{1200-44}{1200-1}}}\right) - \mathrm{P}\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \le \frac{83-85}{\frac{8}{\sqrt{44}}\sqrt{\frac{1200-44}{1200-1}}}\right)$$

$$= \mathrm{P}\left(Z_{\bar{X}} \le \frac{87-85}{\frac{8}{\sqrt{44}}\sqrt{\frac{1200-44}{1200-1}}}\right) - \mathrm{P}\left(Z_{\bar{X}} \le \frac{83-85}{\frac{8}{\sqrt{44}}\sqrt{\frac{1200-44}{1200-1}}}\right)$$



	A	B	C
1		Workings	
2	Part c		
3	Sample size (n) =	44	
4	Population size (N) =	1200	
5	Population mean (μ) =	85	
6	Population variance (σ2) =	64	
7	Population standard deviation (σ) =	8	
8	Sample mean upper limit =	87	
9	Sample mean lower limit =	83	
10	Z X bar upper limit =	1.688873112	
11	Z X bar lower limit =	-1.688873112	
12	Probability (Z X bar ≤ Cell B10) =	0.954378125	
13			
14	Probability(Z X bar ≤ Cell B11) =	0.045621875	
15			
16	Probability (Cell B11 ≤ Z X bar ≤ Cell B10) =	0.908756251	
17			

Figure 9.9 Screenshot for working to compute probability for condition 3 of Example 9.2



	A	B	C	D	E
1		Workings			
2	Part c				
3	Sample size (n) =	44			
4	Population size (N) =	1200			
5	Population mean (μ) =	85			
6	Population variance (σ2) =	64			
7	Population standard deviation (σ) =	=B6^0.5			
8	Sample mean upper limit =	87			
9	Sample mean lower limit =	83			
10	Z X bar upper limit =	=(B8-B5)/((B7/B3^0.5)*((B4-B3)/(B4-1))^0.5)			
11	Z X bar lower limit =	=(B9-B5)/((B7/B3^0.5)*((B4-B3)/(B4-1))^0.5)			
12	Probability (Z X bar ≤ Cell B10) =	=NORM.S.DIST(B10,TRUE)			
13					
14	Probability(Z X bar ≤ Cell B11) =	=NORM.S.DIST(B11,TRUE)			
15					
16	Probability (Cell B11 ≤ Z X bar ≤ Cell B10) =	=B12-B14			
17					

Figure 9.10 Screenshot of formulas for working to compute probability for condition 3 of Example 9.2

The working to obtain the probability of the =equation is shown in Figure 9.9. From Figure 9.9, it is clear that the probability that the sample mean is greater than or equal to 83 and less than or equal to 87 is 0.908756251. The formulas to obtain the probability are shown in Figure 9.10.

9.2.3 *Sampling Distribution of Mean When Normal Population Variance Is Unknown (t Distribution) Using T.DIST, T.INV, T.DIST.2T, T.DIST.RT, and T.INV.2T Functions*

If the variance of the normal population is unknown, then the corresponding sampling distribution is the student's *t*-distribution.

where

X is a random variable, which follows a normal distribution with mean μ and variance σ^2

$Z_{\bar{X}}$ is a standard normal statistic of $\bar{X}$

χ^2 is a random variable which follows a chi-square distribution with degrees of freedom τ

$Z_{\bar{X}}$ and χ^2 are independent from each other

The standard normal statistic for $\bar{X}$ is $\dfrac{\bar{X}-\mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$

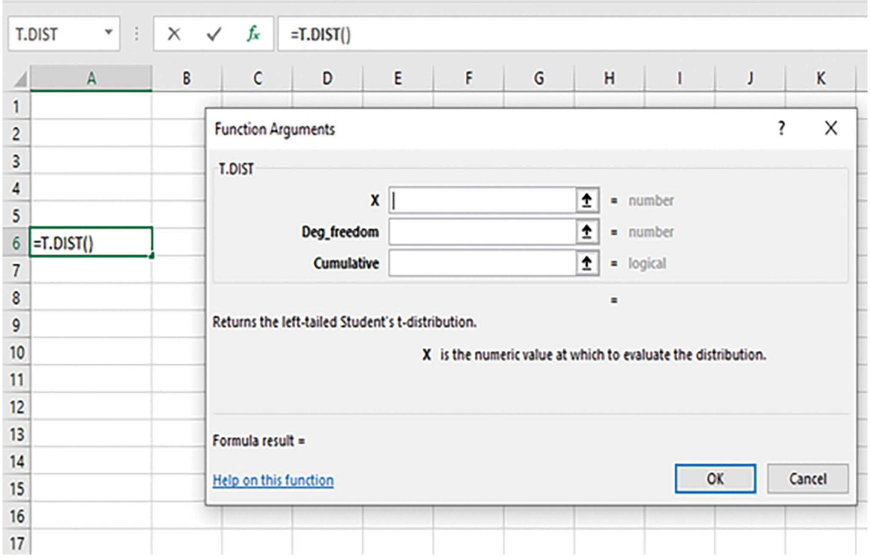
$\chi 2 \;= \dfrac{(n-1)S^2}{\sigma^2}$ with $(n-1)$ degrees of freedom, where S^2 is the variance of the sample

The t distribution is given by the following formula, and its shape is also symmetric, like the normal distribution.

$$t = \frac{Z_{\bar{X}}}{\sqrt{\dfrac{\chi^2}{\tau}}}$$

The final formula for the t distribution is as follows.

$$t = \frac{\bar{X}-\mu}{\left(\dfrac{S}{\sqrt{n}}\right)} \text{with} \left(\text{n}-1\right) \text{ degrees of freedom.}$$



T.DIST

=T.DIST()

Function Arguments

T.DIST

X = number

Deg_freedom = number

Cumulative = logical

=

Returns the left-tailed Student's t-distribution.

X is the numeric value at which to evaluate the distribution.

Formula result =

Help on this function

OK Cancel

Figure 9.11 Screenshot to obtain the cumulative probability for given t and Deg_freedom

The Excel formula to get the value of the cumulative probability for the given *t* and Deg_freedom is as follows, which can be seen in the screenshot in Figure 9.11 [4].

$$= \text{T.DIST}\left(\text{X, Deg\_freedom, TRUE}\right)$$

The screenshot shown in Figure 9.11 is obtained after clicking the sequence of buttons HOME ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ T.DIST.

The Excel formula to get the value of *t* for a given cumulative probability and Deg_freedom is as follows, which can be seen in the screenshot in Figure 9.12.

$$= \text{T.INV}\left(\text{Probability, Deg\_freedom}\right)$$

The screenshot shown in Figure 9.12 is obtained after clicking the sequence of buttons HOME ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ T.INV.

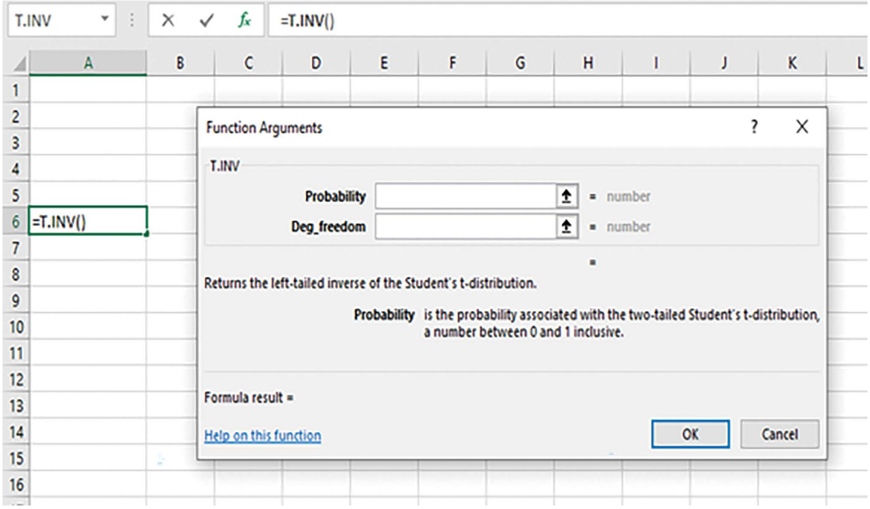
The Excel formula to get the value of the probability at each of the two tails of the *t* distribution for the given *t* and Deg_freedom is as follows, which can be seen in the screenshot in Figure 9.13.

$$= \text{T.DIST.2T}\left(\text{X,Deg\_freedom}\right)$$

The screenshot shown in Figure 9.13 is obtained after the button clicks HOME ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ T.DIS.2T.

The Excel formula to get the value of the probability at the right tail of the *t* distribution for the given *t* and Deg_freedom is as follows, which can be seen in the screenshot in Figure 9.14.

$$= \text{T.DIST.RT}\left(\text{X,Deg\_freedom}\right)$$



T.INV

=T.INV()

6 =T.INV()

Function Arguments

T.INV

Probability = number

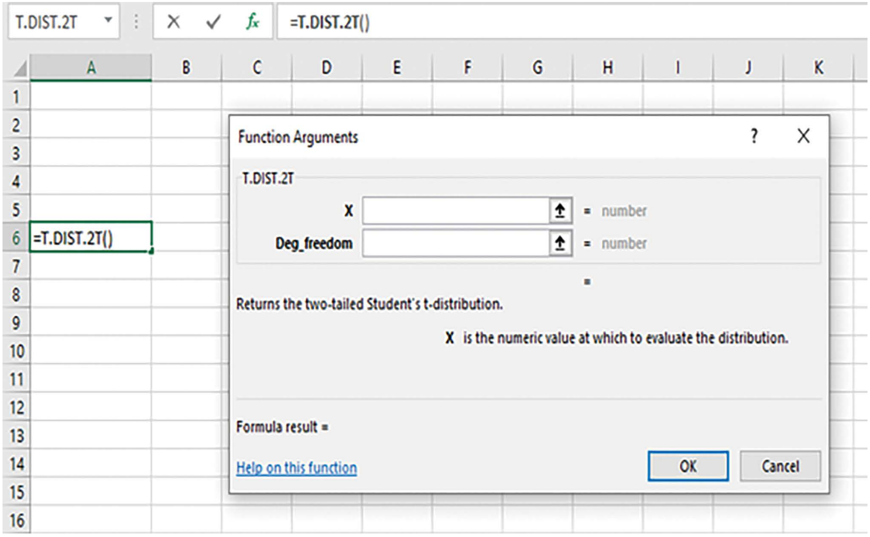
Deg_freedom = number

Returns the left-tailed inverse of the Student's t-distribution.

Probability is the probability associated with the two-tailed Student's t-distribution, a number between 0 and 1 inclusive.

Formula result =

Figure 9.12 Screenshot of obtaining the value of *t* for given cumulative probability and Deg_freedom



T.DIST.2T

=T.DIST.2T()

=T.DIST.2T()

Function Arguments

T.DIST.2T

X = number

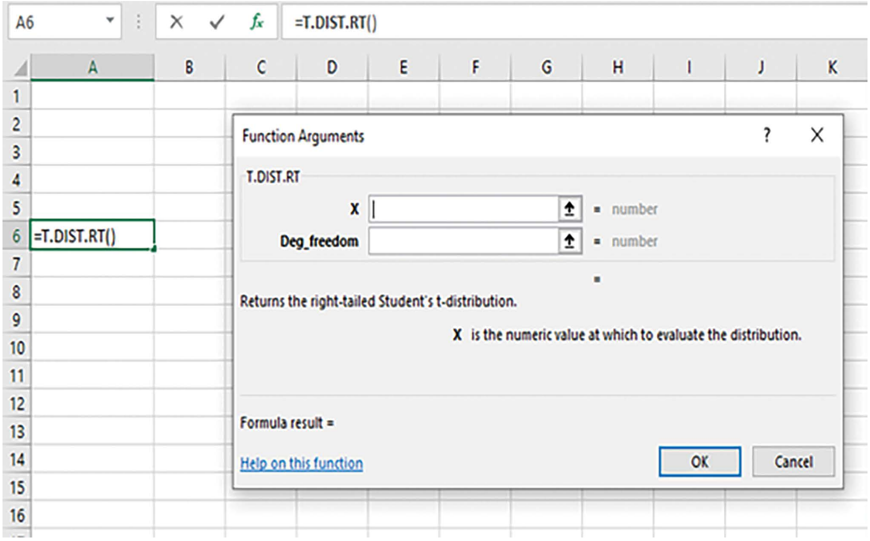
Deg_freedom = number

Returns the two-tailed Student's t-distribution.

X is the numeric value at which to evaluate the distribution.

Formula result =

Figure 9.13 Screenshot for clicking the sequence of buttons, Home ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ T.DIS.2T



A6

=T.DIST.RT()

=T.DIST.RT()

Function Arguments

T.DIST.RT

X = number

Deg_freedom = number

Returns the right-tailed Student's t-distribution.

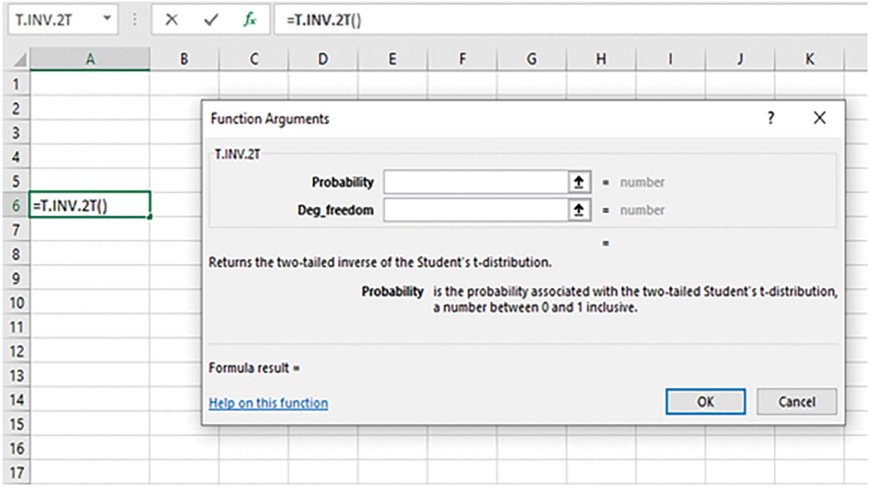
X is the numeric value at which to evaluate the distribution.

Formula result =

Figure 9.14 Screenshot for clicking the sequence of buttons, Home ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ T.DIS.RT

The screenshot shown in Figure 9.14 is obtained after clicking the sequence of buttons HOME ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ T.DIST.RT.

The Excel formula to get the values of t at both tails of the t distribution for a given significance level of α and Deg_freedom is as follows, which can be seen in the screenshot shown in Figure 9.15. In the formula, the probability is α.



=T.INV.2T()

Function Arguments

T.INV.2T

Probability = number

Deg_freedom = number

Returns the two-tailed inverse of the Student's t-distribution.

Probability is the probability associated with the two-tailed Student's t-distribution, a number between 0 and 1 inclusive.

Formula result =

Figure 9.15 Screenshot for clicking the sequence of buttons, Home ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ T.INV.2T

$$= \text{T.INV.2T}(\text{Probability, Deg\_freedom})$$

The screenshot shown in Figure 9.15 is obtained after clicking the sequence of buttons HOME ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ T.INV.2T.

Example 9.3

A random sample of 20 customers of a company is taken from a normal population. The mean of the annual purchases made by the customers of the population is ₹ 75 lakh. The variance of the purchase made by the customers of the sample is ₹ 121 lakhs.

1. Find the probability that the mean annual sales of the sample is less than ₹ 80 lakhs.
2. Find the probability that the mean annual sales of the sample is more than ₹ 80 lakhs.
3. Find the value *t* for a cumulative probability of 0.90.
4. Find the value of *t* for a given level of significance of 0.05 by placing half of it on the right tail.
5. Find the probability at both tails when the annual sales is less than 80 lakhs.

Solution

Size of the random sample, $n = 20$

μ = ₹ 75 lakh and $\sigma^2 = S^2$ = ₹ 121 lakh, S = 11 lakh

$$t = \frac{\bar{X} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} with (n-1) degrees of freedom.$$

$$1. P\left(\bar{X} \le 80\right) = p\left(\frac{\bar{X}-\mu}{\left(\frac{s}{\sqrt{n}}\right)} \le \frac{80-75}{\left(\frac{11}{\sqrt{20}}\right)}\right) = P\left(t \le \frac{80-75}{\left(\frac{11}{\sqrt{20}}\right)}\right) = ?$$

The working of the formula is shown in Figure 9.16. The guidelines for the formulas for the working of part 1 are shown in Figure 9.17. The answer for question 1 is 0.971856, which can be seen in Figure 9.16.

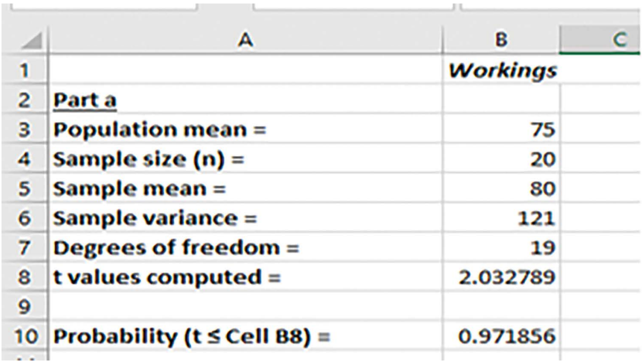
$$2. P\left(\bar{X} \ge 80\right) = p\left(\frac{\bar{X}-\mu}{\left(\frac{s}{\sqrt{n}}\right)} \ge \frac{80-75}{\left(\frac{11}{\sqrt{20}}\right)}\right) = P\left(t \ge \frac{80-75}{\left(\frac{11}{\sqrt{20}}\right)}\right) = 1 - P\left(t \le \frac{80-75}{\left(\frac{11}{\sqrt{20}}\right)}\right)$$

Since the value of $P\left(t \le \frac{80-75}{\left(\frac{11}{\sqrt{20}}\right)}\right)$ was already computed in part 1 of this question, the answer to part 2 is computed using the following formula.

$$P\left(t \ge \frac{80-75}{\left(\frac{11}{\sqrt{20}}\right)}\right) = 1 - P\left(t \le \frac{80-75}{\left(\frac{11}{\sqrt{20}}\right)}\right) = 1 - 0.971856 = 0.028144$$

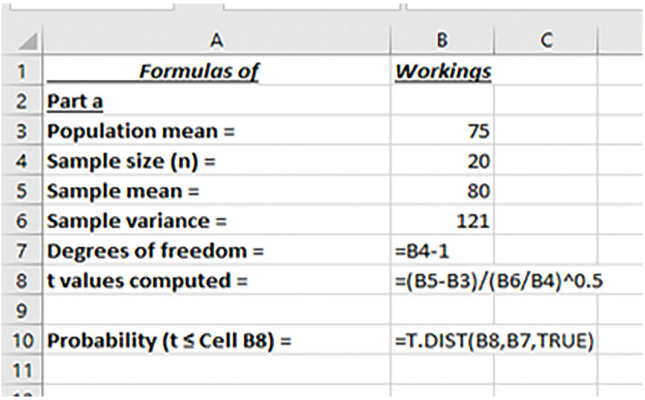
3. The value of t when the level of significance α (0.05) is fully placed on the right tail is computed as follows.

$\alpha = 0.05$
$1 - \alpha = 1–0.05 = 0.95$



	A	B	C
1		**Workings**	
2	**Part a**		
3	**Population mean =**	75	
4	**Sample size (n) =**	20	
5	**Sample mean =**	80	
6	**Sample variance =**	121	
7	**Degrees of freedom =**	19	
8	**t values computed =**	2.032789	
9			
10	**Probability (t ≤ Cell B8) =**	0.971856	

Figure 9.16 Screenshot of working of part 1 of Example 9.3



	A	B	C
1	Formulas of	Workings	
2	Part a		
3	Population mean =	75	
4	Sample size (n) =	20	
5	Sample mean =	80	
6	Sample variance =	121	
7	Degrees of freedom =	=B4-1	
8	t values computed =	=(B5-B3)/(B6/B4)^0.5	
9			
10	Probability (t ≤ Cell B8) =	=T.DIST(B8,B7,TRUE)	
11			

Figure 9.17 Screenshot of guidelines for formulas of working of Example 9.3

When the cumulative probability is 0.95, the value of t is computed using the following Excel formula.

$$= T.INV\left(probability, \deg\_freedom\right)$$

$$= T.INV\left(0.95, 19\right) = 1.729132812$$

4. The value of t when half of the significance level of $0.05(\alpha)$ is placed on the right tail is computed as follows.

When the significance level is 0.05 and half of it (0.025) is placed at the right tail of the t distribution with 19 degrees of freedom, the corresponding value of t is computed using the following Excel formula. In the formula, the probability is α, which is 0.05.

$$= T.INV.RT\left(probability, \deg\_freedom\right)$$

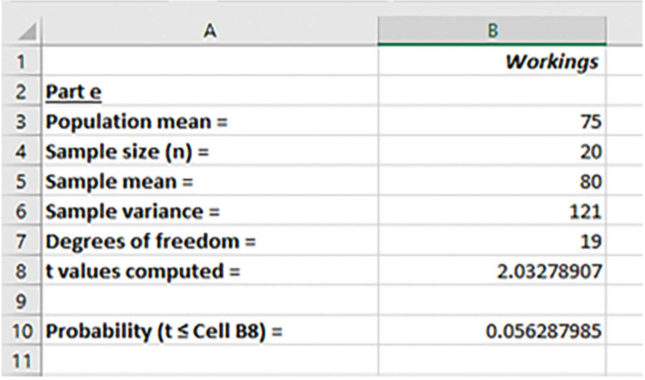
$$= T.INV.RT\left(0.05,\ 19\right) = 2.093024054$$

$$5. P\left(\bar{X} \le 80\right) = p\left(\frac{\bar{X} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} \le \frac{80 - 75}{\left(\frac{11}{\sqrt{20}}\right)}\right) = P\left(t \le \frac{80 - 75}{\left(\frac{11}{\sqrt{20}}\right)}\right) = ?$$

The probability at both tails for the given t value can be computed using the following formula.

$$= \text{T.DIST.2T}\left(\text{X, Deg\_freedom}\right)$$

The working of the formula is shown in Figure 9.18. The guidelines for the formulas for part 5 are shown in Figure 9.19. The answer for part 5 is –0.056287985 and 0.056287985, which can be seen in Figure 9.18. The left tail t value is the mirror image of the right tail t value.



	A	B
1		***Workings***
2	**Part e**	
3	**Population mean =**	**75**
4	**Sample size (n) =**	**20**
5	**Sample mean =**	**80**
6	**Sample variance =**	**121**
7	**Degrees of freedom =**	**19**
8	**t values computed =**	**2.03278907**
9		
10	**Probability (t ≤ Cell B8) =**	**0.056287985**
11		

Figure 9.18 Screenshot of working Example 9.3 part 5

9.3 Sampling Distributions of Variance

The χ^2 distribution and F distribution are the sampling distributions of variance, which are explained in this section.

9.3.1 Chi-Square Distribution Using CHISQ.DIST, CHISQ.DIST.RT, CHISQ.INV, and CHISQ.INV.RT

The formula for the variance of a sample is as follows.

$$S^2 = \frac{\sum_{i=1}^{n}\left(\left(X_i - \bar{X}\right)\right)^2}{n-1}$$

Let
X_i be the i^{th} observation of a sample
$\bar{X}$ be the mean of the sample
n be the size of the sample
S^2 be the variance of a random sample

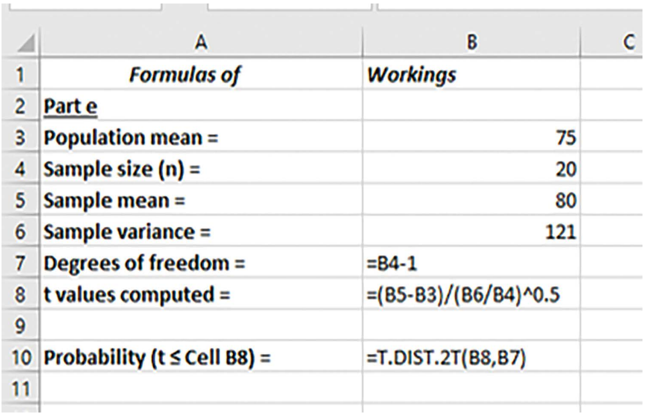
The distribution of S^2 taken from a normal population with variance σ^2 is called a chi-square (χ^2) distribution with (n – 1) degrees of freedom.

$$\chi^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \mathit{with}(n-1)\mathit{degrees\,of\,freedom}$$

The Excel formula to compute the cumulative probability from the left tail for a given value of χ^2 and degrees of freedom is as follows.

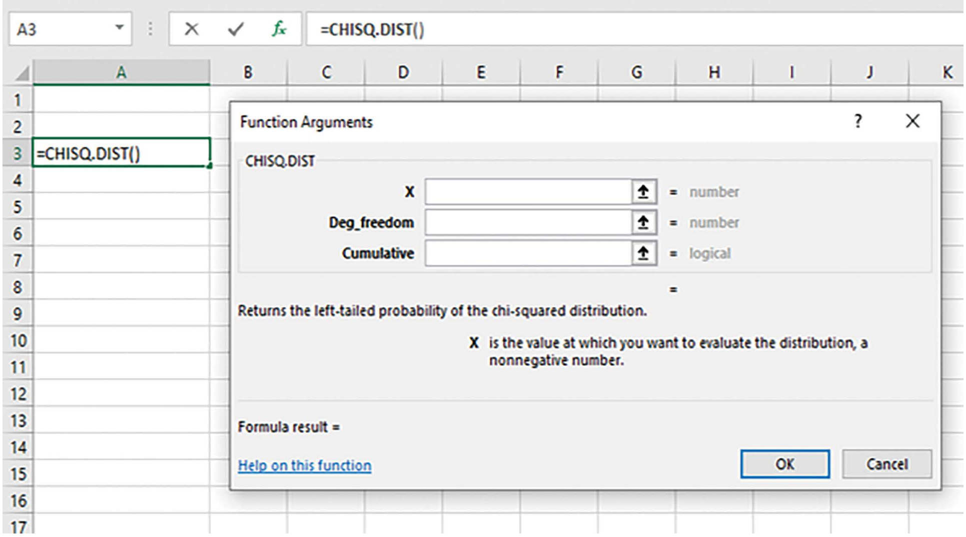
$$= \text{CHISQ.DIST}(\text{X}, \text{Deg\_freedom, TRUE})$$

The screenshot after clicking the buttons HOME ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ CHISQ.DIST is shown in Figure 9.20, which gives the cumulative probability from the left tail for a given chi-square value and degrees of freedom.



	A	B	C
1	***Formulas of***	***Workings***	
2	**Part e**		
3	**Population mean =**	75	
4	**Sample size (n) =**	20	
5	**Sample mean =**	80	
6	**Sample variance =**	121	
7	**Degrees of freedom =**	=B4-1	
8	**t values computed =**	=(B5-B3)/(B6/B4)^0.5	
9			
10	**Probability (t ≤ Cell B8) =**	=T.DIST.2T(B8,B7)	
11			

Figure 9.19 Screenshot for formulas of working Example 9.3 part 5



A3 =CHISQ.DIST()

Function Arguments

CHISQ.DIST

X = number

Deg_freedom = number

Cumulative = logical

Returns the left-tailed probability of the chi-squared distribution.

X is the value at which you want to evaluate the distribution, a nonnegative number.

Formula result =

Figure 9.20 Screenshot of obtaining cumulative probability from left tail for given chi-square value

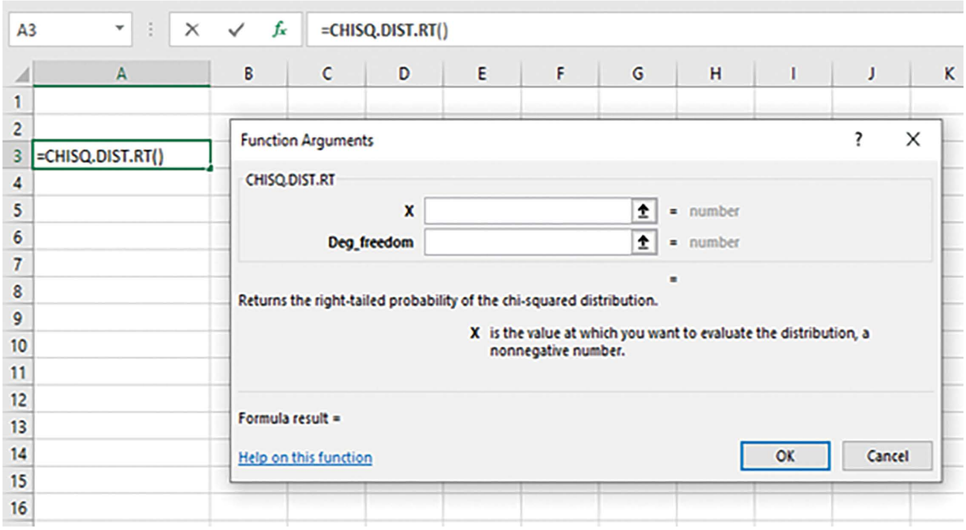
The Excel formula to compute the probability at the right tail for a given value of χ^2 and degrees of freedom is as follows [5].

$$= \text{CHISQ.DIST.RT}(\text{X}, \text{Deg\_freedom})$$

The screenshot after clicking the buttons HOME ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ CHISQ.DIST.RT is shown in Figure 9.21, which gives the probability at the right tail for a given value of χ^2 and degrees of freedom.

The Excel formula to compute the chi-square value for a given cumulative probability from the left tail and degrees of freedom is as follows.

$$= \text{CHISQ.INV}(\text{Probability}, \text{Deg\_freedom})$$



=CHISQ.DIST.RT()

CHISQ.DIST.RT

X = number

Deg_freedom = number

=

Returns the right-tailed probability of the chi-squared distribution.

X is the value at which you want to evaluate the distribution, a nonnegative number.

Formula result =

Help on this function

Figure 9.21 Screenshot of obtaining probability at the right tail (p) for given chi-square value and degrees of freedom

The screenshot after clicking the buttons Formulas $\Longrightarrow$ More Functions $\Longrightarrow$ Statistics $\Longrightarrow$ CHISQ.INV is shown in Figure 9.22, which gives the value of chi-square for a given cumulative probability from the left tail and degrees of freedom.

The Excel formula to compute the chi-square value at the right tail for a significance level and degrees of freedom is as follows [6].

$$= \text{CHISQ.INV.RT}(\text{Probability}, \text{Deg\_freedom})$$

The screenshot after clicking the buttons Formulas $\Longrightarrow$ More Functions $\Longrightarrow$ Statistics $\Longrightarrow$ CHISQ.INV.RT is shown in Figure 9.23, which gives the value of chi-square for a given significance level and degrees of freedom.

Example 9.4

Twenty-five dealers at a company are chosen at random from a normal population. The yearly sales standard deviation of dealers from the normal population and that of a randomly selected sample of 25 dealers are ₹ 70 lakh and ₹ 84 lakh, respectively.

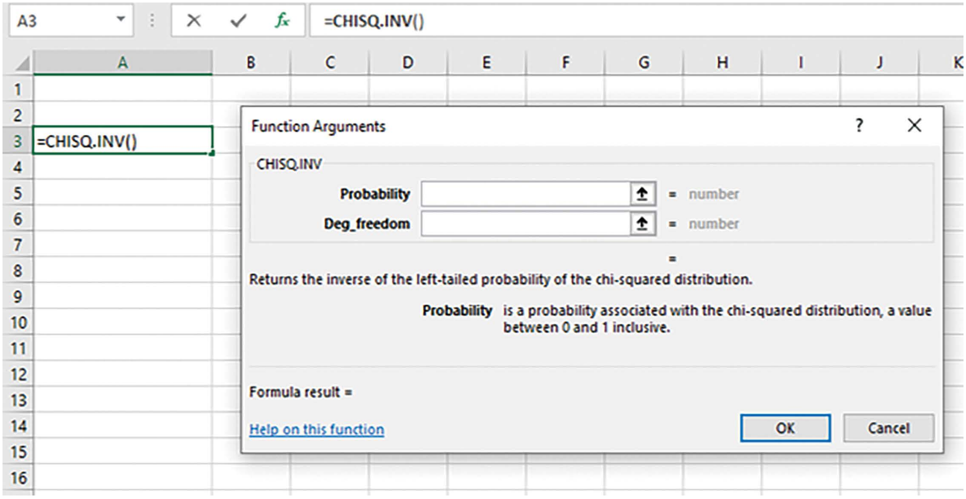
a) Compute the chi-square statistic.

b) Find the probability that the chi-square value is more than the calculated chi-square statistic.

Solution

Sample size, $n = 25$

Standard deviation of the mean annual sales of the population, $\sigma =$ ₹ 70 lakh



=CHISQ.INV()

Function Arguments

CHISQ.INV

Probability = number

Deg_freedom = number

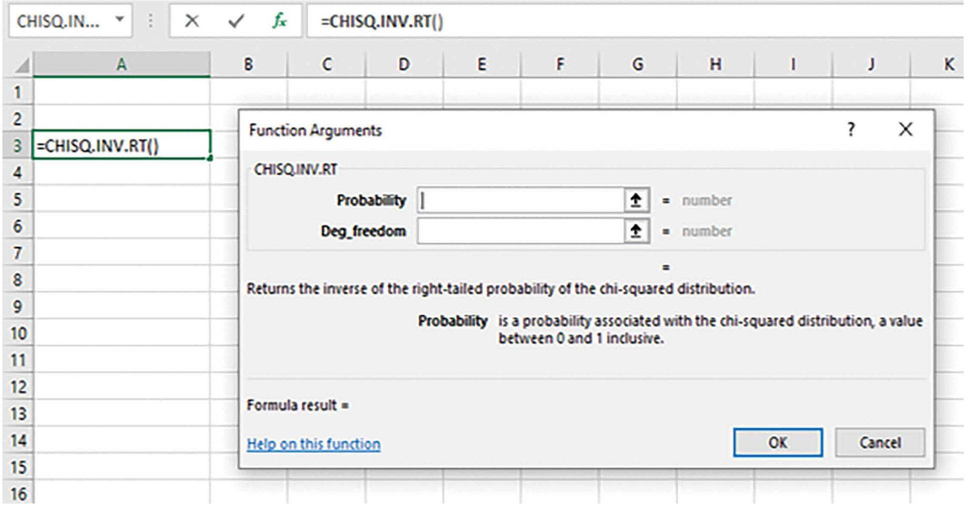
=

Returns the inverse of the left-tailed probability of the chi-squared distribution.

Probability is a probability associated with the chi-squared distribution, a value between 0 and 1 inclusive.

Formula result =

Figure 9.22 Screenshot of obtaining chi-square value for given cumulative probability from left tail and degrees of freedom



=CHISQ.INV.RT()

Function Arguments

CHISQ.INV.RT

Probability = number

Deg_freedom = number

=

Returns the inverse of the right-tailed probability of the chi-squared distribution.

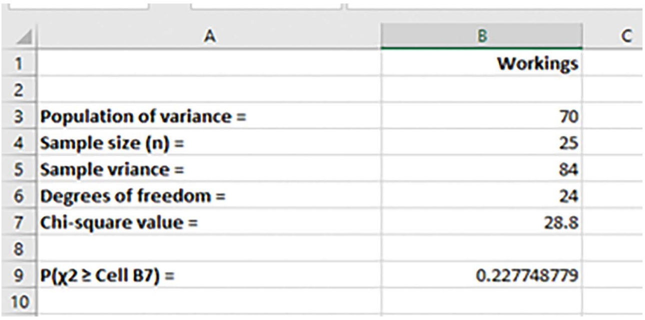
Probability is a probability associated with the chi-squared distribution, a value between 0 and 1 inclusive.

Formula result =

Figure 9.23 Screenshot of obtaining chi-square value at the right tail for given significance value and degrees of freedom

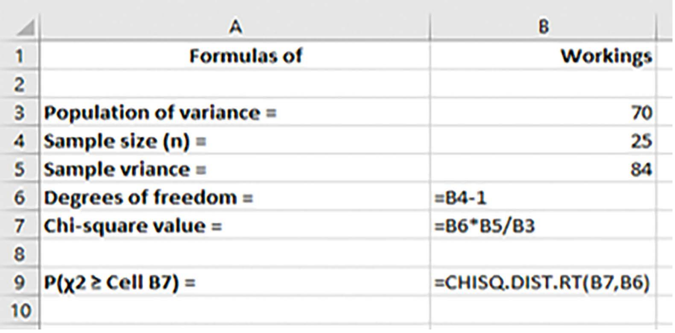
Standard deviation of the mean annual sales of the sample, S = ₹ 84 lakh

$$\chi^2 = \frac{(n-1)\times S^2}{\sigma^2} with (n-1) d.f = \frac{(25-1)\times 84^2}{70^2} \; with 24 d.f$$



	A	B	C
1		Workings	
2			
3	Population of variance =	70	
4	Sample size (n) =	25	
5	Sample vriance =	84	
6	Degrees of freedom =	24	
7	Chi-square value =	28.8	
8			
9	P(χ2 ≥ Cell B7) =	0.227748779	
10			

Figure 9.24 Screenshot of working of Example 9.4



	A	B
1	Formulas of	Workings
2		
3	Population of variance =	70
4	Sample size (n) =	25
5	Sample vriance =	84
6	Degrees of freedom =	=B4-1
7	Chi-square value =	=B6*B5/B3
8		
9	P(χ2 ≥ Cell B7) =	=CHISQ.DIST.RT(B7,B6)
10		

Figure 9.25 Screenshot of guidelines for formulas of working of Example 9.4

a) The answer to the following can be seen from Figure 9.24.

$$\chi^2 = \frac{(n-1)\times S^2}{\sigma^2} with (n-1) d.f = \frac{(25-1)\times 84^2}{70^2} with 24 d.f = ?$$

b) The answer to the following can be seen from Figure 9.24.

$$P\left(\chi^2 \geq \frac{(n-1)\times S^2}{\sigma^2}\right) = P\left(\chi^2 \geq \frac{(25-1)\times 84^2}{70^2}\right) = ?$$

The guidelines for the formulas of the working shown in Figure 9.24 are shown in Figure 9.25.

Example 9.5

Twenty-four respondents are chosen at random to represent a random sample from a normal population. The variance of the annual income of the respondents from the normal population and that of the 24 respondents chosen at random are ₹ 4 lakhs and ₹ 7 lakhs, respectively.

1. Compute the chi-square statistic.
2. Find the probability that the chi-square value is less than the calculated chi-square statistic.
3. Find the chi-square value for a significance level of 0.05 placed at the right tail.

Solution

Sample size, n = 24

Variance of the mean annual sales of the population, σ^2 = ₹ 4 lakhs

Variance of the mean annual sales of the sample, S^2 = ₹ 7 lakhs

$$\chi^2 = \frac{(n-1)\times S^2}{\sigma^2} with (n-1) d.f = \frac{(24-1)\times 7}{4} \; with 23 d.f$$

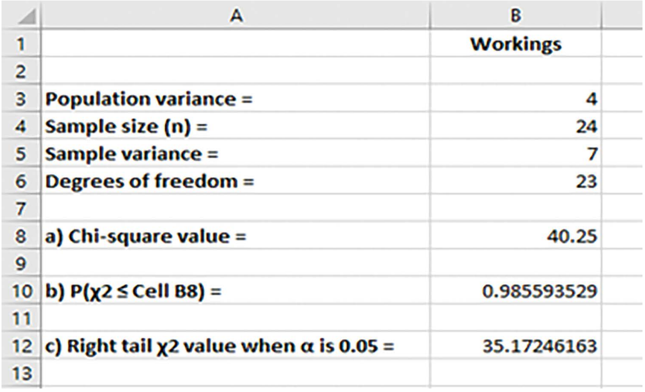
$$P\left(\chi^2 \le \frac{(24-1)\times 7}{4}\right) = ?$$

The answer for the following can be seen from Figure 9.26.

$$\chi^2 = \frac{(n-1)\times S^2}{\sigma^2} with (n-1) d.f = \frac{(24-1)\times 7}{4} = 40.25$$

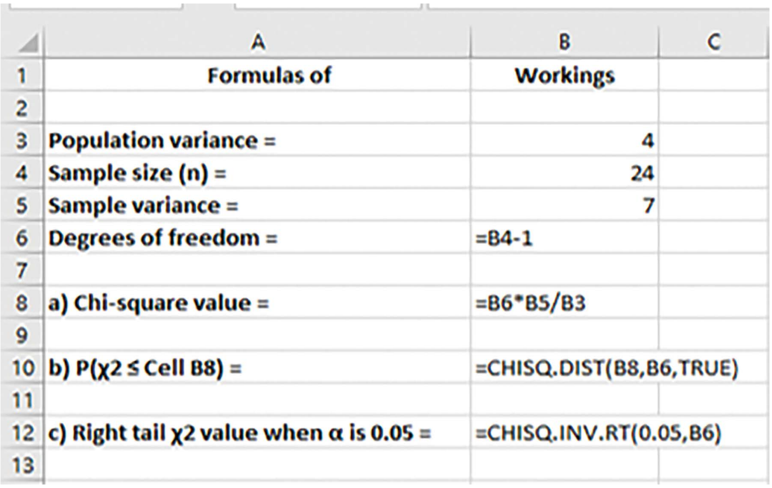
1. The answer for the following can be seen from Figure 9.26.

$$P\left(\chi^2 \le \frac{(n-1)\times S^2}{\sigma^2}\right) = P\left(\chi^2 \le \frac{(24-1)\times 7}{4}\right) = 0.985593529$$



	A	B
1		**Workings**
2		
3	**Population variance =**	**4**
4	**Sample size (n) =**	**24**
5	**Sample variance =**	**7**
6	**Degrees of freedom =**	**23**
7		
8	**a) Chi-square value =**	**40.25**
9		
10	**b) P(χ2 ≤ Cell B8) =**	**0.985593529**
11		
12	**c) Right tail χ2 value when α is 0.05 =**	**35.17246163**
13		

Figure 9.26 Screenshot of working of Example 9.5



	A	B	C
1	Formulas of	Workings	
2			
3	Population variance =	4	
4	Sample size (n) =	24	
5	Sample variance =	7	
6	Degrees of freedom =	=B4-1	
7			
8	a) Chi-square value =	=B6*B5/B3	
9			
10	b) P(χ2 ≤ Cell B8) =	=CHISQ.DIST(B8,B6,TRUE)	
11			
12	c) Right tail χ2 value when α is 0.05 =	=CHISQ.INV.RT(0.05,B6)	
13			

Figure 9.27 Screenshot of guidelines for formulas for working of Example 9.5

2. The answer for the following can be seen from Figure 9.26, which is 35.17246163.

χ^2 value when the significance value is 0.05 placed at the right tail with degrees of freedom 23.

The guidelines for the working in Figure 9.26 are shown in Figure 9.27.

9.3.2 F Distributions Using F.DIST, F.DIST.RT, F.INV, and F.INV.RT Functions

The F distribution is a ratio of two chi-square variables.

Let

n_1 be the size of sample 1
n_2 be the size of sample 2
S_1^2 be the variance of sample 1
S_2^2 by the variance of sample 2

Samples 1 and 2 are independent samples, which are taken from two normal populations with variances σ_1^2 and σ_2^2, respectively.

Then, the *F* distribution is the ratio of the two chi-square variables with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom as follows.

$$F = \frac{\left[\frac{\left(\frac{(n_1-1)\times S_1^2}{\sigma_1^2}\right)}{(n_1-1)}\right]}{\left[\frac{\left(\frac{(n_2-1)\times S_2^2}{\sigma_2^2}\right)}{(n_2-1)}\right]} \mathit{with}(n_1-1)\mathit{and}(n_2-1)\mathit{degrees\,of\,freedom}$$

The previous formula is reduced to:

$$F \quad = \frac{S_1^2}{S_2^2} \text{ with } (n_1 - 1) \text{ and } (n_2 - 1) \text{ degrees of freedom}$$

The Excel formula to compute the cumulative probability from left tail for a given value of F is as follows [7].

$$= \mathrm{F.DIST}\left(\mathrm{X, Deg\_freedom1, Deg\_freedom2.TRUE}\right)$$

The screenshot after clicking the buttons HOME ⟹ Formulas ⟹ More Functions ⟹ Statistic ⟹ F.DIST is shown in Figure 9.28, which gives the cumulative probability from the left tail for a given F value.

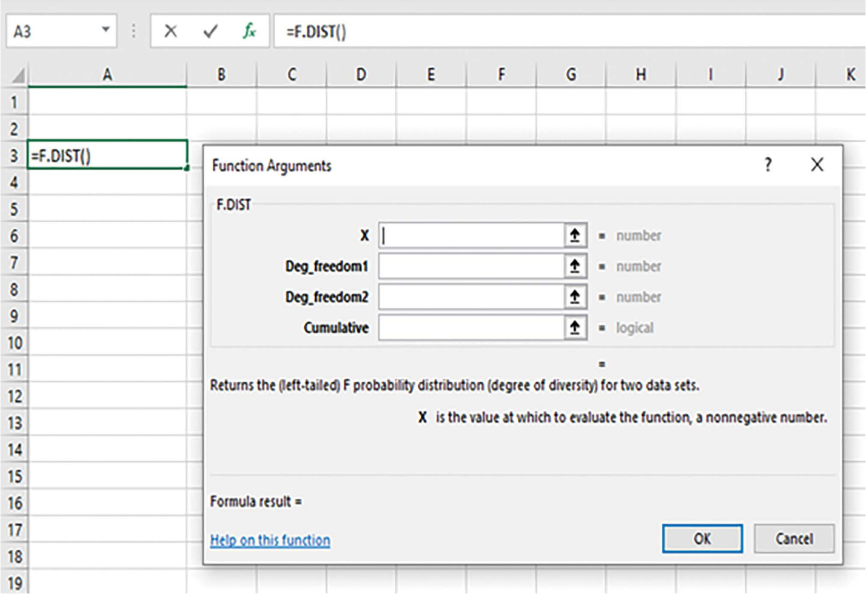
The Excel formula to compute the probability at the right tail for a given value of F is as follows [8].

$$= \mathrm{F.DIST.RT}\left(\mathrm{X, Deg\_freedom1,\ Deg\_freedom2}\right)$$

The screenshot after clicking the buttons HOME ⟹ Formulas ⟹ More Functions ⟹ Statistics ⟹ F.DIST.RT is shown in Figure 9.29, which gives the probability at the right tail for a given value of F.

The Excel formula to compute the F value for a given cumulative probability from the left tail is as follows.

$$= \mathrm{F.INV}\left(\mathrm{Probability, Deg\_freedom1,\ Deg\_freedom2}\right)$$



A3 =F.DIST()

3 =F.DIST()

Function Arguments

F.DIST

X = number

Deg_freedom1 = number

Deg_freedom2 = number

Cumulative = logical

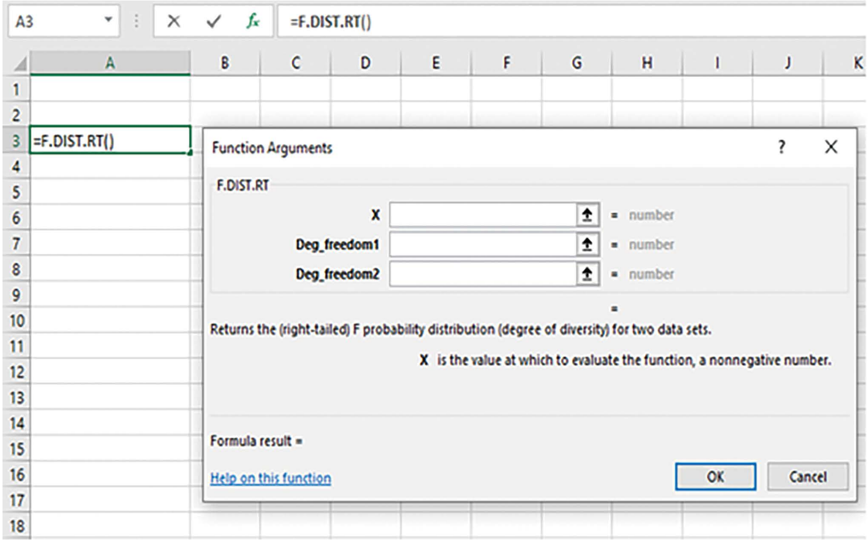
=

Returns the (left-tailed) F probability distribution (degree of diversity) for two data sets.

X is the value at which to evaluate the function, a nonnegative number.

Formula result =

Figure 9.28 Screenshot of clicking the sequence of buttons HOME ⟹ More Functions ⟹ Statistical ⟹ F.DIST



=F.DIST.RT()

Function Arguments

F.DIST.RT

X = number

Deg_freedom1 = number

Deg_freedom2 = number

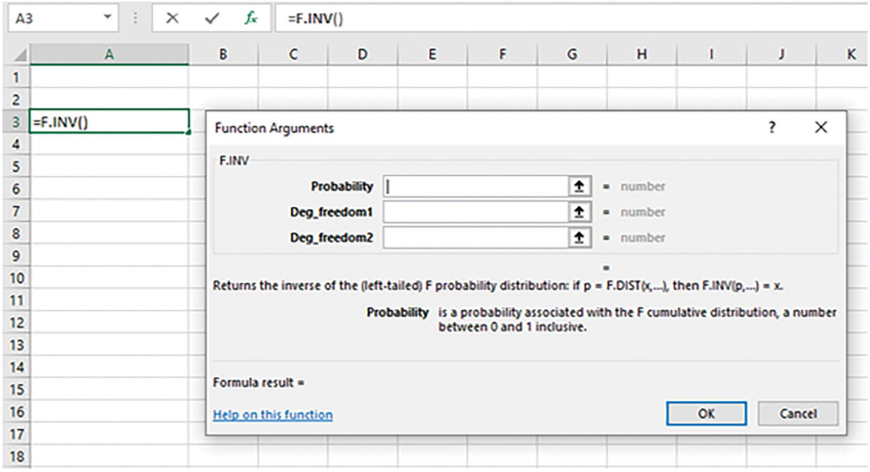
Returns the (right-tailed) F probability distribution (degree of diversity) for two data sets.

X is the value at which to evaluate the function, a nonnegative number.

Formula result =

Help on this function

Figure 9.29 Screenshot of clicking the sequence of buttons HOME ⟹ More Functions ⟹ Statistical ⟹.DIST.RT



=F.INV()

Function Arguments

F.INV

Probability = number

Deg_freedom1 = number

Deg_freedom2 = number

Returns the inverse of the (left-tailed) F probability distribution: if p = F.DIST(x,...), then F.INV(p,...) = x.

Probability is a probability associated with the F cumulative distribution, a number between 0 and 1 inclusive.

Formula result =

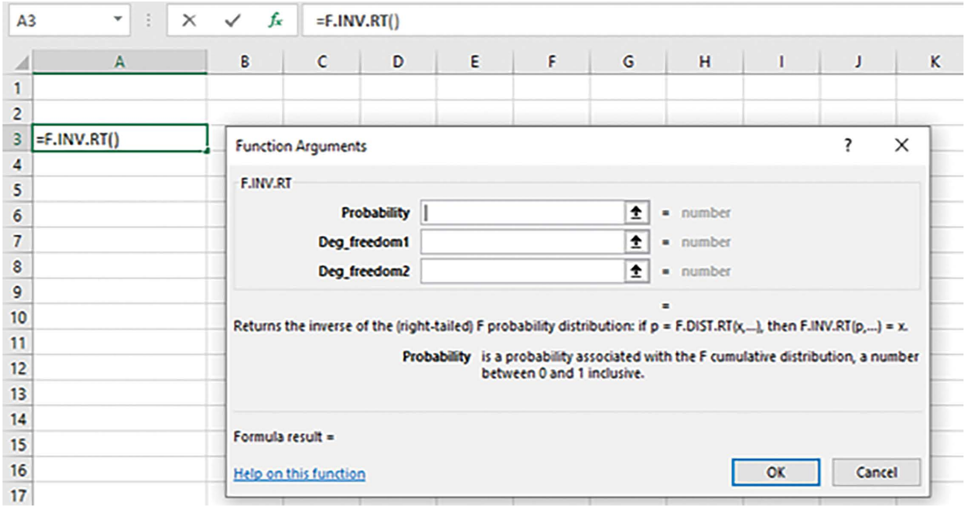
Help on this function

Figure 9.30 Screenshot of clicking the sequence of buttons HOME ⟹ More Functions ⟹ Statistical ⟹ F.INV

The screenshot after clicking the buttons Formulas ⟹ More Functions ⟹ Statistics ⟹ F.INV is shown in Figure 9.30, which gives the value of *F* for a given cumulative probability from the left tail.

The Excel formula to compute the *F* value at the right tail for a significance level (α) is as follows [9, 10].

$$= \text{F.INV.RT}(\text{Probability}, \text{Deg\_freedom1}, \text{Deg\_freedom2})$$



=F.INV.RT()

F.INV.RT

Probability = number

Deg_freedom1 = number

Deg_freedom2 = number

=

Returns the inverse of the (right-tailed) F probability distribution: if p = F.DIST.RT(x,...), then F.INV.RT(p,...) = x.

Probability is a probability associated with the F cumulative distribution, a number between 0 and 1 inclusive.

Formula result =

Figure 9.31 Screenshot of clicking the sequence of buttons HOME ⟹ More Functions ⟹ Statistical ⟹ F.INV.RT

The screenshot after clicking the buttons Formulas ⟹ More Functions ⟹ Statistics ⟹ F.INV.RT is shown in Figure 9.31, which gives the value of *F* at the right tail for a given significance level (α).

Example 9.6

Two independent samples of vendors' goods are drawn from normal populations with the same variance. The first sample has a size of 13, while the second sample has a size of 21. The variances for the first and second samples, respectively, are 100 and 250. What is the probability that the *F* is less than or equal to $S_1^{2/}S_2^{2}$?

Solution

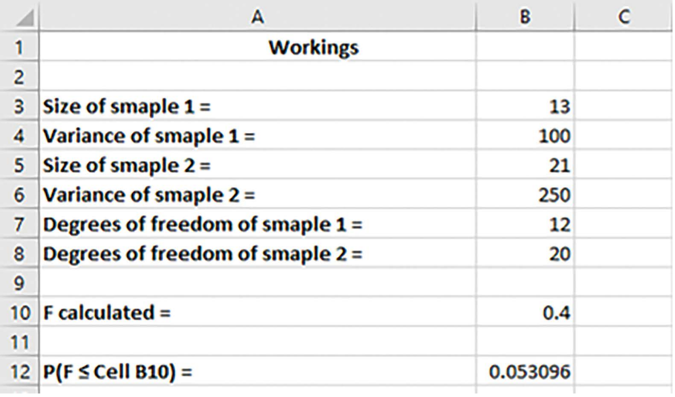
$\sigma_1^2 = \sigma_2^2$
$n_1 = 13$ and $n_2 = 21$
$S_1^2 = 100$
$S_2^2 = 250$

$$F = \frac{S_1^2}{S_2^2} \text{ with } (n_{1\_}1) \text{ and } (n_{2\_}1) \text{ d.f.}$$

$$P\left(F \le \frac{S_1^2}{S_2^2}\right) = ?$$

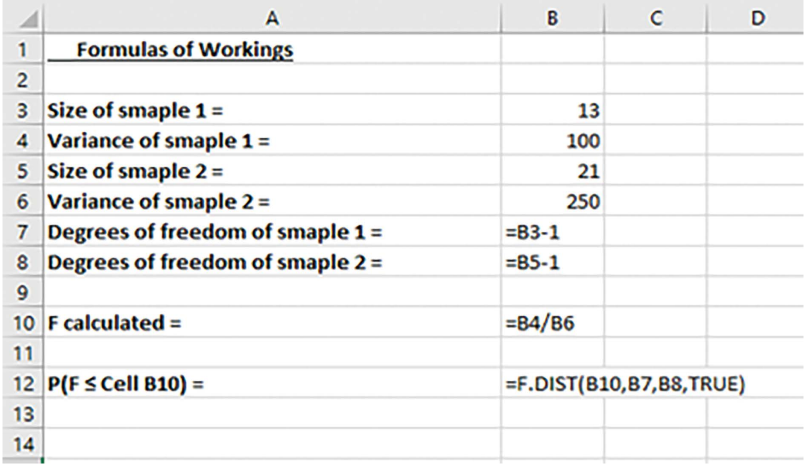
The working of the formula is shown in Figure 9.32. The guidelines for the formulas for the working shown in Figure 9.32 are shown in Figure 9.33

From Figure 9.32, P($F \le 100/250$) = 0.041790339.



	A	B	C
1	**Workings**		
2			
3	Size of smaple 1 =	13	
4	Variance of smaple 1 =	100	
5	Size of smaple 2 =	21	
6	Variance of smaple 2 =	250	
7	Degrees of freedom of smaple 1 =	12	
8	Degrees of freedom of smaple 2 =	20	
9			
10	F calculated =	0.4	
11			
12	P(F ≤ Cell B10) =	0.053096	

Figure 9.32 Screenshot of working of Example 9.6



	A	B	C	D
1	Formulas of Workings			
2				
3	Size of smaple 1 =	13		
4	Variance of smaple 1 =	100		
5	Size of smaple 2 =	21		
6	Variance of smaple 2 =	250		
7	Degrees of freedom of smaple 1 =	=B3-1		
8	Degrees of freedom of smaple 2 =	=B5-1		
9				
10	F calculated =	=B4/B6		
11				
12	P(F ≤ Cell B10) =	=F.DIST(B10,B7,B8,TRUE)		
13				
14				

Figure 9.33 Screenshot of guidelines for formulas of working of Example 9.6

Example 9.7

Two independent samples of diabetics with the same variance are drawn from normal populations. The first sample's size and blood sugar level variance are 10 and 780, respectively. The second sample's size and blood sugar level variation are 21 and 225, respectively.

1. What is the probability that the F-ratio is less than the calculated F statistic?
2. What is the probability that the F-ratio is more than the calculated F statistic?
3. What is the value of *F* when the cumulative probability from the left is 0.85 for the degrees of freedom of this problem?
4. What is the value of *F* when the significance level (α) is 0.05?

Solution

$\sigma_1^2 = \sigma_2^2$

$n_1 = 10, s_1^2 = 780$ & $n_2 = 21, s_2^2 = 225$

$$F_{Cal} = \frac{S_1^2}{S_2^2} \text{ with } (n_{1-}1) \text{ and } (n_{2-}1) \text{ d.f.}$$

$$= \frac{780}{225}$$

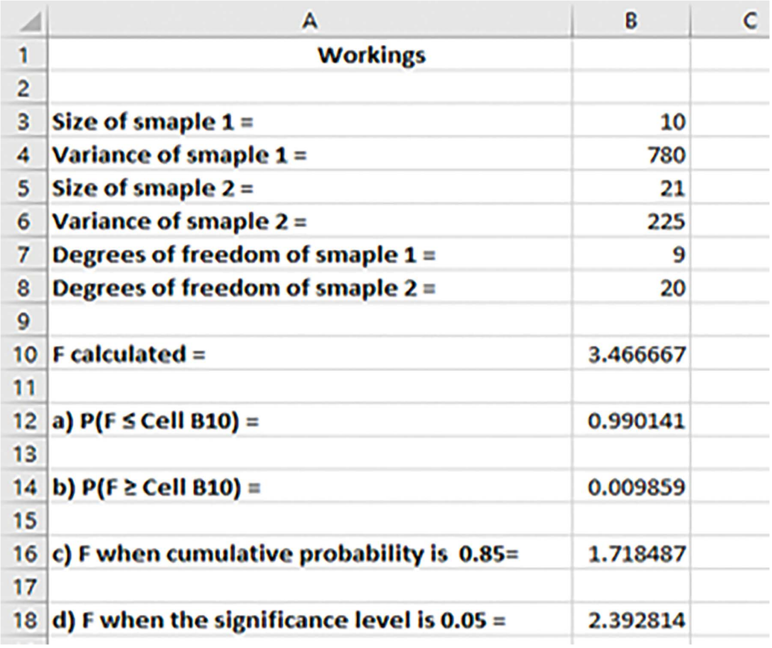
The working of the formula is shown in Figure 9.34. The guidelines for the formulas for the working shown in Figure 9.34 are shown in Figure 9.35.

1. $P(F \le F_{Cal}) = ?$
2. $P(F \ge F_{Cal}) = ?$
3. What is the value of F when the cumulative probability is 0.85?
4. What is the value of F when the significance level is 0.05?

The answers for all these can be seen in Figure 9.34.

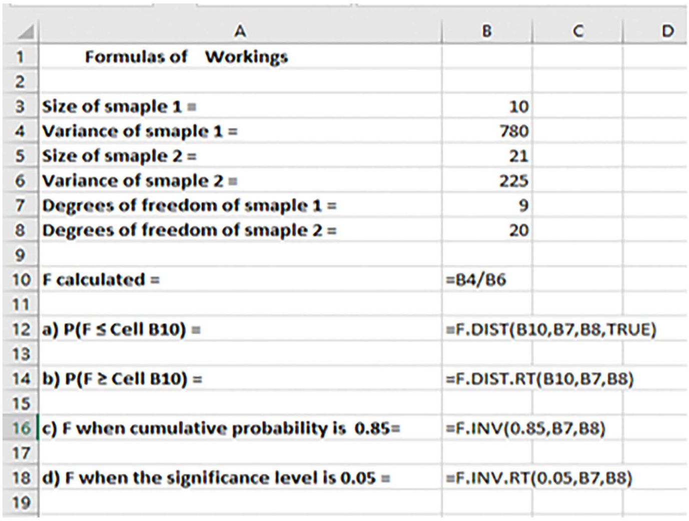
Summary

- If $X \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$



	A	B	C
1	Workings		
2			
3	Size of smaple 1 =	10	
4	Variance of smaple 1 =	780	
5	Size of smaple 2 =	21	
6	Variance of smaple 2 =	225	
7	Degrees of freedom of smaple 1 =	9	
8	Degrees of freedom of smaple 2 =	20	
9			
10	F calculated =	3.466667	
11			
12	a) P(F ≤ Cell B10) =	0.990141	
13			
14	b) P(F ≥ Cell B10) =	0.009859	
15			
16	c) F when cumulative probability is 0.85=	1.718487	
17			
18	d) F when the significance level is 0.05 =	2.392814	

Figure 9.34 Screenshot of working of Example 9.7



	A	B	C	D
1	Formulas of Workings			
2				
3	Size of smaple 1 =	10		
4	Variance of smaple 1 =	780		
5	Size of smaple 2 =	21		
6	Variance of smaple 2 =	225		
7	Degrees of freedom of smaple 1 =	9		
8	Degrees of freedom of smaple 2 =	20		
9				
10	F calculated =	=B4/B6		
11				
12	a) P(F ≤ Cell B10) =	=F.DIST(B10,B7,B8,TRUE)		
13				
14	b) P(F ≥ Cell B10) =	=F.DIST.RT(B10,B7,B8)		
15				
16	c) F when cumulative probability is 0.85=	=F.INV(0.85,B7,B8)		
17				
18	d) F when the significance level is 0.05 =	=F.INV.RT(0.05,B7,B8)		
19				

Figure 9.35 Screenshot of guidelines for formulas of working of Example 9.7

then the standard normal statistic for $\bar{X}$ bar is: $\dfrac{\bar{X}-\mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$

- The sampling distribution of the mean with an infinite population has a variance of σ^2/n. But the variance of a sampling distribution with a finite population will have a different variance, which can be obtained from the variance of the sampling distribution with an infinite population by multiplying by a finite population multiplier
- If $\mathrm{X} \sim \mathrm{N}\left(\mu, \sigma^2\right)$, then $\bar{\mathrm{X}} \sim \mathrm{N}\left(\mu, \dfrac{\sigma^2}{\mathrm{n}}\left(\dfrac{N-n}{N-1}\right)\right)$
- The standard normal statistic $Z_{\bar{X}}$ for $\bar{X}$ is $\dfrac{\bar{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}}$
- If the variance of the normal population is unknown, then the corresponding sampling distribution is the student's *t*-distribution.
- The distribution of S^2 taken from a normal population with variance σ^2 is called the chi-square (χ^2) distribution with (n – 1) degrees of freedom.
- The F distribution is a ratio of two chi-square variables.
- The formula for the F distribution is:

$$F = \frac{S_1^2}{S_2^2} \text{ with } (\mathrm{n}_1 - 1) \text{ and } (n_2 - 1) \text{ degrees of freedom}$$

Keywords

- A sampling distribution of the mean with an infinite population has a variance of σ^2/n. But the variance of the sampling distribution with a finite population will have a different variance, which can be obtained from the variance of the sampling distribution with an infinite population by multiplying by a finite population multiplier.

- T-distribution is for a normal population when its variance is unknown.
- The chi-square distribution is a distribution when the distribution of S^2 is taken from a normal population with variance σ^2, and it has $(n - 1)$ degrees of freedom, where n is the sample size.
- The F distribution is a ratio of two chi-square variables.

Review Questions

1. Distinguish between a sample and population.
2. Give the statistic for the sampling distribution of the mean when the population is infinite, with an explanation.
3. Give the formula for the statistic of the sampling distribution of the mean when the population is infinite to compute the cumulative probability in Excel.
4. From an infinite normal population with a mean and variance of 250 and 800, respectively, a random sample of size 75 is taken. Using Excel, determine the probability that the sample mean is higher than 205.
5. Give the statistic for the sampling distribution of the mean when the population is finite, with an explanation.
6. A random sample of size 54 is taken from a finite normal population of size 1400 which has its mean and variance as 90 and 70, respectively. Answer each of the following questions using Excel.

   a. What is the probability that the sample mean is less than 95?
   b. What is the probability that the sample mean is greater than 85?
   c. What is the probability that the sample mean will be between 84 and 96?

7. Give the statistic for the sampling distribution of the mean when the normal population variance is unknown (*t* distribution), with an explanation.
8. A random sample of 25 customers of a company is taken from a normal population. The mean of the annual purchases made by the customers of the population is ₹ 50 lakh. The variance of the purchase made by the customers of the sample is ₹ 95 lakhs. Using Excel, find the probability that the mean annual sales of the sample is:

   a. Less than ₹ 60 lakhs.
   b. More than ₹ 40 lakhs.
   c. Find the value *t* for a cumulative probability of 0.92.
   d. Find the value of *t* for a given level of significance of 0.10 by placing half of it on the right tail.
   e. Find the probability at both tails when the annual sales are less than 55 lakhs.

9. Twenty prominent academicians from an institution are chosen at random from a normal population to form a random sample. The variance of the annual number of publications in journals from the normal population and that of the random sample of 20 academicians are ₹ 75 lakhs and 90 lakhs, respectively. Answer each of the following questions using Excel.

   a. Compute the chi-square statistic.
   b. Find the probability that the chi-square value is more than the calculated chi-square statistic.

10. Give the statistic of the F distribution, with an explanation.
11. Give the different functions of the F distribution in Excel and explain the arguments in each of them.
12. Two independent samples of vendors' goods are drawn from normal populations with the same variance. The first sample has a size of 15, while the second sample has a size of 22. The variances of the first and second samples, respectively, are 120 and 270. What is the probability that F is less than or equal to S_1^2/S_2^2?

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. https://support.microsoft.com/en-us/office/norm-s-dist-function-1e787282-3832-4520-a9ae-bd2a8d99ba88 [June 27, 2020].
3. www.Excelfunctions.net/Excel-norm-inv-function.html.
4. https://support.microsoft.com/en-us/office/tdist-function-630a7695-4021-4853-9468-4a1f9d-cdd192 [June 27, 2020].
5. https://support.microsoft.com/en-us/office/chisq-dist-rt-function-dc4832e8-ed2b-49ae-8d7c-b28d5804c0f2 [June 27, 2020].
6. https://support.microsoft.com/en-us/office/chisq-inv-rt-function-435b5ed8-98d5-4da6-823f-293e2cbc94fe [June 27, 2020].
7. https://support.microsoft.com/en-us/office/f-dist-function-a887efdc-7c8e-46cb-a74a-f884cd29b25d [June 27, 2020].
8. https://support.microsoft.com/en-us/office/f-dist-rt-function-d74cbb00-6017-4ac9-b7d7-6049badc0520.
9. www.Excelfunctions.net/Excel-f-inv-rt-function.html.
10. https://support.microsoft.com/en-us/office/f-inv-rt-function-d371aa8f-b0b1-40ef-9cc2-496f0693ac00.

10 Testing Hypotheses

Learning Objectives

The study of this chapter will enable the readers to

- Understand the concept of a hypothesis and its importance in sampling.
- Analyse real-life problems through tests concerning a single mean when the mean and variance of the populations are known and the size of the population is finite using Z sampling statistics.
- Understand the process of testing hypotheses concerning the difference between two means when the variances of the populations are known and the sample sizes are large using Z sampling statistics.
- Study the tests of hypotheses concerning a single mean when the variance of the population is unknown and the sample size is small using t tests.
- Analyse real-life problems through tests concerning the difference between two means when the variances of the population are unknown and equal and the sample size is small using t tests.
- Analyse real-life problems through tests concerning the difference between two means when the variances of the population are unknown and unequal and the sample size is small using t tests.
- Study testing hypotheses of real-life problems using the paired t test.

10.1 Introduction

A hypothesis an assumption about a population. The two categories of hypotheses are the null hypothesis and alternate hypothesis. In the null hypothesis ($H_o : \bar{X} \leq k$), an assumption about the population will be made, like the sample mean is less than a specified constant. The alternate hypothesis (H_1: $\bar{X} \leq > k$) is opposite the null hypothesis, say, the sample mean is more than the specified constant. In data analysis, three types of hypothesis testing (≤, ≥, and =) will be carried out. In reality, based on the mean(s) and variance(s) of the sample(s) collected from the population, the investigator has to take a decision to accept or reject the null hypothesis. In this process, a significance level (α) is assumed.

DOI: 10.4324/9781032671703-10

The decision on whether to accept or reject the null hypothesis may be based on the following.

1. The value of the statistic computed will be compared with the critical value of the statistic for a given significance level for taking the decision.

   If the computed value of the statistic is more than the corresponding table value, reject the null hypothesis; otherwise, accept the null hypothesis.

2. The value of the computed *p* (less than 1) will be compared with a given significance level for taking the decision.

   If the *p* value is less than the significance level (α), reject the null hypothesis; otherwise, accept the null hypothesis.

Further, hypothesis testing is classified into hypothesis testing for the mean of a single sample and hypothesis testing for the difference between the means of two samples [1].

10.2 Tests Concerning Single Mean When Mean and Variance of the Populations Are Known and the Size of the Population Is Finite Using the Norm.S.Dist Function

Let

X be a random variable, which follows a normal distribution, representing an independent population

μ and σ^2 be the mean and variance of the population, and these are known for the population parameters

n be the size of the population, which is finite

The hypotheses relating to two different one-tailed tests and a two-tailed test are listed in the following.

Type 1 Test:

H_0: $\mu \leq k$
H_1: $\mu > k$

In this type of hypothesis testing, the significance level (α) is placed at the right tail of the distribution.

Type 2 Test:

H_0: $\mu \geq k$
H_1: $\mu < k$

In this type of hypothesis testing, the significance level (α) is placed at the left tail of the distribution.

Type 3 Test:

H_0: $\mu = k$
H_1: $\mu \neq k$

In this type of hypothesis testing, half of the significance level ($\alpha/2$) is placed at both tails of the distribution.

Let

X be the mean of sample whose size is n.

The values of n be large and greater than 30.

The formulas for the mean and variance of the normal distribution for the sample mean are given as follows.

Mean of the sample mean $\bar{X} \leq \mu$

Variance of the sample mean, $\sigma_{\bar{X}}^2 = \dfrac{\sigma^2}{n}$

Definition of Sampling Distribution for Mean

Let

$X \sim N(\mu, \sigma^2)$

Then the distribution of $\bar{X}$ is as follows.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The standard normal statistic for the sample mean is as follows.

$$Z_{\bar{X}} = \frac{\left(\bar{X} - \mu\right)}{\sigma_{\bar{X}}} = \frac{\left(\bar{X} - \mu\right)}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$$

Example 10.1

The monthly salary of survey respondents is distributed normally with a finite population size of 1000. The average monthly income of the population's responses is estimated to be ₹ 8000. The variance of the population's respondents' monthly incomes is ₹ 3,000,000 lakhs. The researcher believes that the respondents' mean monthly income has decreased from the projected mean of ₹ 8000 in recent years. The average monthly income for the normal population, which is found to be ₹ 8750, is estimated from a random sample of 64 respondents. Using a significance value of 0.01, check whether the mean monthly income has decreased from the predicted mean of ₹ 8000.

Solution

The monthly income of respondents, $X \sim N(\mu, \sigma^2)$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

μ = ₹ 8,000 lakhs, σ^2 = ₹ 3,00,000 lakhs

$n = 64$, $\bar{X}$ = ₹ 8,750 lakhs, $N = 1000$, $\alpha = 0.01$

H_0: $\mu \leq 8000$ & $H_{1:}$ $\mu > 8000$

$$Z_{\bar{X}} = \frac{\left(\bar{X} - \mu\right)}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{8750 - 8000}{\left(\frac{\sqrt{300000}}{\sqrt{64}}\right)}$$

The decision situation is shown in Figure 10.1. If the p value shown at the right tail of Figure 10.1 is less than α, then reject the null hypothesis; otherwise, accept the null hypothesis.

The formula to get the cumulative probability for a given Z is:

= NORM.S.DIST(Z, TRUE)

The Excel screenshot of the output of the Z-test for one sample mean of Example 10.1 is shown in Figure 10.2, and a screenshot of the formulas for the working is shown in Figure 10.3. From Figure 10.2, the probability that Z_X is less than or equal to 10.95445 is 1. This means that p is 0, which is 1 – 1: 0. This is less than α (0.01). Hence, the null hypothesis is to be rejected.

Inference: The mean monthly income has not declined from the expected mean of ₹ 8000.

Example 10.2

A product's daily sales in several stores follows a normal distribution. The product's targeted daily sales mean and population variance are ₹ 10,000 and ₹ 50,000, respectively. The sales manager of the firm that makes the product believes that recent improvements have been made in the product's sales. From the normal population, a random sample of 49 stores is chosen, and it is discovered that their mean daily sales is ₹ 9950. At a

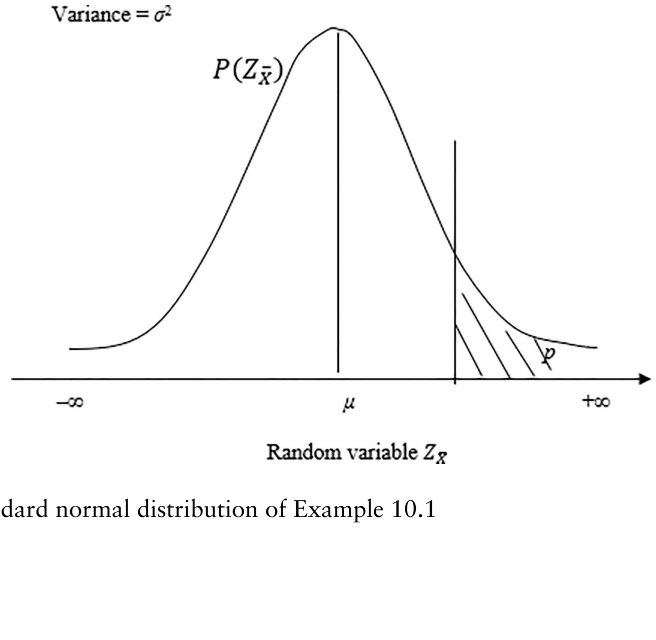
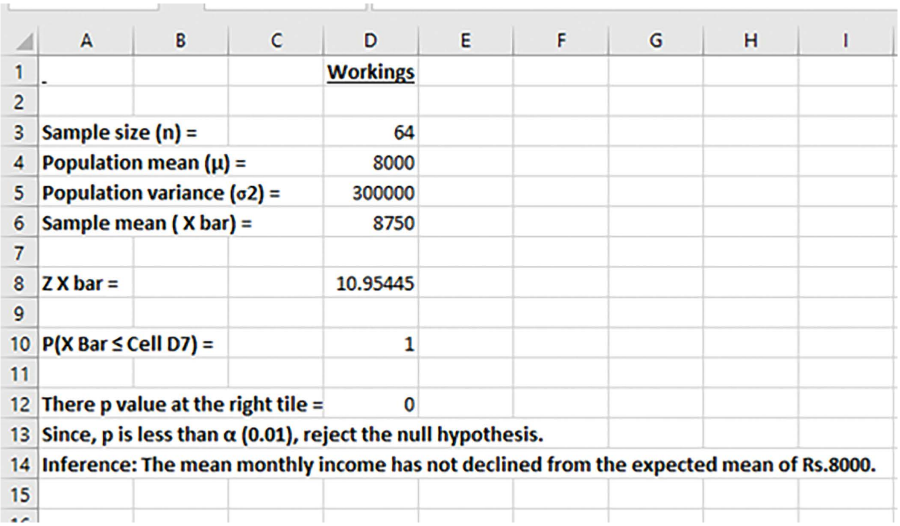
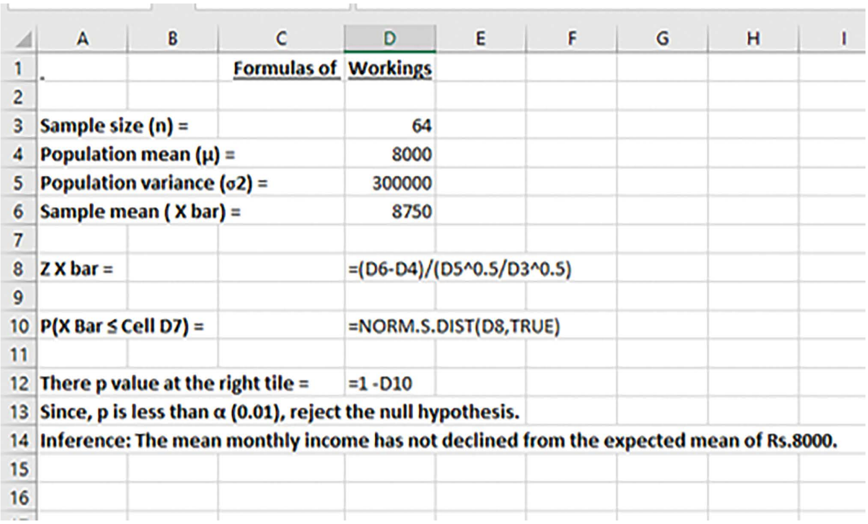


Figure 10.1 Standard normal distribution of Example 10.1



	A	B	C	D	E	F	G	H	I
1	-			**Workings**					
2									
3	**Sample size (n) =**			64					
4	**Population mean (μ) =**			8000					
5	**Population variance (σ2) =**			300000					
6	**Sample mean (X bar) =**			8750					
7									
8	**Z X bar =**			10.95445					
9									
10	**P(X Bar ≤ Cell D7) =**			1					
11									
12	**There p value at the right tile =**			0					
13	**Since, p is less than α (0.01), reject the null hypothesis.**								
14	**Inference: The mean monthly income has not declined from the expected mean of Rs.8000.**								
15									

Figure 10.2 Screenshot of working of Example 10.1



	A	B	C	D	E	F	G	H	I
1	-		**Formulas of**	**Workings**					
2									
3	**Sample size (n) =**			64					
4	**Population mean (μ) =**			8000					
5	**Population variance (σ2) =**			300000					
6	**Sample mean (X bar) =**			8750					
7									
8	**Z X bar =**			=(D6-D4)/(D5^0.5/D3^0.5)					
9									
10	**P(X Bar ≤ Cell D7) =**			=NORM.S.DIST(D8,TRUE)					
11									
12	**There p value at the right tile =**			=1 -D10					
13	**Since, p is less than α (0.01), reject the null hypothesis.**								
14	**Inference: The mean monthly income has not declined from the expected mean of Rs.8000.**								
15									
16									

Figure 10.3 Screenshot of formulas for working of Example 10.1

significance level of 0.05, determine whether the product's sales have really improved in various stores.

Solution

The daily sales of a product, $X \sim N(\mu, \sigma^2)$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

μ = ₹ 10,000, σ^2 = ₹ 50,000
$n = 49$, $\bar{X}$ = ₹ 9, 950, $\alpha = 0.05$
H_0: $\mu \geq 10{,}000$
H_1: $\mu < 10{,}000$

$$Z_{\bar{X}} = \frac{(\bar{X} - \mu)}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{9950 - 10000}{\left(\frac{\sqrt{50000}}{\sqrt{49}}\right)}$$

The decision situation is shown in Figure 10.4. If the *p* value shown at the left tail of Figure 10.4 is less than α, then reject the null hypothesis; otherwise, accept the null hypothesis.

The formula to get the cumulative probability for a given *Z* is: =NORM.S.DIST(Z,TRUE)

The screenshot of the output of the *Z*-test for one sample mean of Example 10.2 is shown in Figure 10.5, and the screenshot of the formulas for the working is shown in Figure 10.6. From Figure 10.5, it can be seen that the *p* value at the left tail is 0.058762434.

The *p* value is more than α (0.05). Hence, the null hypothesis is to be accepted.

Inference: The sales of the product in different shops have improved.

Example 10.3

Small businesses in an industrial park have annual sales that are distributed normally. The industries in the population are expected to generate an average annual revenue of ₹ 75 lakhs. There is a variance of ₹ 400 lakhs for the industries' yearly sales. According to

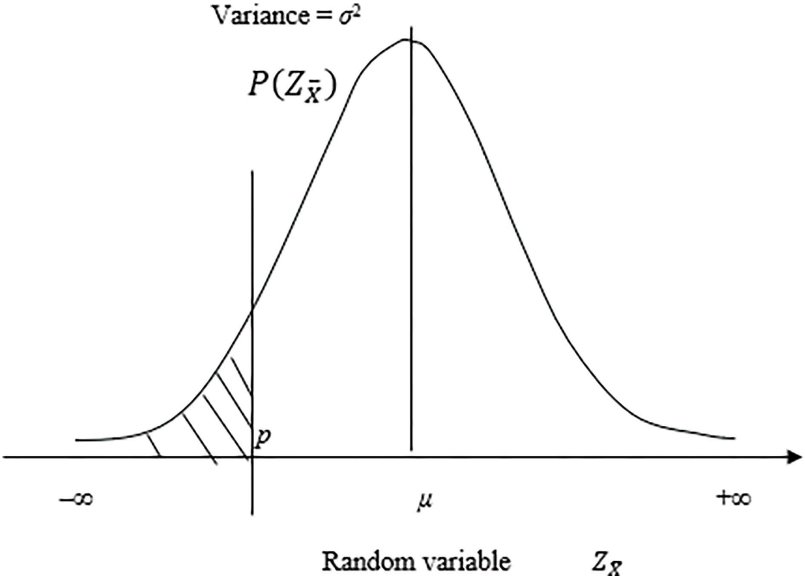
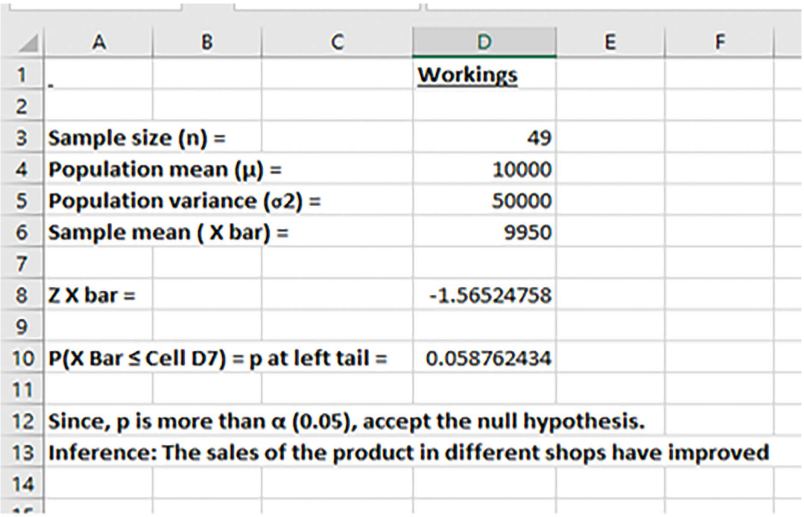
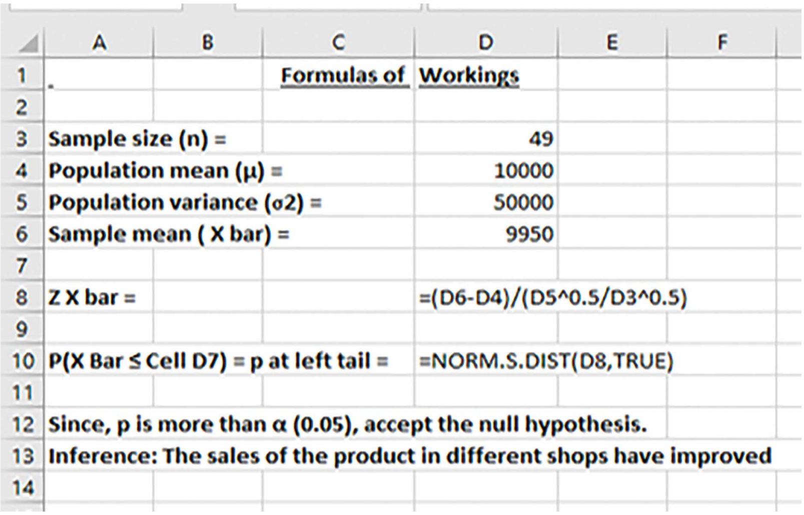


Figure 10.4 Standard normal distribution of Example 10.2



	A	B	C	D	E	F
1	-			Workings		
2						
3	Sample size (n) =			49		
4	Population mean (μ) =			10000		
5	Population variance (σ2) =			50000		
6	Sample mean (X bar) =			9950		
7						
8	Z X bar =			-1.56524758		
9						
10	P(X Bar ≤ Cell D7) = p at left tail =			0.058762434		
11						
12	Since, p is more than α (0.05), accept the null hypothesis.					
13	Inference: The sales of the product in different shops have improved					
14						

Figure 10.5 Screenshot of working of Example 10.2



	A	B	C	D	E	F
1	-		Formulas of	Workings		
2						
3	Sample size (n) =			49		
4	Population mean (μ) =			10000		
5	Population variance (σ2) =			50000		
6	Sample mean (X bar) =			9950		
7						
8	Z X bar =			=(D6-D4)/(D5^0.5/D3^0.5)		
9						
10	P(X Bar ≤ Cell D7) = p at left tail =			=NORM.S.DIST(D8,TRUE)		
11						
12	Since, p is more than α (0.05), accept the null hypothesis.					
13	Inference: The sales of the product in different shops have improved					
14						

Figure 10.6 Screenshot with formulas for working of Example 10.2

the director of the industry department, recent trends in industry performance have not deviated much from the population mean. The mean yearly sales for a random sample of 49 industries from the normal population are found to be ₹ 70 lakhs. Check to see if the industries' sales have not changed from ₹ 75 lakhs at a significance level of 0.05.

Solution

Sample size, n = 49

Annual sales of small-scale industries, $X \sim N(\mu, \sigma^2)$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{\sqrt{n}}\right)$$

μ = ₹ 75 lakhs, σ^2 = ₹ 400 lakhs
$n = 49$, $\bar{X}$ = ₹ 70 lakhs, $\alpha = 0.05$
H_0: $\mu = 75$
H_1: $\mu \neq 75$

$$Z_{\bar{X}} = \frac{\left(\bar{X} - \mu\right)}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{70 - 75}{\left(\frac{\sqrt{400}}{\sqrt{49}}\right)}$$

The decision situation is shown in Figure 10.7. Here, the significance level α is distributed equally to both tails. Hence, the significance level at each tail is $\alpha/2$. If the p value shown at the right tail of Figure 10.7 is less than half of the significance level ($\alpha/2$), then reject the null hypothesis; otherwise, accept the null hypothesis. If the p value shown at the left tail of Figure 10.7 is less than half of the significance level ($\alpha/2$), then reject the null hypothesis; otherwise, accept the null hypothesis.

The formula to get the cumulative probability for given Z is [3]: = NORM.S.DIST(Z, TRUE)

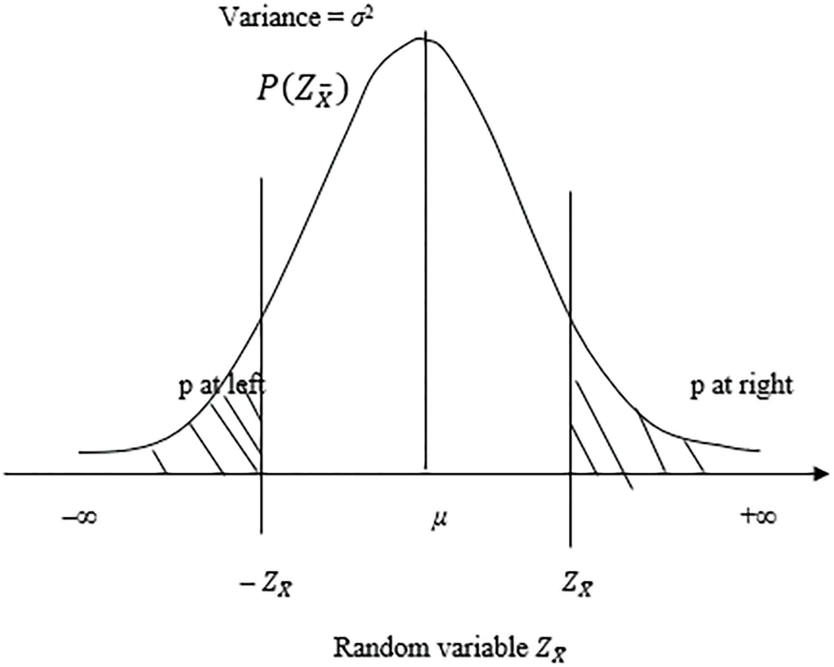
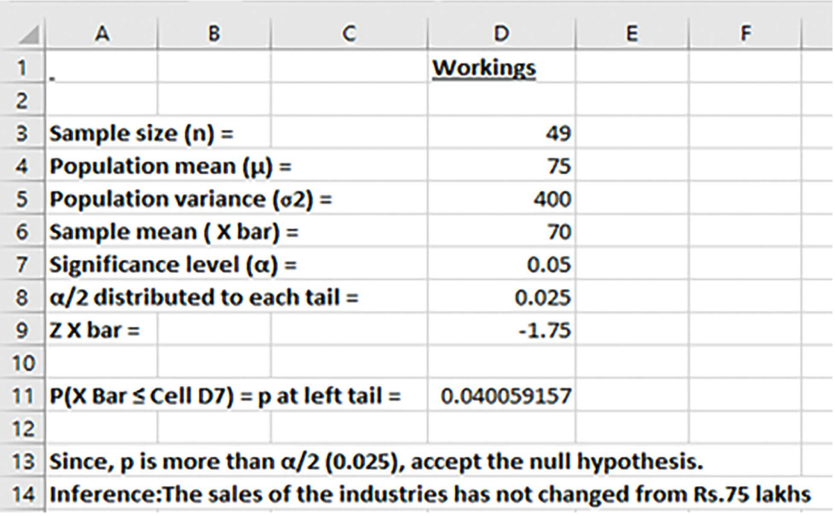
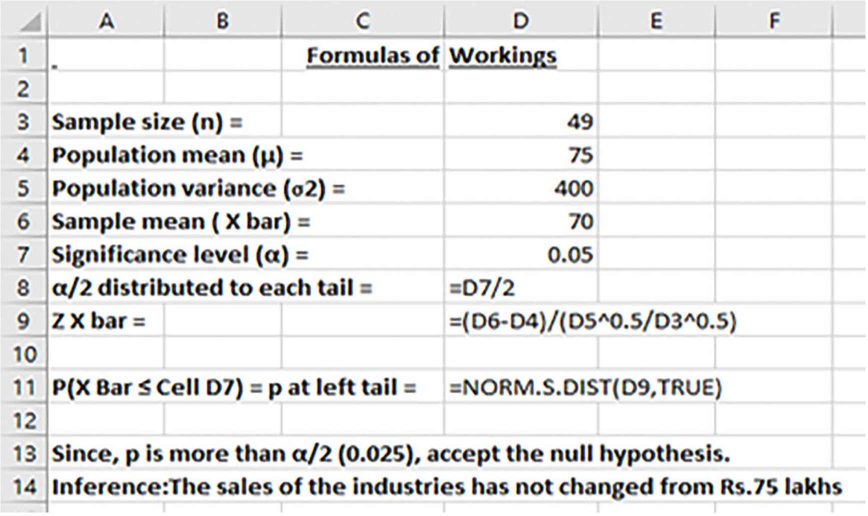


Figure 10.7 Standard normal distribution of Example 10.3



	A	B	C	D	E	F
1	.			Workings		
2						
3	Sample size (n) =			49		
4	Population mean (μ) =			75		
5	Population variance (σ2) =			400		
6	Sample mean (X bar) =			70		
7	Significance level (α) =			0.05		
8	α/2 distributed to each tail =			0.025		
9	Z X bar =			-1.75		
10						
11	P(X Bar ≤ Cell D7) = p at left tail =			0.040059157		
12						
13	Since, p is more than α/2 (0.025), accept the null hypothesis.					
14	Inference:The sales of the industries has not changed from Rs.75 lakhs					

Figure 10.8 Screenshot of working of Example 10.3



	A	B	C	D	E	F
1	.		Formulas of	Workings		
2						
3	Sample size (n) =			49		
4	Population mean (μ) =			75		
5	Population variance (σ2) =			400		
6	Sample mean (X bar) =			70		
7	Significance level (α) =			0.05		
8	α/2 distributed to each tail =			=D7/2		
9	Z X bar =			=(D6-D4)/(D5^0.5/D3^0.5)		
10						
11	P(X Bar ≤ Cell D7) = p at left tail =			=NORM.S.DIST(D9,TRUE)		
12						
13	Since, p is more than α/2 (0.025), accept the null hypothesis.					
14	Inference:The sales of the industries has not changed from Rs.75 lakhs					

Figure 10.9 Screenshot of formulas for working of Example 10.3

The screenshot of the output of the *Z*-test for one sample mean of Example 10.3 is shown in Figure 10.8, and the screenshot of the formulas for the working is shown in Figure 10.9. Since the Z value in Figure 10.8 is negative, the cumulative probability for this value forms the *p* value at the left tail. From Figure 10.8, the *p* value at both ends is 0.04325443, which is more than half of the given significance level 0.025 (0.05/2). Hence, accept the null hypothesis.

Inference: The sales of the small-scale industries have not changed from ₹ 75 lakhs.

10.3 Tests Concerning Difference Between Two Means When the Variances of the Populations Are Known and the Sample Sizes Are Large Using the *Z* Test: Two Sample for Means

Let X_1 and X_2 be two random variables; each follows normal distribution representing two independent populations. They represent two independent normal distributions.

Let μ_1 and σ_1^2 be the mean and variance of population 1 and μ_2 and σ_2^2 be the mean and variance of population 2. Further, it is assumed that the variances of the populations are known.

The hypotheses relating to two different one-tailed tests are as follows.

Type 1 Test:

H_0: $\mu_1 \leq \mu_2$
H_1: $\mu_1 > \mu_2$

In this type of hypothesis testing, the significance level (α) is placed at the right tail of the distribution.

Type 2 Test:

H_0: $\mu_1 \geq \mu_2$
H_1: $\mu_1 < \mu_2$

In this type of hypothesis testing, the significance level (α) is placed at the left tail of the distribution.

Type 3 Test:

H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

In this type of hypothesis testing, half of the significance level ($\alpha/2$) is placed at both tails of the distribution.

The distribution of the difference between the random variables X_1 and X_2 follows normal distribution with mean of $\mu_1 - \mu_2$ and variance of $\sigma_1^2 + \sigma_2^2$.

Let

$\overline{X_1}$ be the mean of sample 1 whose size is n_1
$\overline{X_2}$ be the mean of sample 2 whose size is n_2

The values of n_1 and n_2 are large and greater than 30. Further, these two samples are independent.

The expected values of the means of the two samples, that is, $\overline{X_1}$ and $\overline{X_2}$, are μ_1 and μ_2, respectively.

The formulas for the variances of the two samples are as follows.

$$\text{Variance of sample 1} = \frac{\sigma_1^2}{n_1}$$

$$\text{Variance of sample 2} = \frac{\sigma_2^2}{n_2}$$

The formulas for the mean and variance of the normal distribution for the difference between the sample means are given as follows.

Mean of the difference between the sample means = $\mu_1 - \mu_2$

Variance of the difference between the sample means, $\sigma^2_{\overline{X_1}-\overline{X_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Definition of Sampling Distribution for Difference in Means

Let

$X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$

Then the distribution of the difference between $\overline{X_1} - \overline{X_2}$ is as follows.

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

The standard normal statistic for the difference between the sample means is as follows.

$$Z_{\overline{X_1}-\overline{X_2}} = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sigma_{\overline{X_1}-\overline{X_2}}}$$

Example 10.4

Currently, Vendor 1 sells connecting rods to an automobile company. However, the quality manager of the buying company wants to take into account the supply of connecting rods from Vendor 2 to satisfy the increased demand if the new supplier can offer the connecting rods with good precision on their length on a par with Vendor 1. If the length is greater than the desired value, additional filing time will be required during assembly. Therefore, it ought to be avoided. With variances of 16 mm and 25 mm, respectively, the length of the connecting rods supplied by Vendors 1 and 2 follows a normal distribution.

From Vendor 1, a sample of 32 connecting rods is considered Sample 1, and from Vendor 2, a sample of 32 connecting rods is considered Sample 2. Table 10.1 displays the relevant information regarding the connecting rod length.

Assuming a significance threshold of 0.05, determine if the quality manager can take the supply from Vendor 2 into account.

Solution

The data for Example 10.4 are shown in Table 10.2

The variance of the vendor 1, $\sigma_1^2 = 16$ mm

The variance of the vendor 2, $\sigma_2^2 = 25$ mm

The length of connecting rods from Vendor 1, $X_1 \sim N(\mu_1, \sigma_1^2)$

$$\overline{X_1} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

Table 10.1 Lengths of Connecting Rods in mm

S. No.	*Sample 1*	*Sample 2*
1	450	456
2	451	452
3	453	458
4	448	454
5	447	452
6	446	458
7	452	457
8	447	453
9	451	453
10	452	457
11	448	456
12	454	455
13	447	452
14	451	456
15	452	457
16	449	454
17	452	456
18	451	456
19	448	451
20	455	460
21	452	455
22	449	453
23	451	456
24	447	452
25	450	456
26	452	457
27	453	459
28	451	456
29	452	457
30	453	458
31	453	455
32	453	458

The length of connecting rods from Vendor 2, $X_2 \sim N(\mu_2, \sigma_2^2)$

$$\overline{X_2} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

$\alpha = 0.05$

$$X_1 - X_2 \sim N\left(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2\right)$$

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Check to Consider Supply from Vendor 2

Table 10.2 Data for Example 10.4

S. No.	*Sample 1*	*Sample 2*
1	450	456
2	451	452
3	453	458
4	448	454
5	447	452
6	446	458
7	452	457
8	447	453
9	451	453
10	452	457
11	448	456
12	454	455
13	447	452
14	451	456
15	452	457
16	449	454
17	452	456
18	451	456
19	448	451
20	455	460
21	452	455
22	449	453
23	451	456
24	447	452
25	450	456
26	452	457
27	453	459
28	451	456
29	452	457
30	453	458
31	453	455
32	453	458

The corresponding hypotheses are as follows.

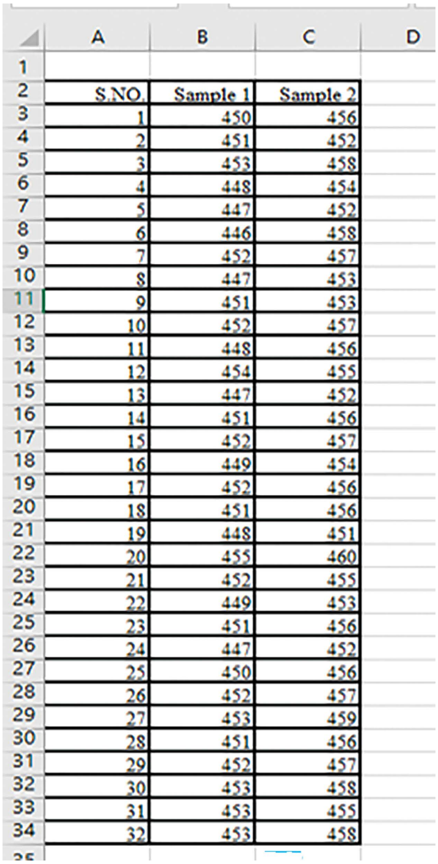
H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

The standard normal statistic for the difference between the sample means is as follows.

$$Z_{\overline{X_1}-\overline{X_2}} = \frac{\left(\overline{X_1}-\overline{X_2}\right)-(\mu_1-\mu_2)}{\sigma_{\overline{X_1}-\overline{X_2}}}$$

The Excel steps applied to solve this problem are explained as follows.

Step1: Input the data for Example in an Excel sheet, as shown in Figure 10.10.



	A	B	C	D
1				
2	S.NO.	Sample 1	Sample 2	
3	1	450	456	
4	2	451	452	
5	3	453	458	
6	4	448	454	
7	5	447	452	
8	6	446	458	
9	7	452	457	
10	8	447	453	
11	9	451	453	
12	10	452	457	
13	11	448	456	
14	12	454	455	
15	13	447	452	
16	14	451	456	
17	15	452	457	
18	16	449	454	
19	17	452	456	
20	18	451	456	
21	19	448	451	
22	20	455	460	
23	21	452	455	
24	22	449	453	
25	23	451	456	
26	24	447	452	
27	25	450	456	
28	26	452	457	
29	27	453	459	
30	28	451	456	
31	29	452	457	
32	30	453	458	
33	31	453	455	
34	32	453	458	

Figure 10.10 Screenshot of input of data for Example 10.4

Step 2: Click the sequence of buttons Home ⟹ Data ⟹ Data Analysis to show the screenshot in Figure 10.11.

Step 3: Click the *Z*-Test: Two sample for Means option from the dropdown menu of Figure 10.11 to show the screenshot in Figure 10.12. One can see the default value of α as 0.05 in the dropdown menu of Figure 10.12.

Step 4: Enter cells B3 to B34 for the variable 1 range and cells C3 to C34 for variable 2 range, click the checkbox for Labels, retain the value of Alpha at 0.05, and click the output option new Worksheet Ply, as shown in the screenshot in Figure 10.13, and then click OK button to show the output in Figure 10.14.

From Figure 10.14, the key results for two tail test when α is 0.05 are shown as follows.

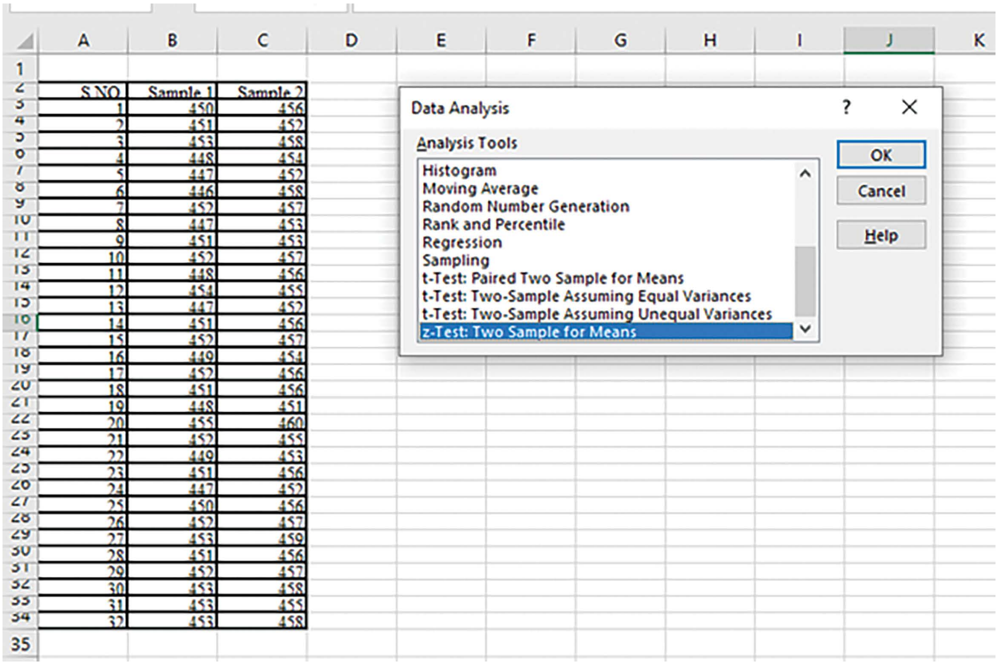
The value of the *Z* statistic is –4.27922.

Critical values of *Z* statistic for two-tailed test = –1.959964 and 1.959964

Since Z_{Cal} (– 4.27922) < Z_{Critical} (–1.959964), Reject H_0, which is $\mu_1 - \mu_2 = 0$.

Inference: The lengths of the connecting rods supplied by Vendor 1 and Vendor 2 differ from each other.

Hence, the quality manager should not consider the connecting rods from Vendor 2.



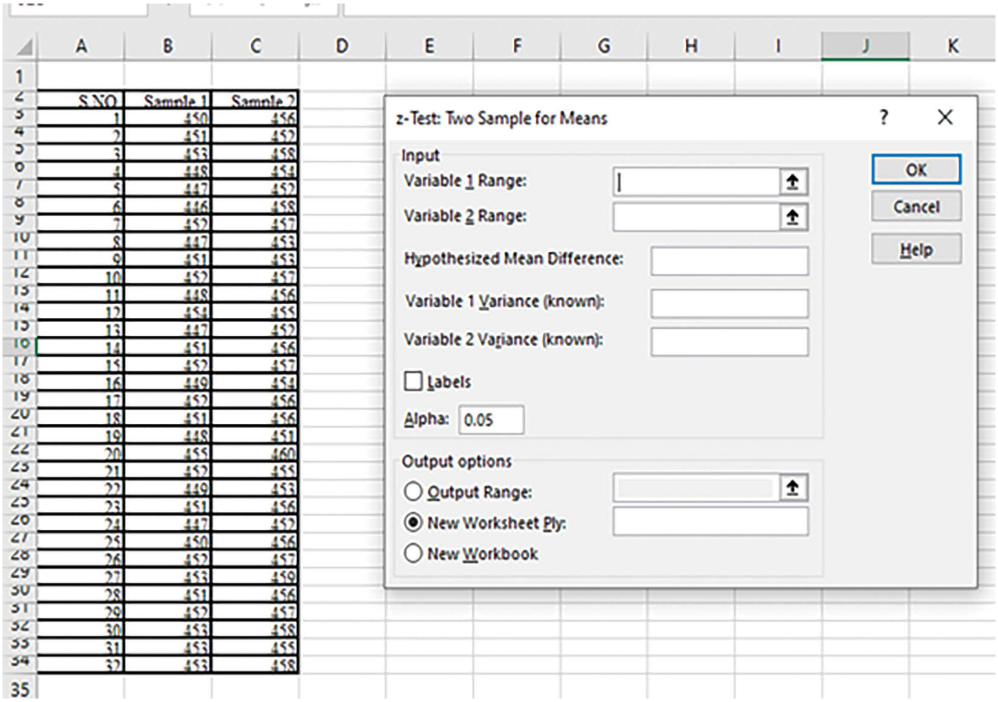
S NO	Sample 1	Sample 2
1	450	456
2	451	452
3	453	458
4	448	454
5	447	452
6	446	458
7	452	457
8	447	453
9	451	453
10	452	457
11	448	456
12	454	455
13	447	452
14	451	456
15	452	457
16	449	454
17	452	456
18	451	456
19	448	451
20	455	460
21	452	455
22	449	453
23	451	456
24	447	452
25	450	456
26	452	457
27	453	459
28	451	456
29	452	457
30	453	458
31	453	455
32	453	458

Data Analysis

Analysis Tools

- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances
- z-Test: Two Sample for Means

Figure 10.11 Screenshot after clicking buttons, Home ⟹ Data button ⟹ Data Analysis



S NO	Sample 1	Sample 2
1	450	456
2	451	452
3	453	458
4	448	454
5	447	452
6	446	458
7	452	457
8	447	453
9	451	453
10	452	457
11	448	456
12	454	455
13	447	452
14	451	456
15	452	457
16	449	454
17	452	456
18	451	456
19	448	451
20	455	460
21	452	455
22	449	453
23	451	456
24	447	452
25	450	456
26	452	457
27	453	459
28	451	456
29	452	457
30	453	458
31	453	455
32	453	458

z-Test: Two Sample for Means

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

Variable 1 Variance (known):

Variable 2 Variance (known):

□ Labels

Alpha: 0.05

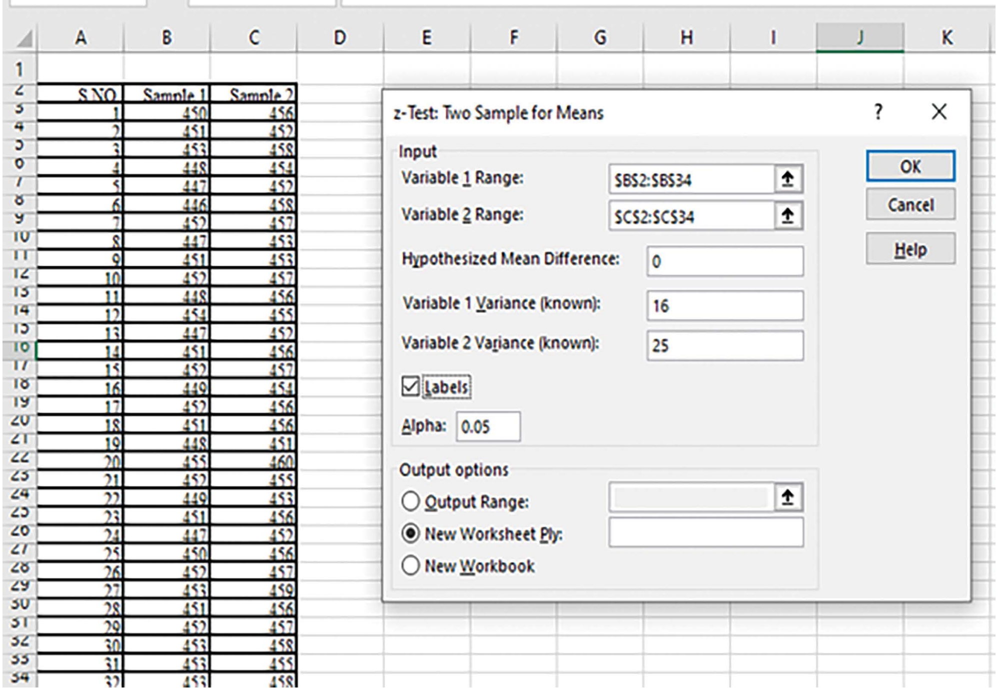
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 10.12 Screenshot after clicking Z-Test: Two sample for Means option in dropdown menu of Figure 10.11



S NO	Sample 1	Sample 2
1	450	456
2	451	452
3	453	458
4	448	454
5	447	452
6	446	458
7	452	457
8	447	453
9	451	453
10	452	457
11	448	456
12	454	455
13	447	452
14	451	456
15	452	457
16	449	454
17	452	456
18	451	456
19	448	451
20	455	460
21	452	455
22	449	453
23	451	456
24	447	452
25	450	456
26	452	457
27	453	459
28	451	456
29	452	457
30	453	458
31	453	455
32	453	458

z-Test: Two Sample for Means

Input

Variable 1 Range: $B$2:$B$34

Variable 2 Range: $C$2:$C$34

Hypothesized Mean Difference: 0

Variable 1 Variance (known): 16

Variable 2 Variance (known): 25

☑ Labels

Alpha: 0.05

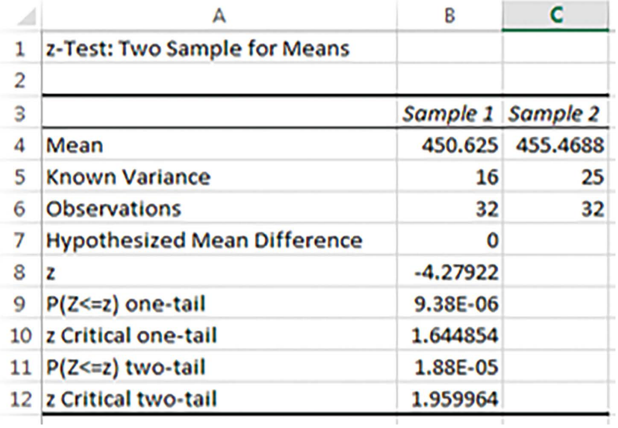
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 10.13 Screenshot of inputting the cell ranges and other options for *Z* test



z-Test: Two Sample for Means

	Sample 1	Sample 2
Mean	450.625	455.4688
Known Variance	16	25
Observations	32	32
Hypothesized Mean Difference	0	
z	-4.27922	
P(Z<=z) one-tail	9.38E-06	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	1.88E-05	
z Critical two-tail	1.959964	

Figure 10.14 Screenshot of output of Z-test: Two samples for Means for Example 10.4

10.4 Tests Concerning Single Mean When the Variance of the Population Is Unknown and the Sample Size Is Small Using T.DIST, T.DIST.RT, and T.DIST.2T Functions

In reality, the population's variance might not be known, and the sample size – less than 30 – could be small. In such a case, a t test will be used to conduct the single mean test. The various tests that could fall under this category are given as follows.

The following list includes the hypotheses for two separate one-tailed tests and a two-tailed test.

Type 1 Test:

H_0: $\mu \leq k$
H_1: $\mu > k$

In this type of hypothesis testing, the significance level (α) is placed at the right tail of the distribution.

Type 2 Test:

$\mathrm{H_0}$: $\mathrm{\mu \geq k}$
$\mathrm{H_1}$: $\mathrm{\mu < k}$

In this type of hypothesis testing, the significance level (α) is placed at the left tail of the distribution.

Type 3 Test:

H_0: $\mu = k$
H_1: $\mu \neq k$

In this type of hypothesis testing, half of the significance level ($\alpha/2$) is placed at both tails of the distribution.

Let X be a random variable with respect to an independent population that follows a normal distribution. Assume that its variance is unknown. The sample size (n) of the population is less than 30.

The statistic of this test is as follows.

$$t = \frac{\left(\bar{\mathrm{X}} - \mu\right)}{\sqrt{\frac{S^2}{n}}}$$

The degree of freedom of this test is $n - 1$.

Example 10.5

The quality manager of a washing machine manufacturer has a gut feeling that the motors it purchases from its suppliers typically last 90 days or less before failing. The quality manager wants to test his intuition on the mean time between failures of the motors. He has therefore chosen a sample of 25 motors, and it has been discovered that the mean period between failures and its variance are, respectively, 93 days and 16 days. Verify the quality manager's perceptions with a significance level of 0.01.

Solution

The mean time between failures, $X \sim N(\mu, \sigma^2)$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

μ = 90 days, σ^2 = unknown
n = 25, $\bar{X}$ = 93 days, S^2 = 16 days, α = 0.01.
H_0: $\mu \leq 90$ days
H_1: $\mu > 90$ days
$\bar{X}$ ~ t distribution with (n–1) $d.f.$

$$t = \frac{\left(\bar{X} - \mu\right)}{\left(\frac{S}{\sqrt{n}}\right)} with\left(n-1\right)d.f.$$

$$= \frac{\left(93-90\right)}{\frac{\sqrt{16}}{\sqrt{25}}}$$

The formula for the p value at the right tail of t distribution is [2]:

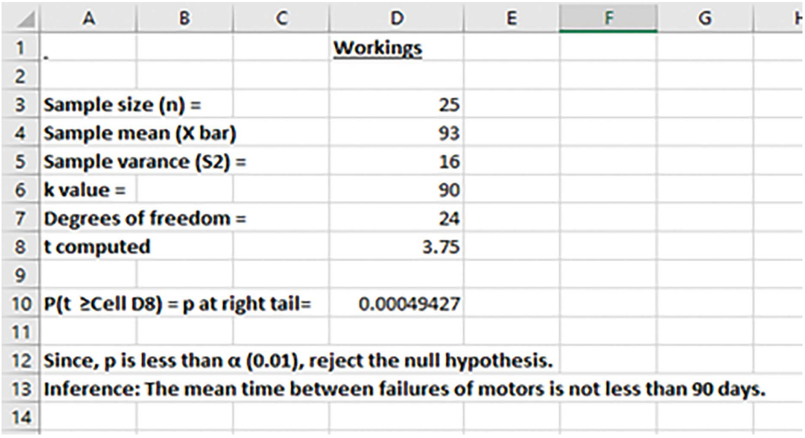
$$= \text{T.DIST.RT}\left(\text{t computed, degrees of freedom}\right)$$

An Excel screenshot of the output of the T-Test for one sample mean of Example 10.5 is shown in Figure 10.15, and the screenshot for the formulas for the working is shown in Figure 10.16. From Figure 10.15, it can be seen that the value of t computed is 3.75. The p value when t is 3.75 is 0.00049427, which is less than the given significance level of 0.01. Hence, the null hypothesis is to be rejected.

Inference: The mean time between failures of motors is not less than 90 days.

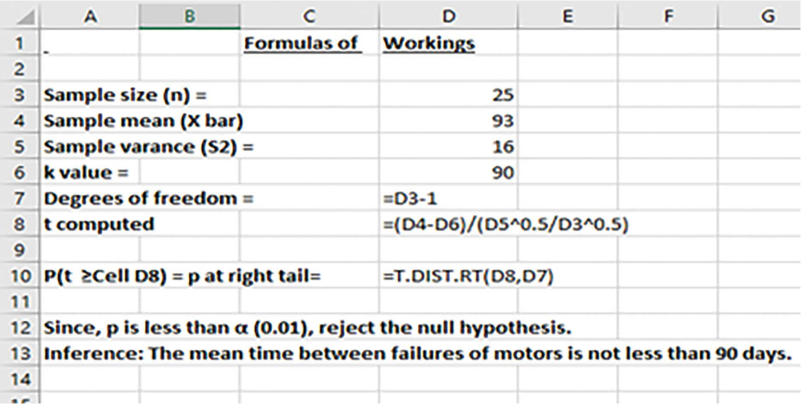
Example 10.6

The weight of electrodes purchased by a foundry follows a normal distribution. The vendor company's sales manager asserts that the electrodes' average weight is at least



	A	B	C	D
1	.			Workings
2				
3	Sample size (n) =			25
4	Sample mean (X bar)			93
5	Sample varance (S2) =			16
6	k value =			90
7	Degrees of freedom =			24
8	t computed			3.75
9				
10	P(t ≥Cell D8) = p at right tail=			0.00049427
11				
12	Since, p is less than α (0.01), reject the null hypothesis.			
13	Inference: The mean time between failures of motors is not less than 90 days.			
14				

Figure 10.15 Screenshot of working of Example 10.5



	A	B	C	D	E	F	G
1	-		Formulas of	Workings			
2							
3	Sample size (n) =			25			
4	Sample mean (X bar)			93			
5	Sample varance (S2) =			16			
6	k value =			90			
7	Degrees of freedom =			=D3-1			
8	t computed			=(D4-D6)/(D5^0.5/D3^0.5)			
9							
10	P(t ≥Cell D8) = p at right tail=			=T.DIST.RT(D8,D7)			
11							
12	Since, p is less than α (0.01), reject the null hypothesis.						
13	Inference: The mean time between failures of motors is not less than 90 days.						
14							

Figure 10.16 Screenshot of formulas for working of Example 10.5

125 grammes. The foundry's quality manager is looking to confirm this assertion. He has therefore collected a sample of 16 electrodes. The electrodes in the sample have a mean weight of 120 gm and a variance of 64 gm, respectively. Verify the sales manager's assertion with a significance level of 0.01.

Solution

The weight of electrodes, $X \sim N(\mu, \sigma^2)$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\mu = 125$ gm
$n = 16$, $\bar{X} = 120$gm, $S^2 = 64$gm, $\alpha = 0.01$.
H_0: $\mu \geq 125$ gm
H_1: $\mu < 125$ gm
$\bar{X} \sim t$ distribution with $(n - 1)$ $d.f.$

$$t = \frac{\left(\bar{X} - \mu\right)}{\left(\frac{S}{\sqrt{n}}\right)} with (n-1) d.f.$$

$$= \frac{(120 - 125)}{\frac{\sqrt{64}}{\sqrt{16}}}$$

$$P\left(t \geq \frac{(120 - 125)}{\frac{\sqrt{64}}{\sqrt{16}}}\right) = ?$$

If the value of *t* is negative, the formula for the cumulative probability of *t* distribution gives the *p* value at the left tail.

The formula for *t* distribution to get the cumulative probability (*p* value in this case) at the left tail:

$$= \text{T.DIST}(\text{X}, \text{deg\_freedom}, \text{TRUE})$$

The screenshot of the output of the T-Test for one sample mean of Example 10.6 is shown in Figure 10.17, and the screenshot for the formulas for the working is shown in Figure 10.18. The probability that *t* is less than *k* (–2.5) is 0.012253, which is the value of *p* at the left tail of the *t* distribution. This is more than the given significance level of 0.01. Hence, accept the null hypothesis.

Inference: The mean weight of the electrodes is significantly more than 125 gm.

Example 10.7

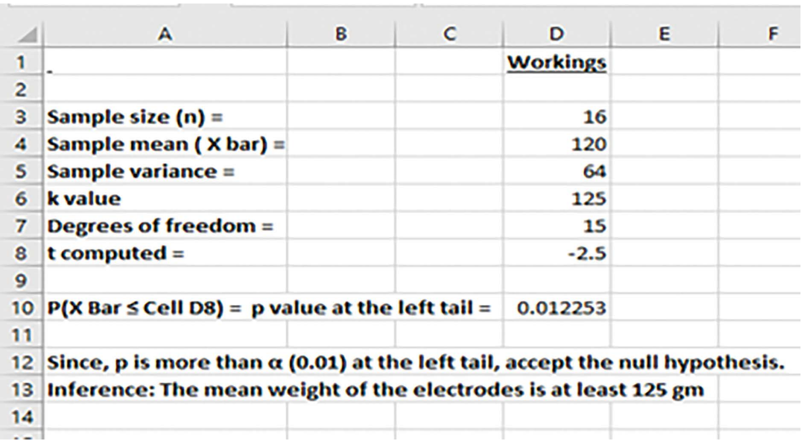
A chemical company's procurement of costly catalyst electrodes has a normal distribution in weight. The vendor company's sales manager asserts that the electrodes typically weigh 200 g. The chemical company's quality manager wants to confirm this assertion. He has therefore collected a sample of 20 electrodes. The sample electrodes' mean and variance are determined to be 205 g and 81 g, respectively. Verify the sales manager's assertion with a significance level of 0.10.

Solution

The weight of electrodes, $X \sim N(\mu, \sigma^2)$

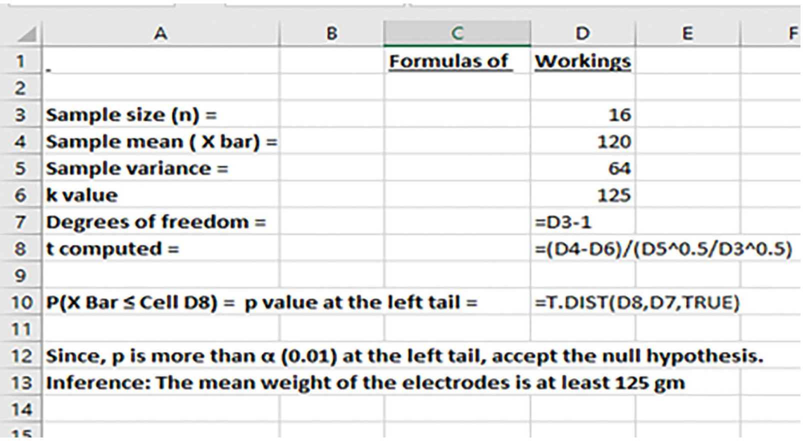
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

μ = 200gm



	A	B	C	D	E	F
1	-			**Workings**		
2						
3	**Sample size (n) =**			16		
4	**Sample mean (X bar) =**			120		
5	**Sample variance =**			64		
6	**k value**			125		
7	**Degrees of freedom =**			15		
8	**t computed =**			-2.5		
9						
10	**P(X Bar ≤ Cell D8) = p value at the left tail =**			0.012253		
11						
12	**Since, p is more than α (0.01) at the left tail, accept the null hypothesis.**					
13	**Inference: The mean weight of the electrodes is at least 125 gm**					
14						

Figure 10.17 Screenshot of working of Example 10.6



	A	B	C	D	E	F
1	.		Formulas of	Workings		
2						
3	Sample size (n) =			16		
4	Sample mean (X bar) =			120		
5	Sample variance =			64		
6	k value			125		
7	Degrees of freedom =			=D3-1		
8	t computed =			=(D4-D6)/(D5^0.5/D3^0.5)		
9						
10	P(X Bar ≤ Cell D8) = p value at the left tail =			=T.DIST(D8,D7,TRUE)		
11						
12	Since, p is more than α (0.01) at the left tail, accept the null hypothesis.					
13	Inference: The mean weight of the electrodes is at least 125 gm					
14						
15						

Figure 10.18 Screenshot of formulas for working of Example 10.6

$n = 20$, $\bar{X} = 205\text{gm}$, $S^2 = 81\text{gm}$
H_0: $\mu = 200$ gm
H_1: $\mu \neq 200$ gm
$\bar{X} \sim t$ distribution with $(n\text{-}1)$ $d.f.$

$$t = \frac{(\bar{X} - \mu)}{\left(\frac{S}{\sqrt{n}}\right)} with(n-1)d.f.$$

$$= \frac{(205 - 200)}{\frac{\sqrt{81}}{\sqrt{20}}}$$

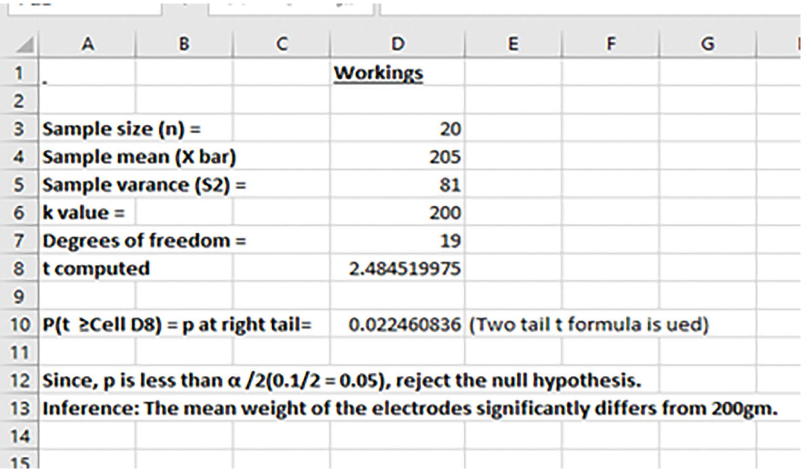
The formula for two-tailed t is: $= T.\text{DIST}.2T(X, \text{deg\_freedom})$

The screenshot of the output of the T-Test for one sample mean of Example 10.7 is shown in Figure 10.19, and the screenshot for the formulas for the working is shown in Figure 10.20. The p value at each end of the t distribution for the t computed is 0.022460836. This is less than half of the significance level (0.1/2 = 0.05). Hence, reject the null hypothesis.

Inference: The mean weight of the electrodes significantly differs from 200 gm.

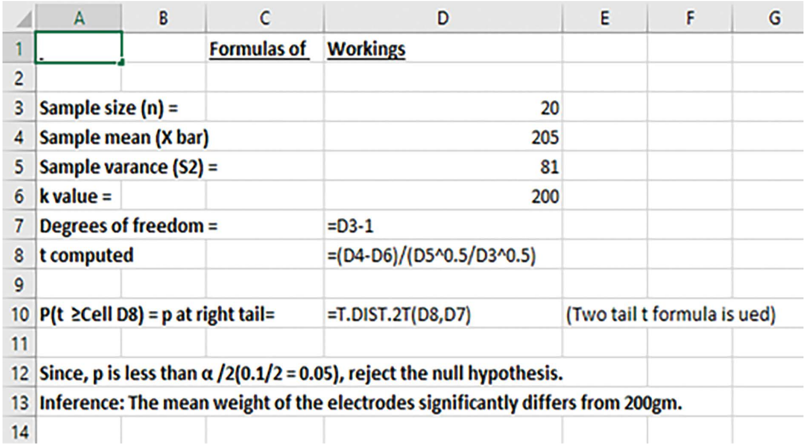
10.5 Tests Concerning Difference Between Two Means When the Variances of the Population Are Unknown and the Sample Size Is Small Using T-Test: Two-Sample Assuming Equal Variance

Let X_1 and X_2 be two random variables with respect to two independent populations that follow normal distribution. Assume that their variances are unknown and equal. The sample size of each population is less than or equal to 30.



	A	B	C	D	E	F	G
1	.			Workings			
2							
3	Sample size (n) =			20			
4	Sample mean (X bar)			205			
5	Sample varance (S2) =			81			
6	k value =			200			
7	Degrees of freedom =			19			
8	t computed			2.484519975			
9							
10	P(t ≥Cell D8) = p at right tail=			0.022460836	(Two tail t formula is ued)		
11							
12	Since, p is less than α /2(0.1/2 = 0.05), reject the null hypothesis.						
13	Inference: The mean weight of the electrodes significantly differs from 200gm.						
14							
15							

Figure 10.19 Screenshot of working of Example 10.7



	A	B	C	D	E	F	G
1	.		Formulas of	Workings			
2							
3	Sample size (n) =			20			
4	Sample mean (X bar)			205			
5	Sample varance (S2) =			81			
6	k value =			200			
7	Degrees of freedom =			=D3-1			
8	t computed			=(D4-D6)/(D5^0.5/D3^0.5)			
9							
10	P(t ≥Cell D8) = p at right tail=			=T.DIST.2T(D8,D7)	(Two tail t formula is ued)		
11							
12	Since, p is less than α /2(0.1/2 = 0.05), reject the null hypothesis.						
13	Inference: The mean weight of the electrodes significantly differs from 200gm.						
14							

Figure 10.20 Screenshot for formulas of working of Example 10.7

In this situation, $\sigma_1^2 = \sigma_2^2$.

The different possible tests of this category are as follows.

Type 1 Test:

$H_0: \mu_1 - \mu_2 \leq k$
$\mathrm{H}_1: \mu_1 - \mu_2 > k$

In this type of hypothesis testing, the significance level (α) is placed at the right tail of the distribution.

Type 2 Test:

$H_0: \mu_1 - \mu_2 \geq k$
$H_1: \mu_1 - \mu_2 < k$

In this type of hypothesis testing, the significance level (α) is placed at the left tail of the distribution.

Type 3 Test:

H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

In this type of hypothesis testing, half of the significance level ($\alpha/2$) is placed at both tails of the distribution.

The standard deviation of the difference in the population means ($\sigma_{\bar{X}1-\bar{X}2}$) cannot be estimated using the sample standard deviations S_1 and S_2. Hence, a polled variance is estimated using the following formula with the assumption that the variances of the populations are equal ($\sigma^2 = \sigma_1{}^2 = \sigma_2{}^2$).

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$$

The formula of the standard deviation for the difference between population means is as follows.

$$\sigma_{\overline{X_1} - \overline{X_2}} = S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The t statistic for this reality is as follows.

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sigma_{\overline{X_1} - \overline{X_2}}} = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with $(n_1 + n_2 - 2)$ degrees of freedom.

Example 10.8

Both the age of the workers in industrial estate 1 and the age of those in industrial estate 2 are distributed normally. According to the researcher who examined these statistics, employees in industrial estate 1 are younger than those in industrial estate 2 on average. As a result, the investigator chose 14 workers from industrial estate 1 and 17 workers from industrial estate 2. Table 10.3 displays the employees' ages at the two industrial estates. Verify the investigator's intuition at a 0.05 level of significance.

Solution

The data for Example 10.8 are shown in Table 10.4.

Let X_1 and X_2 be two random variables with respect to two independent populations that follow normal distribution. Assume that their variances are unknown and the sample size of each population is less than or equal to 30.

The age of the employees working in industrial estate 1, $X_1 \sim N(\mu_1, \sigma_1{}^2)$

$$\overline{X_1} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

Table 10.3 Age of Employees Working in Industrial Estate 1 and Industrial Estate 2

Sample Unit	*Industrial Estate 1*	*Industrial Estate 2*
1	50	30
2	55	40
3	58	57
4	47	45
5	30	50
6	40	42
7	45	53
8	55	50
9	42	60
10	24	47
11	41	30
12	29	46
13	45	49
14	38	53
15	–	40
16	–	25
17	–	45

Table 10.4 Data for Example 10.8

Sample Unit	*Industrial Estate 1*	*Industrial Estate 2*
1	50	30
2	55	40
3	58	57
4	47	45
5	30	50
6	40	42
7	45	53
8	55	50
9	42	60
10	24	47
11	41	30
12	29	46
13	45	49
14	38	53
15	–	40
16	–	25
17	–	45

The age of the employees working in industrial estate 2, $X_2 \sim N(\mu_2, \sigma_2^2)$

$$\overline{X_2} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

$\alpha = 0.05$

The standard deviation of the difference in the population means ($\sigma_{\overline{X_1}-\overline{X_2}}$) cannot be estimated using the sample standard deviations S_1 and S_2. Hence, a polled variance is estimated using the following formula with the assumption that the variances of the populations are equal ($\sigma^2 = \sigma_1{}^2 = \sigma_2{}^2$).

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$$

The formula of the standard deviation for the difference between population means is as follows.

$$\sigma_{\overline{X_1} - \overline{X_2}} = S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The t statistic for this reality is as follows.

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sigma_{\overline{X_1} - \overline{X_2}}} = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with $(n_1 + n_2 - 2)$ degrees of freedom.

H_0: $\mu_1 \leq \mu_2$
H_1: $\mu_{1} > \mu_2$

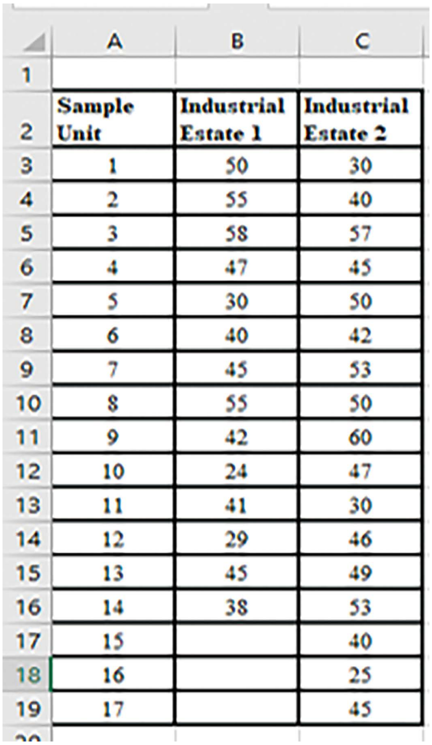
The steps for the working of the T-Test with H_0: $\mu_1 \leq \mu_2$ for Example 10.8 along with Excel screenshots are as follows.

Step1: The screenshot of the input of the data for Example 10.8 in an Excel sheet is shown in Figure 10.21.
Step 2: Click the Data button and then the Data Analysis button in Excel to show the screenshot in Figure 10.22.
Step 3:Click T-Test: Two sample assuming equal variances option from the dropdown menu of Figure 10.22 to show the screenshot in Figure 10.23.
Step 4: Enter the cells for the variable 1 range and the cells for variable 2 range, click the checkbox for Labels, keep the value of Alpha as 0.05, and click the output option new Worksheet Ply, as shown in the screenshot of Figure 10.24, and then click OK button to show the output in Figure 10.25.

From Figure 10.25, the key results are as follows.
Value of the t statistic = –0.571902165
Critical value of Z statistic for right tail t test = 1.699127027
Since t_{Cal} (–0.571902165) < t_{Critical} (1.699127027), accept the H_0, which means $\mu_1 \leq \mu_2$.
Inference: The age of employees in industrial estate 1 is less than that of the employees in industrial estate 2.

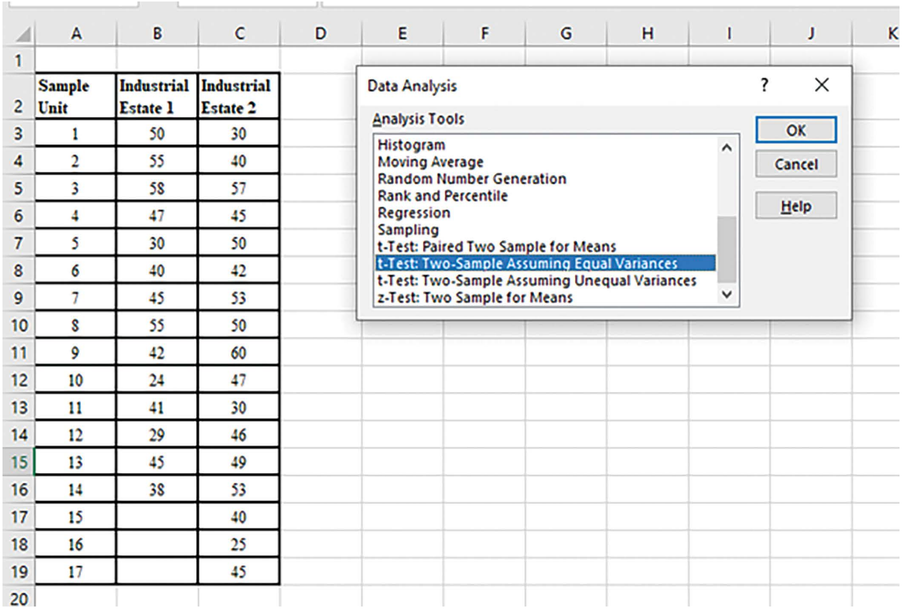
Example 10.9

Both the annual income of employees in industrial estate 1 and that of those in industrial estate 2 are distributed normally. According to the researcher who examined these statistics, the annual income of workers in industrial estate 1 and that in industrial estate



Sample Unit	Industrial Estate 1	Industrial Estate 2
1	50	30
2	55	40
3	58	57
4	47	45
5	30	50
6	40	42
7	45	53
8	55	50
9	42	60
10	24	47
11	41	30
12	29	46
13	45	49
14	38	53
15		40
16		25
17		45

Figure 10.21 Screenshot of input of Example 10.8 in Excel



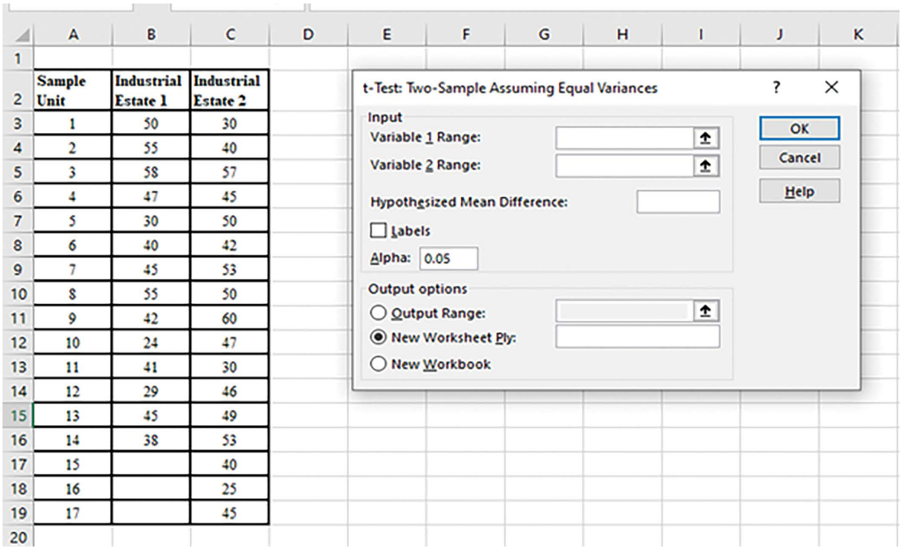
Sample Unit	Industrial Estate 1	Industrial Estate 2
1	50	30
2	55	40
3	58	57
4	47	45
5	30	50
6	40	42
7	45	53
8	55	50
9	42	60
10	24	47
11	41	30
12	29	46
13	45	49
14	38	53
15		40
16		25
17		45

Data Analysis

Analysis Tools: t-Test: Two-Sample Assuming Equal Variances

?
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

Figure 10.22 Screenshot after clicking the Data button and then the Data Analysis button



Sample Unit	Industrial Estate 1	Industrial Estate 2
1	50	30
2	55	40
3	58	57
4	47	45
5	30	50
6	40	42
7	45	53
8	55	50
9	42	60
10	24	47
11	41	30
12	29	46
13	45	49
14	38	53
15		40
16		25
17		45

t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

☐ Labels

Alpha: 0.05

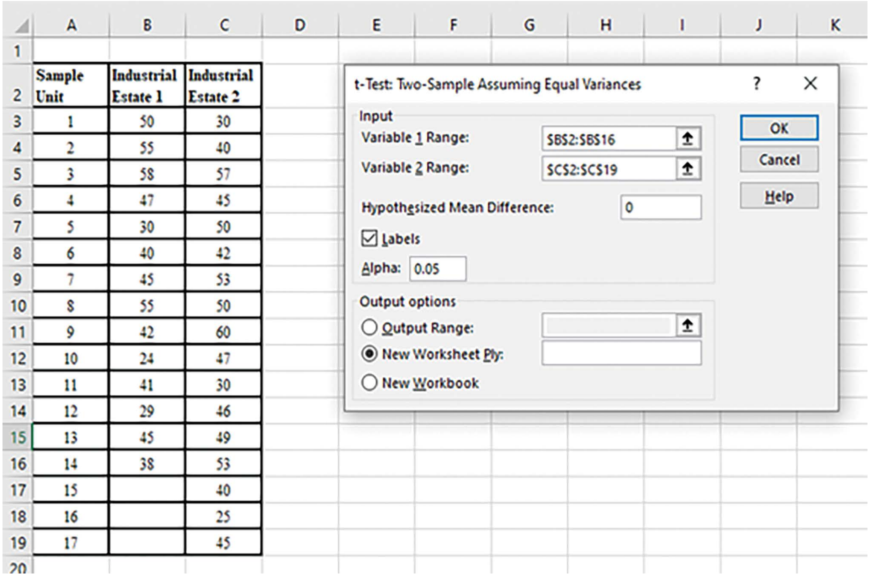
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 10.23 Screenshot after clicking T-Test: Two sample assuming equal variances option in dropdown menu of Figure 10.22



Sample Unit	Industrial Estate 1	Industrial Estate 2
1	50	30
2	55	40
3	58	57
4	47	45
5	30	50
6	40	42
7	45	53
8	55	50
9	42	60
10	24	47
11	41	30
12	29	46
13	45	49
14	38	53
15		40
16		25
17		45

t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range: $B$2:$B$16

Variable 2 Range: $C$2:$C$19

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

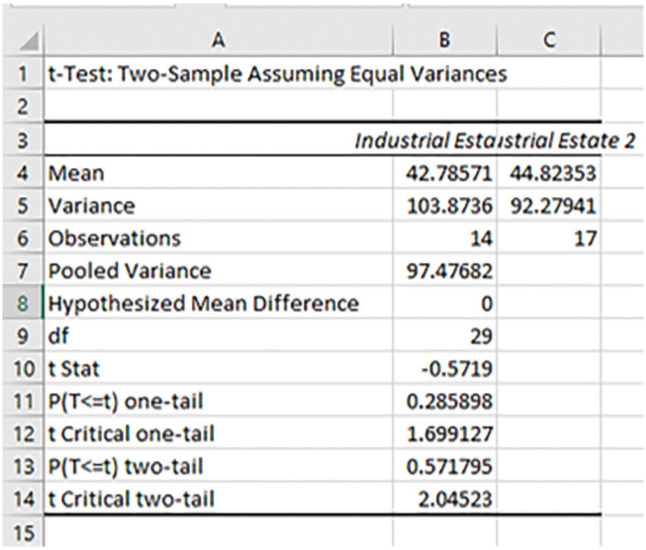
○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 10.24 Screenshot for inputting the cell ranges and other options for *t* test in the dropdown menu of Figure 10.23

2 are not different from one another. Therefore, the investigator chose 15 workers from industrial estate 1 and 18 from industrial estate 2. Table 10.5 displays the annual salaries of the staff at the two industrial parks, expressed in lakhs of rupees. Verify the investigator's intuition at a 0.05 level of significance.



	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
2			
3		*Industrial Esta*	*ıstrial Estate 2*
4	Mean	42.78571	44.82353
5	Variance	103.8736	92.27941
6	Observations	14	17
7	Pooled Variance	97.47682	
8	Hypothesized Mean Difference	0	
9	df	29	
10	t Stat	-0.5719	
11	P(T<=t) one-tail	0.285898	
12	t Critical one-tail	1.699127	
13	P(T<=t) two-tail	0.571795	
14	t Critical two-tail	2.04523	
15			

Figure 10.25 Screenshot of output of T-Test: Two sample assuming equal variances for Example 10.8

Table 10.5 Annual Incomes of Employees of Industrial Estates

Sample Unit	*Industrial Estate 1*	*Industrial Estate 2*
1	100	120
2	105	90
3	108	97
4	97	85
5	150	150
6	140	120
7	145	123
8	111	110
9	140	100
10	124	87
11	90	80
12	120	96
13	108	99
14	98	103
15	84	100
16	–	115
17	–	135
18	–	99

Solution

The data for Example 10.9 are shown in Table 10.6.

The annual income of the employees working in industrial estate 1, $X_1 \sim N(\mu_1, \sigma_1^2)$

$$\overline{X_1} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

Table 10.6 Data for Example 10.9

Sample Unit	*Industrial Estate 1*	*Industrial Estate 2*
1	100	120
2	105	90
3	108	97
4	97	85
5	150	150
6	140	120
7	145	123
8	111	110
9	140	100
10	124	87
11	90	80
12	120	96
13	108	99
14	98	103
15	84	100
16	–	115
17	–	135
18	–	99

The annual income of the employees working in industrial estate 2, $X_2 \sim N(\mu_2, \sigma_2^2)$

$$\overline{X_2} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

The standard deviation of the difference in the population means ($\sigma_{\overline{X_1}-\overline{X_2}}$) cannot be estimated using the sample standard deviations S_1 and S_2. Hence, a polled variance is estimated using the following formula with the assumption that the variances of the populations are equal ($\sigma^2 = \sigma_1^2 = \sigma_2^2$).

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$$

The formula of the standard deviation for the difference between population means is as follows.

$$\sigma_{\overline{X_1}-\overline{X_2}} = S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The t statistic for this reality is as follows.

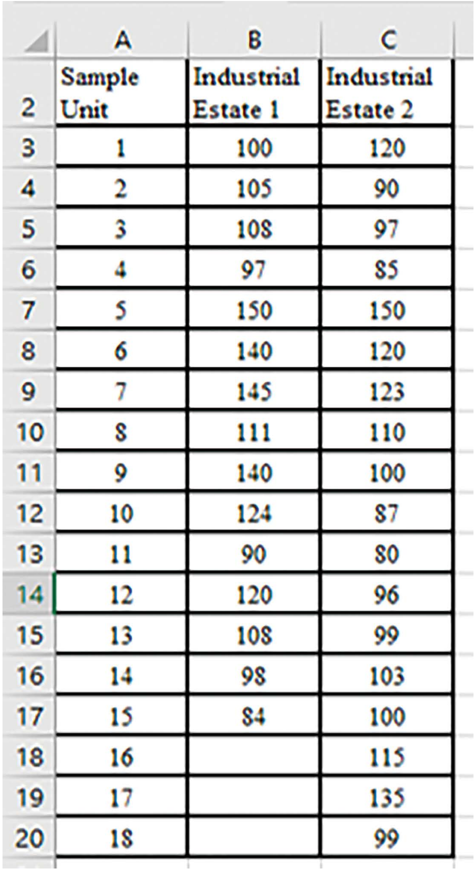
$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sigma_{\overline{X_1}-\overline{X_2}}} = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with $(n_1 + n_2 - 2)$ degrees of freedom.

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

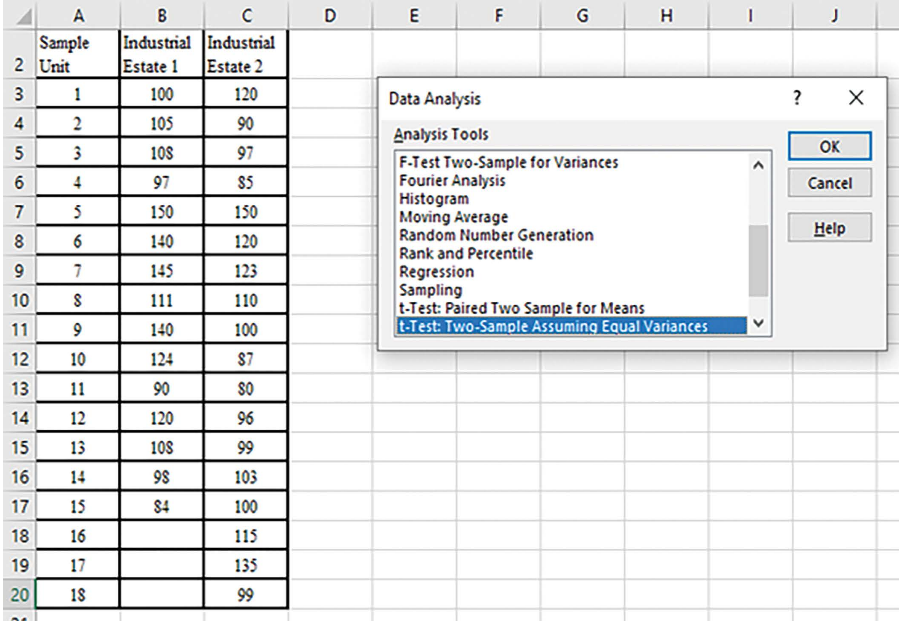
The steps for the working of the T-Test with H_0: $\mu_1 = \mu_2$ for Example 10.9 along with Excel screenshots are as follows.

Step1: Input the data for Example 10.9 in an Excel sheet as shown in Figure 10.26.
Step 2: Click the Data button and then the Data Analysis button in Excel to show the screenshot in Figure 10.27.
Step 3: Click the T-Test: Two sample assuming equal variances option from the drop-down menu of Figure 10.27 to show the screenshot in Figure 10.28.
Step 4: Enter cells B2:B17 for the variable 1 range and cells C2:C20 for variable 2 range, fill in 0 for the hypothesised mean difference, click the checkbox for Labels, retain the value of Alpha as 0.05, and click the output option new Worksheet Ply as shown in the screenshot of Figure 10.29, and then click the OK button to show the output in Figure 10.30.



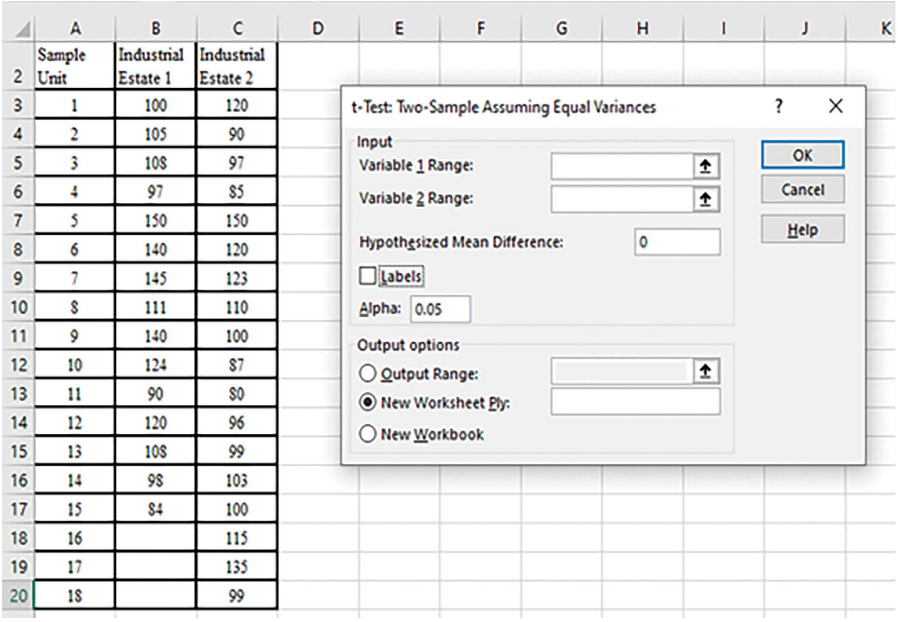
	A	B	C
2	Sample Unit	Industrial Estate 1	Industrial Estate 2
3	1	100	120
4	2	105	90
5	3	108	97
6	4	97	85
7	5	150	150
8	6	140	120
9	7	145	123
10	8	111	110
11	9	140	100
12	10	124	87
13	11	90	80
14	12	120	96
15	13	108	99
16	14	98	103
17	15	84	100
18	16		115
19	17		135
20	18		99

Figure 10.26 Input of Example 10.9 in Excel sheet



	A	B	C
2	Sample Unit	Industrial Estate 1	Industrial Estate 2
3	1	100	120
4	2	105	90
5	3	108	97
6	4	97	85
7	5	150	150
8	6	140	120
9	7	145	123
10	8	111	110
11	9	140	100
12	10	124	87
13	11	90	80
14	12	120	96
15	13	108	99
16	14	98	103
17	15	84	100
18	16		115
19	17		135
20	18		99

Figure 10.27 Screenshot after clicking the Data button and then the Data Analysis button in Excel



	A	B	C
2	Sample Unit	Industrial Estate 1	Industrial Estate 2
3	1	100	120
4	2	105	90
5	3	108	97
6	4	97	85
7	5	150	150
8	6	140	120
9	7	145	123
10	8	111	110
11	9	140	100
12	10	124	87
13	11	90	80
14	12	120	96
15	13	108	99
16	14	98	103
17	15	84	100
18	16		115
19	17		135
20	18		99

t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference: 0

□ Labels

Alpha: 0.05

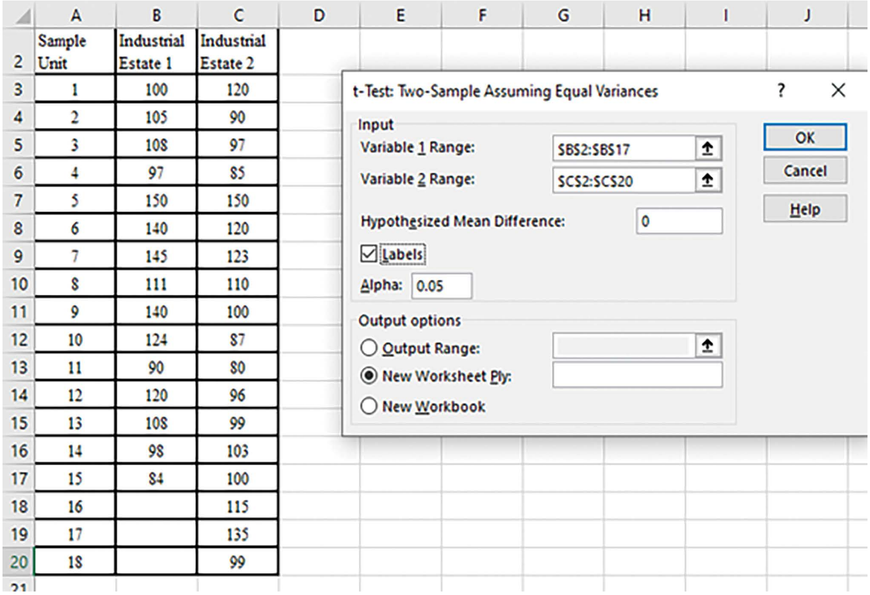
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 10.28 Screenshot after clicking T-Test: Two sample assuming equal variances option in dropdown menu of Figure 10.27



	A	B	C
2	Sample Unit	Industrial Estate 1	Industrial Estate 2
3	1	100	120
4	2	105	90
5	3	108	97
6	4	97	85
7	5	150	150
8	6	140	120
9	7	145	123
10	8	111	110
11	9	140	100
12	10	124	87
13	11	90	80
14	12	120	96
15	13	108	99
16	14	98	103
17	15	84	100
18	16		115
19	17		135
20	18		99

t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range: $B$2:$B$17

Variable 2 Range: $C$2:$C$20

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

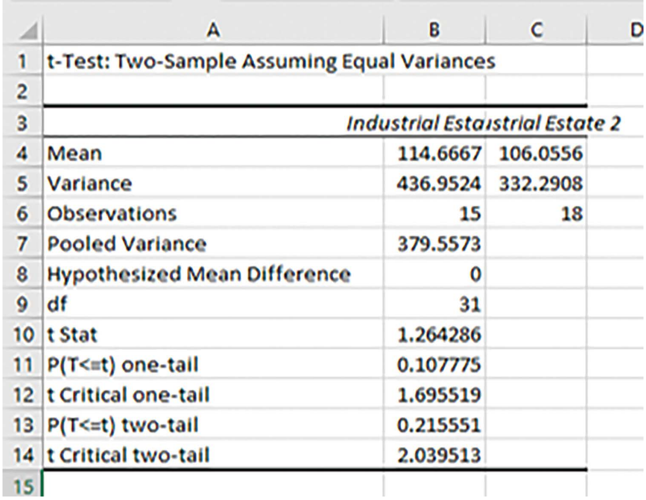
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 10.29 Screenshot of inputting the cell ranges and other options for T-Test: Two sample assuming equal variances



	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
2			
3		*Industrial Esta*	*ıstrial Estate 2*
4	Mean	114.6667	106.0556
5	Variance	436.9524	332.2908
6	Observations	15	18
7	Pooled Variance	379.5573	
8	Hypothesized Mean Difference	0	
9	df	31	
10	t Stat	1.264286	
11	P(T<=t) one-tail	0.107775	
12	t Critical one-tail	1.695519	
13	P(T<=t) two-tail	0.215551	
14	t Critical two-tail	2.039513	

Figure 10.30 Screenshot of output of T-Test: Two sample assuming equal variances for Example 10.9

From Figure 10.30, the key results are as follows.

Value of the t statistic = 1.264286

Critical value of Z statistic for two tail test = –2.039513 (left value) and 2.039513 (right value)

Since t_{Cal} (1.264286) < t_{Critical} (2.039513), accept the H_0, which means $\mu_1 - \mu_2 = 0$.

Inference: There is no significant difference between the mean annual income of the employees of industrial estate 1 and that of the employees of industrial estate 2.

10.6 Tests Concerning Difference Between Two Means When the Variances of the Population Are Unknown and the Sample Size Is Small Using T-Test: Two-Sample Assuming Unequal Variance

Let X_1 and X_2 be two random variables with respect to two independent populations that follow a normal distribution. Assume that their variances are unknown and unequal. The sample size of each population is less than 30.

Let
μ_1 be the mean of population 1
μ_2 be the mean of population 2
σ_1^2 be the variance of population 1
σ_2^2 be the variance of population 2
$\overline{X_1}$ be the mean of sample 1
$\overline{X_2}$ be the mean of sample 2
S_1^2 be the variance of sample 1
S_2^2 be the variance of sample 2
n_1 be the size of sample 1
n_2 be the size of sample 2
$\sigma_1^2 \neq \sigma_2^2$

If X_1 and X_2 follow normal distribution, and n_1 and n_2 are large enough to apply the central limit theorem, the t statistic is given by the following formula.

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

with the degree of freedom given by the nearest integer of the following formula.

$$d.f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}\right]}$$

The different possible tests of this category are as follows.

Type 1 Test:

H_0: $\mu_1 - \mu_2 \leq k$
H_1: $\mu_1 - \mu_2 > k$

In this type of hypothesis testing, the significance level (α) is placed at the right tail of the distribution.

Type 2 Test:

$H_0: \mu_1 - \mu_2 \geq k$

$H_1: \mu_1 - \mu_2 < k$

In this type of hypothesis testing, the significance level (α) is placed at the left tail of the distribution.

Type 3 Test:

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

In this type of hypothesis testing, half of the significance level ($\alpha/2$) is placed at both tails of the distribution.

Example 10.10

Both the employee satisfaction indices for industrial estate 1 employees and industrial estate 2 employees exhibit a normal distribution. According to the researcher who examined the data, the industrial estate 1 employees' employee satisfaction index is not different from that of the industrial estate 2 employees. Therefore, the investigator chose 21 employees from industrial estate 1 and 20 employees from industrial estate 2. Table 10.7 displays the employee satisfaction scores of the two industrial estates' employees. It is

Table 10.7 Satisfaction Indices of Employees

Employee	*Industrial Estate 1*	*Industrial Estate 2*
1	10	4
2	9	3
3	10	5
4	9	4
5	5	6
6	4	5
7	8	4
8	3	7
9	4	4
10	2	5
11	9	6
12	1	7
13	8	5
14	9	8
15	8	4
16	9	5
17	8	6
18	6	4
19	4	6
20	2	7
21	9	–

expected that population 1's variance is greater than population 2's variance. Verify the investigator's intuition at a 0.05 level of significance.

Solution

The data for Example 10.10 are shown in Table 10.8.

Let

X_1 be the satisfaction index of employee in industrial estate 1

X_2 be the satisfaction index of employee in industrial estate 2

X_1 follows $N(\mu_1, \sigma_1^2)$

X_1 follows $N(\mu_2, \sigma_2^2)$

$\sigma_1^2 \neq \sigma_2^2$

$\bar{X}_1$ be the mean of sample 1

$\bar{X}_2$ be the mean of sample 2

S_1^2 be the variance of sample 1

S_2^2 be the variance of sample 2

n_1 be the size of sample 1

n_2 be the size of sample 2

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

Table 10.8 Data for Example 10.10

Employee	*Industrial Estate 1*	*Industrial Estate 2*
1	10	4
2	9	3
3	10	5
4	9	4
5	5	6
6	4	5
7	8	4
8	3	7
9	4	4
10	2	5
11	9	6
12	1	7
13	8	5
14	9	8
15	8	4
16	9	5
17	8	6
18	6	4
19	4	6
20	2	7
21	9	–

with the degree of freedom given by the nearest integer of the following formula.

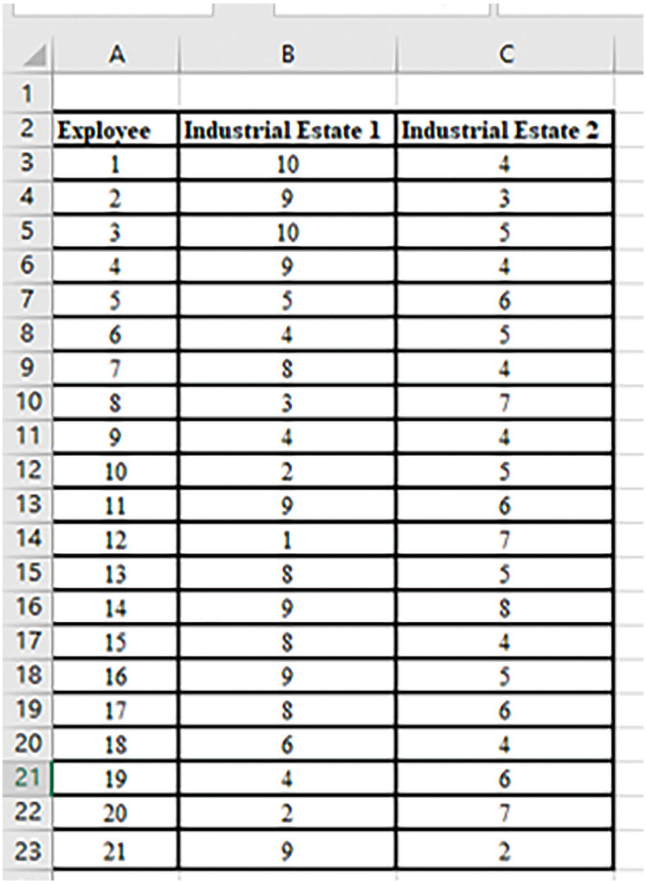
$$d.f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}\right]}$$

The hypotheses of this problem are given as follows.

H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

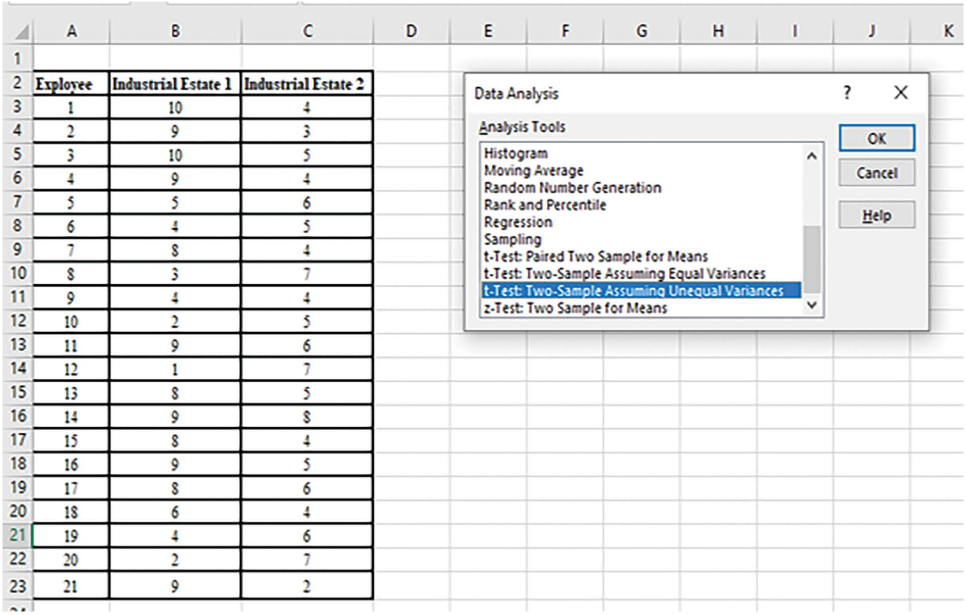
In this type of hypothesis testing, half of the significance level ($\alpha/2$) is placed at both tails of the distribution.

The screenshot of the data for Example 13.10 is shown in Figure 10.31. The screenshot after clicking the sequence of buttons Home ⟹ Data ⟹ Data Analysis is shown in Figure 10.32. The screenshot after selecting the T-Test: Two-Sample Assuming



	A	B	C
1			
2	Explovee	Industrial Estate 1	Industrial Estate 2
3	1	10	4
4	2	9	3
5	3	10	5
6	4	9	4
7	5	5	6
8	6	4	5
9	7	8	4
10	8	3	7
11	9	4	4
12	10	2	5
13	11	9	6
14	12	1	7
15	13	8	5
16	14	9	8
17	15	8	4
18	16	9	5
19	17	8	6
20	18	6	4
21	19	4	6
22	20	2	7
23	21	9	2

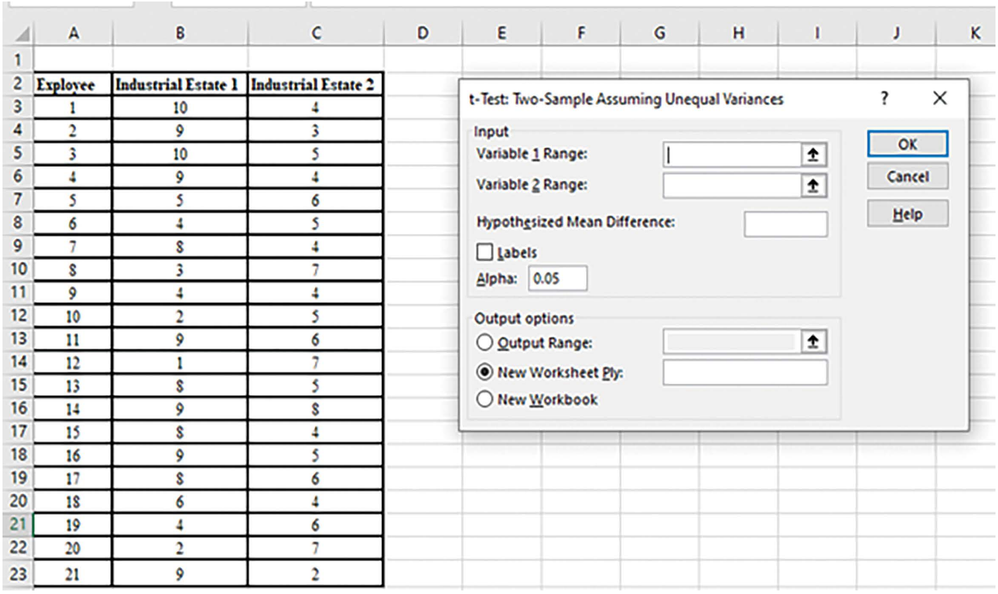
Figure 10.31 Screenshot of data for Example 10.10



Exployee	Industrial Estate 1	Industrial Estate 2
1	10	4
2	9	3
3	10	5
4	9	4
5	5	6
6	4	5
7	8	4
8	3	7
9	4	4
10	2	5
11	9	6
12	1	7
13	8	5
14	9	8
15	8	4
16	9	5
17	8	6
18	6	4
19	4	6
20	2	7
21	9	2

A
B
C
D
E
F
G
H
I
J
K
Data Analysis
?
×
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

Figure 10.32 Screenshot after clicking buttons, Home ⟹ Data ⟹ Data Analysis



Exployee	Industrial Estate 1	Industrial Estate 2
1	10	4
2	9	3
3	10	5
4	9	4
5	5	6
6	4	5
7	8	4
8	3	7
9	4	4
10	2	5
11	9	6
12	1	7
13	8	5
14	9	8
15	8	4
16	9	5
17	8	6
18	6	4
19	4	6
20	2	7
21	9	2

t-Test: Two-Sample Assuming Unequal Variances

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

□ Labels

Alpha: 0.05

Output options

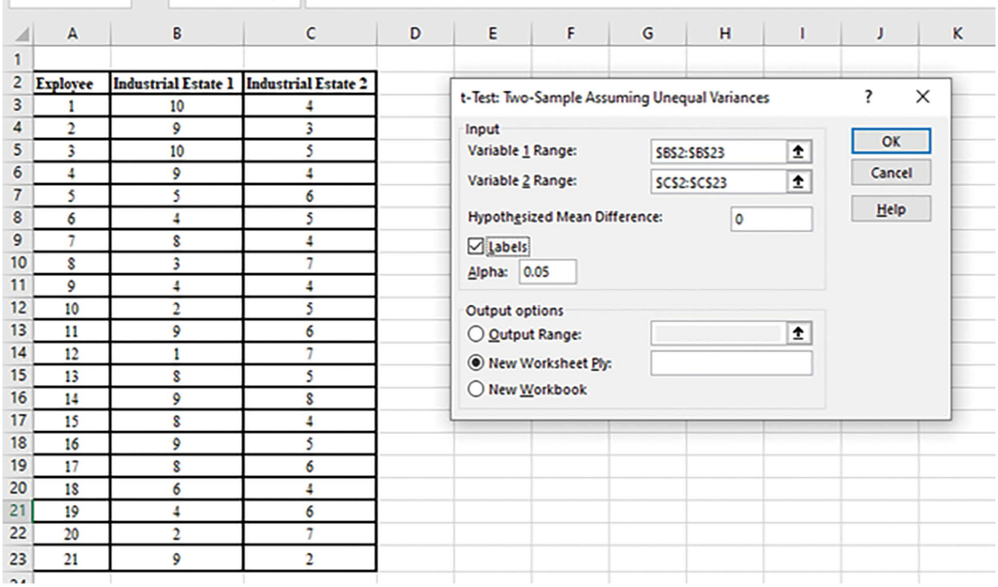
○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 10.33 Screenshot after selecting T-Test: Two Sample Assuming Unequal Variances in the dropdown menu of Figure 10.32 and clicking the OK button

Unequal Variance function in the dropdown menu of Figure 10.32 and clicking the OK button is shown in Figure 10.33. The screenshot after filling the data in the range of cells $B$3:$B$23 in the variable 1 Range and the data in the range $C$2:$C$23 in the Variable 2 Range, entering 0 for Hypothesized Mean Difference, clicking Labels, retaining 0.05 for Alpha in the dropdown menu of Figure 10.33, and clicking the OK button is shown in Figure 10.34. Clicking the OK button in the dropdown menu of Figure 10.34 gives the result shown in the screenshot in Figure 10.35.



	A	B	C
1			
2	Exployee	Industrial Estate 1	Industrial Estate 2
3	1	10	4
4	2	9	3
5	3	10	5
6	4	9	4
7	5	5	6
8	6	4	5
9	7	8	4
10	8	3	7
11	9	4	4
12	10	2	5
13	11	9	6
14	12	1	7
15	13	8	5
16	14	9	8
17	15	8	4
18	16	9	5
19	17	8	6
20	18	6	4
21	19	4	6
22	20	2	7
23	21	9	2

t-Test: Two-Sample Assuming Unequal Variances

Input

Variable 1 Range: $B$2:$B$23

Variable 2 Range: $C$2:$C$23

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

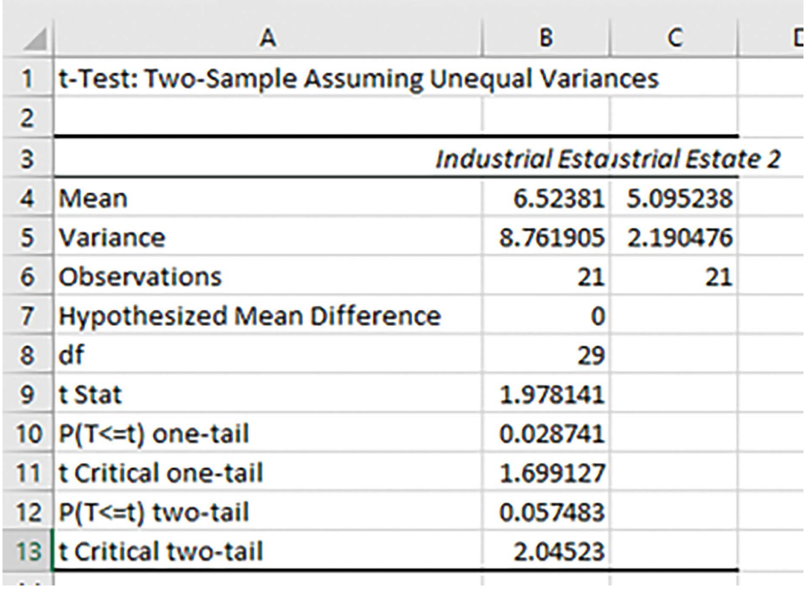
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 10.34 Screenshot after filling data in the dropdown menu of Figure 10.33



	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		*Industrial Esta*	*ıstrial Estate 2*
4	Mean	6.52381	5.095238
5	Variance	8.761905	2.190476
6	Observations	21	21
7	Hypothesized Mean Difference	0	
8	df	29	
9	t Stat	1.978141	
10	P(T<=t) one-tail	0.028741	
11	t Critical one-tail	1.699127	
12	P(T<=t) two-tail	0.057483	
13	t Critical two-tail	2.04523	

Figure 10.35 Screenshot after clicking the OK button in the dropdown menu of Figure 10.34

From Figure 10.35, it can be seen that the value of p for two-tailed test is 0.057483, which is more than the half of the given significance level of 0.025 (0.05/2). Hence, accept the null hypothesis.

Inference: The mean satisfaction index of the employees of industrial estate 1 does not differ significantly from that of the employees of industrial estate 2.

10.7 Paired T Test Using T-Test: Paired Two-Sample for Means Function

There may be a situation in which the value of a random variable at a particular setting may be different from another setting. The first setting may be sales before advertisement, and the second setting may be sales after advertisement of a product sold in n retail outlets.

Now the objective is to test whether there are significance differences among the differences of different pairs of sales (sales without advertisement and sales with advertisements).

Consider the data shown in Table 10.9.

Now the variable d_i follows a t distribution with a degree of freedom of $n - 1$, where n is the number of pairs of observations.

The steps of the paired t test are as follows.

Step 1: Input n pairs of observations, (a_i and b_i), where $i = 1, 2, 3, \ldots, n$.

Step 2: Find the difference of each pair of observations using the following formula.

$$d_i = b_i - a_i, i = 1, 2, 3, \ldots, n$$

Step 3: Find the mean of d_i values, which is given by the following formula.

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

Step 4: Compute the standard deviation of d_i, $i = 1, 2, 3, \ldots, n$.

Step 5: Compute the standard error (SE) of the mean of d_i using the following formula.

$$SE = \frac{S_d}{\sqrt{n}}$$

Step 6: The t statistic is as follows.

$$t = \frac{\bar{d}}{\left(\frac{S_d}{\sqrt{n}}\right)} \text{with } n - 1 \text{ degree of freedom.}$$

Table 10.9 Data for Sales Before and After Advertisement

Retail Outlet Number (i)	*Sales Before Advertisement (a_i)*	*Sales After Advertisement (b_i)*	*Difference Between Sales After Advertisement and Sales Before Advertisement ($d_i = b_i - a_i$)*
1	a_1	b_1	$b_1 - a_1$
2	a_2	b_2	$b_2 - a_2$
3	a_3	b_3	$b_3 - a_3$
4	a_4	b_4	$b_4 - a_4$
5	a_5	b_5	$b_5 - a_5$
6	a_6	b_6	$b_6 - a_6$
7	a_7	b_7	$b_7 - a_7$
8	a_8	b_8	$b_8 - a_8$
9	a_9	b_9	$b_9 - a_9$

The hypotheses of this situation are stated as follows.

$$H_o : \bar{d} \le 0$$

$$\mathrm{H}_1 : \bar{d} > 0$$

Example 10.11

Table 10.10 lists the skill indices for 10 employees on a 0–10 scale before and after a training session. At a significance level of 0.05, determine whether there are any statistically significant differences between the pairs of observations.

Solution

The data for Example 10.11 are shown in Table 10.11.

Table 10.10 Skill Indices of Employees

Employee No.	*Skill Index Before Training*	*Skill Index After Training*
1	5	6
2	8	10
3	4	7
4	7	10
5	1	4
6	4	6
7	6	9
8	3	6
9	5	8
10	6	10

Table 10.11 Data for Example 10.11

Employee No.	*Skill Index Before Training*	*Skill Index After Training*
1	5	6
2	8	10
3	4	7
4	7	10
5	1	4
6	4	6
7	6	9
8	3	6
9	5	8
10	6	10

The observations of the skill indices before (a_i) and after the training programme (b_i) of the employees are paired, and single differences (d_i values) in terms of bi – ai for i = 1, 2, 3, . . . , 10 are obtained.

Then the t test is used for the mean ($\bar{d}$) of such differences, which is known as the paired t test.

The hypotheses of this situation are stated as follows.

$$H_o : \bar{d} \leq 0$$

$$H_1 : \bar{d} > 0$$

The t statistic for this mean of the paired observations is as follows.

$$t = \frac{\bar{d}}{\left(\frac{S_d}{\sqrt{n}}\right)} \text{with } n - 1 \text{ degree of freedom.}$$

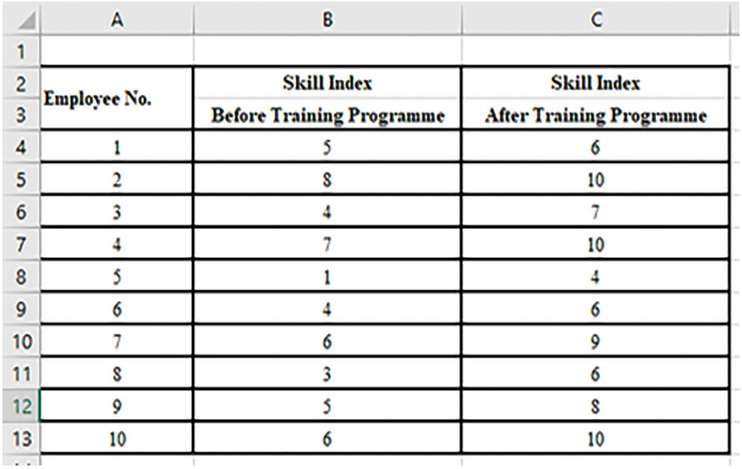
The application of the paired t test to the example problem is illustrated through suitable steps as follows.

Step 1: Input n pairs of observations, (a_i and b_i), where i = 1, 2, 3, . . . , 10 in an Excel sheet as shown in Figure 10.36.

Step 2: Click the Data button and then the Data Analysis button in Excel to show the screenshot in Figure 10.37.

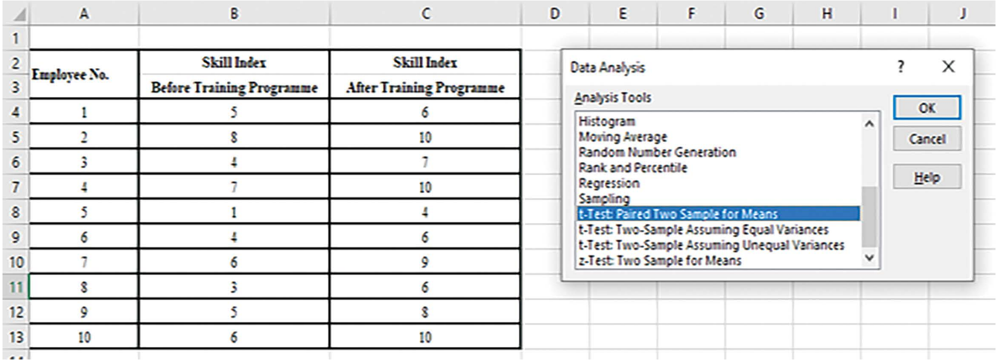
Step 3: Click the T-Test: Paired Two Sample for Means option from the dropdown menu of Figure 10.37 to show the screenshot in Figure 10.38.

Step 4: Enter the cells for the variable 1 range and those for variable 2 range, click the checkbox for Labels, retain the value of Alpha as 0.05, and click the output option



Employee No.	Skill Index Before Training Programme	Skill Index After Training Programme
1	5	6
2	8	10
3	4	7
4	7	10
5	1	4
6	4	6
7	6	9
8	3	6
9	5	8
10	6	10

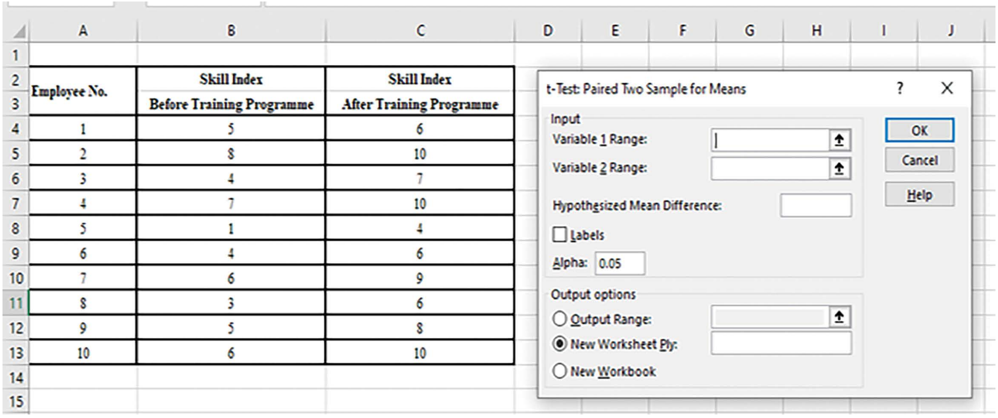
Figure 10.36 Screenshot of input of Example 10.11 in Excel sheet



Employee No.	Skill Index Before Training Programme	Skill Index After Training Programme
1	5	6
2	8	10
3	4	7
4	7	10
5	1	4
6	4	6
7	6	9
8	3	6
9	5	8
10	6	10

A
B
C
D
E
F
G
H
I
J
Data Analysis
?
X
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

Figure 10.37 Screenshot after clicking the Data button and then the Data Analysis button



Employee No.	Skill Index Before Training Programme	Skill Index After Training Programme
1	5	6
2	8	10
3	4	7
4	7	10
5	1	4
6	4	6
7	6	9
8	3	6
9	5	8
10	6	10

t-Test: Paired Two Sample for Means

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

□ Labels

Alpha: 0.05

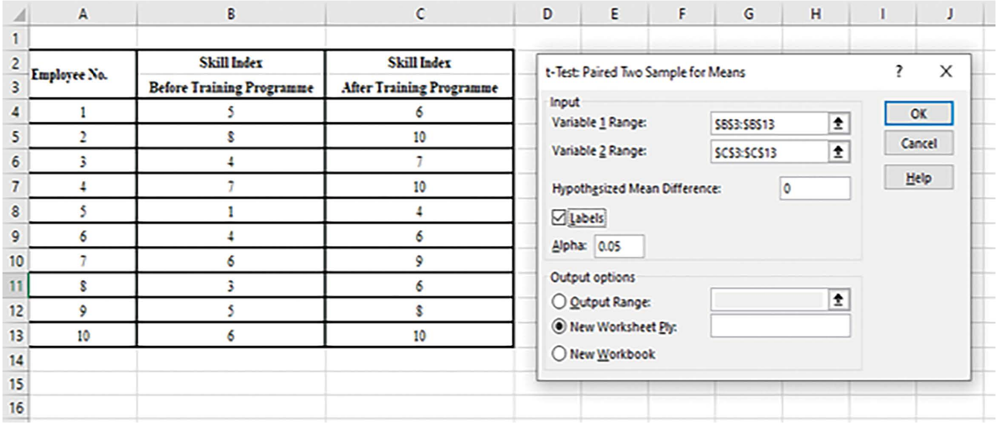
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 10.38 Screenshot after clicking T-Test: Paired Two Sample for Means option in the drop-down menu of Figure 10.37



Employee No.	Skill Index Before Training Programme	Skill Index After Training Programme
1	5	6
2	8	10
3	4	7
4	7	10
5	1	4
6	4	6
7	6	9
8	3	6
9	5	8
10	6	10

t-Test: Paired Two Sample for Means

Input

Variable 1 Range: $B$3:$B$13

Variable 2 Range: $C$3:$C$13

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

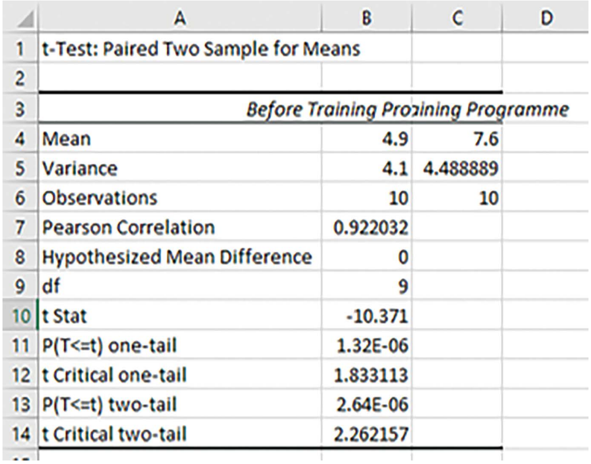
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 10.39 Screenshot of inputting the cell ranges and other options for paired t test



	A	B	C	D
1	t-Test: Paired Two Sample for Means			
2				
3		*Before Training Pro*	*ɔining Programme*	
4	Mean	4.9	7.6	
5	Variance	4.1	4.488889	
6	Observations	10	10	
7	Pearson Correlation	0.922032		
8	Hypothesized Mean Difference	0		
9	df	9		
10	t Stat	-10.371		
11	P(T<=t) one-tail	1.32E-06		
12	t Critical one-tail	1.833113		
13	P(T<=t) two-tail	2.64E-06		
14	t Critical two-tail	2.262157		

Figure 10.40 Screenshot of output of T-Test: Paired two sample for means for Example 10.11

new Worksheet Ply as shown in the screenshot of Figure 10.39, and then click the OK button to show the output in Figure 10.40.

From Figure 10.40, the key results are as follows.
Value of the t statistic = –10.371
Critical value of *t* statistic for one-tailed test = 1.833113
Since t_{Cal} (–10.371) < t_{Critical} (1.833113), accept the H_0, which means that $\bar{d} \leq 0$.
Inference: There are no significant differences among the values of d_i, which is the difference between the skill index after the training programme and the skill index before the training programme of the employees. So the training programme has not increased the skill indices of the employees significantly.

Summary

- A hypothesis is an assumption about a population.
- The hypothesis is classified into two types, the null hypothesis and alternate hypothesis.
- In the null hypothesis, an assumption about the population will be made, for example, the sample mean is less than a specified constant. The alternate hypothesis is the opposite of the null hypothesis, for example, the sample mean is more than the specified constant.
- If the computed value of the statistic is more than the corresponding table value, reject the null hypothesis; otherwise, accept the null hypothesis.
- If the *p* value is less than the significance level (α), reject the null hypothesis; otherwise, accept the null hypothesis.
- The standard normal statistic for the sample mean is $\dfrac{\left(\bar{X}-\mu\right)}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$ for tests concerning a single mean when the mean and variance of the populations are known and the size of the population is finite.

- The standard normal statistic for the difference between the sample means is

$$\frac{\left(\overline{X_1}-\overline{X_2}\right)-(\mu_1-\mu_2)}{\sigma_{\overline{X_1}-\overline{X_2}}}$$

for tests concerning the difference between two means when the variances of the populations are known and the sample sizes are large.

- $t=\dfrac{\left(\overline{X_1}-\overline{X_2}\right)-(\mu_1-\mu_2)}{S_P\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ with (n_1 + n_2 – 2) degrees of freedom is a statistic for tests concerning the difference between two means when the variances of the population are unknown and the sample sizes are small.
- The paired *t* test is used for a situation in which the value of a random variable at a particular setting may be different from another setting.
- The *t* statistic for the paired t test is:

$$t=\frac{\bar{d}}{\left(\dfrac{S_d}{\sqrt{n}}\right)}\text{with } n-1 \text{ degree of freedom.}$$

Keywords

- A hypothesis is an assumption about a population.
- The hypothesis is classified into two types, the null hypothesis and alternate hypothesis.
- In the null hypothesis, an assumption about the population will be made, for example, the sample mean is less than a specified constant. The alternate hypothesis is the opposite of the null hypothesis, for example, the sample mean is more than the specified constant.
- If the computed value of the statistic is more than the corresponding table value, reject the null hypothesis; otherwise, accept the null hypothesis.
- If the *p* value is less than the significance level (α), reject the null hypothesis; otherwise, accept the null hypothesis.
- In reality, the variance of a population may be unknown and the sample size may be small, less than or equal to 30. In such a situation, the test concerning the single mean will be carried out using the *t* test.
- The paired *t* test is used for a situation in which the value of a random variable at a particular setting may be different from another setting.

Review Questions

1. Define hypothesis and explain its types.
2. List and explain the types of decisions taken while testing the hypothesis.
3. List the pair of hypotheses for each of the tests concerning the single mean when the mean and variance of the populations are known and the size of the population is finite.

4. Give the sampling distribution for the mean.
5. With a population size of 1500, the monthly incomes of industrial workers in an industrial estate are distributed normally. The mean monthly income of the workforce is anticipated to be ₹ 50,000. The population's industrial employees' monthly salary has a variance of ₹ 5,00,000. The researcher believes that the workers' average monthly salary has decreased from the predicted mean of ₹ 50,000 in previous years. A sample of 49 industrial workers is randomly selected from the normal population, with a mean monthly income of ₹ 55,000. Using Excel, determine whether the industrial workers' mean monthly income has decreased from the projected mean of ₹ 50,000 at a significance level of 0.05.
6. The number of visitors to a mall each day is distributed normally. The mall's targeted mean daily consumer attendance and population variance are 50,000 and ₹ 2,50,000, respectively. The mall manager believes that recently there has been an increase in the number of visitors to the mall. The mean number of visitors to the mall is found to be 49,925 for a random sample of the normal population over a period of 49 days. Using Excel, determine whether the number of visitors to the mall on different days has indeed increased at a significance level of 0.05.
7. In a metro area, the annual sales of textile stores follow a normal distribution. The population's average annual sales from textile stores are intended to be ₹ 2 crores. The population's textile stores have an annual sales variance of ₹ 10 crores. According to a researcher, the performance of textile stores recently has not deviated from the population mean. A representative sample of 36 stores is chosen at random from the normal population, with a mean annual sales figure of ₹ 2.5 crores. Using Excel, determine if there has been a change in the textile stores' sales from ₹ 2 crores at a significance level of 0.05.
8. List the pair of hypotheses for each of the tests concerning the difference between two means when the variances of the populations are known and the sample sizes are large.
9. The quality manager of a refrigerator manufacturing company has an intuition that a consignment of motors received from a vendor will contain no more than five defective motors. The quality manager wants to put his intuition about how many bad motors are in a shipment to the test. As a result, he has chosen a sample of 20 consignments with a mean number of defective motors of 6 and a variance of 98. Verify the intuition of the quality manager at a significance level of 0.05 using Excel.

S. No.	*Sample 1*	*Sample 2*
1	150	153
2	151	152
3	153	152
4	148	154
5	147	152
6	146	153
7	152	151
8	147	153
9	151	153
10	152	157
11	148	152
12	153	155
13	147	152

S. No.	*Sample 1*	*Sample 2*
14	151	148
15	152	149
16	149	150
17	152	151
18	151	148
19	148	151
20	153	147
21	152	151
22	149	153
23	151	156
24	147	147
25	150	151
26	152	152
27	153	149
28	151	150
29	152	147
30	153	148

10. List the pair of hypotheses for each of the tests concerning the single mean when the variance of the population is unknown and the sample size is small.
11. A machine shop purchases a grease pack, and its weight is distributed normally. The vendor company's sales manager asserts that the grease pack's average weight is at least 510 grammes. The machine shop's quality manager is looking to confirm this assertion. He has therefore collected a sample of 20 grease packs. The grease packets in the sample have a mean weight of 505 gm and a variance of 144 gm. Using Excel, confirm the sales manager's assertion at a significance level of 0.05.
12. A car company's expensive fastener pack's weight is distributed normally. The vendor company's sales manager asserts that the fastener pack's average weight is 800 grams. The car company's quality manager wants to confirm this assertion. He has therefore collected a sample of 25 fastener packs. The fastener packets in the study have a mean and variance of 810 grams and 2500 grams, respectively. Using Excel, confirm the sales manager's assertion at a significance level of 0.05.
13. List the pair of hypotheses for each of the tests concerning the difference between two means when the variances of the population are unknown and the sample size is small.
14. Give the formula for the *t* statistic for the tests concerning the difference between two means when the variances of the population are unknown and the sample size is small.
15. Both the 0 to 10 employee contributions in shop 1 and the 0 to 10 employee contributions in shop 2 for IC engine assembly follow a normal distribution. The researcher who examined the data thinks that the contributions of the employees in Shop 1 are less than those of the employees in Shop 2. Therefore, the investigator chose 15 employees from shop 1 and 18 from shop 2. The following table displays the employee contributions from the two shops. Verify the intuition of the investigator at a significance level of 0.10 using Excel.

Sample Unit	*Shop 1*	*Shop 2*
1	8	7
2	7	8
3	8	7
4	9	9
5	7	5
6	8	8
7	7	7
8	6	5
9	8	6
10	9	7
11	6	9
12	9	6
13	7	9
14	5	5
15	8	8
16		9
17		5
18		8

16. Both companies in industrial estate 1 and industrial estate 2 have yearly revenues in crores of rupees that follow a normal distribution. According to the researcher who examined the data, mean yearly revenue of the companies in industrial estate 1 is different from that of the companies in industrial estate 2. Consequently, the investigator chose 12 companies from industrial estate 1 and 15 companies from industrial estate 2. The following table, which is in crores of rupees, displays the yearly revenues of the companies in the two industrial estates. Check the investigator's intuition at a significance level of 0.05.

Company 1	*Industrial Estate 1*	*Industrial Estate 2*
1	20	25
2	15	35
3	18	40
4	19	30
5	23	50
6	29	25
7	22	22
8	18	39
9	24	45
10	45	34
11	32	45
12	30	25
13	–	34
14	–	36
15	–	44

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. https://statisticsbyjim.com/hypothesis-testing/t-tests-Excel/ [June 25, 2020].
3. https://support.microsoft.com/en-us/office/norm-dist-function-edb1cc14-a21c-4e53-839d-8082074c9f8d [June 27, 2020].

11 Chi-Square Test

Learning Objectives

After reading this chapter, you will be able to

- Apply the chi-square test for checking the independence of categorised data.
- Analyse data using the goodness of fit test for fitting distributions.
- Apply the goodness of fit test for fitting a uniform distribution.
- Analyse the fitting of distribution for a Poisson distribution.
- Apply the goodness of fit test for an exponential distribution.
- Analyse the fitting of a distribution for a normal distribution.

11.1 Introduction

The chi-square test consists of two categories.

- Chi-square test for checking the independence of categorised data
- Goodness of fit test

The first test is used to check whether there is any dependence/independence between two categories of data. The second test is used to check whether a given set of data follows an assumed probability distribution, because the very first step of data analysis demands fitting the right type of probability distribution for the given data [1].

11.2 Chi-Square Test for Checking Independence of Categorised Data Using Excel Sheets and CHISQ.DIST.RT Function

Take a look at two categories, A and B, each with a certain number of levels. Let Category A and Category B have m and n levels, respectively. Under various combinations of their levels, these two categories may or may not have an impact on the observed frequencies. The following variables are used to define the observed frequencies for various combinations of the two groups.

o_{ij} is the observed frequency with respect to the i^{th} level of Category A and the j^{th} level of Category B, where, $i = 1, 2, 3, \ldots, m$ and $j = 1, 2, 3, \ldots, n$.

A generalised format of the categorised data is shown in Table 11.1.

The variables used in Table 11.1 are defined as follows.

$O_{i.}$ is the sum of the frequencies in row i

DOI: 10.4324/9781032671703-11

Table 11.1 Generalised Format of Categorised Data

Category A	*Category B*											Row total $o_{i.}$
	1	2	3	.	.	.	j	.	.	.	n	
1	o_{11}	o_{12}	o_{13}	.	.	.	o_{1j}	.	.	.	o_{1n}	$\boldsymbol{o}_{1.}$
2	o_{21}	o_{22}	o_{23}	.	.	.	o_{2j}	.	.	.	o_{2n}	$\boldsymbol{o}_{2.}$
3	o_{31}	o_{32}	o_{33}	.	.	.	o_{3j}	.	.	.	o_{3n}	$\boldsymbol{o}_{3.}$
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
i	o_{i1}	o_{i2}	o_{i3}	.	.	.	o_{ij}	.	.	.	o_{in}	$\boldsymbol{o}_{i\cdot}$
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
m	o_{m1}	o_{m2}	o_{m3}	.	.	.	o_{mj}	.	.	.	o_{mn}	$\boldsymbol{o}_{m\cdot}$
Column total $o_{.j}$	$\boldsymbol{o}_{.1}$	$\boldsymbol{o}_{.2}$	$\boldsymbol{o}_{.3}$	.	.	.	$\boldsymbol{o}_{\cdot j}$	.	.	.	$\boldsymbol{o}_{\cdot n}$	$\boldsymbol{o}_{..}$

$O_{.j}$ is the sum of the frequencies in column j

$O_{..}$ is the grand total of the frequencies in the entire table

$p_{i.}$ is the marginal probability of row i, $i = 1,2,3, \ldots, m$

$p_{.j}$ is the marginal probability of column j, $j = 1,2,3, \ldots, n$

p_{ij} is the joint probability with respect to row i and column j, where $i = 1,2,3, \ldots, m$ and $j = 1,2,3, \ldots, n$.

The formulas for different probabilities are as follows.

$$p_{i.} = \frac{\text{Sum of the frequencies of the row i}}{\text{Grand total of the frequencies in the entire table}}$$

$$= \frac{O_{i.}}{O_{..}}$$

$$p_{.j} = \frac{\text{Sum of the frequencies of the column j}}{\text{Grand total of the frequencies in the entire table}}$$

$$= \frac{O_{.j}}{O_{..}}$$

$$p_{ij} = \frac{\text{Sum of the frequencies of the row i and column j}}{\text{Grand total of the frequencies in the entire table}}$$

$$= \frac{O_{ij}}{O_{..}}$$

The expected frequency with respect to the i^{th} level of Category A and the j^{th} level of Category B is e_{ij}, which is given by the following formula.

$$e_{ij} = \frac{O_{i.} \times O_{,j}}{O_{..}}$$

The hypotheses of this test are as follows.

$$H_0 : p_{ij} = p_{i.} \times p_{.j}, i = 1,2,3,\ldots,m \ and \ j = 1, 2, 3,\ldots..,n$$

This means that the levels of Category A and the levels of Category B are independent in terms of their frequencies.

$$H_1 : p_{ij} > p_{i.} \times p_{.j}, \ for \ at \ least \ one \ combination \ of \ i \ and \ j$$

$$where, i = 1,2,3,\ldots.,m \ and \ j = 1, 2, 3,\ldots.,n$$

This means that the levels of Category A and the levels of Category B are not independent in terms of their frequencies.

$$The\,computed\,Chi - square\,statistic\left[Chi - square\left(computed\right)\right] = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(O_{ij} - e_{ij}\right)^2}{e_{ij}}$$

11.2.1 Directions for the Test

The degrees of freedom is the product of (m – 1) and (n – 1).

The level of significance placed at the right tail of the chi-square distribution is α.

If the chi-square (computed) value is less than the chi-square table value for the given degrees of freedom and significance level, then accept the null hypothesis H_0; otherwise, reject H_0.

Or:

If the p value for the computed chi-square value is less than the significance level (α), reject the null hypothesis; otherwise, accept the null hypothesis.

Example 11.1

The data summarising the number of respondents in a study under each combination of the level of income and the level of qualification are shown in Table 11.2.

Check whether the income is independent of the qualification while grouping the respondents, at a significance level of 0.05.

Solution

The data for Example 11.1 are shown in Table 11.3.

The level of significance (α) is 0.05.

The degrees of freedom is (3 – 1) (3 – 1), which is 4.

H_0: Row classification is independent of column classification.
H_1: Row classification is dependent on column classification.

Table 11.2 Data for Example 11.1

Level of Income	*Level of Qualification*		
	Diploma	*UG*	*PG*
Low	25	55	20
Medium	60	65	35
High	50	80	75

Table 11.3 Data for Example 11.1

Level of Income	*Level of Qualification*		
	Diploma	*UG*	*PG*
Low	25	55	20
Medium	60	65	35
High	50	80	75

The working of this problem using Excel is shown in Figure 11.1, and the guidelines to formulas for the working shown in Figure 11.1 are shown in Figure 11.2.

p Value of Chi-Square Test

The formula to obtain the *p* value at the right tail of the chi-square distribution is shown as follows [3].

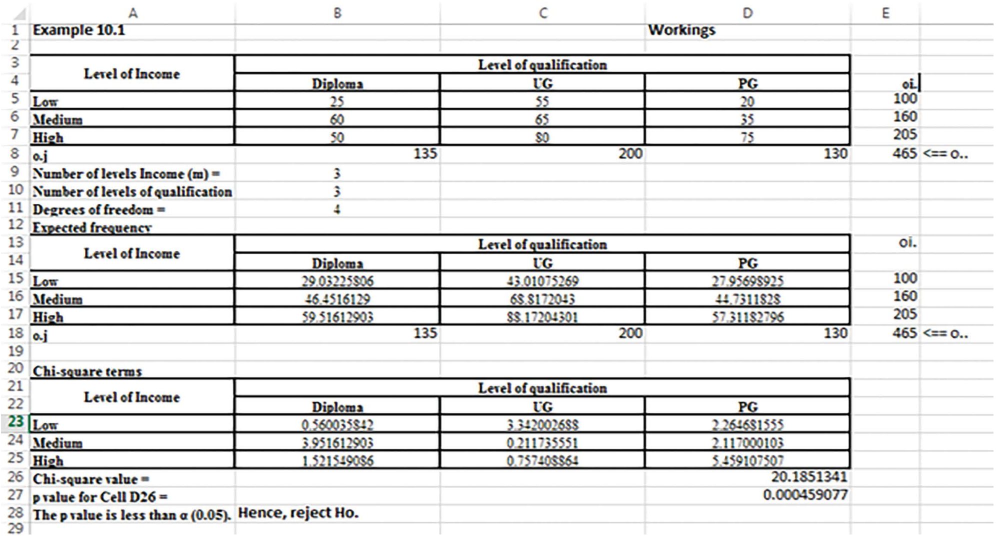
$$Formula := CHI.DIST.RT\left(X, \deg\_freedom\right)$$

In this formula, *X* represents the chi-square(computed) value and deg_freedom represents the degrees of freedom of the given data. The degrees of freedom is given by the formula $(m-1)(n-1)$, where *m* is the number of levels of the income and *n* is the number of levels of qualification. The degrees of freedom is 4.

Since the computed *p* value (0.000459007) is less than the given significance level ($\alpha = 0.05$), reject the null hypothesis. This means that there are dependencies among the levels of income and the levels of qualification.

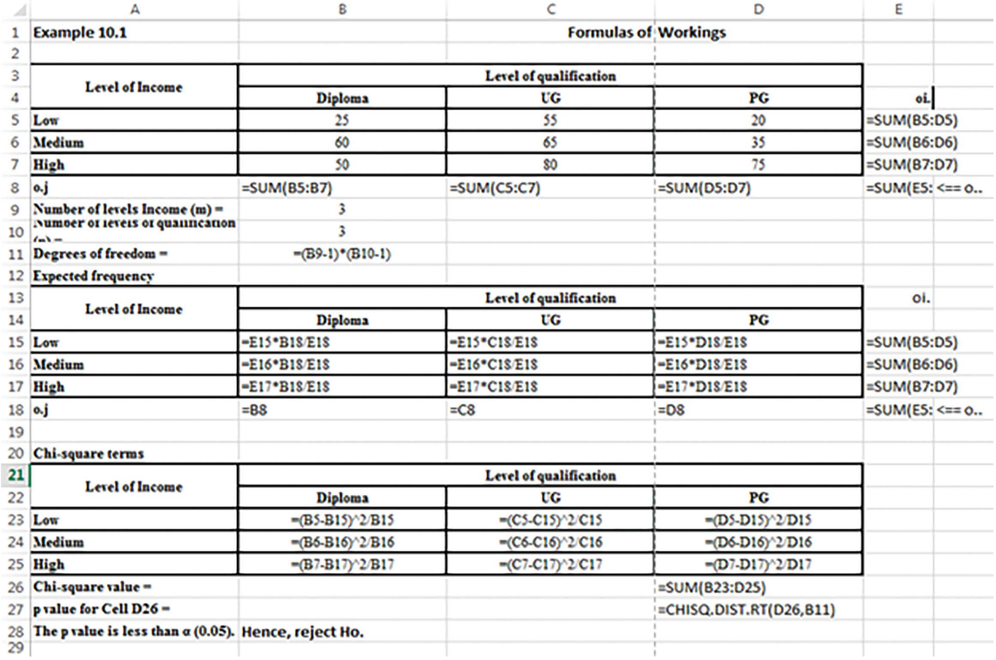
11.3 Goodness of Fit Test Using Excel Sheets and CHISQ.DIST.RT Function

Any statistical study needs to describe the data as a probability distribution in order to conduct additional analytical work. Therefore, the first task in every investigation is to gather pertinent data that are needed for the study. The representation of each sort of data should then take the shape of a useful probability distribution. Finding the appropriate probability distribution for a given set of data can be difficult because not all data will lend themselves to a certain sort of probability distribution. In such a case, the data should be assumed to have a non-standard discrete/continuous distribution as the case may be for further analysis.



	A	B	C	D	E	
1	Example 10.1			Workings		
2						
3	Level of Income	Level of qualification				
4		Diploma	UG	PG	oi.	
5	Low	25	55	20	100	
6	Medium	60	65	35	160	
7	High	50	80	75	205	
8	o.j	135	200	130	465	<== o..
9	Number of levels Income (m) =	3				
10	Number of levels of qualification	3				
11	Degrees of freedom =	4				
12	Expected frequency					
13	Level of Income	Level of qualification			oi.	
14		Diploma	UG	PG		
15	Low	29.03225806	43.01075269	27.95698925	100	
16	Medium	46.4516129	68.8172043	44.7311828	160	
17	High	59.51612903	88.17204301	57.31182796	205	
18	o.j	135	200	130	465	<== o..
19						
20	Chi-square terms					
21	Level of Income	Level of qualification				
22		Diploma	UG	PG		
23	Low	0.560035842	3.342002688	2.264681555		
24	Medium	3.951612903	0.211735551	2.117000103		
25	High	1.521549086	0.757408864	5.459107507		
26	Chi-square value =			20.1851341		
27	p value for Cell D26 =			0.000459077		
28	The p value is less than α (0.05).	Hence, reject Ho.				
29						

Figure 11.1 Working of chi-square test for categorised data for Example 11.1



	A	B	C	D	E	
1	Example 10.1		Formulas of	Workings		
2						
3	Level of Income	Level of qualification				
4		Diploma	UG	PG	oi.	
5	Low	25	55	20	=SUM(B5:D5)	
6	Medium	60	65	35	=SUM(B6:D6)	
7	High	50	80	75	=SUM(B7:D7)	
8	o.j	=SUM(B5:B7)	=SUM(C5:C7)	=SUM(D5:D7)	=SUM(E5:	<== o..
9	Number of levels Income (m) =	3				
10	Number of levels of qualification (n) =	3				
11	Degrees of freedom =	=(B9-1)*(B10-1)				
12	Expected frequency					
13	Level of Income	Level of qualification			oi.	
14		Diploma	UG	PG		
15	Low	=E15*B18/E18	=E15*C18/E18	=E15*D18/E18	=SUM(B5:D5)	
16	Medium	=E16*B18/E18	=E16*C18/E18	=E16*D18/E18	=SUM(B6:D6)	
17	High	=E17*B18/E18	=E17*C18/E18	=E17*D18/E18	=SUM(B7:D7)	
18	o.j	=B8	=C8	=D8	=SUM(E5:	<== o..
19						
20	Chi-square terms					
21	Level of Income	Level of qualification				
22		Diploma	UG	PG		
23	Low	=(B5-B15)^2/B15	=(C5-C15)^2/C15	=(D5-D15)^2/D15		
24	Medium	=(B6-B16)^2/B16	=(C6-C16)^2/C16	=(D6-D16)^2/D16		
25	High	=(B7-B17)^2/B17	=(C7-C17)^2/C17	=(D7-D17)^2/D17		
26	Chi-square value =			=SUM(B23:D25)		
27	p value for Cell D26 =			=CHISQ.DIST.RT(D26,B11)		
28	The p value is less than α (0.05).	Hence, reject Ho.				
29						

Figure 11.2 Screenshot of formulas of working of chi-square test for categorised data for Example 11.1

Curve fitting is the process of adjusting a given collection of data to a suitable probability distribution. The statistical test known as a “goodness of fit test” is performed using the chi-square test to fit a given set of data to a recognised probability distribution.

The null and alternate hypotheses for this goodness of fit test are as follows.

H_0: The given data follow an assumed probability distribution.
H_1: The given data do not follow the assumed probability distribution.

The goodness of fit test has a set of steps to be carried out to draw the inference of whether the given data follow the assumed probability distribution, as listed here.

Step 1: Input the data of a process/entity of a system of interest in the form of observed frequencies for the values of the random variable.
Step 2: Plot the data in the form of a graph and identify a near-probability distribution.
Step 3: Compute the parameters such as mean and, if necessary, variance of the data depending on the type of the probability distribution identified in Step 2.
Step 4: Compute the expected frequency for each value of the random variable using the theoretical probability distribution with parameters such as mean, and if necessary, variance/standard deviation.
Step 4: Compute the chi-square statistic using the following formula.

$$Chi-square\,statistic\,(Computed)=\sum_{i=1}^{n}\frac{(O_i-e_i)^2}{e_i}$$

where
O_i is the observed frequency of the i^{th} value of the random variable
e_i is the expected frequency of the i^{th} value of the random variable
n is the total number of observations of the random variable

Step 5: Find the table value of chi-square [χ^2] for the given degrees of freedom (*d.f.*) and the level of significance (α) by placing the significance level (α) at the right tail of the chi-square distribution.

If the computed chi-square value is more than the table chi-square value, reject the null hypothesis; otherwise, accept the null hypothesis.

Or:

Find the value of p at the right tail for the computed chi-square value and check whether it is less than the given significance level (α). If so, reject the null hypothesis; otherwise, accept the null hypothesis.

If H_0 is accepted, it amounts to an inference that the data follow the assumed probability distribution; otherwise, the data do not follow the assumed probability distribution.

Example 11.2

The daily demand of a product follows a uniform distribution. The observed frequencies of the demand values are summarised in Table 11.4.

Table 11.4 Observed Frequencies of Demand

Demand	26	27	28	29	30	31	32	33	34	35
Observed frequency (o_i)	17	15	13	14	11	15	17	14	19	15

Check whether the given data follow a uniform distribution at a significance level of 0.05.

Solution

The significance level (α) = 0.05

The degrees of freedom (df) = $n - 1 = 10 - 1 = 9$

The data for this problem are shown in Table 11.5.

H_0: The given data follow uniform distribution.
H_1: The given data do not follow uniform distribution.

The working of this problem using Excel to check the hypotheses is shown in Figure 11.3.

The expected frequency of the uniform distribution is given by the following formula.

$$Expected\,frequency\,e_i = \frac{Total\,frequency}{Total\,number\,of\,values\,for\,the\,uniform\,random\,variable}$$

for $i = 1, 2, 3, \ldots, n$, where n is the number of values for the uniform random variable.

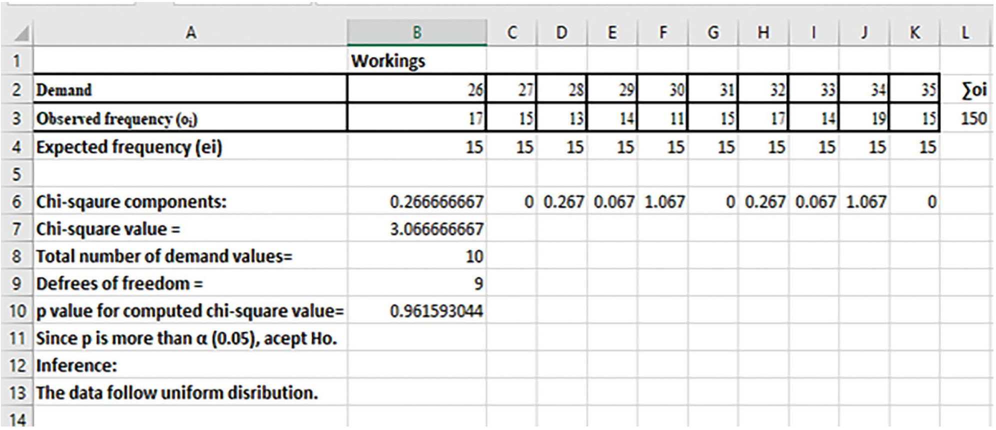
p Value at the Right Tail of Chi-Square Distribution

The formula to obtain the p value at the right tail of the chi-square distribution is shown as follows [3].

$$Formula := CHI.DIST.RT\left(X, \deg\_freedom\right)$$

Table 11.5 Observed Frequencies of Demand

Demand	26	27	28	29	30	31	32	33	34	35
Observed frequency (o_i)	17	15	13	14	11	15	17	14	19	15



	A	B	C	D	E	F	G	H	I	J	K	L
1		Workings										
2	Demand	26	27	28	29	30	31	32	33	34	35	∑oi
3	Observed frequency (o_i)	17	15	13	14	11	15	17	14	19	15	150
4	Expected frequency (ei)	15	15	15	15	15	15	15	15	15	15	
5												
6	Chi-sqaure components:	0.266666667	0	0.267	0.067	1.067	0	0.267	0.067	1.067	0	
7	Chi-square value =	3.066666667										
8	Total number of demand values=	10										
9	Defrees of freedom =	9										
10	p value for computed chi-square value=	0.961593044										
11	Since p is more than α (0.05), acept Ho.											
12	Inference:											
13	The data follow uniform disribution.											
14												

Figure 11.3 Screenshot of working of chi-square test for uniform distribution of Example 11.2

In this formula, X represents the chi-square (computed) value, and deg_freedom represents the degrees of freedom of the given data. The degrees of freedom is given by the formula (n – 1), where n is the number of values for the uniform random variable in the given data.

The working for this problem is shown in Figure 11.3. The guidelines for the formulas of the working in Figure 11.3 are shown in Figure 11.4.

For the given problem, the degrees of freedom is 9, which is 10 – 1. Since the p value (0.961593044) at the right tail is more than the level of significance of (0.05), the null hypothesis is accepted. Hence, the given data follow a uniform distribution.

Example 11.3

Table 11.6 provides a summary of the arrival rates (number of assemblies per 15-minute interval) of an assembled product at the final inspection station of the CPU assembly line of a computer company.

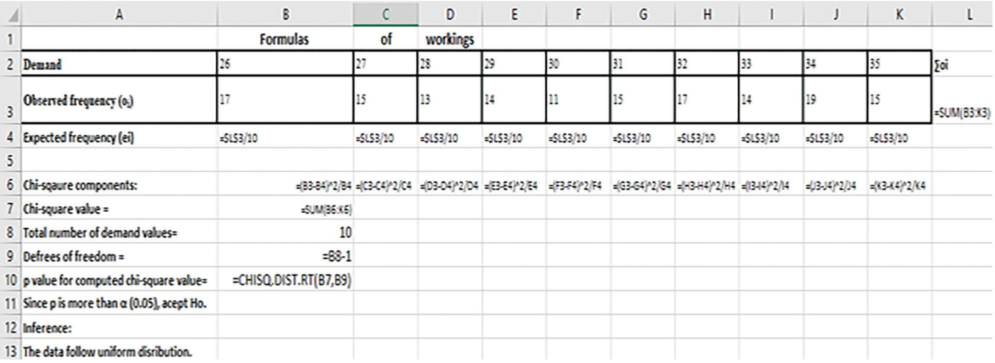
Check whether the given data follow a Poisson distribution at a significance level of 0.05.

Solution

The data for Example 11.3 are shown in Table 11.7.

The hypotheses of the problem are as listed.

H_0: The data for the problem follow a Poisson distribution.
H_1: The data for the problem do not follow a Poisson distribution.



	A	B	C	D	E	F	G	H	I	J	K	L
1		Formulas	of	workings								
2	Demand	26	27	28	29	30	31	32	33	34	35	∑oi
3	Observed frequency (oᵢ)	17	15	13	14	11	15	17	14	19	15	=SUM(B3:K3)
4	Expected frequency (ei)	=$L$3/10	=$L$3/10	=$L$3/10	=$L$3/10	=$L$3/10	=$L$3/10	=$L$3/10	=$L$3/10	=$L$3/10	=$L$3/10	
5												
6	Chi-sqaure components:	=(B3-B4)^2/B4	=(C3-C4)^2/C4	=(D3-D4)^2/D4	=(E3-E4)^2/E4	=(F3-F4)^2/F4	=(G3-G4)^2/G4	=(H3-H4)^2/H4	=(I3-I4)^2/I4	=(J3-J4)^2/J4	=(K3-K4)^2/K4	
7	Chi-square value =	=SUM(B6:K6)										
8	Total number of demand values=	10										
9	Defrees of freedom =	=B8-1										
10	p value for computed chi-square value=	=CHISQ.DIST.RT(B7,B9)										
11	Since p is more than α (0.05), acept Ho.											
12	Inference:											
13	The data follow uniform disribution.											

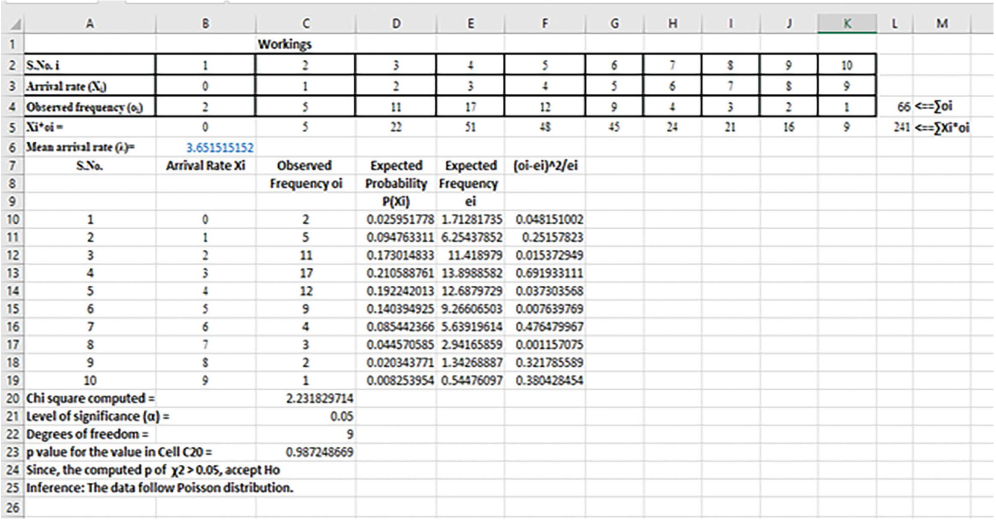
Figure 11.4 Screenshot of formulas for working of chi-square test for uniform distribution of Example 11.2

Table 11.6 Data on Arrival Rates and Frequencies

Serial no. (i)	1	2	3	4	5	6	7	8	9	10
Arrival rate (X_i)	0	1	2	3	4	5	6	7	8	9
Observed frequency (o_i)	2	5	11	17	12	9	4	3	2	1

Table 11.7 Data for Example 11.3

S. No. (i)	1	2	3	4	5	6	7	8	9	10
Arrival rate (X_i)	0	1	2	3	4	5	6	7	8	9
Observed frequency (o_i)	2	5	11	17	12	9	4	3	2	1



	A	B	C	D	E	F	G	H	I	J	K	L	M
1			Workings										
2	S.No. i	1	2	3	4	5	6	7	8	9	10		
3	Arrival rate (Xi)	0	1	2	3	4	5	6	7	8	9		
4	Observed frequency (oi)	2	5	11	17	12	9	4	3	2	1	66	<==∑oi
5	Xi*oi =	0	5	22	51	48	45	24	21	16	9	241	<==∑Xi*oi
6	Mean arrival rate (λ)=	3.651515152											
7	S.No.	Arrival Rate Xi	Observed	Expected	Expected	(oi-ei)^2/ei							
8			Frequency oi	Probability	Frequency								
9				P(Xi)	ei								
10	1	0	2	0.025951778	1.71281735	0.048151002							
11	2	1	5	0.094763311	6.25437852	0.25157823							
12	3	2	11	0.173014833	11.418979	0.015372949							
13	4	3	17	0.210588761	13.8988582	0.691933111							
14	5	4	12	0.192242013	12.6879729	0.037303568							
15	6	5	9	0.140394925	9.26606503	0.007639769							
16	7	6	4	0.085442366	5.63919614	0.476479967							
17	8	7	3	0.044570585	2.94165859	0.001157075							
18	9	8	2	0.020343771	1.34268887	0.321785589							
19	10	9	1	0.008253954	0.54476097	0.380428454							
20	Chi square computed =		2.231829714										
21	Level of significance (α) =		0.05										
22	Degrees of freedom =		9										
23	p value for the value in Cell C20 =		0.987248669										
24	Since, the computed p of χ2 > 0.05, accept Ho												
25	Inference: The data follow Poisson distribution.												
26													

Figure 11.5 Screenshot of working of chi-square test for Poisson distribution of Example 11.3

The formula for the mean arrival rate is as follows.

$$Mean\,arrival\,rate, \lambda = \frac{\sum_{i=1}^{10} X_i \times O_i}{\sum_{i=1}^{10} O_i}$$

The Poisson probability distribution is given by the following formula.

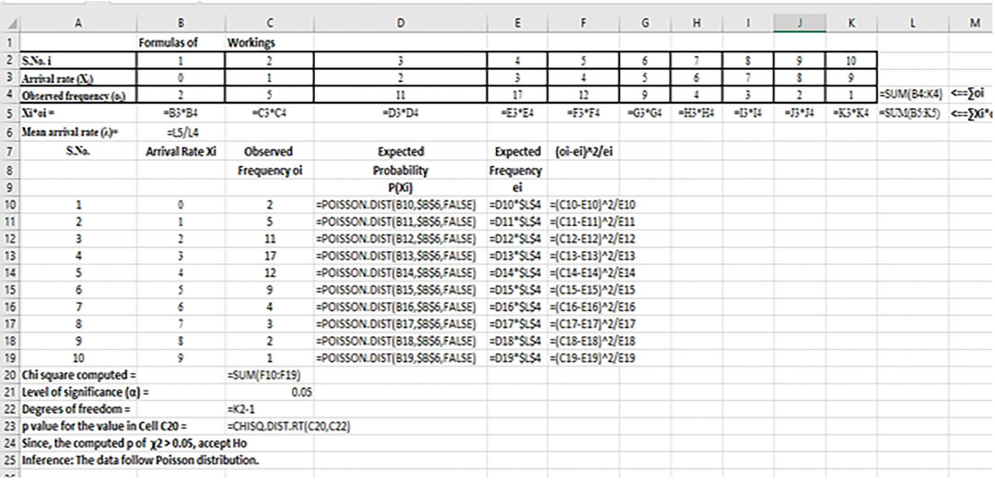
$$P(X) = \frac{\lambda^X e^{-\lambda}}{X!}$$

Where
λ is the arrival rate (number of arrivals per 15-minute interval)
X is the random variable representing arrival rate

The working for the goodness of fit test applied to check the Poisson distribution for the given data is shown in Figure 11.5. The guidelines for the formulas of the working shown in Figure 11.5 are shown in Figure 11.6.

The important formulas used in this example are as listed.

Formula : = POISSON.DIST(Value of random variable X, Mean of Poisson distribution, FALSE)



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Formulas of	Workings										
2	S.No. i	1	2	3	4	5	6	7	8	9	10		
3	Arrival rate (Xi)	0	1	2	3	4	5	6	7	8	9		
4	Observed frequency (oi)	2	5	11	17	12	9	4	3	2	1	=SUM(B4:K4)	<==∑oi
5	Xi*oi =	=B3*B4	=C3*C4	=D3*D4	=E3*E4	=F3*F4	=G3*G4	=H3*H4	=I3*I4	=J3*J4	=K3*K4	=SUM(B5:K5)	<==∑Xi*
6	Mean arrival rate (λ)=	=L5/L4											
7	S.No.	Arrival Rate Xi	Observed	Expected	Expected	(oi-ei)^2/ei							
8			Frequency oi	Probability	Frequency								
9				P(Xi)	ei								
10	1	0	2	=POISSON.DIST(B10,$B$6,FALSE)	=D10*$L$4	=(C10-E10)^2/E10							
11	2	1	5	=POISSON.DIST(B11,$B$6,FALSE)	=D11*$L$4	=(C11-E11)^2/E11							
12	3	2	11	=POISSON.DIST(B12,$B$6,FALSE)	=D12*$L$4	=(C12-E12)^2/E12							
13	4	3	17	=POISSON.DIST(B13,$B$6,FALSE)	=D13*$L$4	=(C13-E13)^2/E13							
14	5	4	12	=POISSON.DIST(B14,$B$6,FALSE)	=D14*$L$4	=(C14-E14)^2/E14							
15	6	5	9	=POISSON.DIST(B15,$B$6,FALSE)	=D15*$L$4	=(C15-E15)^2/E15							
16	7	6	4	=POISSON.DIST(B16,$B$6,FALSE)	=D16*$L$4	=(C16-E16)^2/E16							
17	8	7	3	=POISSON.DIST(B17,$B$6,FALSE)	=D17*$L$4	=(C17-E17)^2/E17							
18	9	8	2	=POISSON.DIST(B18,$B$6,FALSE)	=D18*$L$4	=(C18-E18)^2/E18							
19	10	9	1	=POISSON.DIST(B19,$B$6,FALSE)	=D19*$L$4	=(C19-E19)^2/E19							
20	Chi square computed =		=SUM(F10:F19)										
21	Level of significance (α) =		0.05										
22	Degrees of freedom =		=K2-1										
23	p value for the value in Cell C20 =		=CHISQ.DIST.RT(C20,C22)										
24	Since, the computed p of χ2 > 0.05, accept Ho												
25	Inference: The data follow Poisson distribution.												

Figure 11.6 Screenshot of formulas of working of Poisson distribution of Example 11.3

This is used to compute the Poisson probability for a given value of *X*.

p Value at the Right Tail of Chi-Square Distribution

The formula to obtain the *p* value at the right tail of the chi-square distribution is shown as follows.

$$\text{Formula}: \quad = \text{CHI.DIST.RT}(\text{X, deg\_freedom})$$

In this formula, *X* represents the chi-square (computed) value, and deg_freedom represents the degrees of freedom of the given data. The degrees of freedom is given by the formula ($n - 1$), where n is the number of values for the Poisson random variable in the given data. For the given data of this problem, the degrees of freedom is 9, which is 10 – 1.

In Figure 11.5, the *p* value at the right tail of the chi-square distribution is 0.987248669, which is more than the given significance level of 0.05. Hence, accept the null hypothesis, which means that the given data follow a Poisson distribution.

Example 11.4

Consider the data shown in Table 11.8, which summarises the service times in minutes and their frequencies at a booking counter. Check whether these data follow an exponential distribution at a significance level of 0.1.

Solution

The data for Example 11.4 are shown in Table 11.9.

The hypotheses of the problem are as listed.

H_0: The data for the problem follow an exponential distribution.
H_1: The data for the problem do not follow an exponential distribution

Table 11.8 Service Times and Frequencies of Customers

S. No.	*Random*	*Observed*
	Variable (X_i)	*Frequency*(o_i)
1	10	90
2	11	13
3	12	12
4	14	11
5	15	10
6	16	9
7	17	8
8	18	7
9	19	6
10	20	5
11	21	4
12	22	3
13	23	2
14	24	1

Table 11.9 Data for Example 11.4

S. No.	*Random*	*Observed*
	Variable (X_i) *(Service Time)*	*Frequency* (o_i)
1	10	90
2	11	13
3	12	12
4	14	11
5	15	10
6	16	9
7	17	8
8	18	7
9	19	6
10	20	5
11	21	4
12	22	3
13	23	2
14	24	1

The formula for mean arrival rate is as follows.

$$Mean\,servcie\,time\left(\frac{1}{\mu}\right)=\frac{\sum_{i=1}^{10}(X_i\times O_i)}{\sum_{i=1}^{10}O_i}$$

The expected service rate (μ) is the inverse of the mean service time, which is given by the following formula.

$$\text{Expected service rate }(\mu)=\frac{1}{\left(\frac{1}{\mu}\right)}=\mu$$

The formula for the exponential distribution is as follows.

$$P(X) = \mu e^{-\mu X}$$

where
μ is the service rate, which is more than 0
X is the exponential random variable, which is more than 0

Screenshots of working of the chi-square test and of formulas of working of the chi-square test for Example 11.4 are given in Figures 11.7 and 11.8, respectively.

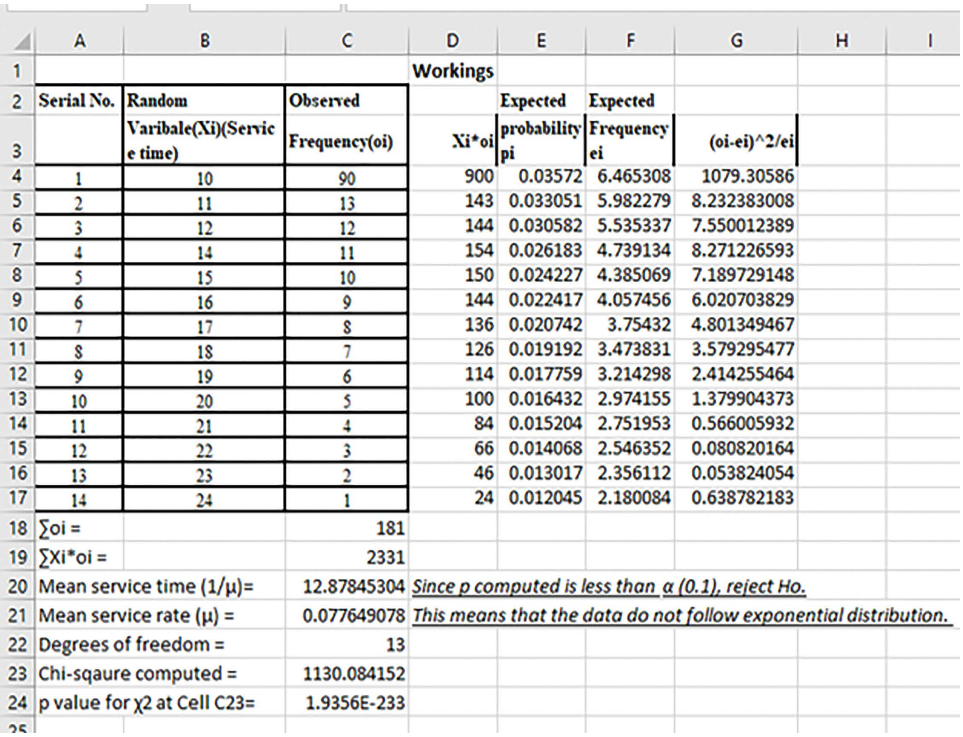
The important formulas used in this example are as listed.

Formula : = EXPON.DIST(Value of random variable X, Mean of exponential distribution, FALSE)

This is used to compute the exponential probability for a given value of X.

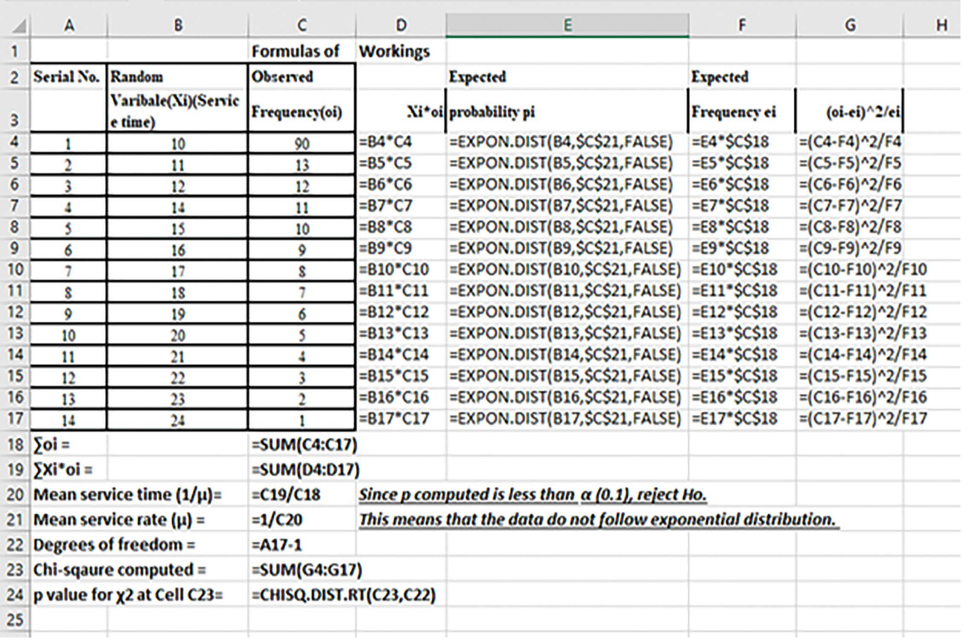
In Figure 11.8, the formula for the exponential distribution that is present in Excel is used to compute the exponential probability mass function.

p Value at the Right Tail of Chi-Square Distribution



	A	B	C	D	E	F	G	H	I
1				Workings					
2	Serial No.	Random	Observed		Expected	Expected			
3		Varibale(Xi)(Service time)	Frequency(oi)	Xi*oi	probability pi	Frequency ei	(oi-ei)^2/ei		
4	1	10	90	900	0.03572	6.465308	1079.30586		
5	2	11	13	143	0.033051	5.982279	8.232383008		
6	3	12	12	144	0.030582	5.535337	7.550012389		
7	4	14	11	154	0.026183	4.739134	8.271226593		
8	5	15	10	150	0.024227	4.385069	7.189729148		
9	6	16	9	144	0.022417	4.057456	6.020703829		
10	7	17	8	136	0.020742	3.75432	4.801349467		
11	8	18	7	126	0.019192	3.473831	3.579295477		
12	9	19	6	114	0.017759	3.214298	2.414255464		
13	10	20	5	100	0.016432	2.974155	1.379904373		
14	11	21	4	84	0.015204	2.751953	0.566005932		
15	12	22	3	66	0.014068	2.546352	0.080820164		
16	13	23	2	46	0.013017	2.356112	0.053824054		
17	14	24	1	24	0.012045	2.180084	0.638782183		
18	∑oi =		181						
19	∑Xi*oi =		2331						
20	Mean service time (1/μ)=		12.87845304	*Since p computed is less than α (0.1), reject Ho.*					
21	Mean service rate (μ) =		0.077649078	*This means that the data do not follow exponential distribution.*					
22	Degrees of freedom =		13						
23	Chi-sqaure computed =		1130.084152						
24	p value for χ2 at Cell C23=		1.9356E-233						

Figure 11.7 Screenshot of working of chi-square test for exponential distribution of Example 11.4



	A	B	C	D	E	F	G	H
1			Formulas of	Workings				
2	Serial No.	Random	Observed		Expected	Expected		
3		Varibale(Xi)(Service time)	Frequency(oi)	Xi*oi	probability pi	Frequency ei	(oi-ei)^2/ei	
4	1	10	90	=B4*C4	=EXPON.DIST(B4,$C$21,FALSE)	=E4*$C$18	=(C4-F4)^2/F4	
5	2	11	13	=B5*C5	=EXPON.DIST(B5,$C$21,FALSE)	=E5*$C$18	=(C5-F5)^2/F5	
6	3	12	12	=B6*C6	=EXPON.DIST(B6,$C$21,FALSE)	=E6*$C$18	=(C6-F6)^2/F6	
7	4	14	11	=B7*C7	=EXPON.DIST(B7,$C$21,FALSE)	=E7*$C$18	=(C7-F7)^2/F7	
8	5	15	10	=B8*C8	=EXPON.DIST(B8,$C$21,FALSE)	=E8*$C$18	=(C8-F8)^2/F8	
9	6	16	9	=B9*C9	=EXPON.DIST(B9,$C$21,FALSE)	=E9*$C$18	=(C9-F9)^2/F9	
10	7	17	8	=B10*C10	=EXPON.DIST(B10,$C$21,FALSE)	=E10*$C$18	=(C10-F10)^2/F10	
11	8	18	7	=B11*C11	=EXPON.DIST(B11,$C$21,FALSE)	=E11*$C$18	=(C11-F11)^2/F11	
12	9	19	6	=B12*C12	=EXPON.DIST(B12,$C$21,FALSE)	=E12*$C$18	=(C12-F12)^2/F12	
13	10	20	5	=B13*C13	=EXPON.DIST(B13,$C$21,FALSE)	=E13*$C$18	=(C13-F13)^2/F13	
14	11	21	4	=B14*C14	=EXPON.DIST(B14,$C$21,FALSE)	=E14*$C$18	=(C14-F14)^2/F14	
15	12	22	3	=B15*C15	=EXPON.DIST(B15,$C$21,FALSE)	=E15*$C$18	=(C15-F15)^2/F15	
16	13	23	2	=B16*C16	=EXPON.DIST(B16,$C$21,FALSE)	=E16*$C$18	=(C16-F16)^2/F16	
17	14	24	1	=B17*C17	=EXPON.DIST(B17,$C$21,FALSE)	=E17*$C$18	=(C17-F17)^2/F17	
18	∑oi =		=SUM(C4:C17)					
19	∑Xi*oi =		=SUM(D4:D17)					
20	Mean service time (1/µ)=		=C19/C18	*Since p computed is less than α (0.1), reject Ho.*				
21	Mean service rate (µ) =		=1/C20	*This means that the data do not follow exponential distribution.*				
22	Degrees of freedom =		=A17-1					
23	Chi-sqaure computed =		=SUM(G4:G17)					
24	p value for χ2 at Cell C23=		=CHISQ.DIST.RT(C23,C22)					
25								

Figure 11.8 Screenshot of formulas for working of chi-square test applied to exponential distribution of Example 11.4

The formula to obtain the p value at the right tail of the chi-square distribution is shown as follows.

$$\text{Formula} := \text{CHI.DIST.RT}\left(X, \text{deg}\_\text{freedom}\right)$$

In this formula, X represents the chi-square (computed) value, and deg_freedom represents the degrees of freedom of the given data. The degrees of freedom is given by the formula $(n - 1)$, where n is the number of values for the exponential random variable in the given data. For the given data of this problem, the degrees of freedom is 9, which is 10 – 1.

From Figure 11.7, it is observed that the p value at the right tail of the chi-square distribution is 1.9356 E-233, which is almost 0. Since it is less than the given significance level of 0.1, the null hypothesis is rejected.

Inference: The given data do not follow an exponential distribution.

Example 11.5

Table 11.10 displays the data on the internal diameter of the bearings manufactured in a production line. At a significance level of 0.05, determine whether the data are normally distributed.

Table 11.10 Data for Internal Diameter of Bearings

S. No.	*Random Variable* (X_i)	*Observed Frequency* (o_i)
1	10	2
2	11	5
3	12	7
4	13	10
5	14	15
6	15	22
7	16	30
8	17	21
9	18	16
10	19	9
11	20	6
12	21	4
13	22	3
14	23	2

Table 11.11 Data of Example 11.5

S. No.	*Random Variable* (X_i)	*Observed Frequency* (o_i)
1	10	2
2	11	5
3	12	7
4	13	10
5	14	15
6	15	22
7	16	30
8	17	21
9	18	16
10	19	9
11	20	6
12	21	4
13	22	3
14	23	2

Solution

The data for Example 11.5 are shown in Table 11.11.

The hypotheses of the problem are as listed.

H_0: The data for the problem follow a normal distribution.
H_1: The data for the problem do not follow a normal distribution

The number of observations (n) = 14

The formula for the mean of the internal diameter of the bearings is as follows.

$$\text{Mean}\left(\bar{X}\right)=\frac{\sum_{i=1}^{14}\left(X_iO_i\right)}{\sum_{i=1}^{14}O_i}$$

The standard deviation of the internal diameter of the bearings is given by the following formula.

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}o_i\times\left(X_i-\bar{X}\right)^2}{\sum_{i=1}^{n}o_i}}$$

where
X_i is the i^{th} value of the random variable, $i = 1, 2, 3, \ldots, n$
$\bar{X}$ is the mean of the values of the random variable
o_i is the i^{th} observed frequency, $i = 1, 2, 3, \ldots, n$
n is the number of values of the random variable

The formula for the normal distribution is as follows.

$$P(X)=\frac{1}{\sigma\times\sqrt{2\pi}}\times e^{\left(-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right)}, where -\infty < X < +\infty$$

where
μ is the mean of the normal distribution
σ is the standard deviation of the normal distribution
X is the normal random variable

The working of the chi-square test and guidelines for formulas are given in Figures 11.9 and 11.10, respectively.

The important formulas used in this example are listed here [2].

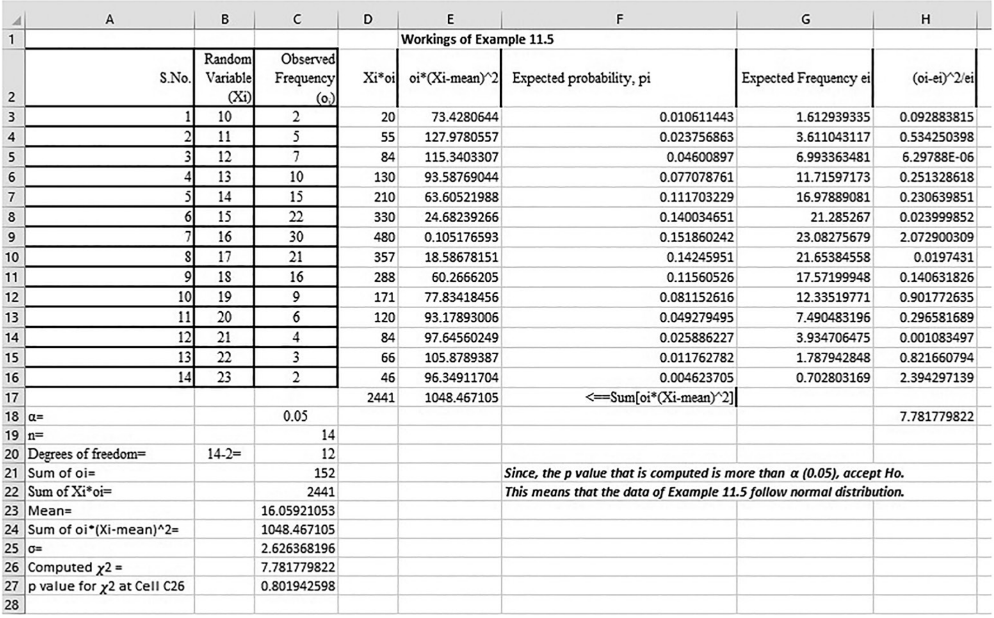
Formula : = NORM.DIST(Value of the random variable X, Mean of normal distribution, Standard deviation of the normal distribution, FALSE)

This is used to compute the normal probability for a given value of *X*.

p Value at the Right Tail of Chi-Square Distribution

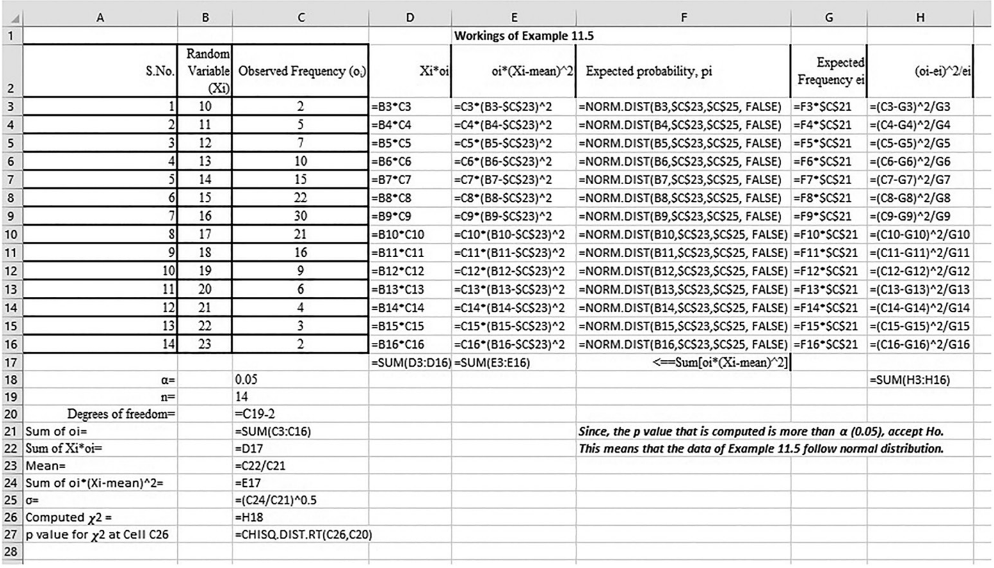
$$\text{Formula : = CHI.DIST.RT}\left(X,\text{deg\_freedom}\right)$$

In this equation, *X* stands for the computed chi-square value, and deg_freedom stands for the degrees of freedom of the given data. The formula (n – 2), where n is the number of values for the normal random variable in the given data, gives the degrees of freedom. The degrees of freedom for the given data of problem is 12, which equals 14 – 2.



	A	B	C	D	E	F	G	H
1					Workings of Example 11.5			
2	S.No.	Random Variable (Xi)	Observed Frequency (oi)	Xi*oi	oi*(Xi-mean)^2	Expected probability, pi	Expected Frequency ei	(oi-ei)^2/ei
3	1	10	2	20	73.4280644	0.010611443	1.612939335	0.092883815
4	2	11	5	55	127.9780557	0.023756863	3.611043117	0.534250398
5	3	12	7	84	115.3403307	0.04600897	6.993363481	6.29788E-06
6	4	13	10	130	93.58769044	0.077078761	11.71597173	0.251328618
7	5	14	15	210	63.60521988	0.111703229	16.97889081	0.230639851
8	6	15	22	330	24.68239266	0.140034651	21.285267	0.023999852
9	7	16	30	480	0.105176593	0.151860242	23.08275679	2.072900309
10	8	17	21	357	18.58678151	0.14245951	21.65384558	0.0197431
11	9	18	16	288	60.2666205	0.11560526	17.57199948	0.140631826
12	10	19	9	171	77.83418456	0.081152616	12.33519771	0.901772635
13	11	20	6	120	93.17893006	0.049279495	7.490483196	0.296581689
14	12	21	4	84	97.64560249	0.025886227	3.934706475	0.001083497
15	13	22	3	66	105.8789387	0.011762782	1.787942848	0.821660794
16	14	23	2	46	96.34911704	0.004623705	0.702803169	2.394297139
17				2441	1048.467105	<==Sum[oi*(Xi-mean)^2]		
18	α=		0.05					7.781779822
19	n=		14					
20	Degrees of freedom=	14-2=	12					
21	Sum of oi=		152			*Since, the p value that is computed is more than α (0.05), accept Ho.*		
22	Sum of Xi*oi=		2441			*This means that the data of Example 11.5 follow normal distribution.*		
23	Mean=		16.05921053					
24	Sum of oi*(Xi-mean)^2=		1048.467105					
25	σ=		2.626368196					
26	Computed χ2 =		7.781779822					
27	p value for χ2 at Cell C26		0.801942598					
28								

Figure 11.9 Screenshot of working of chi-square test applied to normal distribution of Example 11.5



	A	B	C	D	E	F	G	H
1					Workings of Example 11.5			
2	S.No.	Random Variable (Xi)	Observed Frequency (oi)	Xi*oi	oi*(Xi-mean)^2	Expected probability, pi	Expected Frequency ei	(oi-ei)^2/ei
3	1	10	2	=B3*C3	=C3*(B3-$C$23)^2	=NORM.DIST(B3,$C$23,$C$25, FALSE)	=F3*$C$21	=(C3-G3)^2/G3
4	2	11	5	=B4*C4	=C4*(B4-$C$23)^2	=NORM.DIST(B4,$C$23,$C$25, FALSE)	=F4*$C$21	=(C4-G4)^2/G4
5	3	12	7	=B5*C5	=C5*(B5-$C$23)^2	=NORM.DIST(B5,$C$23,$C$25, FALSE)	=F5*$C$21	=(C5-G5)^2/G5
6	4	13	10	=B6*C6	=C6*(B6-$C$23)^2	=NORM.DIST(B6,$C$23,$C$25, FALSE)	=F6*$C$21	=(C6-G6)^2/G6
7	5	14	15	=B7*C7	=C7*(B7-$C$23)^2	=NORM.DIST(B7,$C$23,$C$25, FALSE)	=F7*$C$21	=(C7-G7)^2/G7
8	6	15	22	=B8*C8	=C8*(B8-$C$23)^2	=NORM.DIST(B8,$C$23,$C$25, FALSE)	=F8*$C$21	=(C8-G8)^2/G8
9	7	16	30	=B9*C9	=C9*(B9-$C$23)^2	=NORM.DIST(B9,$C$23,$C$25, FALSE)	=F9*$C$21	=(C9-G9)^2/G9
10	8	17	21	=B10*C10	=C10*(B10-$C$23)^2	=NORM.DIST(B10,$C$23,$C$25, FALSE)	=F10*$C$21	=(C10-G10)^2/G10
11	9	18	16	=B11*C11	=C11*(B11-$C$23)^2	=NORM.DIST(B11,$C$23,$C$25, FALSE)	=F11*$C$21	=(C11-G11)^2/G11
12	10	19	9	=B12*C12	=C12*(B12-$C$23)^2	=NORM.DIST(B12,$C$23,$C$25, FALSE)	=F12*$C$21	=(C12-G12)^2/G12
13	11	20	6	=B13*C13	=C13*(B13-$C$23)^2	=NORM.DIST(B13,$C$23,$C$25, FALSE)	=F13*$C$21	=(C13-G13)^2/G13
14	12	21	4	=B14*C14	=C14*(B14-$C$23)^2	=NORM.DIST(B14,$C$23,$C$25, FALSE)	=F14*$C$21	=(C14-G14)^2/G14
15	13	22	3	=B15*C15	=C15*(B15-$C$23)^2	=NORM.DIST(B15,$C$23,$C$25, FALSE)	=F15*$C$21	=(C15-G15)^2/G15
16	14	23	2	=B16*C16	=C16*(B16-$C$23)^2	=NORM.DIST(B16,$C$23,$C$25, FALSE)	=F16*$C$21	=(C16-G16)^2/G16
17				=SUM(D3:D16)	=SUM(E3:E16)	<==Sum[oi*(Xi-mean)^2]		
18	α=		0.05					=SUM(H3:H16)
19	n=		14					
20	Degrees of freedom=		=C19-2					
21	Sum of oi=		=SUM(C3:C16)			*Since, the p value that is computed is more than α (0.05), accept Ho.*		
22	Sum of Xi*oi=		=D17			*This means that the data of Example 11.5 follow normal distribution.*		
23	Mean=		=C22/C21					
24	Sum of oi*(Xi-mean)^2=		=E17					
25	σ=		=(C24/C21)^0.5					
26	Computed χ2 =		=H18					
27	p value for χ2 at Cell C26		=CHISQ.DIST.RT(C26,C20)					
28								

Figure 11.10 Screenshot of formulas of working of chi-square test applied to normal distribution of Example 11.5

Since the computed value of p (0.801942598) at the right is more than α (0.05), accept the null hypothesis. This means that the given data follow normal distribution.

Summary

- Consider two categories, Category A and Category B, each with a specified number of levels. Let the number of levels in Category A and Category B be m and n, respectively. These two categories may or may not have an effect on the observed frequencies under different combinations of the levels, which can be checked by the chi-square test.
- If the chi-square (computed) value is less than the chi-square table value for the given degrees of freedom and significance level, then accept the null hypothesis H_0; otherwise, reject H_0.
- If the p value for the computed chi-square value is less than the significance level (α), reject the null hypothesis; otherwise, accept the null hypothesis.
- The process of fitting a given set of data to a fitting probability distribution is called curve fitting. The statistical test that is carried out using the chi-square test to fit a given set of data to a recognised probability distribution is called the goodness of fit test.
- The null and alternate hypotheses for the goodness of fit test are as follows.

  H_0: The given data follow an assumed probability distribution.
  H_1: The given data do not follow the assumed probability distribution.

- If H_0 is accepted, it amounts to an inference that the data follow the assumed probability distribution; otherwise, the data do not follow the assumed probability distribution.

Keywords

- The chi-square test for categorised data deals with two categories of data, Category A and Category B, each with a specified number of levels, to check whether there is dependency among the observed frequencies under different combinations of the levels of those categories.
- If the chi-square (computed) value is less than the chi-square table value for the given degrees of freedom and significance level, then accept the null hypothesis H_0; otherwise, reject H_0.
- If the p value for the computed chi-square value is less than the significance level (α), reject the null hypothesis; otherwise, accept the null hypothesis.
- The goodness of fit test is the process of fitting a given set of data to an assumed fitting probability distribution.
- Null and alternate hypotheses for the goodness of fit test are:

  H_0: The given data follow an assumed probability distribution.
  H_1: The given data do not follow the assumed probability distribution.

Review Questions

1. Discuss the generalised aspect of the chi-square test for checking the independence of categorised data.

2. Discuss the direction for the chi-square test for checking the independence of categorised data.
3. The following table displays data summarising the number of respondents in a study for each combination of the level of region and the level of qualification.

Region	*Level of Qualification*		
	Diploma	*UG*	*PG*
North	20	50	20
South	55	60	30
East	30	75	50
West	25	20	30

Check whether the region is independent of the qualification while grouping the respondents at a significance level of 0.10 using Excel.

4. List and explain the steps of goodness of fit test using Excel.
5. The daily production volume of an assembly line follows a uniform distribution. The observed frequency of the daily production volumes are summarised in the table.

Daily Production Volume (Units)	196	197	198	199	200	201	202	203	204	205
Observed Frequency (o_i)	10	9	11	14	7	12	13	7	11	12

Check whether the given data follow a uniform distribution at a significance level of 0.01 using Excel.

6. The arrival rates (number of flights per 15-minute interval) at an airport are summarised in the following table.

Serial No. (i)	1	2	3	4	5	6	7
Arrival rate (X_i)	0	1	2	3	4	5	6
Observed frequency (o_i)	3	6	11	18	12	8	2

Check whether the given data follow a Poisson distribution at a significance level of 0.05 using Excel.

7. Take a look at the information in the table, which lists the frequencies and service times for planes at an airport runway in minutes. Verify whether the data follow an exponential distribution at a significance level of 0.10.
Service times and frequencies of flights

S. No.	*Random*	*Observed*
	Variable (X_i)	*Frequency (o_i)*
1	10	50
2	11	10
3	12	6
4	14	4
5	15	1

8. The following table displays data on the weights of electrodes manufactured by a manufacturer in grams. At a significance level of 0.01, check whether the data follow a normal distribution.

S. No.	*Random Variable* (X_i)	*Observed Frequency* (o_i)
1	96	1
2	97	2
3	98	10
4	99	15
5	100	22
6	101	30
7	102	12
8	103	11
9	104	4
10	105	2

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. https://support.microsoft.com/en-us/office/norm-dist-function-edb1cc14-a21c-4e53–839d-8082074c9f8d [June 27, 2020].
3. https://support.microsoft.com/en-us/office/chisq-dist-function-8486b05e-5c05-4942-a9ea-f6b341518732 [June 27, 2020].

12 Nonparametric Tests

Learning Objectives

After studying this chapter, you will be able to

- Distinguish between parametric and non-parametric tests.
- Analyse problems through a test of the hypothesis for a one-tailed one-sample sign test when the sample size is small.
- Understand the process of testing the hypothesis for a two-tailed one-sample sign test when the sample size is small.
- Study the method of hypothesis testing for the one-tailed one-sample sign test when the sample size is large.
- Analyse problems through hypothesis testing for a two-tailed one-sample sign test when the sample size is large.
- Understand the procedure of performing the Kolmogorov-Smirnov test for fitting a given distribution.
- Study problems using a run test to check if the occurrence of the runs of the given stream of symbols is random for small samples.
- Analyse problems using a run test to check if the occurrence of the runs of the given stream of symbols is random for large samples.
- Analyse the problems using a test of hypothesis for one-tailed two-samples sign test for small samples.
- Understand the process of using a test of hypothesis for two-tailed two-samples sign tests for small samples.
- Study the process of using a test of hypothesis for one-tailed two-samples sign tests for large samples.
- Analyse the problems through a test of hypothesis for two-tailed two-samples sign tests for large samples.
- Study the process of testing the hypothesis for a median test to check whether two samples, which are independent, are drawn from two populations with the same median.
- Understand the test of hypothesis for the Mann-Whitney U test to check whether two samples are drawn from different populations with the same distribution.
- Understand the testing of a hypothesis for the *K*-sample median test to check whether the *K* samples, which are independent and drawn from *K* populations, have the same median.
- Analyse the problem using a test of hypothesis for the Kruskal-Wallis test to check whether the K samples, which are independent, are drawn from K identical populations.

DOI: 10.4324/9781032671703-12

12.1 Introduction

The majority of data from real-world circumstances are expected to follow a normal distribution, whose mean and variance can be estimated from such data. The associated hypotheses can then be tested using the standard tests that were described in the prior chapter. These tests are therefore known as parametric tests.

However, there are some situations where the data will not be normal and/or the parameter(s) cannot be calculated. The test that is performed for this type of data is known as a non-parametric test because it does not require any parameters. Even very tiny samples can be used with the non-parametric test. As a result, it is crucial in pilot studies because there will be a relatively small sample size. Additionally, it can be applied to both ordinal and nominal data. The data that is descriptive in nature such as male and female will form an example of nominal data. It is used to label a variable, which is qualitative. An example of ordinal data is ranked data. Nonparametric tests only require a small number of calculations because the parameters are not calculated. If an investigator fails to identify a suitable parametric test for the data, he can select a suitable nonparametric test [1].

The different non-parametric tests presented in this chapter are as follows.

- One-sample tests, which include the one-sample sign test, Kolmogorov-Smirnov test, and run test for randomness.
- Two-sample tests, which include the two-sample sign test, median test, and Mann-Whitney U test (rank-sum test)
- *K*-sample test, which include the median test and Kruskal-Wallis test (H test).

12.2 One-Sample Sign Tests

Think about a situation where there is a non-normal population with a continuous symmetrical distribution. This population is sampled with a size *n*. The probability (p) that a sample value exceeds the mean value is one-half. In this case, the sample's observations are categorised based on the sample median, with observations greater than or equal to being given a plus (+) sign and those less than or equal to being given a minus (-) sign. The probability that X is greater than the number of plus signs is then calculated by using the number of plus signs as the value of the random variable X of the binomial distribution, with $p = 1/2$ and the number of trials n, to check the following hypotheses [2].

The possible hypotheses testing of this test are as follows.

1. One-tailed one-sample sign test when the sample size is small.
2. Two-tailed one-sample sign test when the sample size is small.
3. One-tailed one-sample sign test when the sample size is large.
4. Two-tailed one-sample sign test when the sample size is large.

12.2.1 One-Tailed One-Sample Sign Test When Sample Size Is Small Using Excel Sheets and BINOM.DIST Function

The hypotheses of the one-tailed one-sample sign test when the sample size is small are as follows.

Test 1

$H_0 p = 1/2$
$H_1: p > ½$

Test 2

H_0: p = 1/2
H_1: p < 1/2

In this test, a small random sample is taken from a non-normal population, and then it is tested against a median value (μ) such that the observations in the sample are more than that median (μ) or less than that median (μ) at a significance level of α using a binomial distribution.

Example 12.1

The final assembly operation in an assembly line producing two-wheelers was the subject of a time study engineer's data collection. Nine observations, 21, 34, 28, 15, 27, 15, 27, 26, and 12, were collected. Using the sign test with a significance level of 0.05, check whether the final assembly operation took 20 minutes (H_0: =20) as opposed to the alternative hypothesis H1: > 20.

Solution

Sample size (n) = 9

Significance level (α = 0.05)

The observations of the final assembly operations time: 21, 34, 28, 15, 27, 15, 27, 26, and 12

Let X be a random variable representing a plus sign when 20 is subtracted from each observation.

H_0: μ = 20 or p = 1/2
H_1: μ > 20 or p > 1/2

The calculations are shown in Figure 12.1 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.1 are shown in Figure 12.2.

12.2.2 Two-Tailed One-Sample Sign Test When Sample Size Is Small Using Excel Sheets and BINOM.DIST

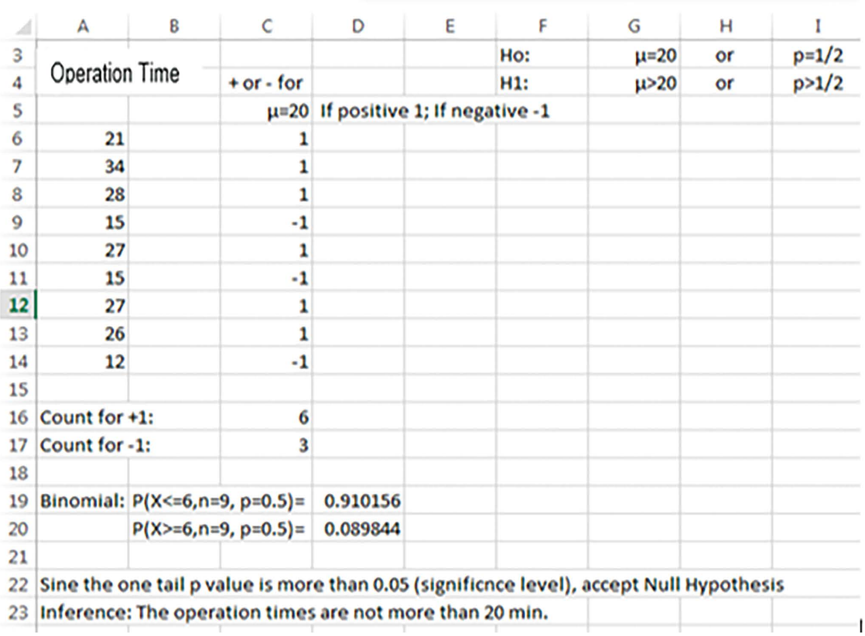
The hypotheses of the two-tailed one-sample sign test when the sample size is small are as follows.

H_0: μ = 75 or p = 1/2
H_1: $\mu \neq$ 75or $p \neq$ 1/2

In this test, a small random sample is taken from a non-normal population, and then it is tested against a median value (μ) such that the observations in the sample are not equal to that median (μ) at a significance level α using binomial distribution.

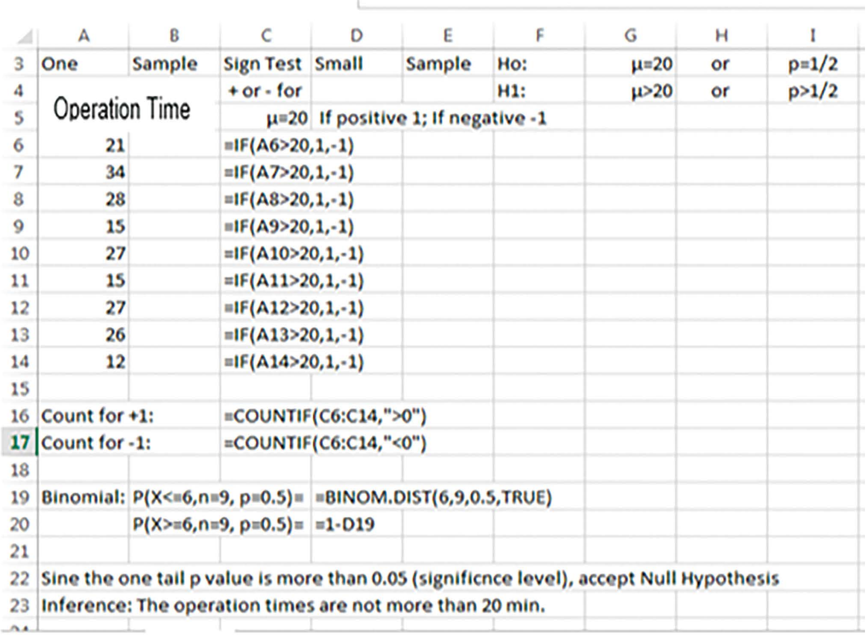
Example 12.2

The CEOs of ten businesses in the automobile industry receive monthly wages of 70, 60, 90, 85, 105, 72, 95, 88, 60, and 100 lakhs of rupees. Using a sign test with a significance



	A	B	C	D	E	F	G	H	I
3	Operation Time					Ho:	$\mu=20$	or	p=1/2
4			+ or - for			H1:	$\mu>20$	or	p>1/2
5			$\mu=20$	If positive 1; If negative -1					
6	21		1						
7	34		1						
8	28		1						
9	15		-1						
10	27		1						
11	15		-1						
12	27		1						
13	26		1						
14	12		-1						
15									
16	Count for +1:		6						
17	Count for -1:		3						
18									
19	Binomial:	P(X<=6,n=9, p=0.5)=		0.910156					
20		P(X>=6,n=9, p=0.5)=		0.089844					
21									
22	Sine the one tail p value is more than 0.05 (significnce level), accept Null Hypothesis								
23	Inference: The operation times are not more than 20 min.								

Figure 12.1 Screenshot of working of one-tailed one-sample sign test for small sample



	A	B	C	D	E	F	G	H	I
3	One	Sample	Sign Test	Small	Sample	Ho:	$\mu=20$	or	p=1/2
4	Operation Time		+ or - for			H1:	$\mu>20$	or	p>1/2
5			$\mu=20$	If positive 1; If negative -1					
6	21		=IF(A6>20,1,-1)						
7	34		=IF(A7>20,1,-1)						
8	28		=IF(A8>20,1,-1)						
9	15		=IF(A9>20,1,-1)						
10	27		=IF(A10>20,1,-1)						
11	15		=IF(A11>20,1,-1)						
12	27		=IF(A12>20,1,-1)						
13	26		=IF(A13>20,1,-1)						
14	12		=IF(A14>20,1,-1)						
15									
16	Count for +1:		=COUNTIF(C6:C14,">0")						
17	Count for -1:		=COUNTIF(C6:C14,"<0")						
18									
19	Binomial:	P(X<=6,n=9, p=0.5)=		=BINOM.DIST(6,9,0.5,TRUE)					
20		P(X>=6,n=9, p=0.5)=		=1-D19					
21									
22	Sine the one tail p value is more than 0.05 (significnce level), accept Null Hypothesis								
23	Inference: The operation times are not more than 20 min.								

Figure 12.2 Screenshot of guidelines for formulas of working of one-tailed one-sample sign test for small sample

level of 0.10, determine whether the CEO's monthly salary is 75 lakhs (H_0: μ = 75 lakh) as against the alternate hypothesis (H_1: $\mu \neq$ 75 lakh).

Solution

Sample size (n) = 10

Significance level (α = 0.10)

The monthly salaries (lakhs of rupees) of CEOs: 70, 60, 90, 85, 105, 72, 95, 88, 60, and 100

Let

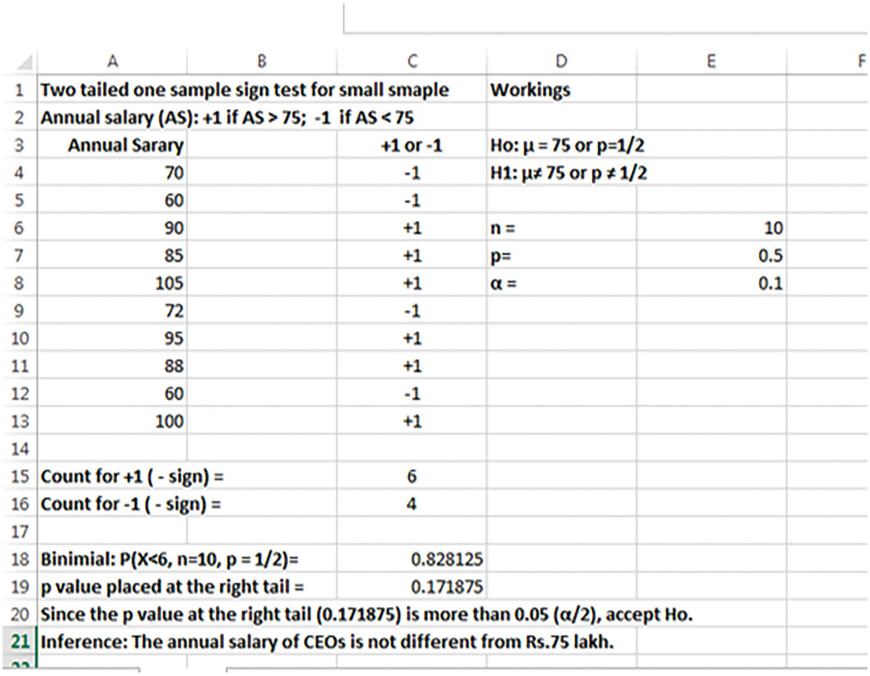
X be a random variable representing plus sign when 75 is subtracted from each observation

H_0: μ = 75 or p = 1/2
H_1: $\mu \neq$ 75 or $p \neq$ 1/2

The calculations are shown in Figure 12.3 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.3 are shown in Figure 12.4.

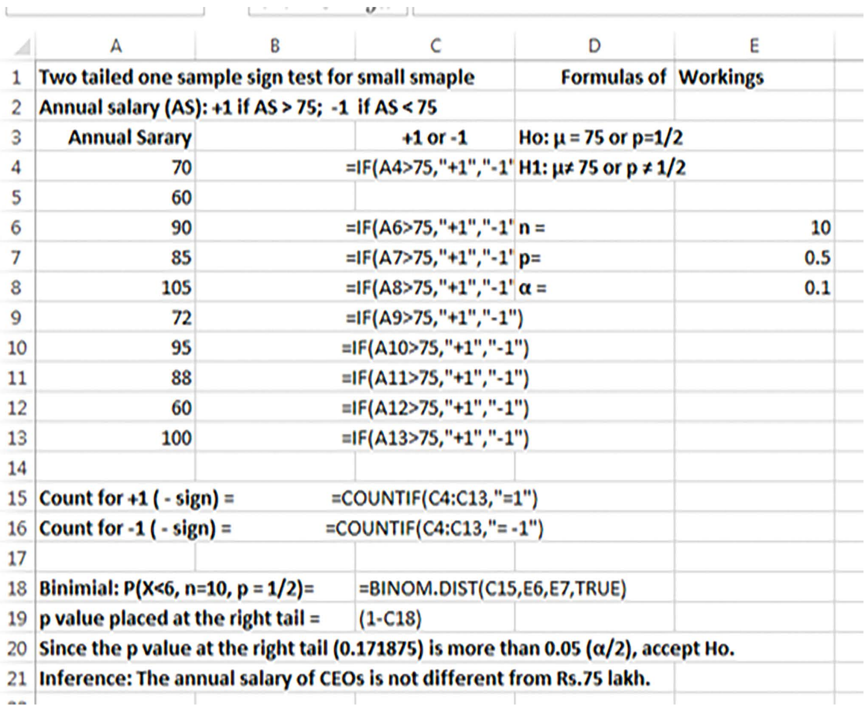
12.2.3 One-Tailed One-Sample Sign Test When Sample Size Is Large Using Excel Sheets and NORM.S.DIST Function

The random sample of n units with the condition that np as well as n (1 – p) is greater than or equal to 5 is selected. A normal approximation to a binomial distribution with



	A	B	C	D	E	F
1	**Two tailed one sample sign test for small smaple**			**Workings**		
2	**Annual salary (AS): +1 if AS > 75; -1 if AS < 75**					
3	**Annual Sarary**		**+1 or -1**	**Ho: μ = 75 or p=1/2**		
4	70		-1	**H1: μ≠ 75 or p ≠ 1/2**		
5	60		-1			
6	90		+1	n =	10	
7	85		+1	p=	0.5	
8	105		+1	α =	0.1	
9	72		-1			
10	95		+1			
11	88		+1			
12	60		-1			
13	100		+1			
14						
15	**Count for +1 (- sign) =**		6			
16	**Count for -1 (- sign) =**		4			
17						
18	**Binimial: P(X<6, n=10, p = 1/2)=**		0.828125			
19	**p value placed at the right tail =**		0.171875			
20	**Since the p value at the right tail (0.171875) is more than 0.05 (α/2), accept Ho.**					
21	**Inference: The annual salary of CEOs is not different from Rs.75 lakh.**					

Figure 12.3 Screenshot of working of two-tailed one-sample sign test for small sample



	A	B	C	D	E
1	Two tailed one sample sign test for small smaple			Formulas of	Workings
2	Annual salary (AS): +1 if AS > 75; -1 if AS < 75				
3	Annual Sarary		+1 or -1	Ho: μ = 75 or p=1/2	
4	70		=IF(A4>75,"+1","-1'	H1: μ≠ 75 or p ≠ 1/2	
5	60				
6	90		=IF(A6>75,"+1","-1'	n =	10
7	85		=IF(A7>75,"+1","-1'	p=	0.5
8	105		=IF(A8>75,"+1","-1'	α =	0.1
9	72		=IF(A9>75,"+1","-1")		
10	95		=IF(A10>75,"+1","-1")		
11	88		=IF(A11>75,"+1","-1")		
12	60		=IF(A12>75,"+1","-1")		
13	100		=IF(A13>75,"+1","-1")		
14					
15	Count for +1 (- sign) =		=COUNTIF(C4:C13,"=1")		
16	Count for -1 (- sign) =		=COUNTIF(C4:C13,"= -1")		
17					
18	Binimial: P(X<6, n=10, p = 1/2)=		=BINOM.DIST(C15,E6,E7,TRUE)		
19	p value placed at the right tail =		(1-C18)		
20	Since the p value at the right tail (0.171875) is more than 0.05 (α/2), accept Ho.				
21	Inference: The annual salary of CEOs is not different from Rs.75 lakh.				

Figure 12.4 Screenshot for guidelines of formulas of the working of two-tailed one-sample sign test for small sample

p = 1/2 is used to test this hypothesis. Here, X is the random variable, which represents the number of plus signs.

The hypotheses of the one-tailed one-sample sign test when the sample size is large are as follows.

Test 1

H_0: p = 1/2
H_1: p > 1/2

Test 2

H_0: p = 1/2
H_1: p < ½

Let
$X \sim B(X, n, p)$

If n is more than 30, this binomial distribution can be approximated to a normal distribution $N\left[np, np(1-p)\right]$.

The standard normal statistic Z of such distribution is as follows.

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

Example 12.3

The Alpha Pharmaceutical Company produces syrup in capsule form, whose volume follows a non-normal distribution. This population's specified syrup volume in each capsule is 15 ml. There will be negative effects if there is too much syrup in the capsule. The volume in the capsule, according to the company's quality engineer, is not more than 15 ml. As a result, the purchase manager of Sigma Hospital, who places an order with the Alpha Pharmaceutical Company for that capsule, chose 32 capsules at random and determined their volumes as follows.

14	16	14	17	16	17	14	16
13.5	17	14	14	17	12	13	16
17	16	16	14	13	12	16	14
12	14.5	16	18	19	16	17	18

Check the claim of the quality engineer of the Alpha Pharmaceutical Company that the volume of the capsule is 15 ml ($H_0 = 15$) against the alternate hypothesis $H_1 > 15$ using a sign test with a significance level of 0.10.

Solution

The data for Example 12.3 are shown in Table 12.1.

The random sample of 32 with the condition that np as well as $n\ (1 - p)$ is greater than or equal to 5 is selected. A normal approximation to the binomial distribution with $p = 1/2$ is used to test this hypothesis.

Here, X is the random variable, which represents the number of plus signs.

The hypotheses of the one-tailed one-sample sign test when the sample size is large are as follows.

Test 1

H_0: $p = 1/2$
H_1: $p > 1/2$

Let

$X \sim B(X, n, p)$

Since n is more than 30, the data can be approximated to a normal distribution $N\left[np, np(1-p)\right]$.

Table 12.1 Data for Example 12.3

14	16	14	17	16	17	14	16
13.5	17	14	14	17	12	13	16
17	16	16	14	13	12	16	14
12	14.5	16	18	19	16	17	18

The standard normal statistic Z of this distribution is as follows.

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

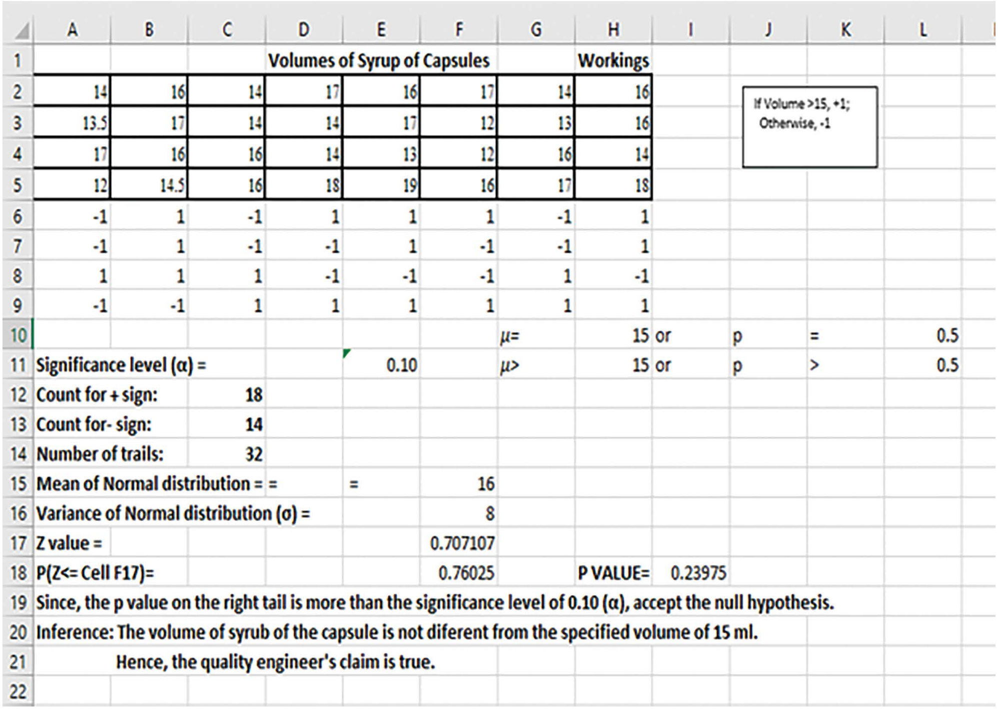
The calculations are shown in Figure 12.5 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.5 are shown in Figure 12.6.

Example 12.4

A certain type of food grain's yield, measured in kilograms, has a non-normal distribution. According to the study, each plot's yield of food grains is greater than 750 kg. The yields of the 36 plots he chose at random to test his intuition are displayed as follows.

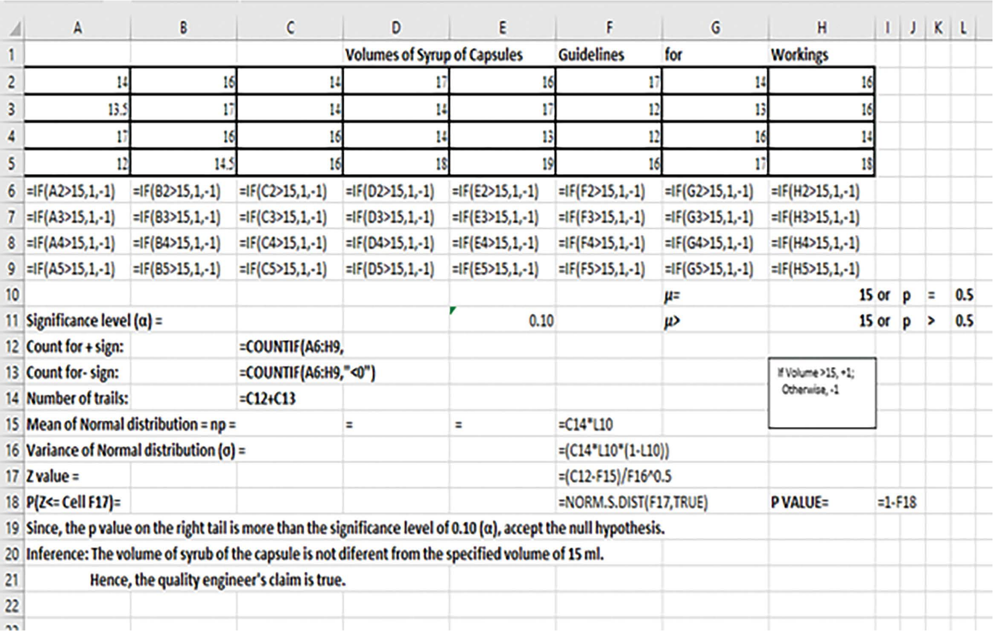
970	700	760	740	600	660	720	850	770
790	745	740	740	740	820	980	700	710
810	850	780	710	890	830	730	860	720
745	700	700	750	740	730	800	710	747

Check the intuition of the researcher, using sign test at a significance level of 0.10.



	A	B	C	D	E	F	G	H	I	J	K	L
1				Volumes of Syrup of Capsules				Workings				
2	14	16	14	17	16	17	14	16		If Volume >15, +1; Otherwise, -1		
3	13.5	17	14	14	17	12	13	16				
4	17	16	16	14	13	12	16	14				
5	12	14.5	16	18	19	16	17	18				
6	-1	1	-1	1	1	1	-1	1				
7	-1	1	-1	-1	1	-1	-1	1				
8	1	1	1	-1	-1	-1	1	-1				
9	-1	-1	1	1	1	1	1	1				
10							μ=	15	or	p	=	0.5
11	Significance level (α) =				0.10		μ>	15	or	p	>	0.5
12	Count for + sign:		18									
13	Count for- sign:		14									
14	Number of trails:		32									
15	Mean of Normal distribution = =				=	16						
16	Variance of Normal distribution (σ) =					8						
17	Z value =					0.707107						
18	P(Z<= Cell F17)=					0.76025		P VALUE=	0.23975			
19	Since, the p value on the right tail is more than the significance level of 0.10 (α), accept the null hypothesis.											
20	Inference: The volume of syrub of the capsule is not diferent from the specified volume of 15 ml.											
21	Hence, the quality engineer's claim is true.											
22												

Figure 12.5 Screenshot of working of Example 12.3



	A	B	C	D	E	F	G	H	I	J	K	L
1				Volumes of Syrup of Capsules		Guidelines	for	Workings				
2	14	16	14	17	16	17	14	16				
3	13.5	17	14	14	17	12	13	16				
4	17	16	16	14	13	12	16	14				
5	12	14.5	16	18	19	16	17	18				
6	=IF(A2>15,1,-1)	=IF(B2>15,1,-1)	=IF(C2>15,1,-1)	=IF(D2>15,1,-1)	=IF(E2>15,1,-1)	=IF(F2>15,1,-1)	=IF(G2>15,1,-1)	=IF(H2>15,1,-1)				
7	=IF(A3>15,1,-1)	=IF(B3>15,1,-1)	=IF(C3>15,1,-1)	=IF(D3>15,1,-1)	=IF(E3>15,1,-1)	=IF(F3>15,1,-1)	=IF(G3>15,1,-1)	=IF(H3>15,1,-1)				
8	=IF(A4>15,1,-1)	=IF(B4>15,1,-1)	=IF(C4>15,1,-1)	=IF(D4>15,1,-1)	=IF(E4>15,1,-1)	=IF(F4>15,1,-1)	=IF(G4>15,1,-1)	=IF(H4>15,1,-1)				
9	=IF(A5>15,1,-1)	=IF(B5>15,1,-1)	=IF(C5>15,1,-1)	=IF(D5>15,1,-1)	=IF(E5>15,1,-1)	=IF(F5>15,1,-1)	=IF(G5>15,1,-1)	=IF(H5>15,1,-1)				
10							μ=	15	or	p	=	0.5
11	Significance level (α) =				0.10		μ>	15	or	p	>	0.5
12	Count for + sign:		=COUNTIF(A6:H9,									
13	Count for- sign:		=COUNTIF(A6:H9,"<0")					If Volume >15, +1; Otherwise, -1				
14	Number of trails:		=C12+C13									
15	Mean of Normal distribution = np =			=	=	=C14*L10						
16	Variance of Normal distribution (σ) =					=(C14*L10*(1-L10))						
17	Z value =					=(C12-F15)/F16^0.5						
18	P(Z<= Cell F17)=					=NORM.S.DIST(F17,TRUE)		P VALUE=	=1-F18			
19	Since, the p value on the right tail is more than the significance level of 0.10 (α), accept the null hypothesis.											
20	Inference: The volume of syrub of the capsule is not diferent from the specified volume of 15 ml.											
21	Hence, the quality engineer's claim is true.											
22												

Figure 12.6 Screenshot of guidelines for the formulas of the working of Example 12.3

Solution

The data for Example 12.4 are shown in Table 12.2.

A random sample of the yields of 36 plots with the condition that np as well as $n(1-p)$ is greater than or equal to 5 is selected. A normal approximation to the binomial distribution with $p = 1/2$ is used to test this hypothesis.

Here, X is the random variable of the yield of the food grain, which represents the number of plus signs.

The hypotheses of the one-tailed one-sample sign test when the sample size is large are as follows.

Test 1

H_0: $\mu = 750$ or $p = 1/2$
H_1: $\mu < 750$ or $p < 1/2$

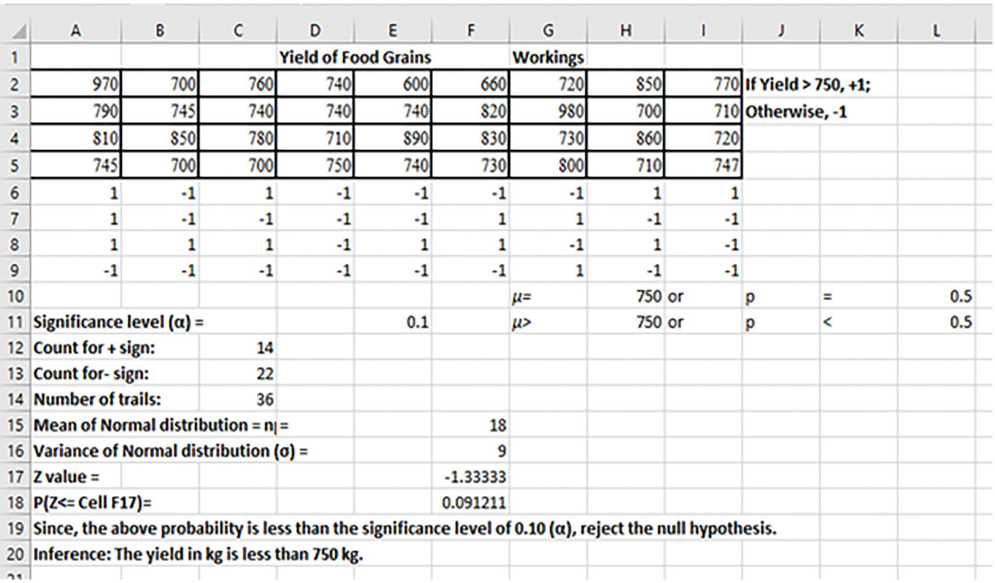
Let

$X \sim B(X, n, p)$

Since the sample size (n) is more than 30, the data can be approximated to a normal distribution $N\left[np, np(1-p)\right]$.

Table 12.2 Data for Example 12.4

970	700	760	740	600	660	720	850	770
790	745	740	740	740	820	980	700	710
810	850	780	710	890	830	730	860	720
745	700	700	750	740	730	800	710	747



	A	B	C	D	E	F	G	H	I	J	K	L
1				Yield of Food Grains			Workings					
2	970	700	760	740	600	660	720	850	770	If Yield > 750, +1;		
3	790	745	740	740	740	820	980	700	710	Otherwise, -1		
4	810	850	780	710	890	830	730	860	720			
5	745	700	700	750	740	730	800	710	747			
6	1	-1	1	-1	-1	-1	-1	1	1			
7	1	-1	-1	-1	-1	1	1	-1	-1			
8	1	1	1	-1	1	1	-1	1	-1			
9	-1	-1	-1	-1	-1	-1	1	-1	-1			
10							μ=	750	or	p	=	0.5
11	Significance level (α) =				0.1		μ>	750	or	p	<	0.5
12	Count for + sign:		14									
13	Count for- sign:		22									
14	Number of trails:		36									
15	Mean of Normal distribution = n =					18						
16	Variance of Normal distribution (σ) =					9						
17	Z value =					-1.33333						
18	P(Z<= Cell F17)=					0.091211						
19	Since, the above probability is less than the significance level of 0.10 (α), reject the null hypothesis.											
20	Inference: The yield in kg is less than 750 kg.											

Figure 12.7 Screenshot of working of Example 12.4

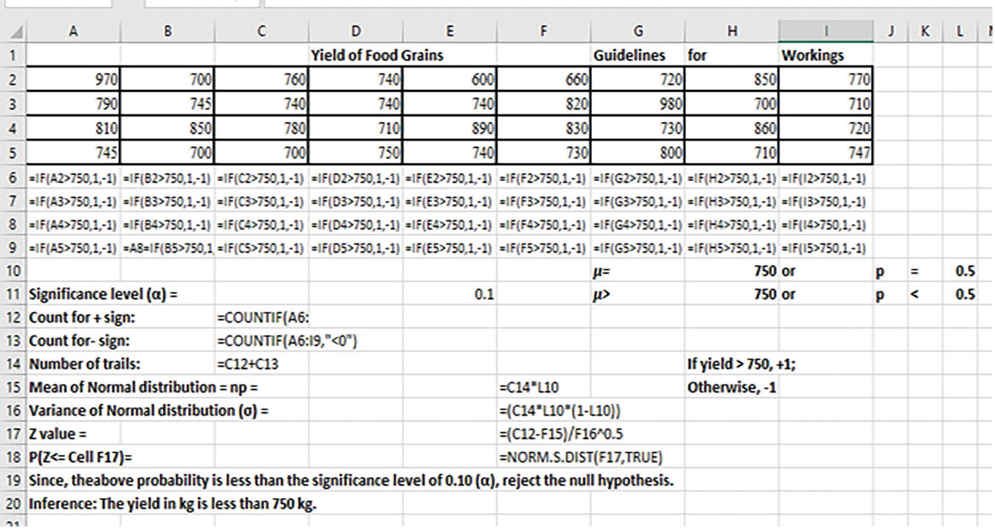
The standard normal statistic Z of this distribution is as follows.

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

The calculations are shown in Figure 12.7 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.7 are shown in Figure 12.8. In Figure 12.7, the value of Z is –1.33333. Hence, the cumulative probability of the normal distribution from its left tail is p value at its left tail.

12.2.4 *Two-Tailed One-Sample Sign Test When Sample Size Is Large Using Excel Sheets and NORM.S.DIST Function*

A random sample of size n with the condition that np as well as $n\ (1 - p)$ is greater than or equal to 5 is selected. A normal approximation to the binomial distribution with $p = 1/2$ is used to test this hypothesis. Here, X is the random variable, which represents the number of plus signs.



	A	B	C	D	E	F	G	H	I	J	K	L
1				Yield of Food Grains			Guidelines	for	Workings			
2	970	700	760	740	600	660	720	850	770			
3	790	745	740	740	740	820	980	700	710			
4	810	850	780	710	890	830	730	860	720			
5	745	700	700	750	740	730	800	710	747			
6	=IF(A2>750,1,-1)	=IF(B2>750,1,-1)	=IF(C2>750,1,-1)	=IF(D2>750,1,-1)	=IF(E2>750,1,-1)	=IF(F2>750,1,-1)	=IF(G2>750,1,-1)	=IF(H2>750,1,-1)	=IF(I2>750,1,-1)			
7	=IF(A3>750,1,-1)	=IF(B3>750,1,-1)	=IF(C3>750,1,-1)	=IF(D3>750,1,-1)	=IF(E3>750,1,-1)	=IF(F3>750,1,-1)	=IF(G3>750,1,-1)	=IF(H3>750,1,-1)	=IF(I3>750,1,-1)			
8	=IF(A4>750,1,-1)	=IF(B4>750,1,-1)	=IF(C4>750,1,-1)	=IF(D4>750,1,-1)	=IF(E4>750,1,-1)	=IF(F4>750,1,-1)	=IF(G4>750,1,-1)	=IF(H4>750,1,-1)	=IF(I4>750,1,-1)			
9	=IF(A5>750,1,-1)	=A8=IF(B5>750,1	=IF(C5>750,1,-1)	=IF(D5>750,1,-1)	=IF(E5>750,1,-1)	=IF(F5>750,1,-1)	=IF(G5>750,1,-1)	=IF(H5>750,1,-1)	=IF(I5>750,1,-1)			
10							μ=	750	or	p	=	0.5
11	Significance level (α) =				0.1		μ>	750	or	p	<	0.5
12	Count for + sign:		=COUNTIF(A6:									
13	Count for- sign:		=COUNTIF(A6:I9,"<0")									
14	Number of trails:		=C12+C13					If yield > 750, +1;				
15	Mean of Normal distribution = np =					=C14*L10		Otherwise, -1				
16	Variance of Normal distribution (σ) =					=(C14*L10*(1-L10))						
17	Z value =					=(C12-F15)/F16^0.5						
18	P(Z<= Cell F17)=					=NORM.S.DIST(F17,TRUE)						
19	Since, theabove probability is less than the significance level of 0.10 (α), reject the null hypothesis.											
20	Inference: The yield in kg is less than 750 kg.											

Figure 12.8 Screenshot of guidelines for the formulas of the working of Example 12.4

The hypotheses of the two-tailed one-sample sign test when the sample size is large are as follows.

Test

H_0: $p = 1/2$
H_1: $p \neq 1/2$

Let

$X \sim B(X, n, p)$

The mean = np
The variance = $np(1 - p)$
If n is more than 30, this can be approximated to a normal distribution $N\left[np, np(1-p)\right]$.

The standard normal statistic Z of this distribution is as follows.

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

Example 12.5

A small-scale industries' annual revenue in lakhs of rupees in a state follows a non-normal distribution. According to a researcher, the small-scale industries' annual revenue is not different from ₹ 150 lakhs. The 32 small-scale industries he chose at random to test his hypothesis are as follows, along with their annual revenues.

170	100	160	150	100	160	100	150
190	100	190	120	140	120	180	120
110	150	180	110	190	130	120	190
120	170	190	180	150	140	130	190

Check the intuition of the researcher using the sign test at a significance level of 0.05.

Solution

The data for Example 12.5 are shown in Table 12.3.

A random sample of 32 with the condition that np as well as $n\ (1 - p)$ is greater than or equal to 5 is selected. A normal approximation to the binomial distribution with $p = 1/2$ is used to test this hypothesis.

Here, X is the random variable, which represents the number of plus signs.

The hypotheses of the one-tailed one-sample sign test when the sample size is large are as follows.

Test: H_0: $p = 1/2$
H_1: $p \neq 1/2$

Let
$X \sim B(X, n, p)$

Since n is more than 30, the data can be approximated to a normal distribution $N\left[np, np(1-p)\right]$.

The standard normal statistic Z of this distribution is as follows.

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

The calculations are shown in Figure 12.9 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.9 are shown in Figure 12.10.

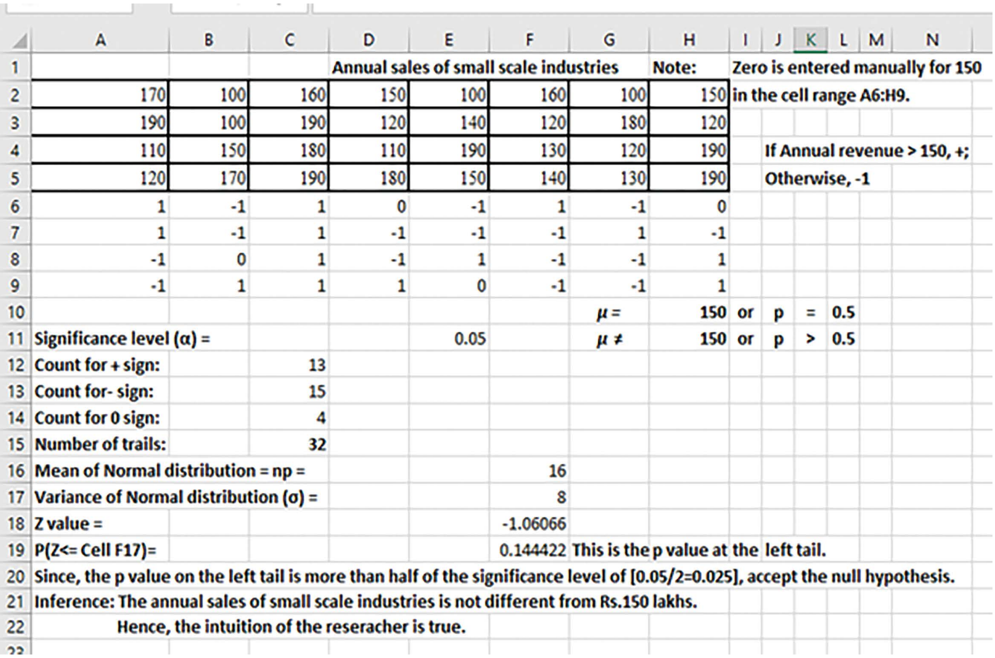
12.3 Test Using Excel Sheets

The *Kolmogorov-Smirnov (K-S)* test is an alternative test to the χ^2 test. The K-S test is a one-tailed test for a small sample, whereas the χ^2 test is also a one-tailed test, but for large sample [3]. The hypotheses for the K-S test are as follows with a significance level of α.

H_0: The given set of data follows an assumed probability distribution.
H_1: The given set of data does not follow an assumed probability distribution.

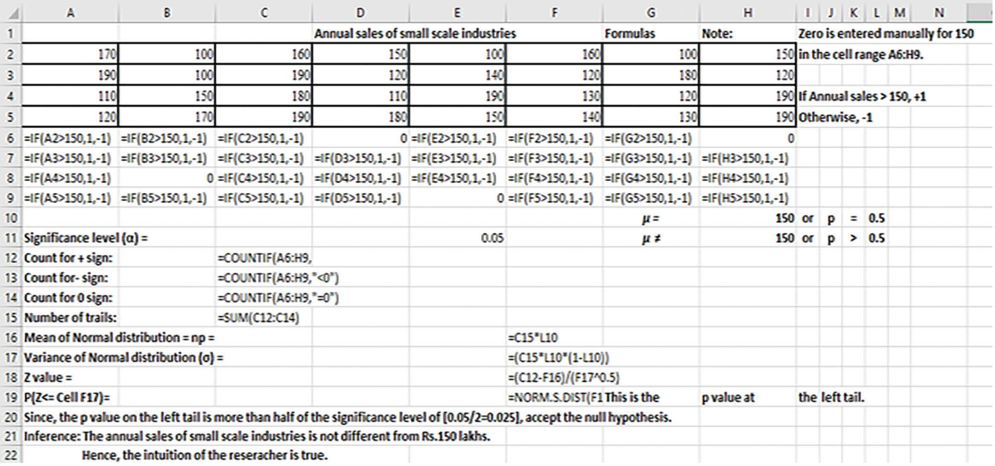
Table 12.3 Data for Example 12.5

170	100	160	150	100	160	100	150
190	100	190	120	140	120	180	120
110	150	180	110	190	130	120	190
120	170	190	180	150	140	130	190



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1				Annual sales of small scale industries				Note:	Zero is entered manually for 150					
2	170	100	160	150	100	160	100	150	in the cell range A6:H9.					
3	190	100	190	120	140	120	180	120						
4	110	150	180	110	190	130	120	190		If Annual revenue > 150, +;				
5	120	170	190	180	150	140	130	190		Otherwise, -1				
6	1	-1	1	0	-1	1	-1	0						
7	1	-1	1	-1	-1	-1	1	-1						
8	-1	0	1	-1	1	-1	-1	1						
9	-1	1	1	1	0	-1	-1	1						
10							μ =	150	or	p	=	0.5		
11	Significance level (α) =				0.05		μ ≠	150	or	p	>	0.5		
12	Count for + sign:		13											
13	Count for- sign:		15											
14	Count for 0 sign:		4											
15	Number of trails:		32											
16	Mean of Normal distribution = np =					16								
17	Variance of Normal distribution (σ) =					8								
18	Z value =					-1.06066								
19	P(Z<= Cell F17)=					0.144422	This is the p value at the left tail.							
20	Since, the p value on the left tail is more than half of the significance level of [0.05/2=0.025], accept the null hypothesis.													
21	Inference: The annual sales of small scale industries is not different from Rs.150 lakhs.													
22	Hence, the intuition of the reseracher is true.													

Figure 12.9 Screenshot of working of Example 12.5



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1				Annual sales of small scale industries			Formulas	Note:	Zero is entered manually for 150					
2	170	100	160	150	100	160	100	150	in the cell range A6:H9.					
3	190	100	190	120	140	120	180	120						
4	110	150	180	110	190	130	120	190	If Annual sales > 150, +1					
5	120	170	190	180	150	140	130	190	Otherwise, -1					
6	=IF(A2>150,1,-1)	=IF(B2>150,1,-1)	=IF(C2>150,1,-1)	0	=IF(E2>150,1,-1)	=IF(F2>150,1,-1)	=IF(G2>150,1,-1)	0						
7	=IF(A3>150,1,-1)	=IF(B3>150,1,-1)	=IF(C3>150,1,-1)	=IF(D3>150,1,-1)	=IF(E3>150,1,-1)	=IF(F3>150,1,-1)	=IF(G3>150,1,-1)	=IF(H3>150,1,-1)						
8	=IF(A4>150,1,-1)	0	=IF(C4>150,1,-1)	=IF(D4>150,1,-1)	=IF(E4>150,1,-1)	=IF(F4>150,1,-1)	=IF(G4>150,1,-1)	=IF(H4>150,1,-1)						
9	=IF(A5>150,1,-1)	=IF(B5>150,1,-1)	=IF(C5>150,1,-1)	=IF(D5>150,1,-1)	0	=IF(F5>150,1,-1)	=IF(G5>150,1,-1)	=IF(H5>150,1,-1)						
10							μ =	150	or	p	=	0.5		
11	Significance level (α) =				0.05		μ ≠	150	or	p	>	0.5		
12	Count for + sign:		=COUNTIF(A6:H9,											
13	Count for- sign:		=COUNTIF(A6:H9,"<0")											
14	Count for 0 sign:		=COUNTIF(A6:H9,"=0")											
15	Number of trails:		=SUM(C12:C14)											
16	Mean of Normal distribution = np =					=C15*L10								
17	Variance of Normal distribution (σ) =					=(C15*L10*(1-L10))								
18	Z value =					=(C12-F16)/(F17^0.5)								
19	P(Z<= Cell F17)=					=NORM.S.DIST(F1	This is the	p value at	the left tail.					
20	Since, the p value on the left tail is more than half of the significance level of [0.05/2=0.025], accept the null hypothesis.													
21	Inference: The annual sales of small scale industries is not different from Rs.150 lakhs.													
22	Hence, the intuition of the reseracher is true.													

Figure 12.10 Screenshot of guidelines for the formulas of the working of Example 12.5

In this test, the cumulative values of the random variable (OF_i) are computed and their absolute differences from respective expected cumulative probabilities (EF_i) calculated using the respective probability distribution.

$$D_i = Absolute\left(OF_i - EF_i\right), i = 1, 2, 3, \ldots, n$$

where
n is the number of observations in the given data set
OF_i is the observed cumulative probability, i = 1, 2, 3, . . ., n
EF_i is the expected cumulative probability, i = 1, 2, 3, . . ., n

The formula for D_{cal} is as follows.

$$D_{cal} = Max\left[Absolute\left(OF_i - EF_i\right)\right], i = 1, 2, 3, \ldots.., n$$

where
n is the number of observations in the given data set
OF_i is the observed cumulative probability, i = 1, 2, 3, . . ., 6
EF_i is the expected cumulative probability, i = 1, 2, 3, . . ., 6
If D_{cal} is more than D from the table given in Annexure 7, then reject the null hypothesis; otherwise, accept the null hypothesis.

When n is large, then formulas are given for different significance levels to compute the D value in Annexure 7.

Example 12.6

The arrival rate of customers (number of customers per hour) at a leading retail shop appears to follow a Poisson distribution. The observed frequencies are given in Table 12.4.

Check whether the given set of data follows a Poisson distribution using the Kolmogorov-Smirnov test at a significance level of 0.10.

Solution

The data for Example 12.6 are shown in Table 12.5.

Sum of the observed frequencies (sample size n) = 25

Significance level (α) = 0.10

Number of values of the random variable X_i = 6

In this test, the cumulative values of the random variable (OF_i) are computed and their absolute differences from respective expected cumulative probabilities (EF_i) calculated using a Poisson probability distribution, which is as follows.

$$P(X) = \frac{\lambda^X e^{-\lambda}}{X!}$$

Table 12.4 Observed Frequencies of Customers

i	*Arrival Rate (X_i)*	*Observed Frequency (O_i)*
1	0	2
2	1	4
3	2	8
4	3	6
5	4	3
6	5	2

Table 12.5 Data for Example 12.6

i	*Arrival Rate* (X_i)	*Observed Frequency* (O_i)
1	0	2
2	1	4
3	2	8
4	3	6
5	4	3
6	5	2

where

λ is the arrival rate of a specific occurrence of an event

X is the random variable representing the occurrence of specific event

$$D_i = Absolute\left(OF_i - EF_i\right), i = 1, 2, 3, \ldots, 6$$

where

OF_i is the observed cumulative probability, $i = 1, 2, 3, \ldots, 6$

EF_i is the expected cumulative probability, $i = 1, 2, 3, \ldots, 6$

The formula for D_{cal} is as follows.

$$D_{cal} = Max\left[Absolute\left(OF_i - EF_i\right)\right], i = 1, 2, 3, \ldots.., 6$$

where

n is the number of observations in the given data set

OF_i is the observed cumulative probability, $i = 1, 2, 3, \ldots, 6$

EF_i is the expected cumulative probability, $i = 1, 2, 3, \ldots, 6$

From Annexure 7, the value of D is 0.24 when the sample size is 25 (total frequency), and the significance level is 0.1.

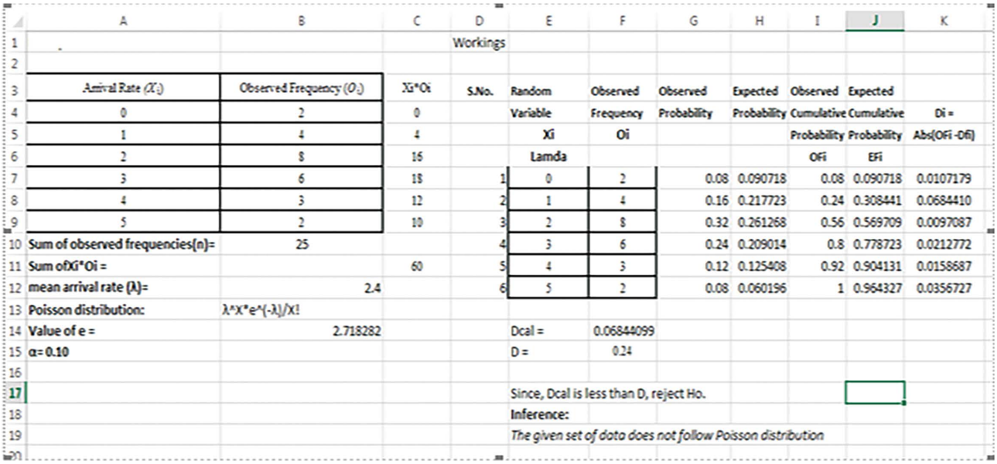
All the calculations are shown in Figure 12.11 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.11 are shown in Figure 12.12.

Example 12.7

The weight of welding rods produced by a factory appears to follow a normal distribution. Table 12.6 displays the weights of the welding rods of a sample's measured frequency. Using the Kolmogorov-Smirnov test with a significance threshold of 0.10, determine whether the provided data follow a normal distribution.

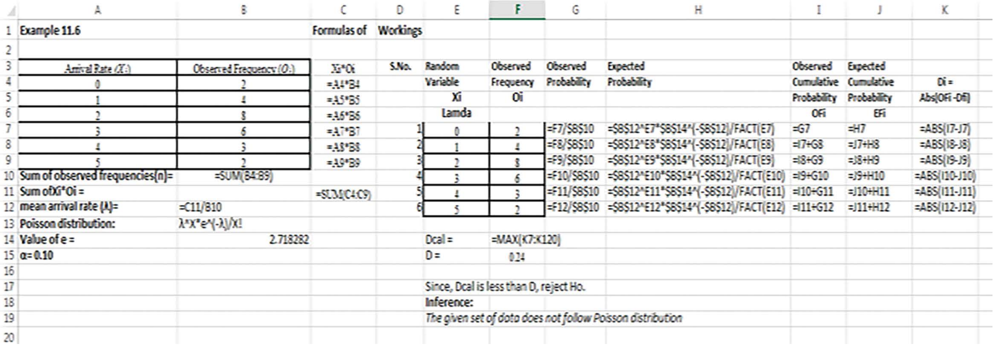
Solution

The data for Example 12.7 are shown in Table 12.7 with mid-points of the class intervals.



	A	B	C	D	E	F	G	H	I	J	K
1				Workings							
2											
3	Arrival Rate (X_i)	Observed Frequency (O_i)	Xi*Oi	S.No.	Random	Observed	Observed	Expected	Observed	Expected	
4	0	2	0		Variable	Frequency	Probability	Probability	Cumulative	Cumulative	Di =
5	1	4	4		Xi	Oi			Probability	Probability	Abs(OFi -Dfi)
6	2	8	16		Lamda				OFi	EFi	
7	3	6	18	1	0	2	0.08	0.090718	0.08	0.090718	0.0107179
8	4	3	12	2	1	4	0.16	0.217723	0.24	0.308441	0.0684410
9	5	2	10	3	2	8	0.32	0.261268	0.56	0.569709	0.0097087
10	Sum of observed frequencies(n)=	25		4	3	6	0.24	0.209014	0.8	0.778723	0.0212772
11	Sum ofXi*Oi =		60	5	4	3	0.12	0.125408	0.92	0.904131	0.0158687
12	mean arrival rate (λ)=	2.4		6	5	2	0.08	0.060196	1	0.964327	0.0356727
13	Poisson distribution:	λ^X*e^(-λ)/X!									
14	Value of e =	2.718282			Dcal =	0.06844099					
15	α= 0.10				D =	0.24					
16											
17					Since, Dcal is less than D, reject Ho.						
18					Inference:						
19					*The given set of data does not follow Poisson distribution*						

Figure 12.11 Screenshot of the working of Example 12.6



	A	B	C	D	E	F	G	H	I	J	K
1	Example 11.6		Formulas of	Workings							
2											
3	Arrival Rate (X_i)	Observed Frequency (O_i)	Xi*Oi	S.No.	Random	Observed	Observed	Expected	Observed	Expected	
4	0	2	=A4*B4		Variable	Frequency	Probability	Probability	Cumulative	Cumulative	Di =
5	1	4	=A5*B5		Xi	Oi			Probability	Probability	Abs(OFi -Dfi)
6	2	8	=A6*B6		Lamda				OFi	EFi	
7	3	6	=A7*B7	1	0	2	=F7/$B$10	=$B$12^E7*$B$14^(-$B$12)/FACT(E7)	=G7	=H7	=ABS(I7-J7)
8	4	3	=A8*B8	2	1	4	=F8/$B$10	=$B$12^E8*$B$14^(-$B$12)/FACT(E8)	=I7+G8	=J7+H8	=ABS(I8-J8)
9	5	2	=A9*B9	3	2	8	=F9/$B$10	=$B$12^E9*$B$14^(-$B$12)/FACT(E9)	=I8+G9	=J8+H9	=ABS(I9-J9)
10	Sum of observed frequencies(n)=	=SUM(B4:B9)		4	3	6	=F10/$B$10	=$B$12^E10*$B$14^(-$B$12)/FACT(E10)	=I9+G10	=J9+H10	=ABS(I10-J10)
11	Sum ofXi*Oi =		=SUM(C4:C9)	5	4	3	=F11/$B$10	=$B$12^E11*$B$14^(-$B$12)/FACT(E11)	=I10+G11	=J10+H11	=ABS(I11-J11)
12	mean arrival rate (λ)=	=C11/B10		6	5	2	=F12/$B$10	=$B$12^E12*$B$14^(-$B$12)/FACT(E12)	=I11+G12	=J11+H12	=ABS(I12-J12)
13	Poisson distribution:	λ^X*e^(-λ)/X!									
14	Value of e =	2.718282			Dcal =	=MAX(K7:K120)					
15	α= 0.10				D =	0.24					
16											
17					Since, Dcal is less than D, reject Ho.						
18					Inference:						
19					*The given set of data does not follow Poisson distribution*						
20											

Figure 12.12 Screenshot of guidelines for the formulas of the working of Example 12.6

Table 12.6 Frequencies of Weights of Welding Rods

Weight in gm		*No. of Welding Rods*
Lower Limit	*Upper Limit*	
100	102	30
102	104	35
104	106	50
106	108	40
108	110	25

The hypotheses of this example are as follows.

H_0: The given set of data follows a normal distribution.
H_1: The given set of data does not follow a normal distribution

Table 12.7 Data for Example 12.7

Class Interval Number i	*Weight in gm*		*Class interval Mid-point (X_i)*	*No. of Welding Rods (O_i)*
	Lower Limit	*Upper Limit*		
1	100	102	101	30
2	102	104	103	35
3	104	106	105	50
4	106	108	107	40
5	108	110	109	25

The formulas for the mean and variance of the normal distribution based on the given frequencies are given as follows.

$$Mean(\mu) = \frac{\sum_{i=1}^{5}(X_i O_i)}{\sum_{i=1}^{5} X_i}$$

where
X_i is the mid-point of the i^{th} class interval, $i = 1, 2, 3, 4, 5$
O_i is the observed frequency of the class interval i, $i = 1, 2, 3, 4, 5$

$$Variance\left(\sigma^2\right) = \frac{\sum_{i=1}^{5} O_i (X_i - \mu)^2}{\sum_{i=1}^{5} O_i}$$

The formula for the standard normal statistics is as follows.

$$Z = \frac{(X - \mu)}{\sigma}$$

where
μ is the mean of the normal distribution
σ is the standard deviation of the normal distribution
X is the normal random variable

The Excel formula to compute the cumulative probability of standard normal distribution is as follows.

= NORM.S.DIST(Z, TRUE)

$$D_i = Absolute(OF_i - EF_i),\ i = 1, 2, 3, \ldots, 5$$

where
OF_i is the observed cumulative probability, $i = 1, 2, 3, 4, 5$
EF_i is the expected cumulative probability using a normal distribution, $i = 1, 2, 3, 4, 5$

The formula for D_{cal} is as follows.

$$D_{cal} = Max\left[Absolute\left(OF_i - EF_i\right)\right],\ i = 1, 2, 3, \ldots.., 5$$

where
n is the number of observations in the given data set
OF_i is the observed cumulative probability, $i = 1, 2, 3, 4, 5$
EF_i is the expected cumulative probability, $i = 1, 2, 3, 4, 5$

From Annexure 7, the value of D is given by the following formula, when the sample size is 180 (n = total frequency) and the significance level is 0.1.

$$D = \frac{1.22}{\sqrt{n}}$$

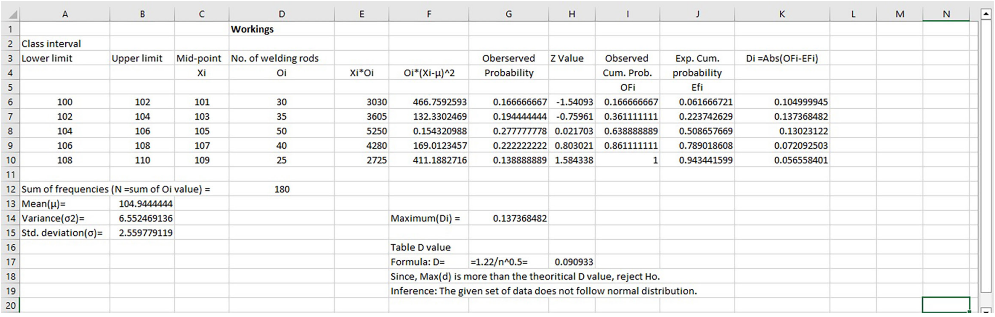
All the calculations are shown in Figure 12.13 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.13 are shown in Figure 12.14.

12.4 Run Test for Randomness

The stream of data that is collected in a system may have certain patterns called runs. Consider the example of customers arriving at the ticket booking counter of a railway station. The customers will be either male (M) or female (F). A sample arrival sequence of the customers in a time interval in terms of their sex is as follows.

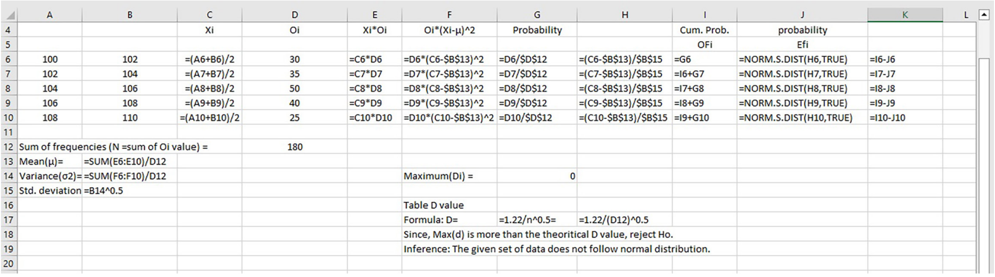
Sequence of Customers (M/F): $\underbrace{MM}_{1}\ \underbrace{FFFF}_{2}\ \underbrace{M}_{3}\ \underbrace{FF}_{4}\ \underbrace{MMMMMM}_{5}\ \underbrace{FFF}_{6}\ \underbrace{MMM}_{7}$

The repetition of the same type of customer until another type of customer arrives at the booking counter is a reality. The collection of the customer code for such a customer type put together is called a run. The stream may have several runs, and the runs will be in random order.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1				Workings										
2	Class interval													
3	Lower limit	Upper limit	Mid-point	No. of welding rods			Oberserved	Z Value	Observed	Exp. Cum.	Di =Abs(OFi-EFi)			
4			Xi	Oi	Xi*Oi	Oi*(Xi-µ)^2	Probability		Cum. Prob.	probability				
5									OFi	Efi				
6	100	102	101	30	3030	466.7592593	0.166666667	-1.54093	0.166666667	0.061666721	0.104999945			
7	102	104	103	35	3605	132.3302469	0.194444444	-0.75961	0.361111111	0.223742629	0.137368482			
8	104	106	105	50	5250	0.154320988	0.277777778	0.021703	0.638888889	0.508657669	0.13023122			
9	106	108	107	40	4280	169.0123457	0.222222222	0.803021	0.861111111	0.789018608	0.072092503			
10	108	110	109	25	2725	411.1882716	0.138888889	1.584338	1	0.943441599	0.056558401			
11														
12	Sum of frequencies (N =sum of Oi value) =			180										
13	Mean(µ)=	104.9444444												
14	Variance(σ2)=	6.552469136				Maximum(Di) =	0.137368482							
15	Std. deviation(σ)=	2.559779119												
16						Table D value								
17						Formula: D=	=1.22/n^0.5=	0.090933						
18						Since, Max(d) is more than the theoritical D value, reject Ho.								
19						Inference: The given set of data does not follow normal distribution.								
20														

Figure 12.13 Screenshot of the working of Example 12.7



	A	B	C	D	E	F	G	H	I	J	K	L
4			Xi	Oi	Xi*Oi	Oi*(Xi-μ)^2	Probability		Cum. Prob.	probability		
5									OFi	Efi		
6	100	102	=(A6+B6)/2	30	=C6*D6	=D6*(C6-$B$13)^2	=D6/$D$12	=(C6-$B$13)/$B$15	=G6	=NORM.S.DIST(H6,TRUE)	=I6-J6	
7	102	104	=(A7+B7)/2	35	=C7*D7	=D7*(C7-$B$13)^2	=D7/$D$12	=(C7-$B$13)/$B$15	=I6+G7	=NORM.S.DIST(H7,TRUE)	=I7-J7	
8	104	106	=(A8+B8)/2	50	=C8*D8	=D8*(C8-$B$13)^2	=D8/$D$12	=(C8-$B$13)/$B$15	=I7+G8	=NORM.S.DIST(H8,TRUE)	=I8-J8	
9	106	108	=(A9+B9)/2	40	=C9*D9	=D9*(C9-$B$13)^2	=D9/$D$12	=(C9-$B$13)/$B$15	=I8+G9	=NORM.S.DIST(H9,TRUE)	=I9-J9	
10	108	110	=(A10+B10)/2	25	=C10*D10	=D10*(C10-$B$13)^2	=D10/$D$12	=(C10-$B$13)/$B$15	=I9+G10	=NORM.S.DIST(H10,TRUE)	=I10-J10	
11												
12	Sum of frequencies (N =sum of Oi value) =			180								
13	Mean(μ)=	=SUM(E6:E10)/D12										
14	Variance(σ2)=	=SUM(F6:F10)/D12				Maximum(Di) =	0					
15	Std. deviation	=B14^0.5										
16						Table D value						
17						Formula: D=	=1.22/n^0.5=	=1.22/(D12)^0.5				
18						Since, Max(d) is more than the theoritical D value, reject Ho.						
19						Inference: The given set of data does not follow normal distribution.						
20												

Figure 12.14 Screenshot of guidelines for the formulas of the working of Example 12.7

Let

n_1 be the number of occurrences of a customer code of male

n_2 be the number of occurrences of a customer code of female

r_{cal} be the number of runs of the whole stream

If n_1 as well as n_2 is less than or equal to 20, then the sample is a small sample. If n_1 or n_2 or both is/are greater than 20, then the sample is a large sample.

The hypotheses for this situation are stated as follows.

H_0: The occurrence of the runs in the given stream of symbols is random.
H_1: The occurrence of the runs in the given stream of symbols is not random.

12.4.1 Run Test for Small Samples Using Excel Sheets and COUNTIF Functions

If the value of n_1 as well as that of n_2 is less than 20, then the sample is regarded as a small sample.

Let

n_1 be the number of occurrences of symbol 1

n_2 be the number of occurrences of symbol 2

r_{cal} be the observed number of runs in the whole stream

From Annexure 8, smaller and larger critical values for n_1 and n_2 with the given significance level α can be obtained.

If r_{cal} is between the smaller critical value and the larger critical value, inclusive, then accept H_0; otherwise, reject H_0.

Example 12.8

The items inspected in the final inspection station of a production line are graded into good (G) and defective (D). The stream of good and defectives items of the production units of the production line is as follows.

$\underline{G\ G}\quad \underline{D\ D}\quad \underline{G\ G\ G}\quad \underline{D}\quad \underline{G\ G\ G\ G\ G}\quad \underline{D}\quad \underline{G\ G\ G\ G}\quad \underline{DDD}$

Check whether the substrings as a result of grading the production units into good/defective are random at a significance level of 0.01.

Solution

The sequence of substrings as a result of grading the production units into good/defective is shown as follows.

$$\underset{1}{\underline{G\ G}}\ \ \underset{2}{\underline{D\ D}}\ \ \underset{3}{\underline{G\ G\ G}}\ \ \underset{4}{\underline{D}}\ \ \underset{5}{\underline{G\ G\ G\ G\ G}}\ \ \underset{6}{\underline{D}}\ \ \underset{7}{\underline{G\ G\ G\ G}}\ \ \underset{8}{\underline{DDD}}$$

The number of G symbols, $n_1 = 14$
Number of D symbols, $n_2 = 7$

From the given stream, the number of runs $(r_{cal}) = 8$

H_0: The occurrence of the runs in the given stream of symbols is random.
H_1: The occurrence of the runs in the given stream of symbols is not random.

From Annexure 8, the smaller and larger critical values when $n_1 = 14$ and $n_2 = 7$ at a significance level of 0.01 are 5 and 15, respectively.

Since r_{cal}, which is 8, falls between these limits (5 to 15), accept H_0.

Inference: The occurrence of the substrings in terms of G/D as a result of grading the production units in the production line is random.

Note: In this test for a small sample, there is not much scope to use an Excel sheet. However, to count the number of Gs and Ds, one can use the COUNTIF command.

The total of each of these can be obtained using the Excel command: =COUNTIF(Range, criteria).

Assume that the data are in the range A1:A21.

The command to count Gs: $= COUNTIF(A1 : A21, "=G")$
The command to count Ds: $= COUNTIF(A1 : A21, "=D")$

12.4.2 Run Test for Large Samples Using Excel Sheets, COUNTIF, and NORM.S.DIST Functions

If the value of n_1 or n_2 or both is/are more than 20, then the sample is regarded as a large sample.

Let
n_1 be the number of occurrences of symbol 1
n_2 be the number of occurrences of symbol 2
r be the number of runs in the whole stream
α be the significance level

In this test, r follows a normal distribution with a mean of μ and a variance of σ^2. The formulas for these are as follows.

$$\mu = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$\sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

The Z statistic is given by the following formula.

$$Z = \frac{(r - \mu)}{\sigma}$$

The hypotheses of this test are as follows.

H_0: The occurrence of the runs of the given stream of symbols is random.
H_1: The occurrence of the runs of the given stream of symbols is not random.

Inference:

1. If Z is positive, find the p at the right tail of the standard normal distribution. If it is more than half of the significance level ($\alpha/2$), accept H_0; otherwise, reject H_0.
2. If Z is negative, find the p at the left tail of the standard normal distribution. If it is more than half of the significance level ($\alpha/2$), accept H_0; otherwise, reject H_0.

Example 12.9

The monthly income of respondents in a survey is classified into high income (H) and low income (L). The sequence of codes for 21 respondents is as follows.

HHH LL HHH L HH LL HH LLL HHH

Check whether the occurrence of H/L is random at a significance level of 0.10.

Solution

The given stream of symbols, that is, high income (H) and low income (L), is as follows.

HHH LL HHH L HH LL HH LLL HHH

The number of runs of the stream (r) = 9
Let
n_1 be the number of occurrences of symbol H, which is 13
n_2 be the number of occurrences of symbol L, which is 8
α be the significance level, which is 0.10
In this test, r follows a normal distribution with a mean of μ and a variance of σ^2
The formulas for these are as follows.

$$\mu = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$\sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

The Z statistic is given by the following formula.

$$Z = \frac{(r - \mu)}{\sigma}$$

The hypotheses of this test are as follows.

H_0: The occurrence of the runs of the given stream of symbols (H/L) is random.
H_1: The occurrence of the runs of the given stream of symbols (H/L) is not random.

All the calculations are shown in Figure 12.15 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.15 are shown in Figure 12.16.

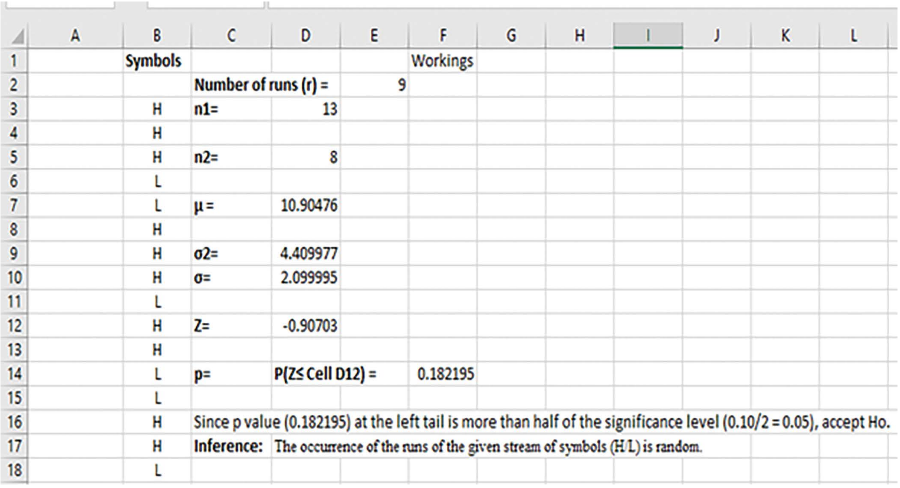
12.5 Two-Sample Tests

There are several instances in real life where a decision will be made based on the findings of two samples. Think about the possibility of hanging banner ads in stores that sell Ganesh Health Care Company products. The company's marketing manager believes that hanging banner advertisements in stores actually boosts sales. The finance manager, who approves the funding for this activity, fears that the sales may not increase after installing advertisement banners in the stores.

Two samples of the same size, each with a non-normal distribution, are used in this study to test an underlying hypothesis. The sample size for the two-sample test could be small or large.

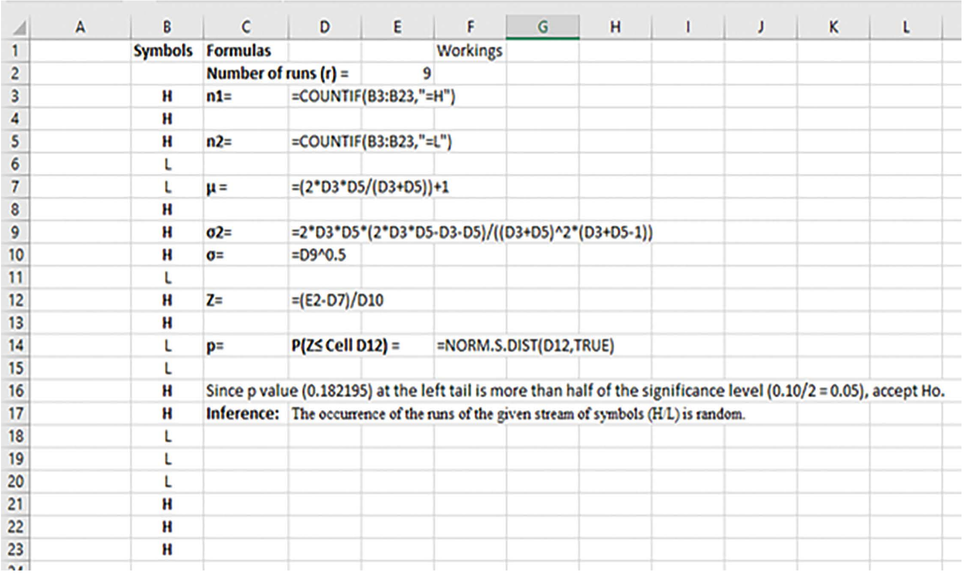
The different nonparametric tests that are performed for two-sample sign tests are as follows.

- Two-sample sign test
- Two-sample median test
- Mann-Whitney U test (rank-sum test)
- Wilcoxon matched-pairs test (rank-sum test)



	A	B	C	D	E	F	G	H	I	J	K	L
1		Symbols				Workings						
2			Number of runs (r) =		9							
3		H	n1=	13								
4		H										
5		H	n2=	8								
6		L										
7		L	μ =	10.90476								
8		H										
9		H	σ2=	4.409977								
10		H	σ=	2.099995								
11		L										
12		H	Z=	-0.90703								
13		H										
14		L	p=	P(Z≤ Cell D12) =		0.182195						
15		L										
16		H	Since p value (0.182195) at the left tail is more than half of the significance level (0.10/2 = 0.05), accept Ho.									
17		H	Inference:	The occurrence of the runs of the given stream of symbols (H/L) is random.								
18		L										

Figure 12.15 Screenshot of workings of Example 12.9



	A	B	C	D	E	F
1		Symbols	Formulas			Workings
2			Number of runs (r) =		9	
3		H	n1=	=COUNTIF(B3:B23,"=H")		
4		H				
5		H	n2=	=COUNTIF(B3:B23,"=L")		
6		L				
7		L	µ =	=(2*D3*D5/(D3+D5))+1		
8		H				
9		H	σ2=	=2*D3*D5*(2*D3*D5-D3-D5)/((D3+D5)^2*(D3+D5-1))		
10		H	σ=	=D9^0.5		
11		L				
12		H	Z=	=(E2-D7)/D10		
13		H				
14		L	p=	P(Z≤ Cell D12) =		=NORM.S.DIST(D12,TRUE)
15		L				
16		H	Since p value (0.182195) at the left tail is more than half of the significance level (0.10/2 = 0.05), accept Ho.			
17		H	Inference:	The occurrence of the runs of the given stream of symbols (H/L) is random.		
18		L				
19		L				
20		L				
21		H				
22		H				
23		H				

Figure 12.16 Screenshot of guidelines for the formulas of the working of Example 12.9

12.5.1 Two-Sample Sign Test

The two-sample sign test deals with two samples with equal sample size, which follow a non-normal distribution. This is further classified into the following.

- Two-sample sign test for small samples
- Two-sample sign test for large samples

In each of the tests, the observations before and after introducing a change in a system will be collected.

Let

n be the size of the sample collected before introducing a change in the system as well as that of the sample collected after introducing that change in that system

X_i be the i^{th} observation before introducing the change in the system, $i = 1, 2, 3, \ldots, n$

Y_i be the i^{th} observation after introducing the change in the system, $i = 1, 2, 3, \ldots, n$

If $X_i - Y_i$ is greater than 0, then it is treated as + (plus sign); if it is less than 0, then it is treated as – sign (minus sign); if they are equal, then it is treated as 0. In the sign test, the 0s should be omitted.

The number of plus signs (+) is treated as a random variable X, which is non-normal.

The two-tailed sign tests are classified into the following.

1. One-tailed two-sample sign tests for small samples
2. Two-tailed two-sample sign tests for small samples
3. One-tailed two-sample sign tests for large samples
4. Two-tailed two-sample sign tests for large samples

12.5.1.1 One-Tailed Two-Sample Sign Test for Small Samples Using Excel Sheets and BINOM.DIST Function

As stated earlier, two samples, which are before and after introducing a change in the system of interest, are taken. These two samples are known to follow a non-normal distribution. For a sample size less than 20, the number of plus signs (positive difference (+) between the i^{th} observation, which has been taken before, and the i^{th} observation, which has been taken after introducing a change in the system) follows a binomial distribution.

The probability that a sample value ($X_i - Y_i$) is more than the mean value (p) is 1/2, and the probability that a sample value ($X_i - Y_i$) is less than the mean value (p) is 1/2.

In this situation, the modified sample value K_i is defined as follows.

$K_i = +$, if $X_i > Y_i$
$= -$, if $X_i < Y_i$
$= 0$, if $X_i = Y_i$

where
X_i is the i^{th} observation of the sample before the change in the system
Y_i is the i^{th} observation of the sample after the change in the system

Then the number of plus signs is taken as the value of the random variable Q of the binomial distribution, with $p = 1/2$ and the number of trials n to compute the probability that Q is more than the number of plus signs to check the following hypotheses. The mean and variance of the samples, that is, np and $np(1 - p)$, should be less than 5.

The hypotheses of the one-tailed two-sample sign test when the sample sizes are small are as follows.

Test 1:

H_0: $\mu_X = \mu_Y$ or $p = 1/2$
$H_{1:}$ $\mu_X > \mu_Y$ or p 1/2
Test 2:
H_0: $\mu_X = \mu_Y$ or $p = 1/2$
$H_{1:}$ $\mu_X < \mu_Y$ or $p < 1/2$

Let

Q be the number of plus signs representing the random variable

$Q \sim B(\mathrm{Q}, n, p)$

If $P(\mathrm{Q} \geq$ Number of plus signs) is less than the significance level at the right tail, then reject the null hypothesis; otherwise, accept the null hypothesis.

If $P(\mathrm{Q} \leq$ Number of plus signs) is less than the significance level at the left tail, then reject the null hypothesis; otherwise, accept the null hypothesis.

Example 12.10

The shop floor manager believed that providing the operators on a production line with appropriate training would result in a decrease in the number of defective assemblies produced each day. So, over the course of eight days, he gathered data on the number of defective assemblies made each day. Additionally, an arrangement was established for a training programme, after which comparable data were gathered for 8 days. These results are displayed in Table 12.8.

Check the null hypothesis that the training program has decreased the number of defective assemblies produced per day (H_0; $\mu_X = \mu_Y$) against the alternate hypothesis H_1: $\mu_X > \mu_Y$ using a sign test at a significance level of 0.05.

Solution

The data for Example 12.10 are shown in Table 12.9.

Sample size (n) = 8

Significance level ($\alpha = 0.05$)

Let

X_i be the number of defective assemblies produced on the i^{th} day before the training program

Y_i be the number of defective assemblies produced on the i^{th} day after the training program

If $X_i - Y_i > 0$, then the sign is +, which is indicated by +1

If $X_i - Y_i < 0$, then the sign is –, which is indicated by –1

If $X_i - Y_i = 0$, then the sign is 0

Let Q be a random variable representing number of plus signs when after observation is subtracted from before observation ($X_i - Y_i$), which follows non-normal distribution, the binomial distribution.

$$H_o : \mu_X = \mu_Y \; orp = 1/2$$

$$H_{1:}\, \mu_X > \mu_Y \; orp > 1/2$$

$$np = 8 \times 0.5 = 4 < 5$$

$$np(1-p) = 8 \times 0.5 \times 0.5 = 2 < 5$$

Hence, the two samples are small samples.

Table 12.8 Data for Example 12.10

	Defective Assemblies per Day							
Day	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Before training	112	110	110	107	104	107	122	108
After training	110	113	100	106	102	107	120	105

Table 12.9 Before and After Data for Example 12.10

	Defective Assemblies per Day							
Day	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Before training	112	110	110	107	104	107	122	108
After training	110	113	100	106	102	107	120	105

Q *follows* $B(Q, n, p)$

$$P\left(Q \geq Nunber\,of\,plus\,signs, n = 8, p = 0.5\right) = 1 - P\left(Q \leq Nunber\,of\,plus\,signs, n = 8, p = 0.5\right)$$

$$= p\,value\,at\,the\,right\,tail$$

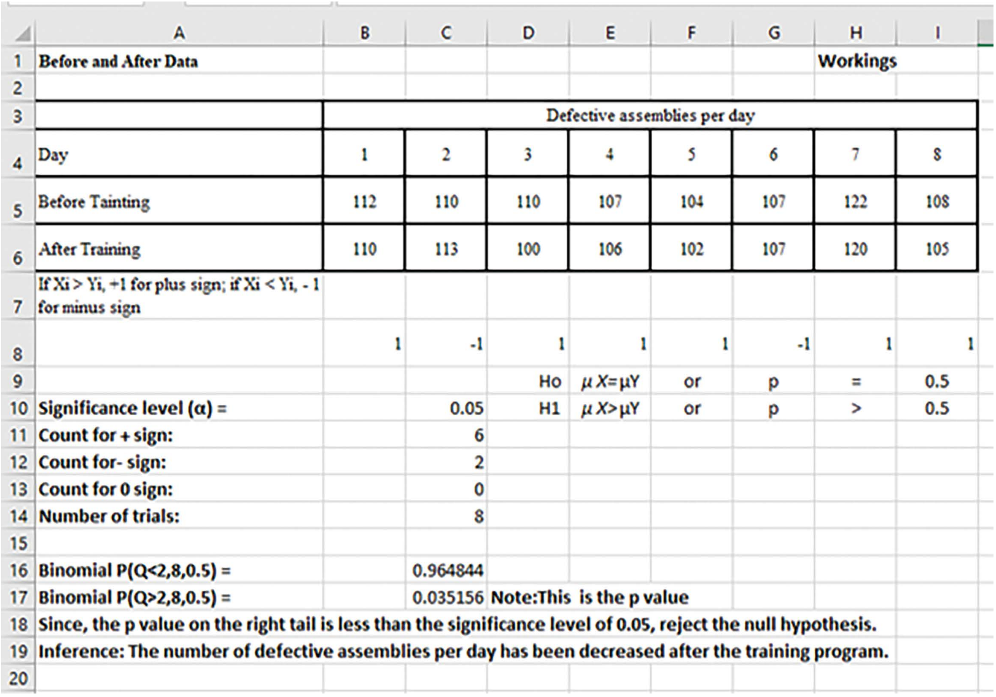
All the calculations are shown in Figure 12.17 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.17 are shown in Figure 12.18.

12.5.1.2 *Two-Tailed Two-Sample Sign Test for Small Samples Using Excel Sheets and BINOM.DIST Function*

Two samples, which are before and after introducing a change in the system of interest, are taken. These two samples are known to follow a non-normal distribution. For a sample size less than 20, the number of plus signs (number of positive difference (+) between the i^{th} observation taken before a change and the i^{th} observation taken after introducing a change in the system) follows a binomial distribution.

In this situation, the modified sample value K_i is defined as follows.

K_i = +, if $X_i > Y_i$
= −, if $X_i < Y_i$
= 0, if $X_i = Y_i$



	A	B	C	D	E	F	G	H	I
1	Before and After Data							Workings	
2									
3		Defective assemblies per day							
4	Day	1	2	3	4	5	6	7	8
5	Before Tainting	112	110	110	107	104	107	122	108
6	After Training	110	113	100	106	102	107	120	105
7	If Xi > Yi, +1 for plus sign; if Xi < Yi, - 1 for minus sign								
8		1	-1	1	1	1	-1	1	1
9				Ho	µ X=µY	or	p	=	0.5
10	Significance level (α) =		0.05	H1	µ X>µY	or	p	>	0.5
11	Count for + sign:		6						
12	Count for- sign:		2						
13	Count for 0 sign:		0						
14	Number of trials:		8						
15									
16	Binomial P(Q<2,8,0.5) =		0.964844						
17	Binomial P(Q>2,8,0.5) =		0.035156	Note:This is the p value					
18	Since, the p value on the right tail is less than the significance level of 0.05, reject the null hypothesis.								
19	Inference: The number of defective assemblies per day has been decreased after the training program.								
20									

Figure 12.17 Screenshot of the working of Example 12.10

where
X_i is the i^{th} observation of the sample before the change in the system
Y_i is the i^{th} observation of the sample after the change in the system

Then the number of plus signs is taken as the value of the random variable Q of the binomial distribution, with $p = 1/2$ and the number of trials n to compute the probability that Q is more than the number of plus signs to check the following hypotheses. The mean and variance of the samples, that is, np and $np(1 - p)$ should be less than 5.

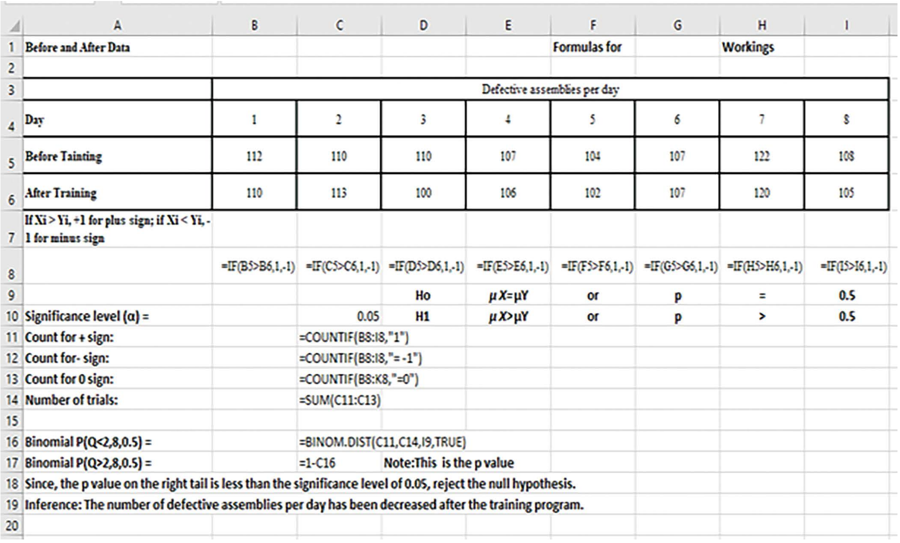
The hypotheses of the two-tailed two-sample sign test when the sample size is small are as follows.

H_0: $\mu_X = \mu_Y$ or $p = 1/2$
$H_{1:}$ $\mu_{X \neq} \mu_Y$ or $p \neq 1/2$

Let
Q be the number of plus signs representing the random variable
$Q \sim B(Q, n, p)$
If $P(Q \geq$ Number of plus signs) is less than the significance level at the right tail, then reject the null hypothesis; otherwise, accept the null hypothesis.
If $P(Q \leq$ Number of plus signs) is less than the significance level at the left tail, then reject the null hypothesis; otherwise, accept the null hypothesis.

Example 12.11

A new device has been developed in the medical field to regulate people's bio-parameters. According to the purchasing manager of a prestigious hospital, the new device serves the same purpose as an existing one. So, using both the old and the new gadgets, he collected



	A	B	C	D	E	F	G	H	I
1	Before and After Data					Formulas for		Workings	
2									
3		Defective assemblies per day							
4	Day	1	2	3	4	5	6	7	8
5	Before Tainting	112	110	110	107	104	107	122	108
6	After Training	110	113	100	106	102	107	120	105
7	If Xi > Yi, +1 for plus sign; if Xi < Yi, -1 for minus sign								
8		=IF(B5>B6,1,-1)	=IF(C5>C6,1,-1)	=IF(D5>D6,1,-1)	=IF(E5>E6,1,-1)	=IF(F5>F6,1,-1)	=IF(G5>G6,1,-1)	=IF(H5>H6,1,-1)	=IF(I5>I6,1,-1)
9				Ho	μX=μY	or	p	=	0.5
10	Significance level (α) =		0.05	H1	μX>μY	or	p	>	0.5
11	Count for + sign:		=COUNTIF(B8:I8,"1")						
12	Count for- sign:		=COUNTIF(B8:I8,"=-1")						
13	Count for 0 sign:		=COUNTIF(B8:K8,"=0")						
14	Number of trials:		=SUM(C11:C13)						
15									
16	Binomial P(Q<2,8,0.5) =		=BINOM.DIST(C11,C14,I9,TRUE)						
17	Binomial P(Q>2,8,0.5) =		=1-C16	Note:This is the p value					
18	Since, the p value on the right tail is less than the significance level of 0.05, reject the null hypothesis.								
19	Inference: The number of defective assemblies per day has been decreased after the training program.								
20									

Figure 12.18 Screenshot of guidelines for the formulas of the working of Example 12.10

data on a composite bio-index from 10 different patients, and the results are shown in Table 12.10.

Check the hypothesis that the composite bio-index obtained from the new device does not differ from that of the existing device at a significance level of 0.05.

Solution

The data for Example 12.11 are shown in Table 12.11.

Sample size (n) = 9

Significance level ($\alpha = 0.05$)

Let Q be a random variable representing the number of plus signs [when after observation is subtracted from before observation ($X_i - Y_i$)], which follows non-normal distribution, the binomial distribution.

The resulting hypotheses for two-tailed test are stated as follows.

H_0: $\mu_X = \mu_Y$ or $p = 1/2$
H_1: $\mu_X \neq \mu_Y$ or $p \neq 1/2$
$np = 9 \times 0.5 = 4.5 < 5$
$np(1 - p) = 9 \times 0.5 \times 0.5 = 2.25 < 5$

Hence, the two samples are small samples.

Let

X_i be the composite bio-index of the i^{th} patient using the existing device
Y_i be the composite bio-index of the i^{th} patient using the new device
If $X_i - Y_i > 0$, then the sign is +, which is indicated by +1
If $X_i - Y_i < 0$, then the sign is –, which is indicated by –1
If $X_i - Y_i = 0$, then the sign is 0

The number of plus signs is assumed to follow a binomial distribution.

Q follows B(Q, n, p)

$$P\left(Q \geq Number\,of\,plus\,signs, n = 9, p = 1/2\right) = 1 - P(Q \leq Number\,of\,plus\,signs,\; n = 9, p = 1/2)$$

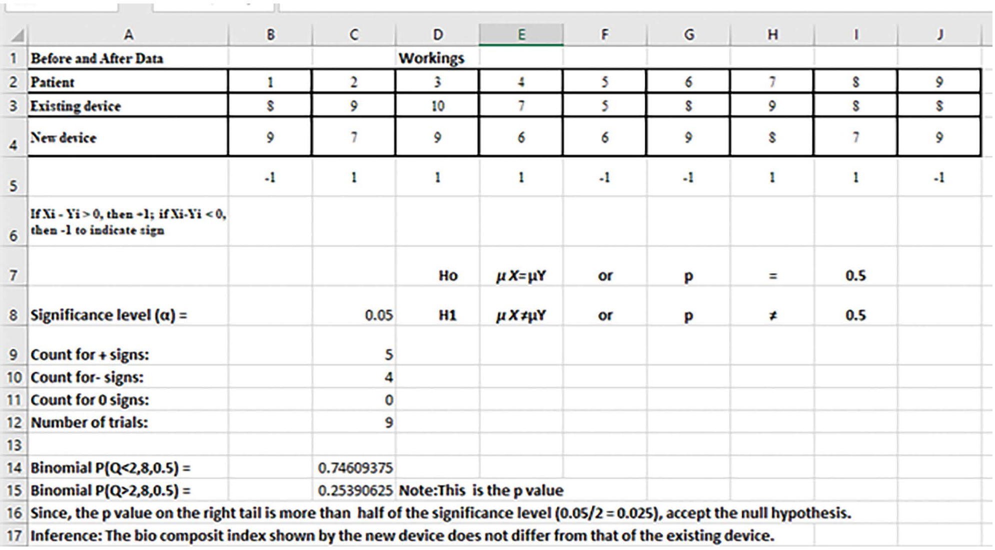
$$= p\,value\,at\,the\,right\,tail$$

Table 12.10 Data for Example 12.11

Patient	1	2	3	4	5	6	7	8	9
Existing device	8	9	10	7	5	8	9	8	8
New device	9	7	9	6	6	9	8	7	9

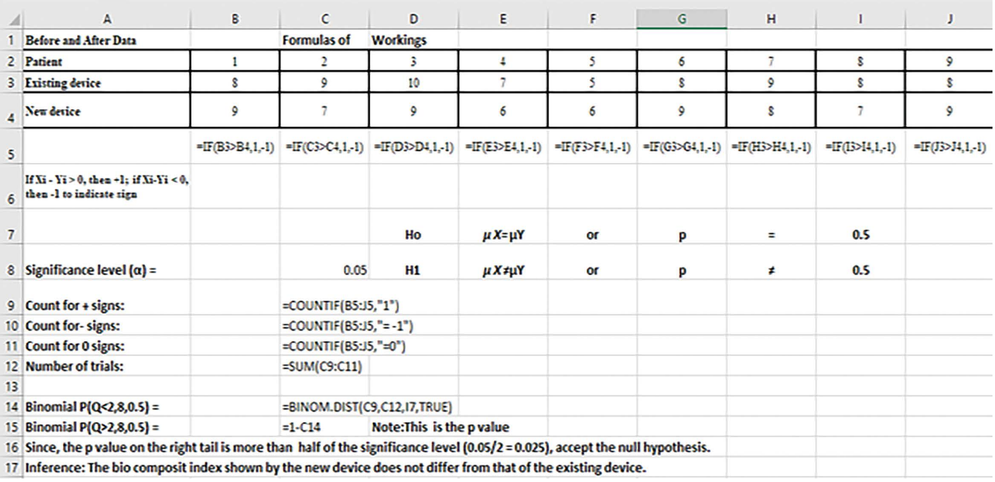
Table 12.11 Data for Example 12.11

Patient	1	2	3	4	5	6	7	8	9
Existing device	8	9	10	7	5	8	9	8	8
New device	9	7	9	6	6	9	8	7	9



	A	B	C	D	E	F	G	H	I	J
1	Before and After Data			Workings						
2	Patient	1	2	3	4	5	6	7	8	9
3	Existing device	8	9	10	7	5	8	9	8	8
4	New device	9	7	9	6	6	9	8	7	9
5		-1	1	1	1	-1	-1	1	1	-1
6	If Xi - Yi > 0, then +1; if Xi-Yi < 0, then -1 to indicate sign									
7				Ho	µX=µY	or	p	=	0.5	
8	Significance level (α) =		0.05	H1	µX≠µY	or	p	≠	0.5	
9	Count for + signs:		5							
10	Count for- signs:		4							
11	Count for 0 signs:		0							
12	Number of trials:		9							
13										
14	Binomial P(Q<2,8,0.5) =		0.74609375							
15	Binomial P(Q>2,8,0.5) =		0.25390625	Note:This is the p value						
16	Since, the p value on the right tail is more than half of the significance level (0.05/2 = 0.025), accept the null hypothesis.									
17	Inference: The bio composit index shown by the new device does not differ from that of the existing device.									

Figure 12.19 Screenshot of the working of Example 12.11



	A	B	C	D	E	F	G	H	I	J
1	Before and After Data		Formulas of	Workings						
2	Patient	1	2	3	4	5	6	7	8	9
3	Existing device	8	9	10	7	5	8	9	8	8
4	New device	9	7	9	6	6	9	8	7	9
5		=IF(B3>B4,1,-1)	=IF(C3>C4,1,-1)	=IF(D3>D4,1,-1)	=IF(E3>E4,1,-1)	=IF(F3>F4,1,-1)	=IF(G3>G4,1,-1)	=IF(H3>H4,1,-1)	=IF(I3>I4,1,-1)	=IF(J3>J4,1,-1)
6	If Xi - Yi > 0, then +1; if Xi-Yi < 0, then -1 to indicate sign									
7				Ho	µX=µY	or	p	=	0.5	
8	Significance level (α) =		0.05	H1	µX≠µY	or	p	≠	0.5	
9	Count for + signs:		=COUNTIF(B5:J5,"1")							
10	Count for- signs:		=COUNTIF(B5:J5,"= -1")							
11	Count for 0 signs:		=COUNTIF(B5:J5,"=0")							
12	Number of trials:		=SUM(C9:C11)							
13										
14	Binomial P(Q<2,8,0.5) =		=BINOM.DIST(C9,C12,I7,TRUE)							
15	Binomial P(Q>2,8,0.5) =		=1-C14	Note:This is the p value						
16	Since, the p value on the right tail is more than half of the significance level (0.05/2 = 0.025), accept the null hypothesis.									
17	Inference: The bio composit index shown by the new device does not differ from that of the existing device.									

Figure 12.20 Screenshot of guidelines for the formulas of the working of Example 12.11

All the calculations are shown in Figure 12.19 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.19 are shown in Figure 12.20.

12.5.2 Two-Sample Sign Test for Large Samples

Two random samples, each with size *n* with the condition that *np* as well as *n* (1 – *p*) is greater than or equal to 5, are selected.

The probability that the number of plus signs of the combined sample (sum of plus signs when $X_i - Y_i$ is positive) is more than the mean value (*p*) is 1/2 and the probability that number of plus signs of the combined sample (sum of plus signs when $X_i - Y_i$ is positive) is less than the mean value (p) is1/2. In this situation, the observations ($X_i - Y_i$) of the sample are classified based on the sample median μ such that the observations ($X_i - Y_i$) of the sample which are more than μ are assigned the plus (+) sign and the observations which are less than μ are assigned the minus (–) sign. Then the number of plus signs is taken as the value of the random variable Q of the binomial distribution, with $p = 1/2$ and the number of trials n to compute the probability that Q is more than the number of plus signs to compute p at the right tail. A normal approximation to a binomial distribution with $p = 1/2$ is used to test this hypothesis.

12.5.2.1 One-Tailed Two-Sample Sign Test for Large Samples Using Excel Sheets and NORM.S.DIST Function

The hypotheses of the one-tailed two-sample sign test when the sample sizes are large are as follows.

Test 1: H_0: $\mu_X = \mu_Y$ or $p = 1/2$ and $H_{1:}$ $\mu_X > \mu_Y$ or $p > 1/2$
Test 2: H_0: $\mu_X = \mu_Y$ or $p = ½$ and $H_{1:}$ $\mu_X < \mu_Y$ or $p < ½$
Let $Q \sim B(Q, n, p)$
$Mean = \mu_Q = np$

$$Variance = \sigma_Q^2 = np(1-p)$$

If n is greater than or equal to 20, then Q can be approximated to a normal distribution $N[np, np(1 - p)]$.

The combined standard normal statistic Z is as follows.

$$Z = \frac{Q - \mu_Q}{\sigma_Q} = \frac{Q - np}{\sqrt{np(1-p)}}$$

In the first test on the combined statistic, find K, which is the value of Z, and then find $P(Z \geq K)$, which is the p value at the right tail. Verify whether it is less than the given significance level, α kept at the right tail of the distribution. If so, reject the null hypothesis; otherwise, accept the null hypothesis.

In the second test on the combined statistic, first find K, which is the value of Z, and then find $P(Z \leq K)$, and it is the p computed value at the left tail. If this p value is less than the significance level (α) kept at the left tail of the distribution, reject the null hypothesis; otherwise, accept the null hypothesis.

Example 12.12

The fail percentage of students in 20 randomly selected management departments of affiliated colleges in a university before and after introducing a trimester system are summarised in Table 12.12.

Check the hypothesis that the fail percentage after introducing the trimester system increased against the alternate hypothesis that it dropped at a significance level of 0.05 using a sign test.

Table 12.12 Details of Percentage of Failures

Management Department	*Before Introducing Trimester*	*After Introducing Trimester*
1	40	21
2	35	46
3	3	2
4	46	35
5	22	11
6	51	58
7	55	60
8	8	1
9	14	20
10	31	12
11	12	15
12	33	11
13	58	34
14	30	34
15	25	13
16	27	35
17	16	14
18	13	10
19	27	40
20	28	20

Solution

The data for Example 12.12 are shown in Table 12.13.

The hypotheses of the one-tailed two-sample sign test when the sample sizes are large are as follows.

H_0: $\mu_X = \mu_Y$ or $p = 1/2$
$H_{1:}$ $\mu_X > \mu_Y$ or $p = 1/2$
Let

*X*i be the percentage of failures of the i^{th} department before implementing the trimester system

*Y*i be the percentage of failures of the i^{th} department after implementing the trimester system

If $X_i - Y_i > 0$, then the sign is +, which is indicated by +1

If $X_i - Y_i < 0$, then the sign is –, which is indicated by −1

If $X_i - Y_i = 0$, then the sign is 0

The number of plus signs is assumed to follow a binomial distribution.

Let $Q \sim B(Q, n, p)$

$$Mean = \mu_Q = np$$

$$Variance = \sigma_Q^2 = np(1-p)$$

Table 12.13 Details of Percentage Failures in Departments

Management Department	*Before*	*After*
1	40	21
2	35	46
3	3	2
4	46	35
5	22	11
6	51	58
7	55	60
8	8	1
9	14	20
10	31	12
11	12	15
12	33	11
13	58	34
14	30	34
15	25	13
16	27	35
17	16	14
18	13	10
19	27	40
20	28	20

If n is greater than or equal to 20, then Q can be approximated to a normal distribution $N[np, np(1-p)]$.

The combined standard normal statistic Z is as follows.

$$Z = \frac{Q - \mu_Q}{\sigma_Q} = \frac{Q - np}{\sqrt{np(1-p)}}$$

All the calculations are shown in Figure 12.21 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.21 are shown in Figure 12.22.

Example 12.13

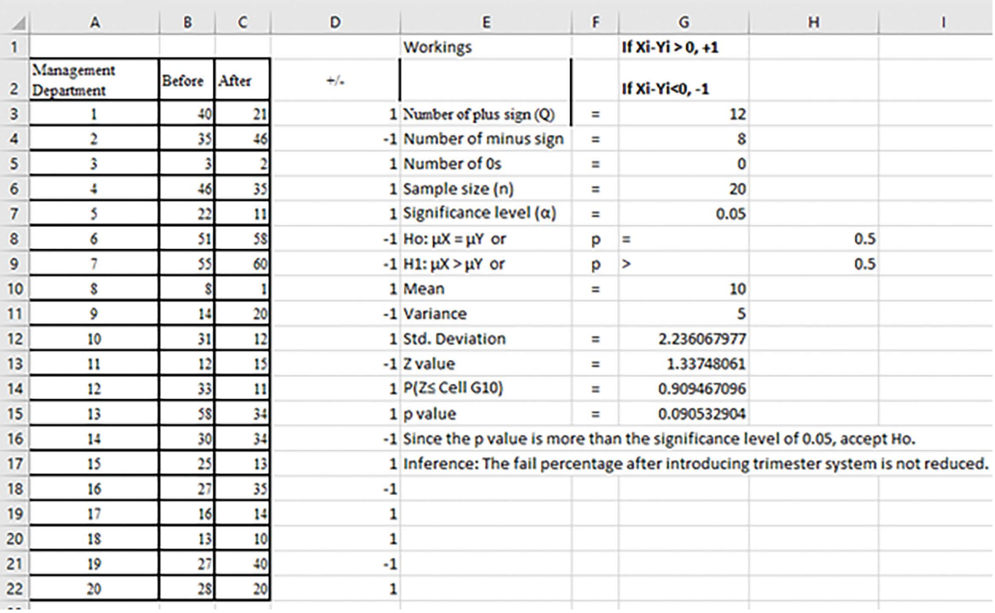
A state government has provided small-scale industries with a stimulus package. The director of the state's industry department gathered data on the annual sales in lakhs of rupees of 20 companies before and after the stimulus package was announced. The data are presented in Table 12.14. Using a sign test with a significance threshold of 0.05, compare the null hypothesis that the annual sales have not been improved against the alternative hypothesis that the annual sales have been improved.

Solution

The data for Example 12.13 are shown in Table 12.15.

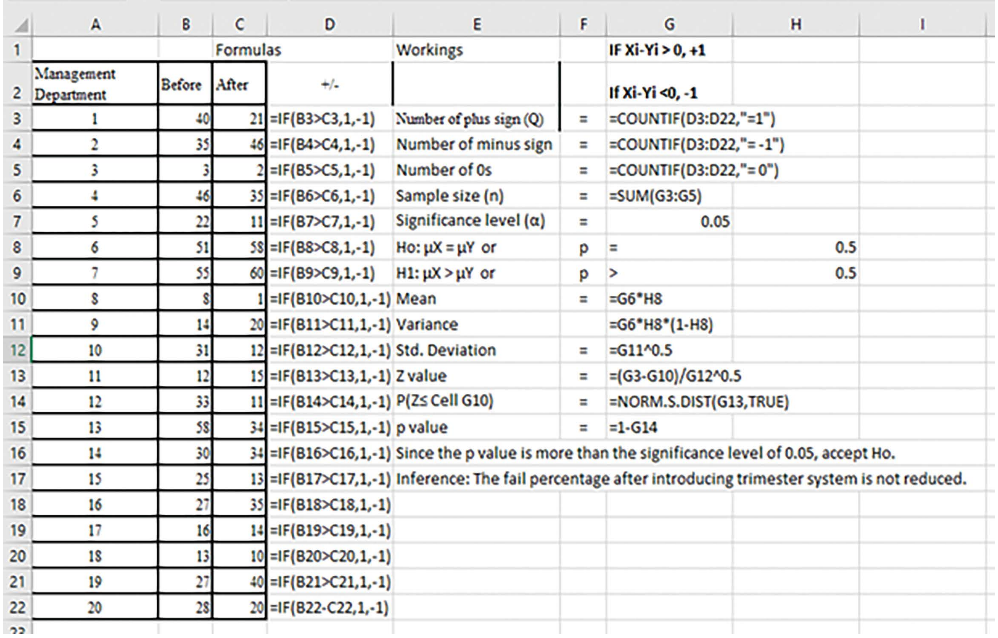
The hypotheses of the one-tailed two-sample sign test when the sample size is large are as follows.

H_0: $\mu_X = \mu_Y$ or $p = 1/2$
$H_{1:}$ $\mu_X < \mu_Y$ or $p = 1/2$



	A	B	C	D	E	F	G	H	I
1					Workings		If Xi-Yi > 0, +1		
2	Management Department	Before	After	+/-			If Xi-Yi<0, -1		
3	1	40	21	1	Number of plus sign (Q)	=	12		
4	2	35	46	-1	Number of minus sign	=	8		
5	3	3	2	1	Number of 0s	=	0		
6	4	46	35	1	Sample size (n)	=	20		
7	5	22	11	1	Significance level (α)	=	0.05		
8	6	51	58	-1	Ho: µX = µY or	p	=	0.5	
9	7	55	60	-1	H1: µX > µY or	p	>	0.5	
10	8	8	1	1	Mean	=	10		
11	9	14	20	-1	Variance		5		
12	10	31	12	1	Std. Deviation	=	2.236067977		
13	11	12	15	-1	Z value	=	1.33748061		
14	12	33	11	1	P(Z≤ Cell G10)	=	0.909467096		
15	13	58	34	1	p value	=	0.090532904		
16	14	30	34	-1	Since the p value is more than the significance level of 0.05, accept Ho.				
17	15	25	13	1	Inference: The fail percentage after introducing trimester system is not reduced.				
18	16	27	35	-1					
19	17	16	14	1					
20	18	13	10	1					
21	19	27	40	-1					
22	20	28	20	1					

Figure 12.21 Screenshot of the working of Example 12.12



	A	B	C	D	E	F	G	H	I
1			Formulas		Workings		IF Xi-Yi > 0, +1		
2	Management Department	Before	After	+/-			If Xi-Yi <0, -1		
3	1	40	21	=IF(B3>C3,1,-1)	Number of plus sign (Q)	=	=COUNTIF(D3:D22,"=1")		
4	2	35	46	=IF(B4>C4,1,-1)	Number of minus sign	=	=COUNTIF(D3:D22,"= -1")		
5	3	3	2	=IF(B5>C5,1,-1)	Number of 0s	=	=COUNTIF(D3:D22,"= 0")		
6	4	46	35	=IF(B6>C6,1,-1)	Sample size (n)	=	=SUM(G3:G5)		
7	5	22	11	=IF(B7>C7,1,-1)	Significance level (α)	=	0.05		
8	6	51	58	=IF(B8>C8,1,-1)	Ho: µX = µY or	p	=	0.5	
9	7	55	60	=IF(B9>C9,1,-1)	H1: µX > µY or	p	>	0.5	
10	8	8	1	=IF(B10>C10,1,-1)	Mean	=	=G6*H8		
11	9	14	20	=IF(B11>C11,1,-1)	Variance		=G6*H8*(1-H8)		
12	10	31	12	=IF(B12>C12,1,-1)	Std. Deviation	=	=G11^0.5		
13	11	12	15	=IF(B13>C13,1,-1)	Z value	=	=(G3-G10)/G12^0.5		
14	12	33	11	=IF(B14>C14,1,-1)	P(Z≤ Cell G10)	=	=NORM.S.DIST(G13,TRUE)		
15	13	58	34	=IF(B15>C15,1,-1)	p value	=	=1-G14		
16	14	30	34	=IF(B16>C16,1,-1)	Since the p value is more than the significance level of 0.05, accept Ho.				
17	15	25	13	=IF(B17>C17,1,-1)	Inference: The fail percentage after introducing trimester system is not reduced.				
18	16	27	35	=IF(B18>C18,1,-1)					
19	17	16	14	=IF(B19>C19,1,-1)					
20	18	13	10	=IF(B20>C20,1,-1)					
21	19	27	40	=IF(B21>C21,1,-1)					
22	20	28	20	=IF(B22-C22,1,-1)					

Figure 12.22 Screenshot of guidelines for the formulas of the working of Example 12.12

Table 12.14 Annual Sales Before and After Stimulus Package

Company	*Before*	*After*
1	17	19
2	15	14
3	17	18
4	14	15
5	18	19
6	19	17
7	14	17
8	12	17
9	16	19
10	19	18
11	18	17
12	17	19
13	12	16
14	18	17
15	15	16
16	13	16
17	14	19
18	17	19
19	20	19
20	20	21

Table 12.15 Annual Sales Before and After Stimulus Package

Company	*Before*	*After*
1	17	19
2	15	14
3	17	18
4	14	15
5	18	19
6	19	17
7	14	17
8	12	17
9	16	19
10	19	18
11	18	17
12	17	19
13	12	16
14	18	17
15	15	16
16	13	16
17	14	19
18	17	19
19	20	19
20	20	21

Let
*X*i be the annual sales of the i^{th} company before implementing the stimulus package
*Y*i be the annual sales of the i^{th} company after implementing the stimulus package
If $X_i - Y_i > 0$, then the sign is +, which is indicated by +1
If $X_i - Y_i < 0$, then the sign is –, which is indicated by –1
If $X_i - Y_i = 0$, then the sign is 0

The number of plus signs is assumed to follow a binomial distribution.

$$\text{Let } Q \sim B(Q, n, p)$$

$$Mean = \mu_Q = np$$

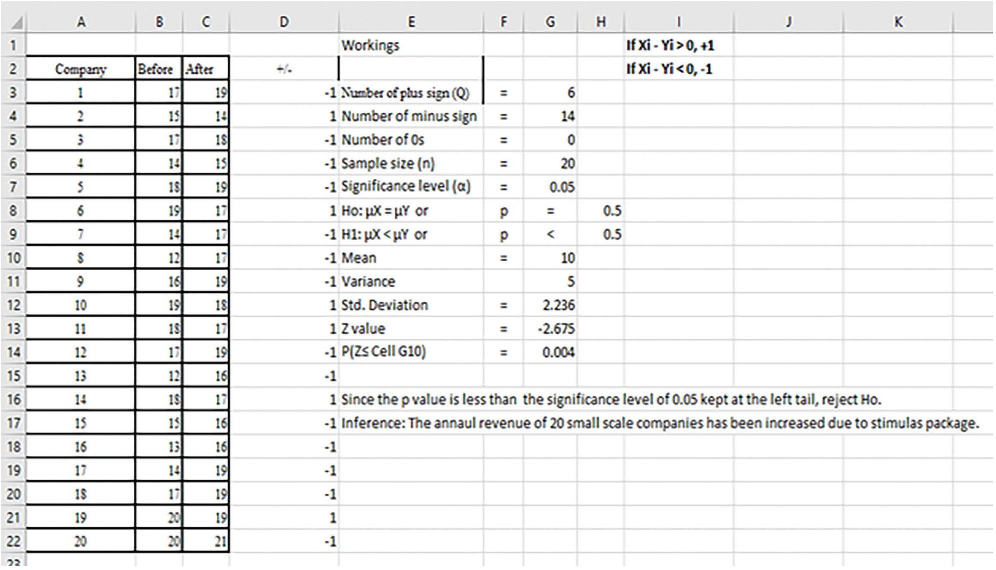
$$Variance = \sigma_Q^2 = np(1-p)$$

If is n is greater than or equal to 20, then Q can be approximated to a normal distribution $N(np, np(1-p)$.

The combined standard normal statistic Z is as follows.

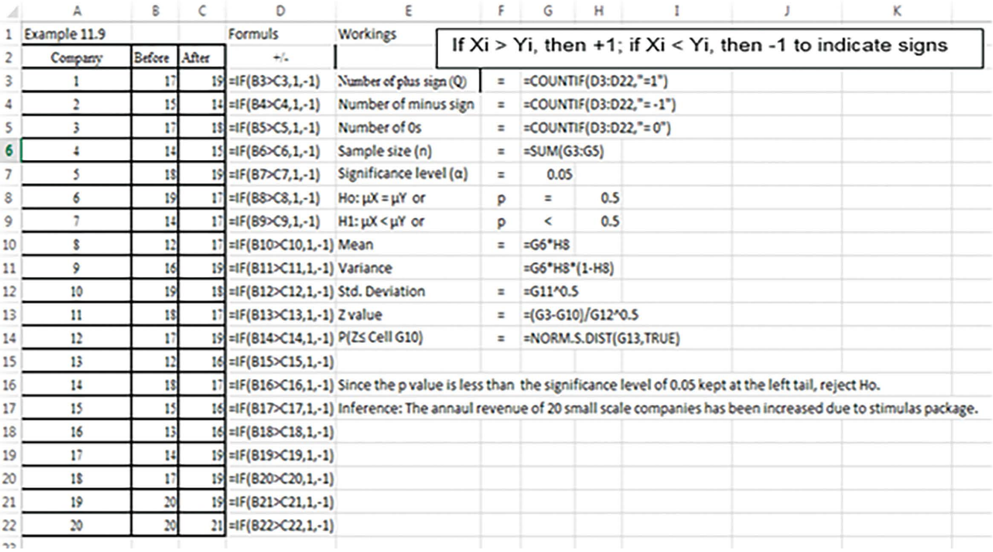
$$Z = \frac{Q - \mu_Q}{\sigma_Q} = \frac{Q - np}{\sqrt{np(1-p)}}$$

All the calculations are shown in Figure 12.23 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.23 are shown in Figure 12.24.



	A	B	C	D	E	F	G	H	I	J	K
1					Workings				If Xi - Yi > 0, +1		
2	Company	Before	After	+/-					If Xi - Yi < 0, -1		
3	1	17	19	-1	Number of plus sign (Q)	=	6				
4	2	15	14	1	Number of minus sign	=	14				
5	3	17	18	-1	Number of 0s	=	0				
6	4	14	15	-1	Sample size (n)	=	20				
7	5	18	19	-1	Significance level (α)	=	0.05				
8	6	19	17	1	Ho: μX = μY or	p	=	0.5			
9	7	14	17	-1	H1: μX < μY or	p	<	0.5			
10	8	12	17	-1	Mean	=	10				
11	9	16	19	-1	Variance		5				
12	10	19	18	1	Std. Deviation	=	2.236				
13	11	18	17	1	Z value	=	-2.675				
14	12	17	19	-1	P(Z≤ Cell G10)	=	0.004				
15	13	12	16	-1							
16	14	18	17	1	Since the p value is less than the significance level of 0.05 kept at the left tail, reject Ho.						
17	15	15	16	-1	Inference: The annaul revenue of 20 small scale companies has been increased due to stimulas package.						
18	16	13	16	-1							
19	17	14	19	-1							
20	18	17	19	-1							
21	19	20	19	1							
22	20	20	21	-1							

Figure 12.23 Screenshot of the working of Example 12.13



	A	B	C	D	E	F	G	H
1	Example 11.9			Formuls	Workings			
2	Company	Before	After	+/-				
3	1	17	19	=IF(B3>C3,1,-1)	Number of plus sign (Q)	=	=COUNTIF(D3:D22,"=1")	
4	2	15	14	=IF(B4>C4,1,-1)	Number of minus sign	=	=COUNTIF(D3:D22,"=-1")	
5	3	17	18	=IF(B5>C5,1,-1)	Number of 0s	=	=COUNTIF(D3:D22,"=0")	
6	4	14	15	=IF(B6>C6,1,-1)	Sample size (n)	=	=SUM(G3:G5)	
7	5	18	19	=IF(B7>C7,1,-1)	Significance level (α)	=	0.05	
8	6	19	17	=IF(B8>C8,1,-1)	Ho: μX = μY or	p	=	0.5
9	7	14	17	=IF(B9>C9,1,-1)	H1: μX < μY or	p	<	0.5
10	8	12	17	=IF(B10>C10,1,-1)	Mean	=	=G6*H8	
11	9	16	19	=IF(B11>C11,1,-1)	Variance		=G6*H8*(1-H8)	
12	10	19	18	=IF(B12>C12,1,-1)	Std. Deviation	=	=G11^0.5	
13	11	18	17	=IF(B13>C13,1,-1)	Z value	=	=(G3-G10)/G12^0.5	
14	12	17	19	=IF(B14>C14,1,-1)	P(Z≤ Cell G10)	=	=NORM.S.DIST(G13,TRUE)	
15	13	12	16	=IF(B15>C15,1,-1)				
16	14	18	17	=IF(B16>C16,1,-1)	Since the p value is less than the significance level of 0.05 kept at the left tail, reject Ho.			
17	15	15	16	=IF(B17>C17,1,-1)	Inference: The annaul revenue of 20 small scale companies has been increased due to stimulas package.			
18	16	13	16	=IF(B18>C18,1,-1)				
19	17	14	19	=IF(B19>C19,1,-1)				
20	18	17	19	=IF(B20>C20,1,-1)				
21	19	20	19	=IF(B21>C21,1,-1)				
22	20	20	21	=IF(B22>C22,1,-1)				

If Xi > Yi, then +1; if Xi < Yi, then -1 to indicate signs

Figure 12.24 Screenshot of guidelines for the formulas of the working of Example 12.13

12.5.2.2 *Two-Tailed Two-Sample Sign Test for Large Samples Using Excel Sheets and NORM.S.DIST Function*

Two samples, which are before and after introducing a change in the system of interest, are taken. These two samples are known to be non-normal. If the sample size is greater than or equal to 20, then the number of plus signs (sum of the positive difference (+) between the i^{th} observation before a change and the i^{th} observation taken after a change in the system) follows a binomial distribution. The number of plus signs is taken as the value of the random variable Q of the binomial distribution, with $p = 1/2$ and the number of trials n to compute the probability that Q is more than the number of plus signs to check the following hypotheses. The mean and variance of the samples, that is, np and $np(1 - p)$, are at least 5.

H_0: $\mu_X = \mu_Y$ or $p = 1/2$
$H_{1:}$ $\mu_X \neq \mu_Y$ or $p \neq 1/2$
Let $Q \sim B(Q, n, p)$

$$Mean = \mu_Q = np$$

$$Variance = \sigma_Q^2 = np(1-p)$$

If n is greater than or equal to 20, then Q can be approximated to a normal distribution $N[np, np(1 - p)]$.

The combined standard normal statistic Z is as follows.

$$Z = \frac{Q - \mu_Q}{\sigma_Q} = \frac{Q - np}{\sqrt{np(1-p)}}$$

In this test on the combined statistic, find K, which is the value of Z, and then find $P(Z \geq K)$, which is the p value at the right tail. Verify whether it is less than half of the given significance level, $\alpha/2$ kept at the right tail of the distribution. If so, reject the null hypothesis; otherwise, accept the null hypothesis.

Example 12.14

Table 12.16 provides a summary of the monthly revenues (in lakhs of rupees) of a product in 21 randomly chosen retail stores before and after the introduction of a point-of-sale advertisement. At a significance level of 0.10, determine whether the sales revenues before and after the point-of-sale advertisement differ.

Solution

The data for Example 12.14 are shown in Table 12.17.

The hypotheses of the two-tailed two-sample sign test when the sample size is large are as follows.

H_0: $\mu_X = \mu_Y$ or $p = 1/2$

$H_{1:}$ $\mu_X \neq \mu_Y$ or $p \neq 1/2$

Table 12.16 Data for Example 12.14

Retail Store	*Sales Revenue in Lakhs of Rupees*	
	Before Introducing Point-of-Sale Advertisement	*After Introducing Point-of-Sale Advertisement*
1	17	11
2	22	21
3	18	16
4	16	15
5	18	12
6	13	20
7	17	15
8	14	22
9	18	20
10	18	16
11	16	15
12	17	22
13	15	14
14	16	14
15	14	16
16	13	12
17	17	15
18	15	14
19	16	18
20	18	20
21	19	17

Table 12.17 Data for Example 12.14

Retail Store	*Sales Revenue in Lakhs of Rupees*	
	Before Introducing Point-of-Sale Advertisement	*After Introducing Point-of-Sale Advertisement*
1	17	11
2	22	21
3	18	16
4	16	15
5	18	12
6	13	20
7	17	15
8	14	22
9	18	20
10	18	16
11	16	15
12	17	22
13	15	14
14	16	14
15	14	16
16	13	12
17	17	15
18	15	14
19	16	18
20	18	20
21	19	17

Let

Xi be the sales revenue of the i^{th} retail store room before putting the point of sales advertisement

Yi be the sales revenue of the i^{th} retail store room after putting the point of sales advertisement

If $X_i - Y_i > 0$, then the sign is +, which is indicated by +1

If $X_i - Y_i < 0$, then the sign is –, which is indicated by –1

If $X_i - Y_i = 0$, then the sign is 0

The number of plus signs is assumed to follow a binomial distribution.

Let $Q \sim B(Q, n, p)$

$$Mean = \mu_Q = np$$

$$Variance = \sigma_Q^2 = np(1-p)$$

If is n is greater than or equal to 20, then Q can be approximated to a normal distribution $N[np, np(1-p)]$.

The combined standard normal statistic Z is as follows.

$$Z = \frac{Q-\mu_Q}{\sigma_Q} = \frac{Q-np}{\sqrt{np(1-p)}}$$

All the calculations are shown in Figure 12.25 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.25 are shown in Figure 12.26.

12.6 Median Test Using Excel Sheets, COUNTIF, COMBIN, and CHISQ. DIST.RT Functions

In the two-sample sign test, which was presented in the earlier sections, the sample sizes are equal. This constraint is not present in the two-sample median test. The objective of the median test is to check whether the two samples which are independent have been drawn from two populations with the same median. The associated hypotheses are as follows.

H_0: The two samples, which are independent, are drawn from the two populations which have the same median.

H_1: The two samples, which are independent, are not drawn from the two populations which have the same median.

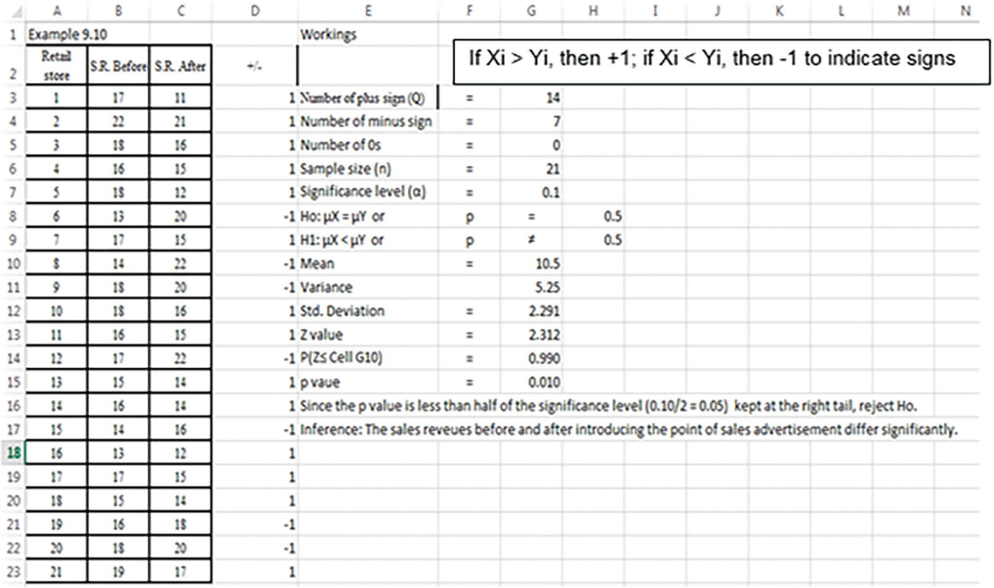
Steps of Median Test

The steps that are required to perform the median test are as follows.

Step1: Input the following.

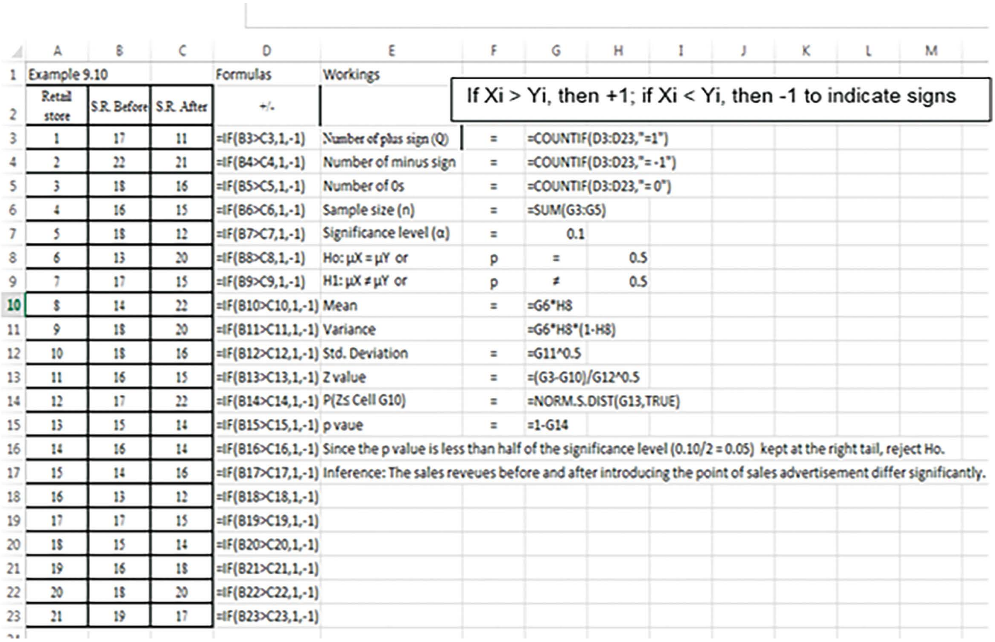
a. Size of sample 1 (n_1) and that of sample 2 (n_2).
b. Observations of sample 1 $A(i)$, $i = 1, 2, 3, \ldots, n_1$.
c. Observation of sample 2, $B(i)$, $i = 1, 2, 3, \ldots, n_2$.

Step 2: Combine the observations of the two samples and sort them from low to high.



	A	B	C	D	E	F	G	H
1	Example 9.10				Workings			
2	Retail store	S.R. Before	S.R. After	+/-		If Xi > Yi, then +1; if Xi < Yi, then -1 to indicate signs		
3	1	17	11	1	Number of plus sign (Q)	=	14	
4	2	22	21	1	Number of minus sign	=	7	
5	3	18	16	1	Number of 0s	=	0	
6	4	16	15	1	Sample size (n)	=	21	
7	5	18	12	1	Significance level (α)	=	0.1	
8	6	13	20	-1	Ho: μX = μY or	p	=	0.5
9	7	17	15	1	H1: μX < μY or	p	≠	0.5
10	8	14	22	-1	Mean	=	10.5	
11	9	18	20	-1	Variance		5.25	
12	10	18	16	1	Std. Deviation	=	2.291	
13	11	16	15	1	Z value	=	2.312	
14	12	17	22	-1	P(Z≤ Cell G10)	=	0.990	
15	13	15	14	1	p vaue	=	0.010	
16	14	16	14	1	Since the p value is less than half of the significance level (0.10/2 = 0.05) kept at the right tail, reject Ho.			
17	15	14	16	-1	Inference: The sales reveues before and after introducing the point of sales advertisement differ significantly.			
18	16	13	12	1				
19	17	17	15	1				
20	18	15	14	1				
21	19	16	18	-1				
22	20	18	20	-1				
23	21	19	17	1				

Figure 12.25 Screenshot of the working of Example 12.14



	A	B	C	D	E	F	G	H
1	Example 9.10			Formulas	Workings			
2	Retail store	S.R. Before	S.R. After	+/-				
3	1	17	11	=IF(B3>C3,1,-1)	Number of plus sign (Q)	=	=COUNTIF(D3:D23,"=1")	
4	2	22	21	=IF(B4>C4,1,-1)	Number of minus sign	=	=COUNTIF(D3:D23,"= -1")	
5	3	18	16	=IF(B5>C5,1,-1)	Number of 0s	=	=COUNTIF(D3:D23,"= 0")	
6	4	16	15	=IF(B6>C6,1,-1)	Sample size (n)	=	=SUM(G3:G5)	
7	5	18	12	=IF(B7>C7,1,-1)	Significance level (α)	=	0.1	
8	6	13	20	=IF(B8>C8,1,-1)	Ho: μX = μY or	p	=	0.5
9	7	17	15	=IF(B9>C9,1,-1)	H1: μX ≠ μY or	p	≠	0.5
10	8	14	22	=IF(B10>C10,1,-1)	Mean	=	=G6*H8	
11	9	18	20	=IF(B11>C11,1,-1)	Variance		=G6*H8*(1-H8)	
12	10	18	16	=IF(B12>C12,1,-1)	Std. Deviation	=	=G11^0.5	
13	11	16	15	=IF(B13>C13,1,-1)	Z value	=	=(G3-G10)/G12^0.5	
14	12	17	22	=IF(B14>C14,1,-1)	P(Z≤ Cell G10)	=	=NORM.S.DIST(G13,TRUE)	
15	13	15	14	=IF(B15>C15,1,-1)	p vaue	=	=1-G14	
16	14	16	14	=IF(B16>C16,1,-1)	Since the p value is less than half of the significance level (0.10/2 = 0.05) kept at the right tail, reject Ho.			
17	15	14	16	=IF(B17>C17,1,-1)	Inference: The sales reveues before and after introducing the point of sales advertisement differ significantly.			
18	16	13	12	=IF(B18>C18,1,-1)				
19	17	17	15	=IF(B19>C19,1,-1)				
20	18	15	14	=IF(B20>C20,1,-1)				
21	19	16	18	=IF(B21>C21,1,-1)				
22	20	18	20	=IF(B22>C22,1,-1)				
23	21	19	17	=IF(B23>C23,1,-1)				

If Xi > Yi, then +1; if Xi < Yi, then -1 to indicate signs

Figure 12.26 Screenshot of guidelines for the formulas of the working of Example 12.14

Step 3: Find the median of the combined sorted observations obtained in Step 2.
Step 4: Find the frequency of the observations of sample 1 for the condition Observation > Median.
Let it be *a*.
Step 5: Find the frequency of the observations of sample 2 for the condition Observation > Median.
Let it be *b*.
Step 6: Find the frequency of the observations of sample 1 for the condition Observation ≤ Median.
Let it be *c*.
Step 7: Find the frequency of the observations of sample 2 for the condition Observation ≤ Median.
Let it be *d*.
Step 8: Summarise the results of steps from 4 to 7 in the format shown in Table 12.18.
Step 9: If the size of the pooled sample is small (≤ 30), the *p* is given by the following formula; otherwise, go to Step 10.

$$p=\frac{\left(n_{1_{C_a}}\times n_{2_{C_b}}\right)}{(n_1+n_2)_{C_{(a+b)}}}$$

where
n_1 is the size of sample 1
n_2 is the size of sample 2
a is the frequency from sample 1 for the condition Observation > Median

b is the frequency from sample 2 for the condition Observation > Median

If the value of *p* is less than the given significance level, then reject the null hypothesis; otherwise, accept the null hypothesis.

Step 10: If the size of the pooled sample is large (>30), then use the following formula to obtain χ^2 statistic [4].

$$\chi^2 = \frac{N\left(|ad - bc| - \frac{N}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)} with 1 degree of freedom$$

Where

n_1 is the size of sample 1

n_2 is the size of sample 2

a is the frequency from sample 1 for the condition Observation > Median

b is the frequency from sample 2 for the condition Observation > Median

c is the frequency from sample 1 for the condition Observation ≤ Median

d is the frequency from sample 2 for the condition Observation ≤ Median

N is the sum of n_1 and n_2.

If the χ^2 calculated is more than the table χ^2 with 1 degree of freedom and a significance level of α, then reject the null hypothesis; otherwise, accept the null hypothesis.

Or:

Find P($\chi^2 \geq$ Computed χ^2 value), and if it is less than the given significance level α, then reject H_0; otherwise, accept H_0.

End of the step is reached.

Example 12.15

In Table 12.19, two alternative productivity enhancement strategies (increased incentive, enhanced welfare measure) are compared in terms of the percentage of incremental efficiency of various shops in a factory. At a significance level of 0.05, determine whether the two samples came from populations with the same median.

Solution

The data for Example 12.15 are shown in Table 12.20.

Table 12.18 Intermediate Calculations of Two-Sample Median Test

	Sample 1	*Sample 2*	*Row Total*
Above median	*a*	*b*	*a* + *b*
Below median	*c*	*d*	*c* + *d*
Column total	*a* + *c*	*b* + *d*	$\boldsymbol{N = n_1 + n_2}$

Table 12.19 Percentage Incremental Efficiencies of Shops

Shop	*Increased Incentive*	*Increased Welfare Measure*
1	30	35
2	35	40
3	51	47
4	45	60
5	70	60
6	50	70
7	38	

Table 12.20 Data for Example 12.15

Shop	*Increased Incentive*	*Increased Welfare Measure*
1	30	35
2	35	40
3	51	47
4	45	60
5	70	60
6	50	70
7	38	

The populations of this example are as follows.

1. Percentage incremental efficiencies of the shops due to the strategy Increased incentive
2. Percentage incremental efficiencies of the shops due to the strategy Increased welfare measure

The size of sample 1 (n_1) from population 1 is 7, and that of sample 2 (n_2) is 6
The size of the combined population ($n_1 + n_2$) is 13
The hypotheses of this example are as follows.

H_0: The two samples, which are independent, are drawn from the two populations which have the same median.
H_1: The two samples, which are independent, are not drawn from the two populations which have the same median.

Significance level (α) = 0.05

Since the size of the combined sample is less than 30, the formula to compute the value of p is as follows.

$$p = \frac{\left(n_{1_{C_a}} \times n_{2_{C_b}}\right)}{(n_1 + n_2)_{C_{(a+b)}}}$$

where
n_1 is the size of sample 1
n_2 is the size of sample 2

a is the frequency from sample 1 for the condition Observation > Median
b is the frequency from sample 2 for the condition Observation > Median

If the value of p is less than the given significance level (α) of 0.05, then reject the null hypothesis; otherwise, accept the null hypothesis.

All the calculations are shown in Figure 12.27 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.27 are shown in Figure 12.28.

Example 12.16

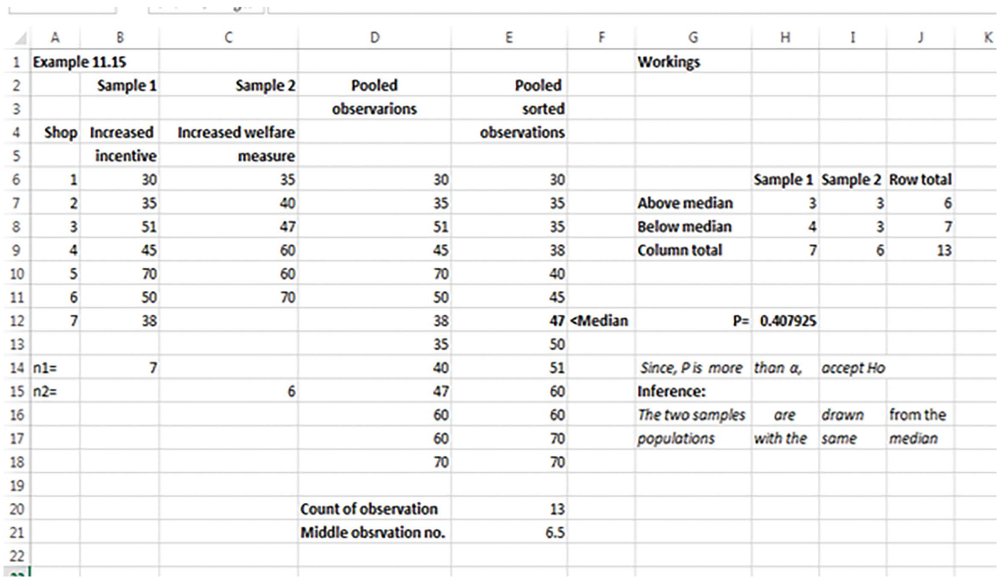
A marketing manager tried two strategies of sales promotion, price discount and door delivery. He collected data on the increase of daily sales from 16 shops using the price discount strategy and collected data on the increase of daily sales from 15 shops using the door delivery strategy. The increase in the daily sales is in thousands of rupees, which are summarised in Table 12.21. Check whether these two independent samples are drawn from populations with the same median at a significance level of 0.01.

Solution

The data for Example 12.16 are shown in Table 12.22.

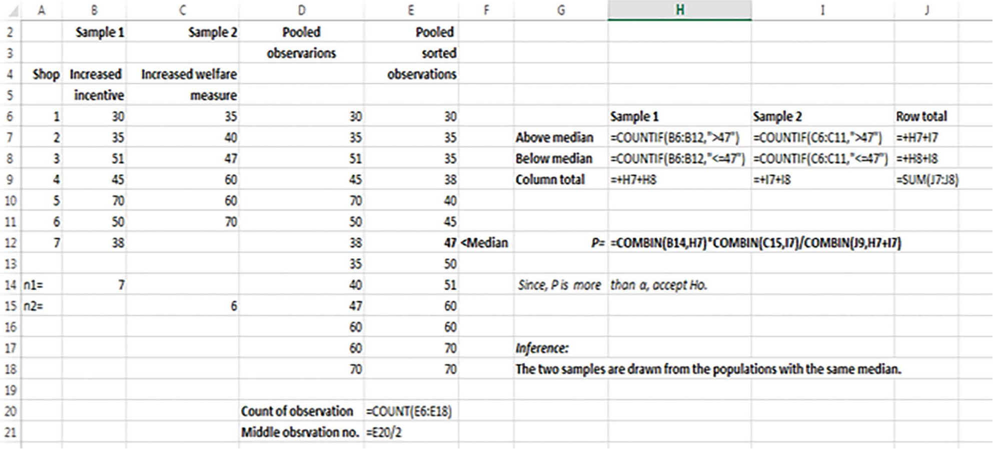
The populations of this example are as follows.

1. The increase in the daily sales of shops due to the strategy Price Discount
2. The increase in the daily sales of shops due to the strategy Door Delivery



	A	B	C	D	E	F	G	H	I	J	K
1	Example 11.15						Workings				
2		Sample 1	Sample 2	Pooled	Pooled						
3				observarions	sorted						
4	Shop	Increased	Increased welfare		observations						
5		incentive	measure								
6	1	30	35	30	30			Sample 1	Sample 2	Row total	
7	2	35	40	35	35		Above median	3	3	6	
8	3	51	47	51	35		Below median	4	3	7	
9	4	45	60	45	38		Column total	7	6	13	
10	5	70	60	70	40						
11	6	50	70	50	45						
12	7	38		38	47	<Median	P=	0.407925			
13				35	50						
14	n1=	7		40	51		*Since, P is more*	*than α,*	*accept Ho*		
15	n2=		6	47	60		Inference:				
16				60	60		*The two samples*	*are*	*drawn*	from the	
17				60	70		*populations*	*with the*	*same*	*median*	
18				70	70						
19											
20				Count of observation	13						
21				Middle obsrvation no.	6.5						
22											

Figure 12.27 Screenshot of the working of Example 12.15



	A	B	C	D	E	F	G	H	I	J
2		Sample 1	Sample 2	Pooled	Pooled					
3				observarions	sorted					
4	Shop	Increased	Increased welfare		observations					
5		incentive	measure							
6	1	30	35	30	30			Sample 1	Sample 2	Row total
7	2	35	40	35	35		Above median	=COUNTIF(B6:B12,">47")	=COUNTIF(C6:C11,">47")	=+H7+I7
8	3	51	47	51	35		Below median	=COUNTIF(B6:B12,"<=47")	=COUNTIF(C6:C11,"<=47")	=+H8+I8
9	4	45	60	45	38		Column total	=+H7+H8	=+I7+I8	=SUM(J7:J8)
10	5	70	60	70	40					
11	6	50	70	50	45					
12	7	38		38	47	<Median	P=	=COMBIN(B14,H7)*COMBIN(C15,I7)/COMBIN(J9,H7+I7)		
13				35	50					
14	n1=	7		40	51		Since, P is more	than a, accept Ho.		
15	n2=		6	47	60					
16				60	60					
17				60	70		Inference:			
18				70	70		The two samples are drawn from the populations with the same median.			
19										
20				Count of observation	=COUNT(E6:E18)					
21				Middle obsrvation no.	=E20/2					

Figure 12.28 Screenshot of guidelines for the formulas of the working of Example 12.15

Table 12.21 Data for Example 12.16 on Increase in Daily Sales

Retail Shop	*Strategy*	
	Price Discount	*Door Delivery*
1	35	40
2	40	42
3	56	52
4	50	57
5	75	72
6	55	63
7	43	39
8	45	44
9	51	64
10	61	67
11	74	73
12	76	80
13	87	90
14	92	88
15	86	83
16	90	-

The size of sample 1 (n_1) from population 1 is 16, and that of sample 2 (n_2) from population 2 is 15.

The size of the combined population ($n_1 + n_2$) is 31.

The hypotheses of this example are as follows.

H_0: The two samples, which are independent, are drawn from two populations with the same median.

H_1: The two samples, which are independent, are not drawn from two populations with the same median.

Table 12.22 Increase in Daily Sales of Shops

Retail Shop	*Strategy*	
	Price Discount	*Door Delivery*
1	35	40
2	40	42
3	56	52
4	50	57
5	75	72
6	55	63
7	43	39
8	45	44
9	51	64
10	61	67
11	74	73
12	76	80
13	87	90
14	92	88
15	86	83
16	90	

Significance level (α) = 0.01

Degree of freedom = 1

Since the size of the combined sample is more than 30, the formula to compute the value of χ^2 is as follows.

$$\chi^2 = \frac{N\left(|ad - bc| - \frac{N}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)} with 1 degree of freedom$$

where

n_1 is the size of sample 1

n_2 is the size of sample 2

a is the frequency from sample 1 for the condition Observation > Median

b is the frequency from sample 2 for the condition Observation > Median

c is the frequency from sample 1 for the condition Observation ≤ Median

d is the frequency from sample 2 for the condition Observation ≤ Median

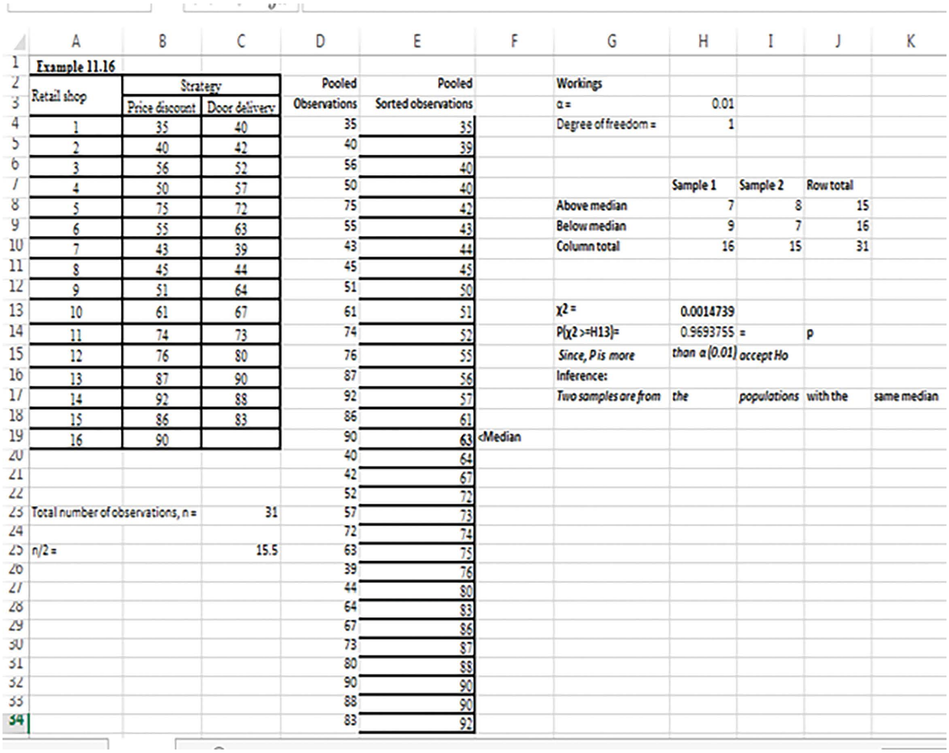
N is the sum of n_1 and n_2

Find P($\chi^2 \geq$ Computed χ^2 value), and if it is less than the given significance level α, then reject H_0; otherwise, accept H_0.

The chi-square formula used to compute probability of p is as follows.

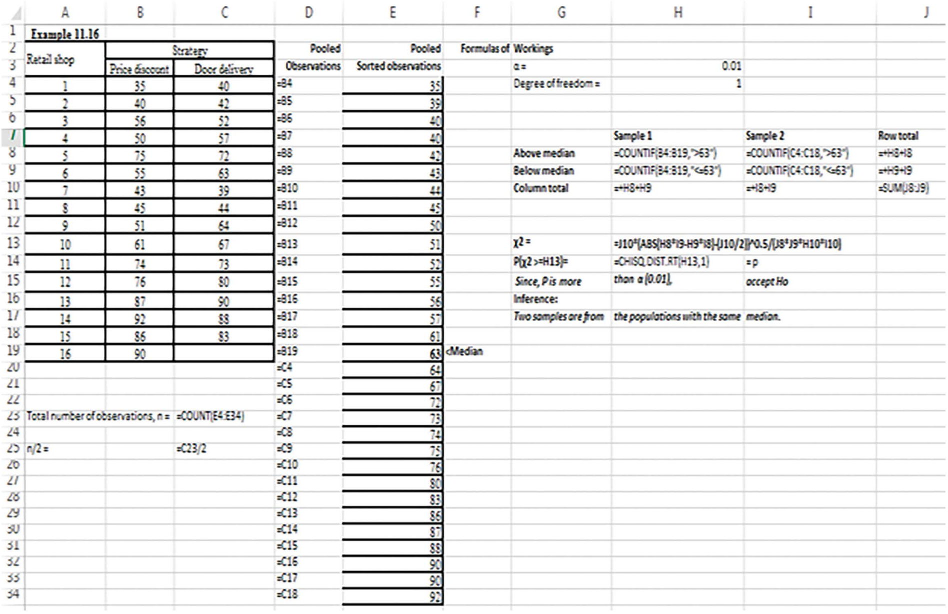
$$= CHISQ.DIST.RT\,(Computed\ Chi.square\ value,\ Degree\ of\ freedom\ of\ 1)$$

All the calculations are shown in Figure 12.29 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.29 are shown in Figure 12.30.



	A	B	C	D	E	F	G	H	I	J	K
1	**Example 11.16**										
2	Retail shop	Strategy		Pooled	Pooled		Workings				
3		Price discount	Door delivery	Observations	Sorted observations		α =	0.01			
4	1	35	40	35	35		Degree of freedom =	1			
5	2	40	42	40	39						
6	3	56	52	56	40						
7	4	50	57	50	40			Sample 1	Sample 2	Row total	
8	5	75	72	75	42		Above median	7	8	15	
9	6	55	63	55	43		Below median	9	7	16	
10	7	43	39	43	44		Column total	16	15	31	
11	8	45	44	45	45						
12	9	51	64	51	50						
13	10	61	67	61	51		χ2 =	0.0014739			
14	11	74	73	74	52		P(χ2 >=H13)=	0.9693755	=	p	
15	12	76	80	76	55		*Since, P is more*	*than α (0.01)*	*accept Ho*		
16	13	87	90	87	56		Inference:				
17	14	92	88	92	57		*Two samples are from*	*the*	*populations*	with the	same median
18	15	86	83	86	61						
19	16	90		90	63	<Median					
20				40	64						
21				42	67						
22				52	72						
23	Total number of observations, n =		31	57	73						
24				72	74						
25	n/2 =		15.5	63	75						
26				39	76						
27				44	80						
28				64	83						
29				67	86						
30				73	87						
31				80	88						
32				90	90						
33				88	90						
34				83	92						

Figure 12.29 Screenshot of the working of Example 12.16



	A	B	C	D	E	F	G	H	I	J
1	**Example 11.16**									
2	Retail shop	Strategy		Pooled	Pooled	Formulas of	Workings			
3		Price discount	Door delivery	Observations	Sorted observations		α =	0.01		
4	1	35	40	=B4	35		Degree of freedom =	1		
5	2	40	42	=B5	39					
6	3	56	52	=B6	40					
7	4	50	57	=B7	40			Sample 1	Sample 2	Row total
8	5	75	72	=B8	42		Above median	=COUNTIF(B4:B19,">63")	=COUNTIF(C4:C18,">63")	=+H8+I8
9	6	55	63	=B9	43		Below median	=COUNTIF(B4:B19,"<=63")	=COUNTIF(C4:C18,"<=63")	=+H9+I9
10	7	43	39	=B10	44		Column total	=+H8+H9	=+I8+I9	=SUM(J8:J9)
11	8	45	44	=B11	45					
12	9	51	64	=B12	50					
13	10	61	67	=B13	51		χ2 =	=J10*(ABS(H8*I9-H9*I8)-(J10/2))^0.5/(J8*J9*H10*I10)		
14	11	74	73	=B14	52		P(χ2 >=H13)=	=CHISQ.DIST.RT(H13,1)	= p	
15	12	76	80	=B15	55		*Since, P is more*	*than α (0.01),*	*accept Ho*	
16	13	87	90	=B16	56		Inference:			
17	14	92	88	=B17	57		*Two samples are from*	*the populations with the same*	*median.*	
18	15	86	83	=B18	61					
19	16	90		=B19	63	<Median				
20				=C4	64					
21				=C5	67					
22				=C6	72					
23	Total number of observations, n =		=COUNT(E4:E34)	=C7	73					
24				=C8	74					
25	n/2 =		=C23/2	=C9	75					
26				=C10	76					
27				=C11	80					
28				=C12	83					
29				=C13	86					
30				=C14	87					
31				=C15	88					
32				=C16	90					
33				=C17	90					
34				=C18	92					

Figure 12.30 Screenshot of guidelines for the formulas of the working of Example 12.16

12.7 Mann-Whitney U Test Using Excel Sheets and NORM.S.DIST Function

The Mann-Whitney U test is an alternative to the two-sample *t* test, and it is powerful. This test is based on the ranks of the combined observations of the samples. This is also called a rank-sum test.

Let
n_1 be the size of sample 1
n_2 be the size of sample 2

$$N = n_1 + n_2$$

The null and alternate hypotheses of this test are as follows.
Test 1 (Less-than type)

H_0: The two samples are drawn from different populations with the same distribution.
H_1: Population 1 is stochastically less than population 2.

Test 2 (Greater-than type)

H_0: The two samples are drawn from different populations with the same distribution.
H_1: Population 1 is stochastically larger than population 2.

Test 3 (Not-equal-to type)

H_0: The two samples are drawn from different populations with the same distribution.
H_1: Population 1 stochastically differs from Population 2.

The rank of the combined observations of the two samples are obtained first. Later, the sum of the ranks of the observations of sample 1 (R_1) and that of the observations of sample 2 (R_2) is obtained.
Then the value of *U* is obtained using the following formula.

$$U = n_1 n_2 + \left[\frac{n_1(n_1+1)}{2}\right] - R_1$$

$$Mean\,\mu_U = \frac{n_1 n_2}{2}$$

$$Variance\,\sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

This *U* follows a normal distribution with a mean of μ_U and variance of σ^2_U.

That is, $U \sim N(\mu_U, \sigma^2_{U)}$

The standard normal statistic of U is as follows.

$$Z_U = \frac{(U - \mu_U)}{\sigma_U}$$

The steps of the Mann-Whitney U test are as follows.

Step 1: Input the following.

Size of sample 1 (n_1)
Size of sample 2 (n_2)
Observations of sample 1 (A_i, i = 1,2, 3, . . ., n_1)
Observations of sample 2 (B_i, i = 1, 2, 3, . . ., n_2)

Step 2: $N = n_1 + n_2$.
Step 3: Combine the observations of the two samples and sort them from low to high.
Step 4: Assign the ranks to the sorted observations from low to high from top to bottom of the sorted column.
Step 5: If an observation, say, k in the sorted column, repeats q times, then sum up the ranks of those repeated observations and find the average rank of the repeated observations by dividing that sum by q. Then assign the average rank to those repeated observations.
Step 6: Find the sum of the ranks of the observations of sample 1, and let it be R_1
Find the sum of the ranks of the observations of sample 2, and let it be $R2$.
Step 7: Find U using the following formula.

$$U = n_1 n_2 + \left[\frac{n_1(n_1 + 1)}{2}\right] - R_1$$

Step 8: Find the mean (μ_U) and the variance (σ^2_U) using the following formulas.

$$Mean\, \mu_U = \frac{n_1 n_2}{2}$$

$$Variance\, \sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Step 9: Find the standard normal statistic Z_U using the following formula.

$$Z_U = \frac{(U - \mu_U)}{\sigma_U}$$

Step 10: Find the probability that $P\left(Z_U \le \frac{(U - \mu_U)}{\sigma_U}\right)$.

a. If Z_U is negative, then $P\left(Z_U \le \frac{(U - \mu_U)}{\sigma_U}\right)$ gives the p value at the left tail of the standard normal distribution.

b. If Z_U is positive, then $1 - P\left(Z_U \le \frac{(U - \mu_U)}{\sigma_U}\right)$ gives the p value at the right tail of the standard normal distribution.

For test 1, if the p value obtained in the Step 10.b is less than the given significance level, then reject the null hypothesis; otherwise, accept the null hypothesis.

For test 2, if the p value obtained in the Step 10.a is less than the given significance level, then reject the null hypothesis; otherwise, accept the null hypothesis.

For test 3, if the p value obtained in the Step 10.a when Z_U is negative or if the p value obtained in the Step 10.b when Z_U is positive is less than half of the significance level, then reject the null hypothesis; otherwise, accept the null hypothesis.

Example 12.17

An investigator is conducting a survey and has designated two enumerators (enumerator 1 and enumerator 2). Table 12.23 summarises the number of randomly selected respondents who were enumerated by each of the enumerators on each of the days.

Check whether the two samples shown in Table 12.23 are drawn from identical populations against the alternate hypothesis that population 1 stochastically differs from population 2 using the Mann Whitney U test at a significance level of 0.01.

Solution

The data for Example 12.17 are shown in Table 12.24.

Let

n_1 be the size of sample 1
n_2 be the size of sample 2

$$N = n_1 + n_2$$

H_0: The two samples are drawn from different populations with the same distribution.
H_1: Population 1 stochastically differs from population 2.

Table 12.23 Number of Respondents Covered by Enumerators

Day	*Enumerator 1*	*Enumerator 2*
1	24	30
2	17	20
3	34	15
4	28	22
5	15	31
6	35	24
7	25	12
8	13	16
9	22	21
10	28	8
11	27	22
12	15	14
13	13	16
14	17	

Table 12.24 Data for Example 12.17

Day	*Enumerator 1*	*Enumerator 2*
1	24	30
2	17	20
3	34	15
4	28	22
5	15	31
6	35	24
7	25	12
8	13	16
9	22	21
10	28	8
11	27	22
12	15	14
13	13	16
14	17	

The rank of the combined observations of the two samples are obtained first. Later, the sum of the ranks of the observations of sample 1 (R_1) and that of the observations of sample 2 (R_2) is obtained.

Then the value of U is obtained using the following formula.

$$U = n_1 n_2 + \left[\frac{n_1(n_1+1)}{2}\right] - R_1$$

$$Mean\, \mu_U = \frac{n_1 n_2}{2}$$

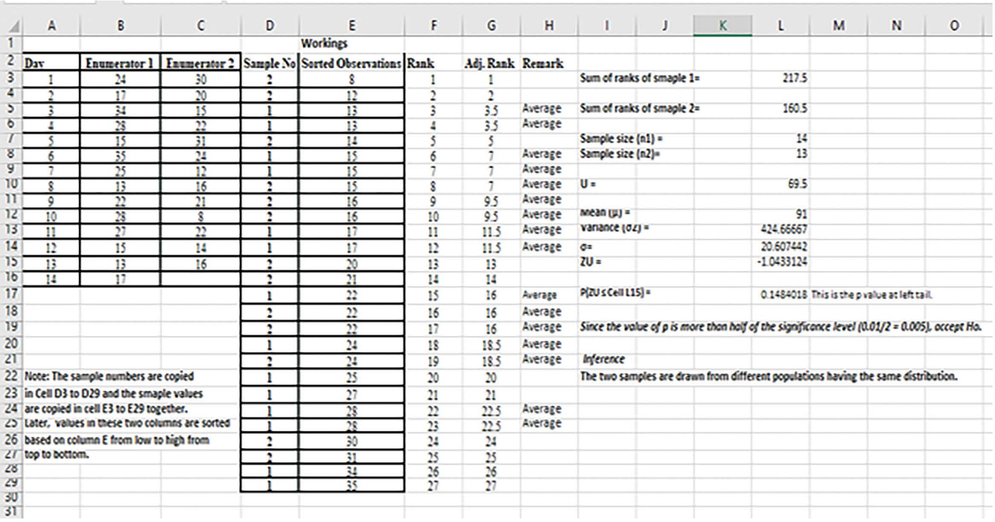
$$Variance\, \sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

The significance level (α) = 0.01
This U follows a normal distribution with a mean of μ_U and variance of $\sigma^2{}_U$.
That is, $U \sim N(\mu_U, \sigma^2{}_{U)}$
The standard normal statistic of U is as follows.

$$Z_U = \frac{(U - \mu_U)}{\sigma_U}$$

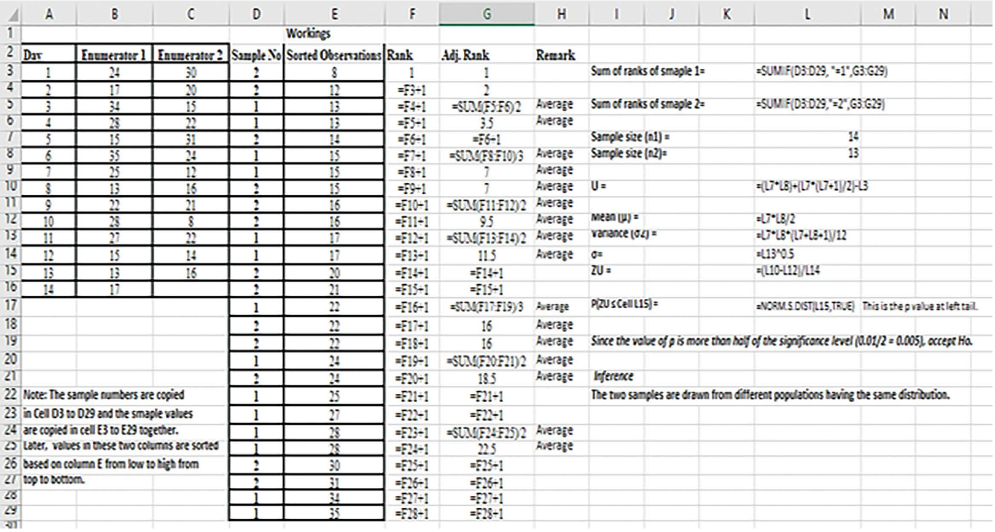
$$P\left(Z_U \le \frac{(U-\mu_U)}{\sigma_U}\right) = ?$$

All the calculations are shown in Figure 12.31 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.31 are shown in Figure 12.32.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1					Workings								
2	Dav	Enumerator 1	Enumerator 2	Sample No	Sorted Observations	Rank	Adj. Rank	Remark					
3	1	24	30	2	8	1	1		Sum of ranks of smaple 1=			217.5	
4	2	17	20	2	12	2	2						
5	3	34	15	1	13	3	3.5	Average	Sum of ranks of smaple 2=			160.5	
6	4	28	22	1	13	4	3.5	Average					
7	5	15	31	2	14	5	5		Sample size (n1) =			14	
8	6	35	24	1	15	6	7	Average	Sample size (n2)=			13	
9	7	25	12	1	15	7	7	Average					
10	8	13	16	2	15	8	7	Average	U =			69.5	
11	9	22	21	2	16	9	9.5	Average					
12	10	28	8	2	16	10	9.5	Average	Mean (μ) =			91	
13	11	27	22	1	17	11	11.5	Average	Variance (σ2) =			424.66667	
14	12	15	14	1	17	12	11.5	Average	σ=			20.607442	
15	13	13	16	2	20	13	13		ZU =			-1.0433124	
16	14	17		2	21	14	14						
17				1	22	15	16	Average	P(ZU ≤ Cell L15) =			0.1484018	This is the p value at left tail.
18				2	22	16	16	Average					
19				2	22	17	16	Average	*Since the value of p is more than half of the significance level (0.01/2 = 0.005), accept Ho.*				
20				1	24	18	18.5	Average					
21				2	24	19	18.5	Average	*Inference*				
22	Note: The sample numbers are copied			1	25	20	20		The two samples are drawn from different populations having the same distribution.				
23	in Cell D3 to D29 and the smaple values			1	27	21	21						
24	are copied in cell E3 to E29 together.			1	28	22	22.5	Average					
25	Later, values in these two columns are sorted			1	28	23	22.5	Average					
26	based on column E from low to high from			2	30	24	24						
27	top to bottom.			2	31	25	25						
28				1	34	26	26						
29				1	35	27	27						
30													
31													

Figure 12.31 Screenshot of the working of Example 12.17



	A	B	C	D	E	F	G	H	I	J	K	L	M
1					Workings								
2	Day	Enumerator 1	Enumerator 2	Sample No	Sorted Observations	Rank	Adj. Rank	Remark					
3	1	24	30	2	8	1	1		Sum of ranks of smaple 1=			=SUMIF(D3:D29, "=1",G3:G29)	
4	2	17	20	2	12	=F3+1	2						
5	3	34	15	1	13	=F4+1	=SUM(F5:F6)/2	Average	Sum of ranks of smaple 2=			=SUMIF(D3:D29,"=2",G3:G29)	
6	4	28	22	1	13	=F5+1	3.5	Average					
7	5	15	31	2	14	=F6+1	=F6+1		Sample size (n1) =			14	
8	6	35	24	1	15	=F7+1	=SUM(F8:F10)/3	Average	Sample size (n2)=			13	
9	7	25	12	1	15	=F8+1	7	Average					
10	8	13	16	2	15	=F9+1	7	Average	U =			=(L7*L8)+(L7*(L7+1)/2)-L3	
11	9	22	21	2	16	=F10+1	=SUM(F11:F12)/2	Average					
12	10	28	8	2	16	=F11+1	9.5	Average	Mean (μ) =			=L7*L8/2	
13	11	27	22	1	17	=F12+1	=SUM(F13:F14)/2	Average	Variance (σ2) =			=L7*L8*(L7+L8+1)/12	
14	12	15	14	1	17	=F13+1	11.5	Average	σ=			=L13^0.5	
15	13	13	16	2	20	=F14+1	=F14+1		ZU =			=(L10-L12)/L14	
16	14	17		2	21	=F15+1	=F15+1						
17				1	22	=F16+1	=SUM(F17:F19)/3	Average	P(ZU ≤ Cell L15) =			=NORM.S.DIST(L15,TRUE)	This is the p value at left tail.
18				2	22	=F17+1	16	Average					
19				2	22	=F18+1	16	Average	*Since the value of p is more than half of the significance level (0.01/2 = 0.005), accept Ho.*				
20				1	24	=F19+1	=SUM(F20:F21)/2	Average					
21				2	24	=F20+1	18.5	Average	*Inference*				
22	Note: The sample numbers are copied			1	25	=F21+1	=F21+1		The two samples are drawn from different populations having the same distribution.				
23	in Cell D3 to D29 and the smaple values			1	27	=F22+1	=F22+1						
24	are copied in cell E3 to E29 together.			1	28	=F23+1	=SUM(F24:F25)/2	Average					
25	Later, values in these two columns are sorted			1	28	=F24+1	22.5	Average					
26	based on column E from low to high from			2	30	=F25+1	=F25+1						
27	top to bottom.			2	31	=F26+1	=F26+1						
28				1	34	=F27+1	=F27+1						
29				1	35	=F28+1	=F28+1						

Figure 12.32 Screenshot of guidelines for the formulas of the working of Example 12.17

12.8 K-Sample Tests

In many real-life situations, investigators will be dealing with more than two samples, say, *K* samples, where *K* is 3 or above. In such a situation, the following tests can be used.

- *K*-sample median test
- Kruskal-Wallis test

These tests are presented in the following subsections.

12.8.1 K-Sample Median Test Using Excel Sheets, COUNTIF, COMBIN, and CHISQ. DIST.RT Functions

As stated earlier, if the number of samples is 3 or above, say, *K* samples, one can use the *K*-sample median test. In Section 12.6, the two-sample median test was presented. The *K*-sample median test is an extension of the two-sample median test.

The objective of the *K*-sample median test is to check whether the *K* samples, which are independent, have been drawn from *K* populations with the same median. The associated hypotheses are as follows.

H_0: The *K* samples, which are independent, are drawn from *K* populations with the same median.

H_1: The *K* samples, which are independent, are not drawn from *K* populations with the same median.

Steps of K-Sample Median Test

The steps that are required to perform the *K*-sample median test are as follows.

Step1: Input the following.

a. Number of samples *K*.
b. Size of the sample *j* (n_{j}), $j = 1, 2, 3, \ldots, K$.
c. Observations of the sample *j* $A(j,\mathrm{q})$, $j = 1, 2, 3, \ldots, K$ and $\mathrm{q} = 1, 2, 3, \ldots, n_{\mathrm{j.}}$

Step 2: Combine the observations of the *K* samples and sort them from low to high.
Step 3: Find the median of the combined sorted observations obtained in step 2.
Step 4: Find the frequency $o_{1\mathrm{j}}$ of the observations of sample *j* for the condition Observation > Median for $j = 1, 2, 3, \ldots, K$.
Step 5: Find the frequency $o_{2\mathrm{j}}$ of the observations of sample *j* for the condition Observation ≤ Median. for $j = 1, 2, 3, \ldots, K$.
Step 8: Summarise the results of the previous steps in the format shown in Table 12.25.
Step 9: If the size of the pooled sample is small (≤ 30), *p* is given by the following formula; otherwise, go to Step 10.

$$p = \frac{\left(n_{1_{C_{o11}}} \times n_{2_{C_{o12}}} \times n_{3_{C_{o13}}} \times \ldots.. \times n_{k_{C_{o1k}}}\right)}{(n_1 + n_{2+} n_3 + \cdots + n_k)_{C_{(o11+o12+o13+\ldots+o1k)}}}$$

Table 12.25 Intermediate Calculations of *K*-Sample Median Test

	Sample											*Row total*
	1	*2*	*3*	.	.	.	*j*	.	.	.	*K*	
Above median	o_{11}	o_{12}	o_{13}	.	.	.	$o_{1\mathrm{j}}$	.	.	.	o_{1K}	K $\Sigma\, o_{1j}$ $j=1$
Below median	o_{21}	o_{22}	o_{23}	.	.	.	o_{2j}	.	.	.	o_{2K}	K $\Sigma\, o_{2j}\, j=1$
Column total	n_1	n_2	n_3	.	.	.	n_j	.	.	.	n_K	K $\mathrm{N} = \Sigma\, n_j$ $j=1$

where n_j is the size of sample j, $j = 1, 2, 3, \ldots, K$

o_{1j} is the frequency of the observations of sample j for the condition Observation > Median for $j = 1, 2, 3, \ldots, K$

o_{2j} is the frequency of the observations of sample j for the condition Observation ≤ Median for $j = 1, 2, 3, \ldots, K$

If the value of p is less than the given significance level, then reject the null hypothesis; otherwise, accept the null hypothesis.

Step 10: If the size of the pooled sample is large (>30), then use the following formula to obtain the χ^2 statistic.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{K}\left(\frac{o_{ij} - e_{ij}}{e_{ij}}\right)^2$$

where

o_{1j} is the frequency of the observations of sample j for the condition Observation > Median for $j = 1, 2, 3, \ldots, K$

o_{2j} is the frequency of the observations of sample j for the condition Observation ≤ Median for $j = 1, 2, 3, \ldots, K$

e_{ij} is the expected frequency of row i and column j of Table 12.25, which is computed using the following formula

$$e_{ij} = \frac{Row\,i\,Total \times Column\,j\,Total}{N} = \frac{\sum_{j=1}^{K} o_{.j} \sum_{i=1}^{2} o_{i.}}{N}$$

If the χ^2 calculated is more than the table χ^2 with $(K - 1)$ degrees of freedom and a significance level of α, then reject the null hypothesis; otherwise, accept the null hypothesis.

Or:

Find P(χ^2 ≥Computed χ^2 value), and if it is more than the given significance level α, then reject H_0; otherwise, accept H_0.

The formula for this probability, which is the value of p at the right tail, is as follows.

$$= \text{CHISQ.DIST.RT(X, Degrees of freedom)}$$

Example 12.18

In a company, the capacity utilisations (in %) of each of the months of the first quarter for the past seven years are summarised in Table 12.26. Check whether there are

Table 12.26 Capacity Utilisations of Three Months of First Quarter

Year	*January*	*February*	*March*
1	90	80	70
2	88	75	95
3	90	85	70
4	60	78	50
5	80	90	69
6	65	74	68
7	85	90	86

significant differences among the capacity utilisations (in %) of the three months of the first quarter at a significance level of 0.05.

Solution

The data for Example 12.18 are shown in Table 12.27.

The populations of this example are as follows.

1. Capacity utilisations (in %) during January
2. Capacity utilisations (in %) during February
3. Capacity utilisations (in %) during March

The size of sample 1 (n_1) from population 1, the size of sample 2(n_2) from population 2, and the size of sample 3 (n_3) from population 3 are 7, 7, and 7, respectively.

The size of the combined population ($n_1 + n_{2\,+}\, n_3$) is 21.

The hypotheses of this example are as follows.

H_0: The three samples, which are independent, are drawn from three populations with the same median.

H_1: The three samples, which are independent, are not drawn from three populations with the same median.

Since the size of the combined sample is less than 30, the formula to compute the value of p is as follows.

$$p = \frac{\left(n_{1_{C_{o11}}} \times n_{2_{C_{o12}}} \times n_{3_{C_{o13}}} \times \ldots.. \times n_{k_{C_{o1k}}}\right)}{(n_1 + n_{2+} n_3 + \cdots + n_k)_{C_{(o11+o12+o13+\ldots+o1k)}}}$$

where n_i is the size of sample j, j = 1, 2, 3, . . . , K

o_{1j} is the frequency of the observations of the sample i for the condition Observation > Median for j = 1, 2, 3, . . ., K

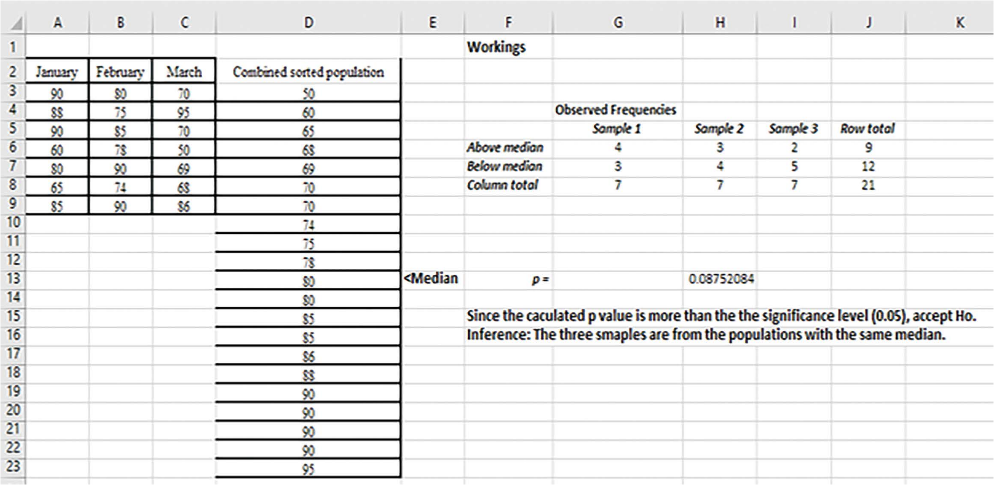
o_{2j} is the frequency of the observations of the sample i for the condition Observation ≤ Median for j = 1, 2, 3, . . ., K

If the value of p is less than the given significance level, then reject the null hypothesis; otherwise, accept the null hypothesis.

All the calculations are shown in Figure 12.33 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.33 are shown in Figure 12.34.

Table 12.27 Data for Example 12.18

Year	*January*	*February*	*March*
1	90	80	70
2	88	75	95
3	90	85	70
4	60	78	50
5	80	90	69
6	65	74	68
7	85	90	86



January	February	March	Combined sorted population	
90	80	70	50	
88	75	95	60	
90	85	70	65	
60	78	50	68	
80	90	69	69	
65	74	68	70	
85	90	86	70	
			74	
			75	
			78	
			80	<Median
			80	
			85	
			85	
			86	
			88	
			90	
			90	
			90	
			90	
			95	

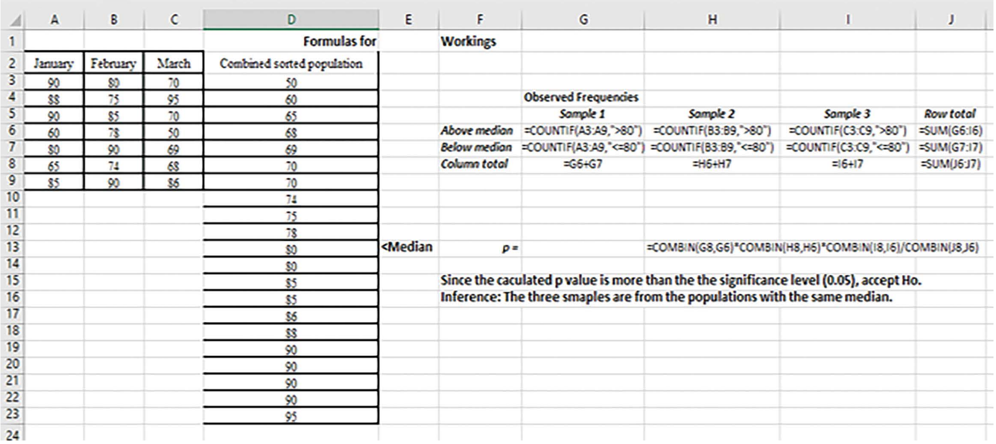
Workings

	Observed Frequencies			
	Sample 1	*Sample 2*	*Sample 3*	*Row total*
Above median	4	3	2	9
Below median	3	4	5	12
Column total	7	7	7	21

p = 0.08752084

Since the caculated p value is more than the the significance level (0.05), accept Ho.
Inference: The three smaples are from the populations with the same median.

Figure 12.33 Screenshot of the working of Example 12.18



January	February	March	Formulas for Combined sorted population	
90	80	70	50	
88	75	95	60	
90	85	70	65	
60	78	50	68	
80	90	69	69	
65	74	68	70	
85	90	86	70	
			74	
			75	
			78	
			80	<Median
			80	
			85	
			85	
			86	
			88	
			90	
			90	
			90	
			90	
			95	

Workings

	Observed Frequencies			
	Sample 1	*Sample 2*	*Sample 3*	*Row total*
Above median	=COUNTIF(A3:A9,">80")	=COUNTIF(B3:B9,">80")	=COUNTIF(C3:C9,">80")	=SUM(G6:I6)
Below median	=COUNTIF(A3:A9,"<=80")	=COUNTIF(B3:B9,"<=80")	=COUNTIF(C3:C9,"<=80")	=SUM(G7:I7)
Column total	=G6+G7	=H6+H7	=I6+I7	=SUM(J6:J7)

p = =COMBIN(G8,G6)*COMBIN(H8,H6)*COMBIN(I8,I6)/COMBIN(J8,J6)

Since the caculated p value is more than the the significance level (0.05), accept Ho.
Inference: The three smaples are from the populations with the same median.

Figure 12.34 Screenshot of guidelines for the formulas of the working of Example 12.18

Example 12.19

In order to improve daily sales, a marketing manager tested three sales promotion strategies in 11 shops: price discount, extended warranty, and door delivery. She also gathered data on the increase in daily sales in thousands of rupees. Table 12.28 provides an overview of these. Verify that the three separate samples were taken from populations with the same median at the significance level of 0.01.

Solution

The data for Example 12.19 are shown in Table 12.29.

Table 12.28 Data for Example 12.19 on Increase in Daily Sales (Thousands of Rupees)

Retail Shop	*Strategy*		
	Price Discount	*Extended Warranty*	*Door Delivery*
1	55	63	35
2	43	39	40
3	45	44	56
4	51	64	50
5	61	67	75
6	74	73	40
7	76	80	42
8	87	90	52
9	92	88	57
10	86	83	72
11	85	84	71

Table 12.29 Data for Example 12.19

Retail Shop	*Strategy*		
	Price Discount	*Extended Warranty*	*Door Delivery*
1	55	63	35
2	43	39	40
3	45	44	56
4	51	64	50
5	61	67	75
6	74	73	40
7	76	80	42
8	87	90	52
9	92	88	57
10	86	83	72
11	85	84	71

The populations of this example are as follows.

1. Increased sales due to price discount
2. Increased sales due to extended warranty
3. Increased sales due to door delivery

The size of sample 1 (n_1) from population 1, the size of sample 2(n_2) from population 2, and the size of sample 3 (n_3) from population 3 are 11, 11, and 11, respectively.

The size of the combined populations ($n_1 + n_2 + n_3$) is 33.

The hypotheses of this example are as follows.

H_0: The three samples, which are independent, are drawn from three populations with the same median.

H_1: The three samples, which are independent, are not drawn from three populations with the same median.

Since the size of the combined sample is 33, the formula to compute the value of χ^2 is as follows.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\left(\frac{o_{ij} - e_{ij}}{e_{ij}}\right)^2$$

where

o_{1j} is the frequency of the observations of sample j for the condition Observation > Median for $j = 1, 2, 3$

o_{2j} is the frequency of the observations of sample j for the condition Observation ≤ Median for $j = 1, 2, 3$

e_{ij} is the expected frequency of row i and column j, which is computed using the following formula

$$e_{ij} = \frac{\sum_{j=1}^{3} o_{ij} \sum_{i=1}^{2} o_{ij}}{N}$$

Find $P(\chi^2 \geq \text{Computed } \chi^2 \text{ value})$ and if it is less than the given significance level α, then reject H_0; otherwise, accept H_0.

The formula to compute the probability, which is the value of p at the right tail, is as follows.

$$= \text{CHISQ.DIST.RT}(\text{X, Degrees of freedom})$$

All the calculations are shown in Figure 12.35 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.35 are shown in Figure 12.36

12.8.2 Kruskal–Wallis Test (H Test) Using Excel Sheets

The Kruskal-Wallis test is like the K-sample median test in which the objective is to test whether K samples are from K identical populations. This test is alternatively called the H test, and the statistic that is computed for this test is denoted by H. This test is an alternative approach for ANOVA with a single factor.

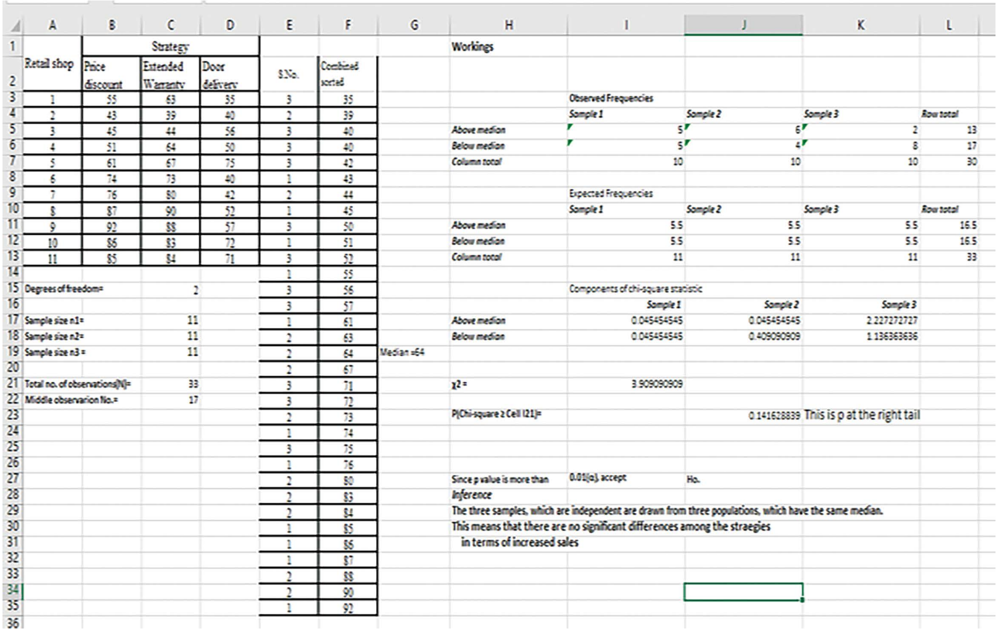
The hypotheses of the H test are listed as follows.

H_0: K samples, which are independent, are drawn from K identical populations.
H_1: K samples, which are independent, are not drawn from K identical populations.

The steps of H test are as follows.

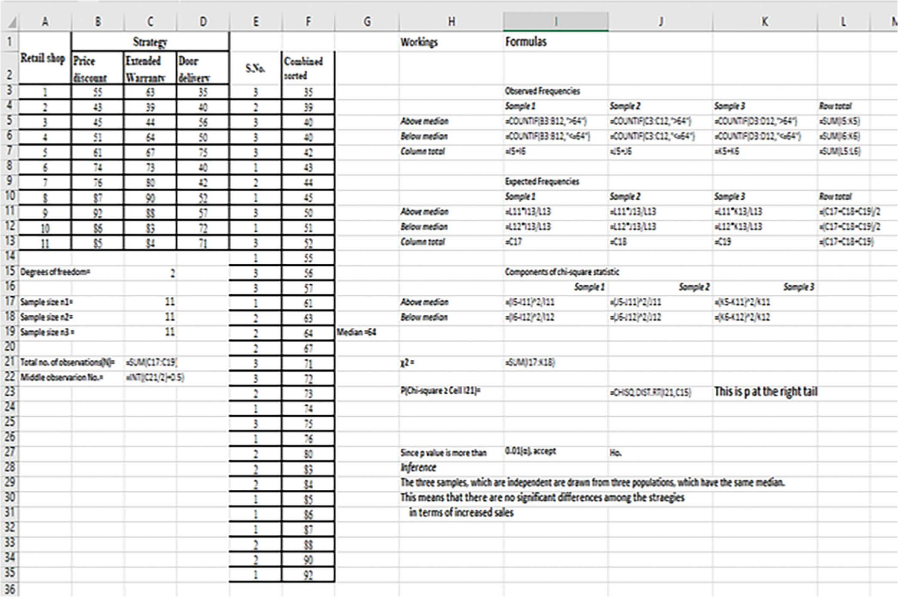
Step 1: Input the following.

a. Number of samples, K.
b. Size of the sample i (n_i), $i = 1, 2, 3, \ldots, K$.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Retail shop	Strategy						Workings				
2		Price discount	Extended Warranty	Door delivery	S.No.	Combined sorted						
3	1	55	63	35	3	35			Observed Frequencies			
4	2	43	39	40	2	39			*Sample 1*	*Sample 2*	*Sample 3*	*Row total*
5	3	45	44	56	3	40		*Above median*	5	6	2	13
6	4	51	64	50	3	40		*Below median*	5	4	8	17
7	5	61	67	75	3	42		*Column total*	10	10	10	30
8	6	74	73	40	1	43						
9	7	76	80	42	2	44			Expected Frequencies			
10	8	87	90	52	1	45			*Sample 1*	*Sample 2*	*Sample 3*	*Row total*
11	9	92	88	57	3	50		*Above median*	5.5	5.5	5.5	16.5
12	10	86	83	72	1	51		*Below median*	5.5	5.5	5.5	16.5
13	11	85	84	71	3	52		*Column total*	11	11	11	33
14					1	55						
15	Degrees of freedom=		2		3	56			Components of chi-square statistic			
16					3	57			*Sample 1*	*Sample 2*	*Sample 3*	
17	Sample size n1=		11		1	61		*Above median*	0.045454545	0.045454545	2.227272727	
18	Sample size n2=		11		2	63		*Below median*	0.045454545	0.409090909	1.136363636	
19	Sample size n3 =		11		2	64	Median =64					
20					2	67						
21	Total no. of observations(N)=		33		3	71		χ2 =	3.909090909			
22	Middle observarion No.=		17		3	72						
23					2	73		P(Chi-square ≥ Cell I21)=		0.141628839	This is p at the right tail	
24					1	74						
25					3	75						
26					1	76						
27					2	80		Since p value is more than	0.01(α), accept	Ho.		
28					2	83		*Inference*				
29					2	84		The three samples, which are independent are drawn from three populations, which have the same median.				
30					1	85		This means that there are no significant differences among the straegies				
31					1	86		in terms of increased sales				
32					1	87						
33					2	88						
34					2	90						
35					1	92						
36												

Figure 12.35 Screenshot of the working of Example 12.19



	A	B	C	D	E	F	G	H	I	J	K	L
1	Retail shop	Strategy						Workings	Formulas			
2		Price discount	Extended Warranty	Door delivery	S.No.	Combined sorted						
3	1	55	63	35	3	35			Observed Frequencies			
4	2	43	39	40	2	39			*Sample 1*	*Sample 2*	*Sample 3*	*Row total*
5	3	45	44	56	3	40		*Above median*	=COUNTIF(B3:B12,">64")	=COUNTIF(C3:C12,">64")	=COUNTIF(D3:D12,">64")	=SUM(I5:K5)
6	4	51	64	50	3	40		*Below median*	=COUNTIF(B3:B12,"<=64")	=COUNTIF(C3:C12,"<=64")	=COUNTIF(D3:D12,"<=64")	=SUM(I6:K6)
7	5	61	67	75	3	42		*Column total*	=I5+I6	=J5+J6	=K5+K6	=SUM(L5:L6)
8	6	74	73	40	1	43						
9	7	76	80	42	2	44			Expected Frequencies			
10	8	87	90	52	1	45			*Sample 1*	*Sample 2*	*Sample 3*	*Row total*
11	9	92	88	57	3	50		*Above median*	=L11*I13/L13	=L11*J13/L13	=L11*K13/L13	=(C17+C18+C19)/2
12	10	86	83	72	1	51		*Below median*	=L12*I13/L13	=L12*J13/L13	=L12*K13/L13	=(C17+C18+C19)/2
13	11	85	84	71	3	52		*Column total*	=C17	=C18	=C19	=(C17+C18+C19)
14					1	55						
15	Degrees of freedom=		2		3	56			Components of chi-square statistic			
16					3	57			*Sample 1*	*Sample 2*	*Sample 3*	
17	Sample size n1=		11		1	61		*Above median*	=(I5-I11)^2/I11	=(J5-J11)^2/J11	=(K5-K11)^2/K11	
18	Sample size n2=		11		2	63		*Below median*	=(I6-I12)^2/I12	=(J6-J12)^2/J12	=(K6-K12)^2/K12	
19	Sample size n3 =		11		2	64	Median =64					
20					2	67						
21	Total no. of observations(N)=	=SUM(C17:C19)			3	71		χ2 =	=SUM(I17:K18)			
22	Middle observarion No.=	=INT((C21/2)+0.5)			3	72						
23					2	73		P(Chi-square ≥ Cell I21)=		=CHISQ.DIST.RT(I21,C15)	This is p at the right tail	
24					1	74						
25					3	75						
26					1	76						
27					2	80		Since p value is more than	0.01(α), accept	Ho.		
28					2	83		*Inference*				
29					2	84		The three samples, which are independent are drawn from three populations, which have the same median.				
30					1	85		This means that there are no significant differences among the straegies				
31					1	86		in terms of increased sales				
32					1	87						
33					2	88						
34					2	90						
35					1	92						
36												

Figure 12.36 Screenshot of guidelines for the formulas of the working of Example 12.19

c. Observations of the sample i $A(i,j)$, $i = 1, 2, 3, \ldots, K$ and $j = 1, 2, 3, \ldots, n_i$.

Step 2: Combine the observations of the K samples and sort them from low to high.
Step 3: $N = n_1 + n_{2\,+}\, \mathrm{n}_{3\,+\,.\,.\,.\,.\,.\,.\,+}\, n_K$.

Step 4: Assign ranks to the sorted observations from low to high from top to bottom of the sorted column.

Step 5: If an observation, say, *k* in the sorted column, repeats *q* times, then sum up the ranks of those repeated observations and find the average rank of the repeated observations by dividing that sum by *q*. Then assign the average rank to those repeated observations.

Step 6: Find the sum of the ranks of the observations of sample *i*, and let it be R_i for $i = 1, 2, 3, \ldots, K$.

Step 7: Find *H* using the following formula.

$$H = \frac{12}{N(N+1)} \times \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(N+1)$$

This *H* value equals a χ^2 variable with $(K - 1)$ degrees of freedom.

Step 8: Find $P(\chi^2 \geq H)$ with $(K - 1)$ degrees of freedom. This is the *p* value [4].

Step 9: If the computed *p* value is less than the significance level α, then reject H_0; otherwise, accept H_0.

Example 12.20

A stock market analyst is interested in examining how the type of the company affects its earnings per share (EPS). He therefore gathered EPS data from a secondary source for the previous four years of four different companies, which is summarised in Table 12.30. Check whether there are significant differences among the companies in terms of yearly EPS data at a significance level of 0.05.

Solution

The data for Example 12.20 are shown in Table 12.31.

Number of samples is 4

Size of each of the samples is 4 ($n_1 = n_2 = n_2 = n_4$)

$N = n_1 + n_{2\,+}\, n_{3\,+}\, n_4 = 4 + 4 + 4 + 4 = 16$

Significance level (α) = 0.05

H_0: The four samples, which are independent, are drawn from four identical populations.

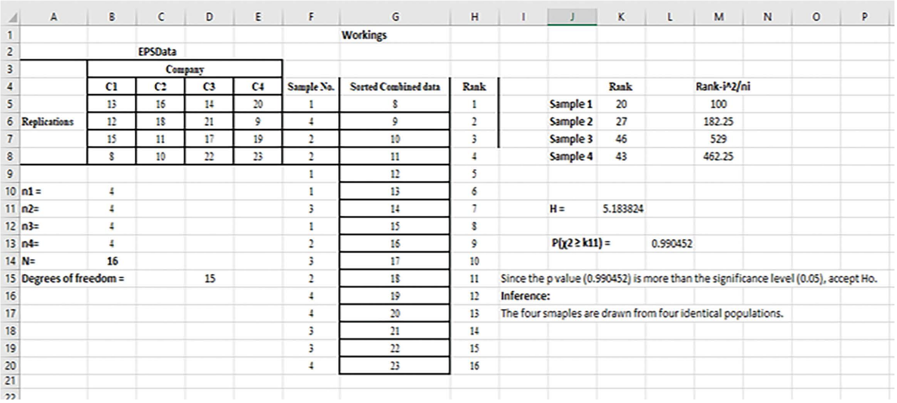
H_1: The four samples, which are independent, are not drawn from four identical populations.

Table 12.30 Yearly EPS Data for Companies

Replications	*Company*			
	C1	*C2*	*C3*	*C4*
	12	16	9	13
	8	18	22	8
	15	11	15	5
	18	10	25	20

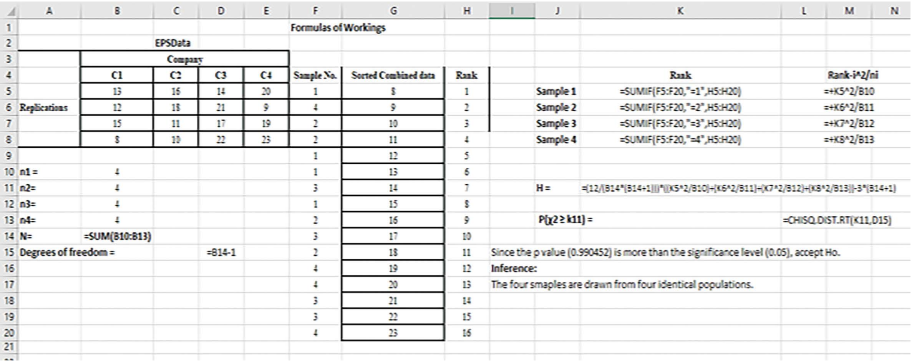
Table 12.31 Data for Example 12.20

	Company			
	C1	*C2*	*C3*	*C4*
Replications	12	16	9	13
	8	18	22	8
	15	11	15	5
	18	10	25	20



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1							Workings									
2			EPSData													
3		Company														
4		C1	C2	C3	C4	Sample No.	Sorted Combined data	Rank			Rank		Rank-i^2/ni			
5		13	16	14	20	1	8	1		Sample 1	20		100			
6	Replications	12	18	21	9	4	9	2		Sample 2	27		182.25			
7		15	11	17	19	2	10	3		Sample 3	46		529			
8		8	10	22	23	2	11	4		Sample 4	43		462.25			
9						1	12	5								
10	n1 =	4				1	13	6								
11	n2=	4				3	14	7		H =	5.183824					
12	n3=	4				1	15	8								
13	n4=	4				2	16	9		P(χ2 ≥ k11) =		0.990452				
14	N=	16				3	17	10								
15	Degrees of freedom =			15		2	18	11	Since the p value (0.990452) is more than the significance level (0.05), accept Ho.							
16						4	19	12	Inference:							
17						4	20	13	The four smaples are drawn from four identical populations.							
18						3	21	14								
19						3	22	15								
20						4	23	16								
21																

Figure 12.37 Screenshot of the working of Example 12.20



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1						Formulas of Workings								
2			EPSData											
3		Company												
4		C1	C2	C3	C4	Sample No.	Sorted Combined data	Rank			Rank		Rank-i^2/ni	
5		13	16	14	20	1	8	1		Sample 1	=SUMIF(F5:F20,"=1",H5:H20)		=+K5^2/B10	
6	Replications	12	18	21	9	4	9	2		Sample 2	=SUMIF(F5:F20,"=2",H5:H20)		=+K6^2/B11	
7		15	11	17	19	2	10	3		Sample 3	=SUMIF(F5:F20,"=3",H5:H20)		=+K7^2/B12	
8		8	10	22	23	2	11	4		Sample 4	=SUMIF(F5:F20,"=4",H5:H20)		=+K8^2/B13	
9						1	12	5						
10	n1 =	4				1	13	6						
11	n2=	4				3	14	7		H =	=(12/(B14*(B14+1)))*((K5^2/B10)+(K6^2/B11)+(K7^2/B12)+(K8^2/B13))-3*(B14+1)			
12	n3=	4				1	15	8						
13	n4=	4				2	16	9		P(χ2 ≥ k11) =		=CHISQ.DIST.RT(K11,D15)		
14	N=	=SUM(B10:B13)				3	17	10						
15	Degrees of freedom =			=B14-1		2	18	11	Since the p value (0.990452) is more than the significance level (0.05), accept Ho.					
16						4	19	12	Inference:					
17						4	20	13	The four smaples are drawn from four identical populations.					
18						3	21	14						
19						3	22	15						
20						4	23	16						
21														

Figure 12.38 Screenshot of guidelines for the formulas of the working of Example 12.20

The formula for H is presented as follows.

$$H = \frac{12}{N(N+1)} \times \sum_{i=1}^{4} \frac{R_i^2}{n_i} - 3(N+1)$$

Where

n_i is the size of sample i, $i = 1, 2, 3, 4$

$N = n_1 + n_{2\,+}\, n_{3\,+}\, n_4$

R_i is the sum of the ranks of the observations of sample i, $i = 1, 2, 3, 4$

This *H* value equals a χ^2 variable with (4 – 1) degrees of freedom.

All the calculations are shown in Figure 12.37 along with the inference of the hypothesis testing. The formulas for the calculations in Figure 12.37 are shown in Figure 12.38.

Summary

- In a non-parametric test, the parameters of the distribution are not required.
- In a one-sample sign test, the probability that a sample value is more than the mean value (p) is 1/2.
- In a one-tailed one-sample sign test when the sample size is small, a small random sample is taken from a non-normal population, and then it is tested against a median value (μ) such that the observations in the sample are more than that median (μ) or less than that median (μ) at a significance level of α using a binomial distribution.
- In two-tailed one-sample sign test when the sample size is small, a small random sample is taken from a non-normal population, and then it is tested against a median value (μ) such that the observations in the sample are not equal to that median (μ) at a significance level α using a binomial distribution.
- The Kolmogorov-Smirnov (K-S) test is an alternative to the χ^2 test. It is a one-tailed test for a small sample, whereas the χ^2 test is also a one-tailed test, but for large samples.
- The stream of data that is collected in a system may have certain patterns called runs.
- In a run test, if the value of n_1 as well as that of n_2 is less than 20, then the sample is regarded as a small sample.
- In run test, if the value of n_1 or n_2 or both is/are more than 20, then the sample is regarded as a large sample.
- In one-tailed/two-tailed sample sign test for small samples, for a sample size less than 20, the number of plus signs (positive difference (+) between the i^{th} observation taken before and the i^{th} observation taken after introducing a change in the system) follows a binomial distribution.
- In the two-sample sign test for large samples, two random samples, each with size n with the condition that np as well as $n\ (1 - p)$ is greater than or equal to 5, are selected.
- In the two-tailed two-sample sign test for large samples, when the sample size is greater than or equal to 20, then the number of plus signs (sum of the positive difference (+) between the i^{th} observation taken before and the i^{th} observation taken after introducing a change in the system) follows a binomial distribution.
- The objective of the median test is to check whether the two samples, which are independent, have been drawn from two populations with the same median.
- The Mann-Whitney U test is an alternative to the two-sample *t* test, and it is powerful. It is also known as a rank-sum test.
- If the number of samples is 3 or above, say *K* samples, one can use the *K*-sample median test.
- The objective of the *K*-sample median test is to check whether the *K* samples, which are independent, have been drawn from *K* populations with the same median.
- The Kruskal-Wallis (*H*) test is like the *K*-sample median test in which the objective is to test whether the *K* samples are from *K* identical populations. This test is an alternative approach for ANOVA with a single factor.

Keywords

- A parametric test is applied to data which have the estimate(s) of parameter(s).
- A non-parametric test is applied to data which do not have the estimate(s) of parameter(s).
- A one-sample sign test assumes the number of plus signs as the value of the random variable X of the binomial distribution, with $p = 1/2$ and the number of trials n to compute the probability that X is more than the number of plus signs to check different hypotheses.
- The one-tailed one-sample sign test when the sample size is small assumes a small random sample from a non-normal population, and then it is tested against a median value (μ) such that the observations in the sample are more than that median (μ) or less than that median (μ) at a significance level of α using a binomial distribution.
- The two-tailed one-sample sign test when the sample size is small assumes a small random sample taken from a non-normal population, and then it is tested against a median value (µ) such that the observations in the sample are not equal to that median (μ) at a significance level α using a binomial distribution.
- A large sample size in a one-tailed/two-tailed sign test considers a random sample of n units with the condition that np as well as $n\ (1 - p)$ is greater than or equal to 5.
- The Kolmogorov-Smirnov (K-S) test is an alternative to the χ^2 test.
- A run means a stream of data that is collected in a system with certain patterns.
- A large sample in a run test means that the value of n_1 or n_2 or both is/are more than 20.
- Large samples in a two-sample sign test consider two random samples, each with size n, with the condition that np as well as $n\ (1 - p)$ is greater than or equal to 5.
- The median test checks whether two samples, which are independent, have been drawn from two populations with the same median.
- The Mann-Whitney U test is an alternative to the two-sample t test, and it is powerful. It is also known as a rank-sum test.
- A K-sample median test is used if the number of samples is 3 or above, say, K samples.
- The Kruskal-Wallis (H) test is like the K-sample median test in which the objective is to test whether the K samples are drawn from K identical populations. This test is an alternative approach for ANOVA with a single factor.

Review Questions

1. Distinguish between parametric and nonparametric tests.
2. List different non-parametric tests.
3. List the pairs of hypotheses for one-tailed one-sample sign test when the sample size is small, with suitable examples.
4. At the end of a biscuit production and packing line, the quality assistant of a biscuit manufacturing unit gathered data on the weight of a biscuit packet in kg. Ten observations total, 51, 54, 48, 53, 55, 46, 56, 54, 49, and 52, were made. Using a sign test with a significance level of 0.10, determine whether the weight of the biscuit packet is 50 gm (H_0: = 50) in comparison to the alternative hypothesis H_1: > 50 through Excel.
5. Ten pharmaceutical companies' annual revenues (in crores of rupees) are 174, 160, 195, 185, 205, 172, 195, 188, 160, and 200. Check whether the annual revenue of the pharma companies is 175 lakhs (H_0: μ = 175 crores) against the alternate hypothesis H_1:$\mu \neq$ 175 crores using a sign test with a significance level of 0.05 through Excel.

6. List the pairs of hypotheses for a one-tailed one-sample sign test when the sample size is large, with suitable examples.
7. The weight of medicine produced by Beta Pharmaceutical Company in capsule form exhibits a non-normal distribution. For this population, the prescribed weight of the medicine in the capsule is 100 mg. Any extra medicine in the capsule will have adverse effects. The company's quality engineer asserts that the weight of the capsule is no more than 100 mg. As a result, the purchase manager of Gamma Hospital, who places an order with Beta Pharmaceutical Company for that capsule, has chosen a random sample of 32 capsules and determined their weights as follows.

104	98	102	101	98	107	100	106
97	107	104	104	103	102	103	106
107	106	106	96	103	102	106	95
102	104	106	108	99	106	107	108

Check the claim of the quality engineer of the Beta Pharmaceutical Company that the weight of the capsule is 100 mg ($H_0 = 100$) against the alternate hypothesis $H_1 > 100$ using a sign test with a significance level of 0.05 using Excel.

8. The percentage of marks of the students of a degree program in a college follows a non-normal distribution. A researcher feels that the percentage of the marks of the students in the degree is more than 75%. To test his intuition, he has selected a random sample of 36 students, and their mark percentages are shown as follows.

97	70	76	74	60	66	72	85	77
79	74	74	74	74	82	98	70	71
81	85	78	71	89	83	73	86	72
74	70	70	75	74	73	80	71	74

Check the intuition of the researcher using a sign test at a significance level of 0.05 using Excel.

9. Discuss the essentials of the Kolmogorov-Smirnov test.
10. The arrival rate of customers (number of customers per 15-minute interval) at a leading mall appears to follow a Poisson distribution. The observed frequencies are given in the following table.

i	*Arrival Rate* (X_i)	*Observed Frequency* (O_i)
1	0	3
2	1	5
3	2	12
4	3	9
5	4	7
6	5	4
7	6	3
8	7	1

Check whether the given set of data follows a Poisson distribution using the Kolmogorov-Smirnov test at a significance level of 0.05 through Excel.

11. Explain the essentials of the run test for randomness of data.
12. Discuss the essentials for a run test of a small sample.

13. The vehicles arriving at a toll gate are heavy (H) and light (L). The stream of heavy and light vehicles passing the toll gate is as follows.
HHLLLHHLLLHHHHLLHHHLL
Check whether the occurrence of the substrings in terms of H and L are random at a significance level of 0.05 using Excel.
14. List and briefly explain different two-sample non-parametric tests.
15. Give the different pairs of hypotheses under a one-tailed two-sample sign test for small samples, with suitable examples.
16. A shop floor manager believed that implementing the Six Sigma technique would reduce the number of defective assemblies produced each day in a line. Before adopting the Six Sigma technique in the assembly line, he thus collected data on the number of defective assemblies produced each day for eight days. These data are given in the following table. This shows the number of defective assemblies that were gathered over the course of eight days after implementing the Six Sigma technique on the assembly line.

	Defective Assemblies per Day							
Day	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Before Training	12	10	10	17	14	17	12	18
After Training	10	13	10	16	12	17	10	15

Check the null hypothesis that the implementation of Six Sigma approach has decreased the number of defective assemblies produced per day (H_0; $\mu_X = \mu_Y$) against the alternate hypothesis H_1: $\mu_{X <} \mu_Y$ using the sign test at a significance level of 0.01 through Excel.
17. In machining, fixtures play a major role in improving manufacturing productivity. A new fixture is available in the market. Hence, the line manager wants to use it in a costly machine under his control. The purchase manager of the company, who will be initiating buying the new fixture, believes that the new fixture is same as the existing fixture for the purpose stated. Hence, he collected data on the number of units produced per shift by that machine during nine different shifts by using the existing fixture as well as using the new fixture, and the data are as given in the following table.

Day	1	2	3	4	5	6	7	8	9
Existing Device	800	900	1000	700	500	800	900	800	800
New Device	900	700	900	600	600	900	800	700	900

Check the hypothesis that the number of units produced per shift using the new fixture does not differ from that of the existing fixture at a significance level of 0.01 using Excel.
18. List the pairs of hypotheses for the one-tailed two-sample sign test for large samples, with suitable examples.
19. The pass percentages of students in 20 randomly selected technology departments of a university before and after introducing online classes are summarised in the following table.

Technology Department	*Before Using ICT Facilities*	*After Using ICT Facilities*
1	65	80
2	70	74
3	96	98
4	70	70
5	80	89
6	65	72
7	80	80
8	92	99
9	85	80
10	78	82
11	88	85
12	70	89
13	70	76
14	80	65
15	84	87
16	73	65
17	84	86
18	87	90
19	84	70
20	78	80

Check the hypothesis that the pass percentage after introducing online classes has been reduced at a significance level of 0.05 using the sign test.

20. The following table summarises the semi-annual income (in thousands of rupees) of 21 randomly chosen farmers in a district before and after the introduction of a government stimulus package.

Company	*Sales Revenue in Lakhs of Rupees*	
	Before Introducing Stimulus Package	*After Introducing Stimulus Package*
1	170	110
2	200	210
3	180	160
4	160	150
5	180	160
6	130	200
7	170	150
8	140	220
9	180	200
10	180	160
11	160	150
12	170	220
13	120	140
14	160	140
15	140	160
16	130	140
17	170	150
18	145	140
19	160	180
20	180	200
21	190	180

Use Excel to test the null hypothesis that there is no difference between the semi-annual income before and after the stimulus package was introduced at a significance level of 0.05.

21. List and explain the steps of the median test.
22. The percentages of incremental weekly sales of different shops against two different sales promotion strategies (strategy X and strategy Y) are summarised in the following table. Check whether the two samples were drawn from populations with the same median at a significance level of 0.01 using Excel.

Shop	*Strategy X*	*Strategy Y*
1	35	37
2	37	45
3	52	49
4	47	62
5	70	64
6	52	72
7	40	38
8	60	

23. Explain the steps of the Mann-Whitney U test.
24. In a machine shop, two operators are assigned to two different machines of the same specification. The numbers of work pieces machined per shift by the operators during randomly selected shifts are summarised in the following table.

Shift	*Operator 1*	*Operator 2*
1	28	34
2	21	24
3	34	15
4	32	26
5	15	31
6	32	26
7	25	12
8	17	120
9	22	21
10	30	28
11	29	24
12	16	15
13	15	18
14	20	

Check whether the two samples shown in the table are drawn from identical populations against the alternate hypothesis that population 1 stochastically differs from population 2 using the Mann-Whitney U test at a significance level of 0.05 through Excel.

25. The following table provides an overview of a company's four-quarter EPS values over the last seven years. Using Excel, determine whether the four quarters' EPS values differ significantly from one another at a significance threshold of 0.01.

Year	*Quarter 1*	*Quarter 2*	*Quarter 3*	*Quarter 4*
1	80	82	75	60
2	85	77	85	85
3	95	86	82	95
4	62	80	73	88
5	6575	92	95	78
6	6788	76	56	60
7	82	90	74	75

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. www.csun.edu/~mr31841/documents/thesigntest.pdf [June 25, 2020].
3. https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test [June 25, 2020].
4. https://support.microsoft.com/en-us/office/chisq-dist-function-8486b05e-5c05-4942-a9ea-f6b341518732 [June 27, 2020].

13 Correlation and Covariance

Learning Objectives

After going through this chapter, you will be able to

- Understand the importance of the correlation coefficient.
- Analyse ungrouped data using correlation coefficient (Pearson's coefficient of correlation.)
- Apply computing the correlation coefficient of grouped data (Pearson's coefficient of correlation).
- Understand the working of the rank correlation coefficient with illustrations.
- Analyse data using the auto-correlation coefficient.
- Study the theory and applications of covariance.

13.1 Introduction

In reality, there are numerous circumstances in which two variables may be related, such as country's economic growth and the growth in its stock market index, government spending and economic growth, and so on. Both covariance and the correlation coefficient, which are discussed in this chapter, can be used to estimate these relationships [1].

13.2 Correlation

The correlation coefficient is a metric used to express the strength of association between two variables. The correlation coefficient has a range of values from –1 to +1. When two variables have a correlation coefficient of –1, their inverse correlation is at its highest. If it is 0, there is absolutely no relationship between the two variables and a zero degree of association between them. The two variables have the highest possible positive correlation if it is 1 [4].

The different types of correlation coefficient are listed here.

- Correlation coefficient of ungrouped data (Pearson's coefficient of correlation)
- Correlation coefficient of grouped data (Pearson's coefficient of correlation)
- Rank correlation coefficient
- Auto-correlation coefficient
- Covariance

DOI: 10.4324/9781032671703-13

13.3 Correlation Coefficient of Ungrouped Data Using Excel Sheets and Correl Function

In reality, the degree of correlation between two variables is a matter of interest. Consider the cost of living and the urbanisation index as two variables. Generally speaking, it is believed that as urbanisation increases, so will the cost of life. Such a notion may not hold true in direct proportion if the area of study is entirely surrounded by farming communities and other cities are located far away from it. Whatever the situation, the government or organisations will be motivated to do studies that will aid in relocating upcoming businesses so that the rise in the expense of living will be kept within a tolerable range.

The formula for the correlation coefficient (r) is shown as follows.

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n\sigma_X\sigma_Y}$$

where

n is the total number of observations of each of the variables

X_i is the i^{th} observation of the variable X, for $i = 1, 2,3, \ldots, n$

Y_i is the i^{th} observation of the variable Y for $i = 1, 2,3 \ldots, n$

$\bar{X}$ is the mean of the observations of the variable X

$\bar{Y}$ is the mean of the observations of the variable Y

σ_X is the standard deviation of the variable X

σ_Y is the standard deviation of the variable Y

r is the correlation coefficient

Details of the formula are shown as follows.

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

The significance of the correlation coefficient can be tested using the t test, whose test statistic is as follows.

$$t = \frac{r}{\sqrt{\left(1-r^2\right)\left(n-2\right)}}$$

where

t is the t test statistic to test the significance of the correlation coefficient

r is the correlation coefficient

n is the number of observations for the variable X as well as for the variable Y

The hypotheses, that are, the null hypothesis (H_0) and alternate hypothesis (H_1) to test the significance of the correlation coefficient, are stated as follows.

H_0: $r = 0$, which means that the variables are not associated

H_1: $r \neq 0$, which means that the variables are associated

13.3.1 Testing Guidelines

If the calculated t value is more than the table t value for $n - 2$ degrees of freedom at a given significance level α, then reject H_0; otherwise, accept H_0.

Or

If the p value at the right tail is less than the given significance level, then reject H_0; otherwise, accept H_0.

The correlation coefficient for ungrouped data can be obtained in two ways as listed as follows.

1. Using the CORREL function under the sequence of button clicks Home ⟹ Formulas ⟹ More Functions ⟹ Statistical.
2. Using the Correlation function under the sequence of buttons Home ⟹ Data ⟹ Data Analysis.

13.3.2 Using CORREL Function

The correlation coefficient between two variables can be determined by clicking the sequence of buttons Home ⟹ Formulas ⟹ More Functions ⟹ Statistical ⟹ CORREL, which is explained using an example in this section.

Example 13.1

The annual training expenditure (lakhs of rupees) and the corresponding labour productivity index (0 to100) for the past eight years of a company are shown in Table 13.1.

Find the correlation coefficient between annual training expenditure and labour productivity index:

1. Using your own Excel sheet.
2. Using an inbuilt function of Excel.

Table 13.1 Annual Training Expenditure and Labour Productivity Index

Year i	*Annual Training Expenditure X_i*	*Labour Productivity Index Y_i*
1	5	80
2	7	90
3	9	75
4	10	85
5	12	95
6	15	70
7	18	95
8	20	60

Solution

The data for Example 13.1 are presented in Table 13.2.

Table 13.2 Data for Example 13.1

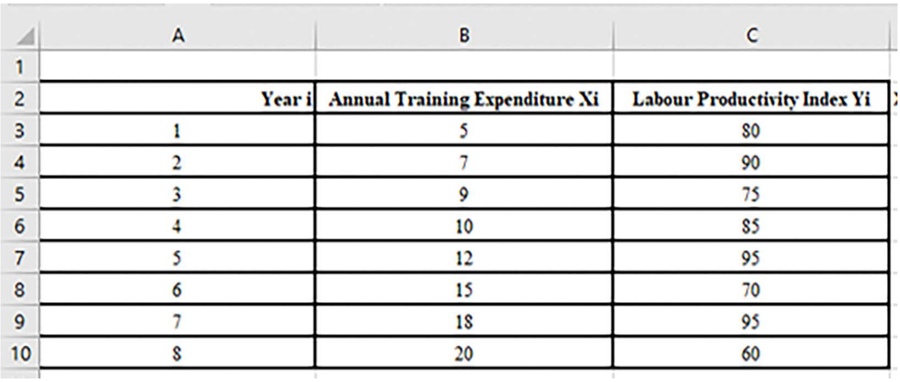
Year i	*Annual Training Expenditure* X_i	*Labour Productivity Index* Y_i
1	5	80
2	7	90
3	9	75
4	10	85
5	12	95
6	15	70
7	18	95
8	20	60

The formula to obtain the correlation coefficient between the variable *X* (annual training expenditure) and the variable *Y* (labour productivity index) is as follows.

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

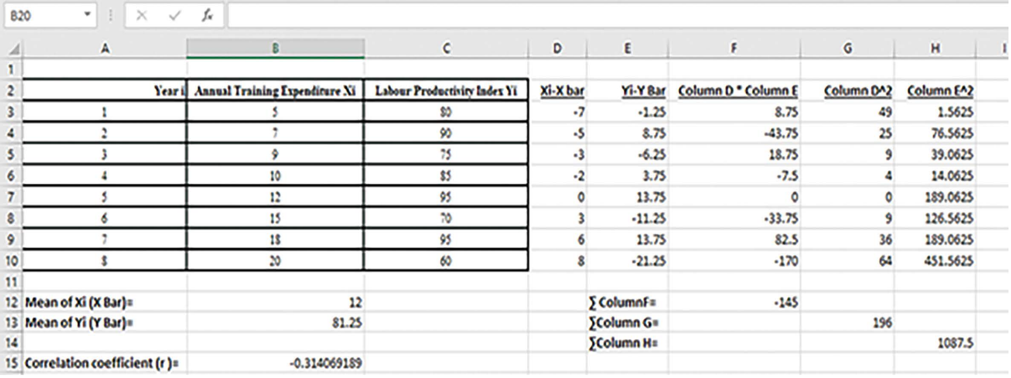
1. Computation of Correlation Coefficient using Custom Designed Excel Sheet
   The screenshot of the input of the data for Example 13.1 is shown in Figure 13.1. The Excel working to obtain the correlation coefficient is shown in Figure 13.2. The corresponding formulas in the Excel sheet are shown in Figure 13.3. From Figure 13.2, one can see that the correlation coefficient between the annual training expenditure and the labour productivity index is –0.314069189. This means that the two variables are negatively correlated.
2. Computation of Correlation Coefficient Using Inbuilt Function in Excel

The screenshot of the input of the data of Example 13.1 has already been shown in Figure 13.1. The screenshot of the sequence of button clicks Formulas ⟹ More Functions ⟹ Statistical is shown in Figure 13.4. The screenshot after clicking the CORREL function in the dropdown menu of Figure 13.4 is shown in Figure 13.5. The dropdown menu in Figure 13.5 has array 1 and array 2. The range of cells containing the data of variable 1 (annual training expenditure) should be entered in array 1, and the range of cells containing the data of variable 2 should be entered in array 2 (labour productivity index). The values should be numbers, names, arrays, or references that contain numbers. The screenshot after entering the range of cells for each variable in the dropdown menu of Figure 13.5 is shown in Figure 13.6. Clicking the OK button in the dropdown menu of Figure 13.6 gives the result of the correlation coefficient, as shown in Figure 13.7. From Figure 13.7, one can see that the correlation coefficient between the annual training expenditure and the labour productivity index is –0.314069189.



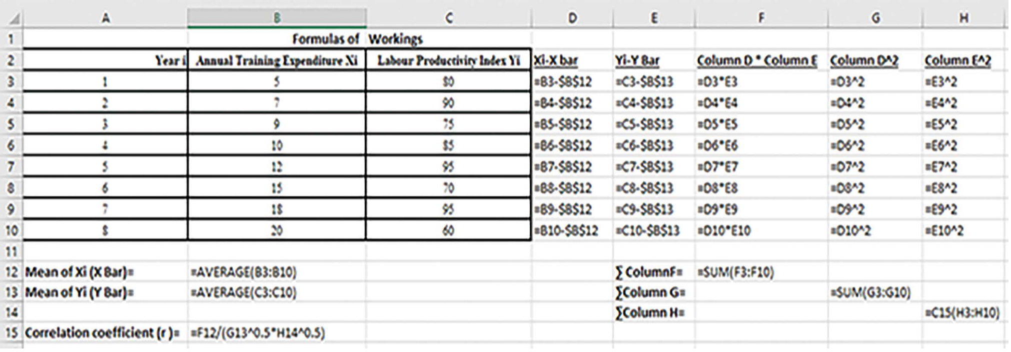
	A	B	C
1			
2	Year i	Annual Training Expenditure Xi	Labour Productivity Index Yi
3	1	5	80
4	2	7	90
5	3	9	75
6	4	10	85
7	5	12	95
8	6	15	70
9	7	18	95
10	8	20	60

Figure 13.1 Screenshot of input of Example 13.1



	A	B	C	D	E	F	G	H
1								
2	Year i	Annual Training Expenditure Xi	Labour Productivity Index Yi	Xi-X bar	Yi-Y Bar	Column D * Column E	Column D^2	Column E^2
3	1	5	80	-7	-1.25	8.75	49	1.5625
4	2	7	90	-5	8.75	-43.75	25	76.5625
5	3	9	75	-3	-6.25	18.75	9	39.0625
6	4	10	85	-2	3.75	-7.5	4	14.0625
7	5	12	95	0	13.75	0	0	189.0625
8	6	15	70	3	-11.25	-33.75	9	126.5625
9	7	18	95	6	13.75	82.5	36	189.0625
10	8	20	60	8	-21.25	-170	64	451.5625
11								
12	Mean of Xi (X Bar)=	12			∑ColumnF=	-145		
13	Mean of Yi (Y Bar)=	81.25			∑Column G=		196	
14					∑Column H=			1087.5
15	Correlation coefficient (r)=	-0.314069189						

Figure 13.2 Screenshot of working of correlation coefficient for Example 13.1

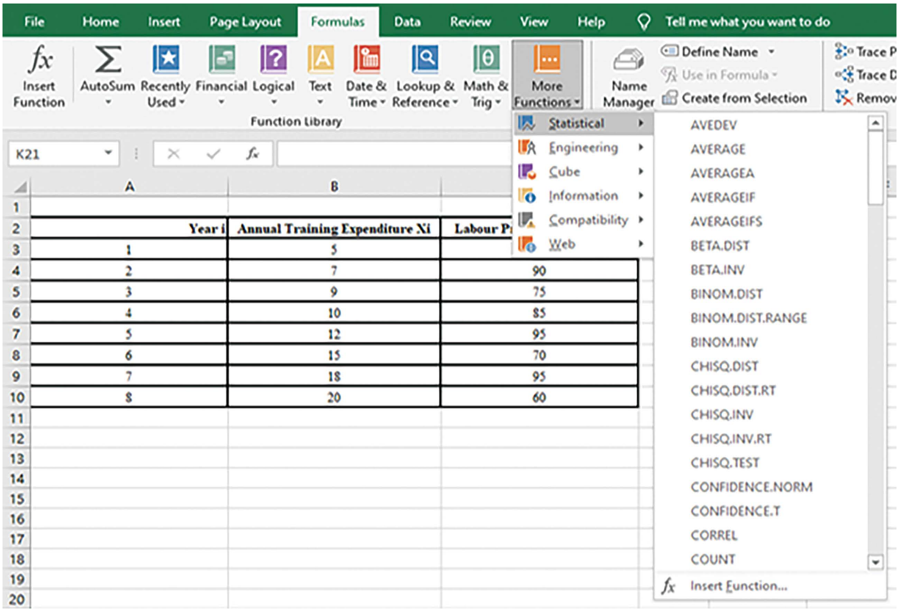


	A	B	C	D	E	F	G	H
1		Formulas of	Workings					
2	Year i	Annual Training Expenditure Xi	Labour Productivity Index Yi	Xi-X bar	Yi-Y Bar	Column D * Column E	Column D^2	Column E^2
3	1	5	80	=B3-$B$12	=C3-$B$13	=D3*E3	=D3^2	=E3^2
4	2	7	90	=B4-$B$12	=C4-$B$13	=D4*E4	=D4^2	=E4^2
5	3	9	75	=B5-$B$12	=C5-$B$13	=D5*E5	=D5^2	=E5^2
6	4	10	85	=B6-$B$12	=C6-$B$13	=D6*E6	=D6^2	=E6^2
7	5	12	95	=B7-$B$12	=C7-$B$13	=D7*E7	=D7^2	=E7^2
8	6	15	70	=B8-$B$12	=C8-$B$13	=D8*E8	=D8^2	=E8^2
9	7	18	95	=B9-$B$12	=C9-$B$13	=D9*E9	=D9^2	=E9^2
10	8	20	60	=B10-$B$12	=C10-$B$13	=D10*E10	=D10^2	=E10^2
11								
12	Mean of Xi (X Bar)=	=AVERAGE(B3:B10)			∑ColumnF=	=SUM(F3:F10)		
13	Mean of Yi (Y Bar)=	=AVERAGE(C3:C10)			∑Column G=		=SUM(G3:G10)	
14					∑Column H=			=C1S(H3:H10)
15	Correlation coefficient (r)=	=F12/(G13^0.5*H14^0.5)						

Figure 13.3 Screenshot of formulas of working of correlation coefficient for Example 13.1

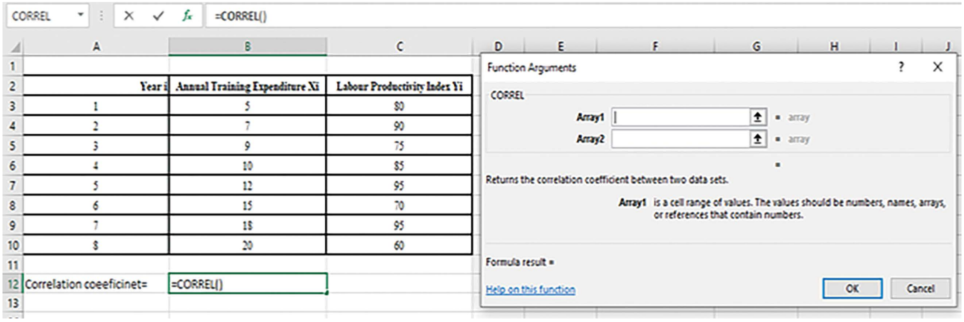
13.3.3 Correlation Coefficient Using Correlation Function

This section presents the determination of the correlation coefficient between a pair of variables using the correlation function under the sequence of button clicks Home ⟹ Data ⟹ Data Analysis. The screenshots of different steps of the correlation function under data analysis are explained while solving a problem.



	A	B	C
1			
2	Year i	Annual Training Expenditure Xi	Labour P
3	1	5	
4	2	7	90
5	3	9	75
6	4	10	85
7	5	12	95
8	6	15	70
9	7	18	95
10	8	20	60

Figure 13.4 Screenshot of for the sequence of button clicks, Formulas ⟹ More Functions ⟹ Statistical


CORREL   =CORREL()

	A	B	C
1			
2	Year i	Annual Training Expenditure Xi	Labour Productivity Index Yi
3	1	5	80
4	2	7	90
5	3	9	75
6	4	10	85
7	5	12	95
8	6	15	70
9	7	18	95
10	8	20	60
11			
12	Correlation coeeficinet=	=CORREL()	
13			

Function Arguments

CORREL

Array1 = array

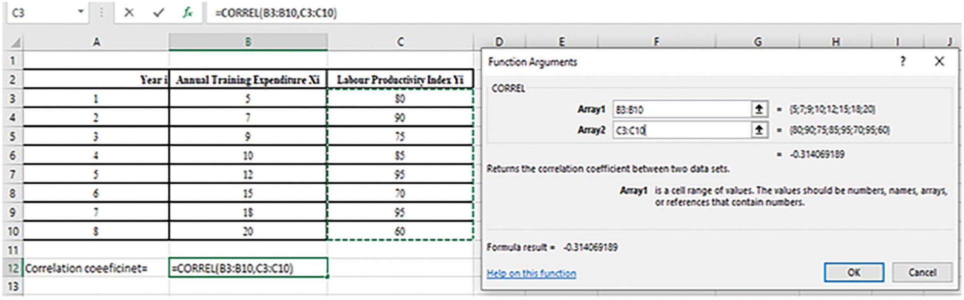
Array2 = array

Returns the correlation coefficient between two data sets.

Array1 is a cell range of values. The values should be numbers, names, arrays, or references that contain numbers.

Formula result =

Figure 13.5 Screenshot after clicking “CORREL” function in the dropdown menu of Figure 13.4


C3   =CORREL(B3:B10,C3:C10)

	A	B	C
1			
2	Year i	Annual Training Expenditure Xi	Labour Productivity Index Yi
3	1	5	80
4	2	7	90
5	3	9	75
6	4	10	85
7	5	12	95
8	6	15	70
9	7	18	95
10	8	20	60
11			
12	Correlation coeeficinet=	=CORREL(B3:B10,C3:C10)	
13			

Function Arguments

CORREL

Array1 B3:B10 = {5;7;9;10;12;15;18;20}

Array2 C3:C10 = {80;90;75;85;95;70;95;60}

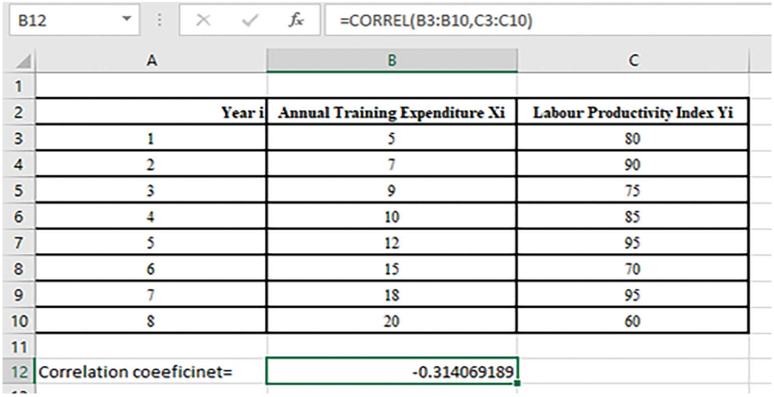
= -0.314069189

Returns the correlation coefficient between two data sets.

Array1 is a cell range of values. The values should be numbers, names, arrays, or references that contain numbers.

Formula result = -0.314069189

Figure 13.6 Screenshot after entering the range of cells containing data for each variable in the dropdown menu of Figure 13.5



B12 =CORREL(B3:B10,C3:C10)

	A	B	C
1			
2	Year i	Annual Training Expenditure Xi	Labour Productivity Index Yi
3	1	5	80
4	2	7	90
5	3	9	75
6	4	10	85
7	5	12	95
8	6	15	70
9	7	18	95
10	8	20	60
11			
12	Correlation coeeficinet=	-0.314069189	

Figure 13.7 Screenshot after clicking the OK button in the dropdown menu of Figure 13.6

Example 13.2

Table 13.3 displays a company's R&D spending (in lakhs of rupees) and yearly sales revenue (in crores of rupees) for the previous ten years.

Find the correlation coefficient between R&D expenditure and yearly sales revenue using an inbuilt function of Excel under Data Analysis.

Solution

The screenshot of the data for Example 13.2 is shown in Figure 13.8. The sequence of button clicks to obtain the correlation function under the Data Analysis button is as follows, which gives the screenshot in Figure 13.9.

Home ⟹ Data ⟹ Data Analysis

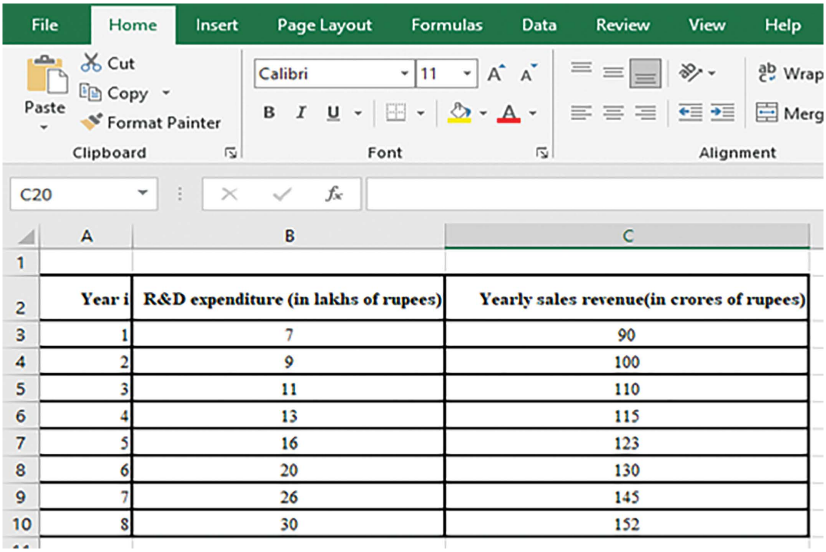
The selection of the correlation button in the dropdown menu of Figure 13.9 and clicking OK gives the screenshot in Figure 13.10. The screenshot after filling the range of cells $B$2 to $C$10 in the box against Input Range, clicking Columns against Grouped By, and clicking Labels in First Row in the dropdown menu of Figure 13.10 is shown in Figure 13.11. Clicking the OK button in the dropdown menu of Figure 13.11 gives the correlation matrix as shown in Figure 13.12. From Figure 13.12, one can see the correlation coefficient between the R&D expenditure and yearly sales revenue as 0.985986, which is very high. This means that the more the R&D expenditure, the more the annual sales of the company.

Example 13.3

The annual sales revenue (in crores of rupees), the sales force, and the annual advertising expenditures(lakhs of rupees) are the three variables that the industrial engineer of a company is particularly interested in examining for correlation coefficients. As shown

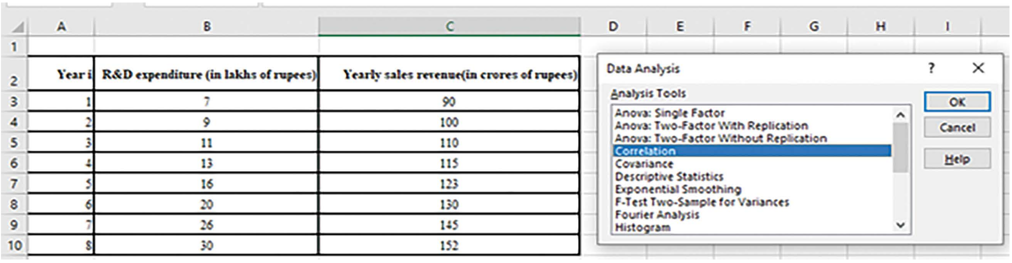
Table 13.3 Data for Example 13.2

Year i	*R&D Expenditure (Lakhs of Rupees)*	*Yearly Sales Revenue (Crores of Rupees)*
1	7	90
2	9	100
3	11	110
4	13	115
5	16	123
6	20	130
7	26	145
8	30	152



Year i	R&D expenditure (in lakhs of rupees)	Yearly sales revenue(in crores of rupees)
1	7	90
2	9	100
3	11	110
4	13	115
5	16	123
6	20	130
7	26	145
8	30	152

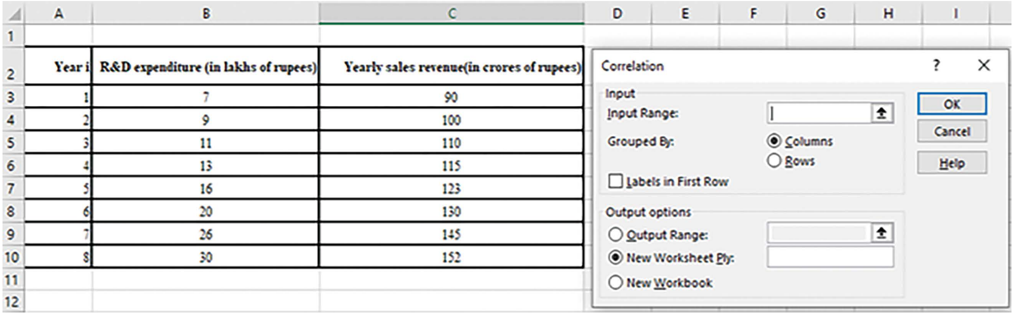
Figure 13.8 Screenshot of data for Example 13.2



Year i	R&D expenditure (in lakhs of rupees)	Yearly sales revenue(in crores of rupees)
1	7	90
2	9	100
3	11	110
4	13	115
5	16	123
6	20	130
7	26	145
8	30	152

Figure 13.9 Screenshot for the sequence of buttons, Home $\Longrightarrow$ Data $\Longrightarrow$ Data Analysis

in Table 13.4, he therefore has data for these variables over the previous eight years. In Excel, use the Regression tool under the Data Analysis button to find the correlation matrix among these variables.



Year i	R&D expenditure (in lakhs of rupees)	Yearly sales revenue(in crores of rupees)
1	7	90
2	9	100
3	11	110
4	13	115
5	16	123
6	20	130
7	26	145
8	30	152

Correlation

Input

Input Range:

Grouped By: ● Columns ○ Rows

□ Labels in First Row

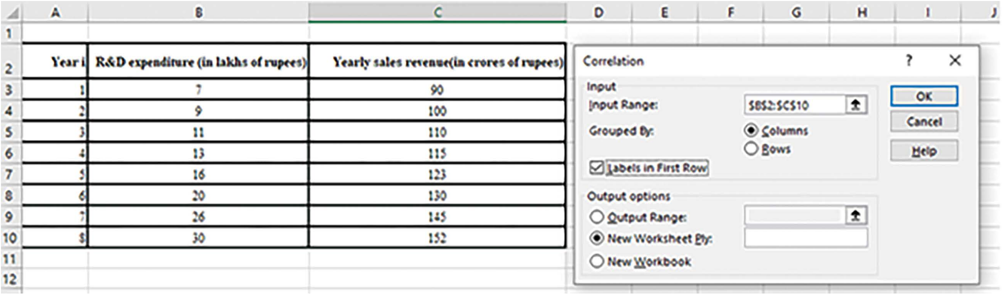
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 13.10 Screenshot after selecting Correlation function in the dropdown menu of Figure 13.9 and clicking the OK button



Year i	R&D expenditure (in lakhs of rupees)	Yearly sales revenue(in crores of rupees)
1	7	90
2	9	100
3	11	110
4	13	115
5	16	123
6	20	130
7	26	145
8	30	152

Correlation

Input

Input Range: $B$2:$C$10

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

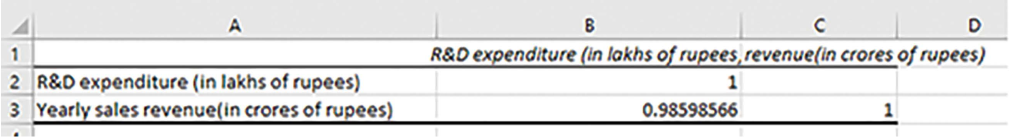
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 13.11 Screenshot after filling the data in the dropdown menu of Figure 13.10



	R&D expenditure (in lakhs of rupees	revenue(in crores of rupees)
R&D expenditure (in lakhs of rupees)	1	
Yearly sales revenue(in crores of rupees)	0.98598566	1

Figure 13.12 Screenshot after clicking the OK button in the dropdown menu of Figure 13.11

Table 13.4 Data for Example 13.3

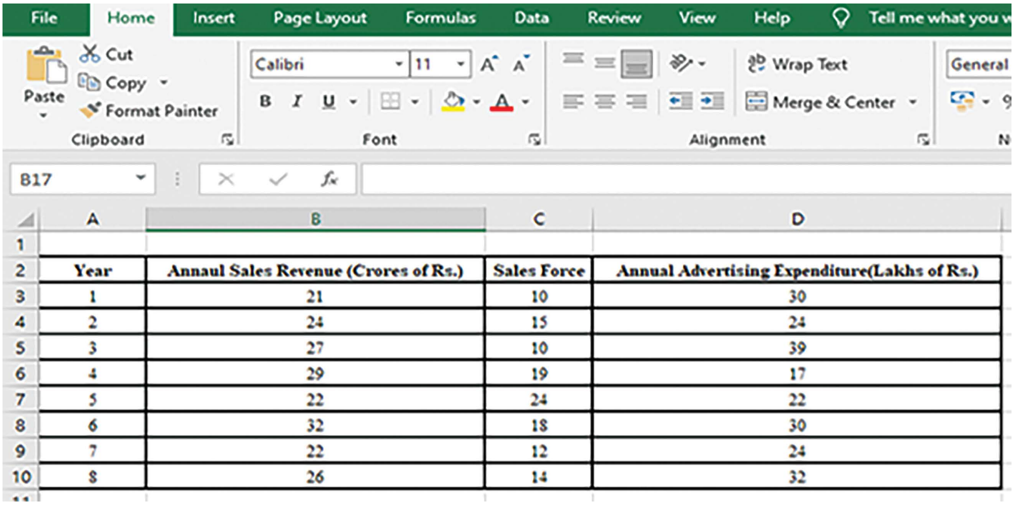
Year	*Annual Sales Revenue (Crores of Rupees)*	*Salesforce*	*Annual Advertising Expenditure (Lakhs of Rupees)*
1	21	10	30
2	24	15	24
3	27	10	39
4	29	19	17
5	22	24	22
6	32	18	30
7	22	12	24
8	26	14	32

Solution

The screenshot of the data for Example 13.3 is shown in Figure 13.13. The sequence of button clicks to obtain the correlation function under the Data Analysis button is as follows, which gives the screenshot in Figure 13.14

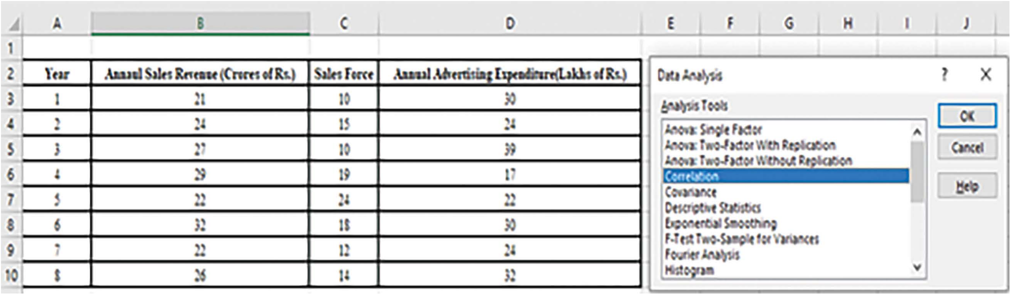
Home ⟹ Data ⟹ Data Analysis

Selecting the correlation button in the dropdown menu of Figure 13.14 and clicking OK gives the screenshot in Figure 13.15. The screenshot after filling the range of cells $B$2 to $D$10 in the box against Input Range, clicking Columns against Grouped By, and clicking Labels in First Row in the dropdown menu of Figure 13.15 is shown in Figure 13.16. Clicking the OK button in the dropdown menu of Figure 13.16 gives the correlation matrix among the variables, that is, annual sales revenue, sales force, and annual advertising expenditure, as shown in Figure 13.17.



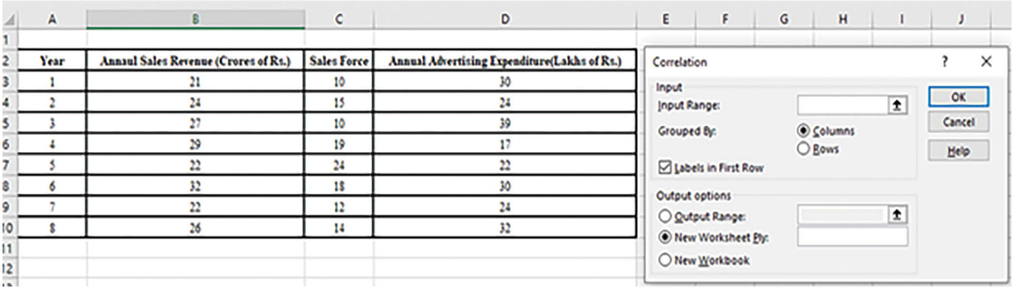
Year	Annaul Sales Revenue (Crores of Rs.)	Sales Force	Annual Advertising Expenditure(Lakhs of Rs.)
1	21	10	30
2	24	15	24
3	27	10	39
4	29	19	17
5	22	24	22
6	32	18	30
7	22	12	24
8	26	14	32

Figure 13.13 Screenshot of data for Example 13.3



Year	Annaul Sales Revenue (Crores of Rs.)	Sales Force	Annual Advertising Expenditure(Lakhs of Rs.)
1	21	10	30
2	24	15	24
3	27	10	39
4	29	19	17
5	22	24	22
6	32	18	30
7	22	12	24
8	26	14	32

Figure 13.14 Screenshot for the sequence of button clicks, Home ⟹ Data ⟹ Data Analysis



Year	Annaul Sales Revenue (Crores of Rs.)	Sales Force	Annual Advertising Expenditure(Lakhs of Rs.)
1	21	10	30
2	24	15	24
3	27	10	39
4	29	19	17
5	22	24	22
6	32	18	30
7	22	12	24
8	26	14	32

Correlation

Input

Input Range:

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

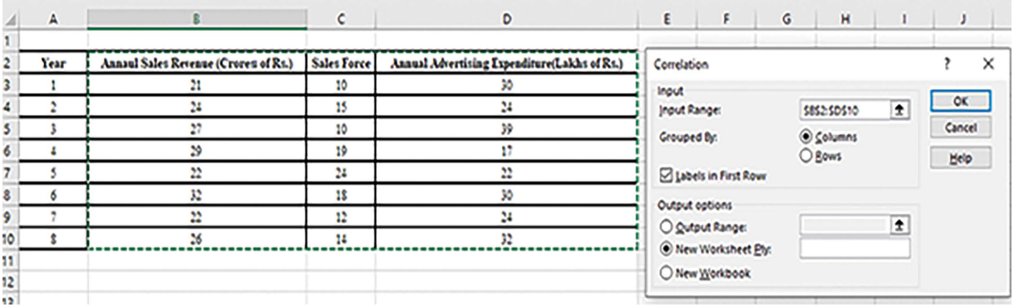
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 13.15 Screenshot after selecting Correlation function in the dropdown menu of Figure 13.14 and clicking the OK button



Year	Annaul Sales Revenue (Crores of Rs.)	Sales Force	Annual Advertising Expenditure(Lakhs of Rs.)
1	21	10	30
2	24	15	24
3	27	10	39
4	29	19	17
5	22	24	22
6	32	18	30
7	22	12	24
8	26	14	32

Correlation

Input

Input Range: $B$2:$D$10

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

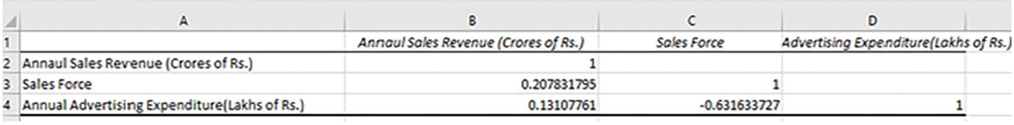
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 13.16 Screenshot after filling the data in the dropdown menu of Figure 13.15



	Annaul Sales Revenue (Crores of Rs.)	Sales Force	Advertising Expenditure(Lakhs of Rs.)
Annaul Sales Revenue (Crores of Rs.)	1		
Sales Force	0.207831795	1	
Annual Advertising Expenditure(Lakhs of Rs.)	0.13107761	-0.631633727	1

Figure 13.17 Screenshot after clicking the OK button in the dropdown menu of Figure 13.16

From the correlation matrix shown in Figure 13.17, the detailed results are as follows.

The correlation coefficient between the annual sales revenue and the sales force is 0.207832.

The correlation coefficient between the annual sales revenue and the annual advertising expenditure is 0.131078.

The correlation coefficient between the sales force and the annual advertising expenditure is –0.63163.

13.4 Correlation Coefficient of Grouped Data Using Excel Sheets

Table 13.5 displays a sample data set containing grouped data. An industrial estate contains 108 companies. In this example, increasing sales revenue is taken into account as variable *X* and increasing R&D expenses as variable *Y*. Table 13.5 lists the number of firms under each combination of percentage growth in sales revenue and R&D

Table 13.5 Sample Data Set of Grouped Data

Growth in Sales Revenue (%)	*R&D Expenditure (Crores of Rupees)*		
	4–8	*8–12*	*12–16*
10–12	2	5	8
12–14	3	7	12
14–16	4	9	15
16–18	8	15	20

expenditure. There are intervals for every variable. Find the correlation coefficient between these two variables using Excel.

The formula to compute the correlation coefficient of the grouped data is presented as follows.

$$r_g = \frac{N\sum_{i=1}^{m}\sum_{j=1}^{n}F_{ij} - \sum_{i=1}^{m}f_{i.}X_i\sum_{j=1}^{n}f_{.j}Y_j}{\sqrt{N\sum_{i=1}^{m}f_{i.}X_i^2 - \left(\sum_{i=1}^{m}f_{i.}X_i\right)^2}\ \sqrt{N\sum_{j=1}^{n}f_{.j}Y_j^2 - \left(\sum_{j=1}^{n}f_{.j}Y_j\right)^2}}$$

where

r_g is the correlation coefficient of grouped data with two variables

m is the number of intervals of the variable *X*

n is the number of intervals of the variable *Y*

X_i is the mid-point of the i^{th} interval of the variable *X* for *i* = 1, 2, 3, . . . , *m*

Y_j is the mid-point of the j^{th} interval of the variable *Y* for *j* = 1, 2, 3, . . . , *n*

f_{ij} is the frequency of the i^{th} midpoint of the interval of the variable *X* and j^{th} midpoint of the interval of the variable *Y*

$f_{i.}$ is the sum of the frequencies for all the values of *j* of the variable *Y* for a given *i* of the variable *X*

$f._j$ is the sum of the frequencies for all the values of *i* of the variable *X* for a given *j* of the variable *Y*

F_{ij} is the product of f_{ij}, X_i and Y_j $\left(\text{that is, } F_{ij} = f_{ij} \times X_i \times Y_j\right)$)

N is the sum of the frequencies for all the values of *i* of the variable *i* and for all the values of *j* of the variable *Y*, which is given by the following formula

$$N = \sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}$$

Example 13.4

As shown in Table 13.6, a consultant gathered data summarising the number of projects completed by a company for each combination of project cost (in crores of rupees) and rate of return (%). Discover the data's correlation coefficient using Excel.

Table 13.6 Frequencies for Rate of Return and Project Cost Combinations

Rate of Return (%)	*Project Cost (Crores of Rupees)*		
	4–8	*8–12*	*12–16*
10–12	1	4	1
12–14	2	6	1
14–16	3	8	7
16-–18	1	3	5

Table 13.7 Data for Example 13.4 With Mid-Values of Variables X_i and Y_i

Mid Rate of Return (%; X_i)	*Mid Project Cost (Crores of Rupees; Y_i)*		
	6	*10*	*14*
11	1	4	1
13	2	6	1
15	3	8	7
17	1	3	5

Solution

The data for Example 13.4 are shown in Table 13.7.

The formula to compute the correlation coefficient of the grouped data is presented as follows.

$$r_g = \frac{N\sum_{i=1}^{m}\sum_{j=1}^{n}F_{ij} - \sum_{i=1}^{m}f_{i.}X_i\sum_{j=1}^{n}f_{.j}Y_j}{\sqrt{N\sum_{i=1}^{m}f_{i.}X_i^2 - \left(\sum_{i=1}^{m}f_{i.}X_i\right)^2}\ \sqrt{N\sum_{j=1}^{n}f_{.j}Y_j^2 - \left(\sum_{j=1}^{n}f_{.j}Y_j\right)^2}}$$

where

r_g is the correlation coefficient of grouped data of two variables

m is the number of intervals of the variable X

n is the number of intervals of the variable Y

X_i is the mid-point of the ith interval of the variable X for $i = 1, 2, 3, \ldots, m$

Y_j is the mid-point of jth interval of the variable Y for $j = 1, 2, 3, \ldots, n$

f_{ij} is the frequency of the ith midpoint of the interval of the variable X and jth midpoint of the interval of the variable Y

$f_{i.}$ is the sum of the frequencies for all the values of j of the variable Y for a given i of the variable X

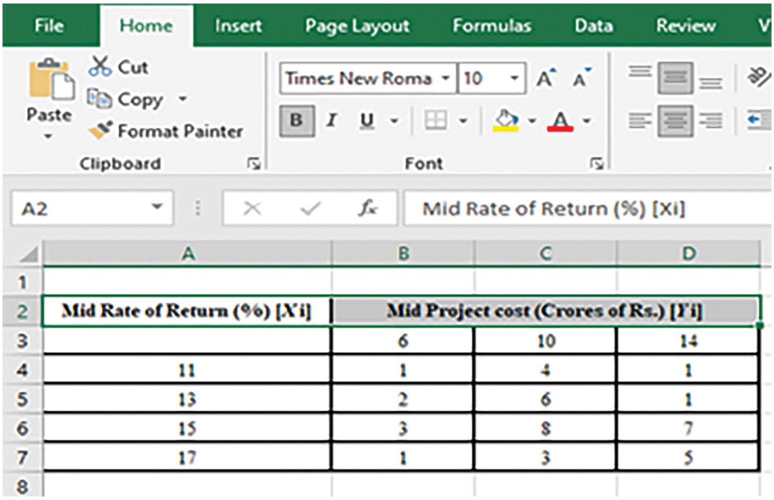
$f._j$ is the sum of the frequencies for all the values of i of the variable X for a given j of the variable Y

F_{ij} is the product of f_{ij}, X_i and Y_j (that is, $F_{ij} = f_{ij} \times X_i \times Y_j$)

N is the sum of the frequencies for all the values of *i* of the variable *i* and for all the values of *j* of the variable *Y*, which is given by the following formula

$$N = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}$$

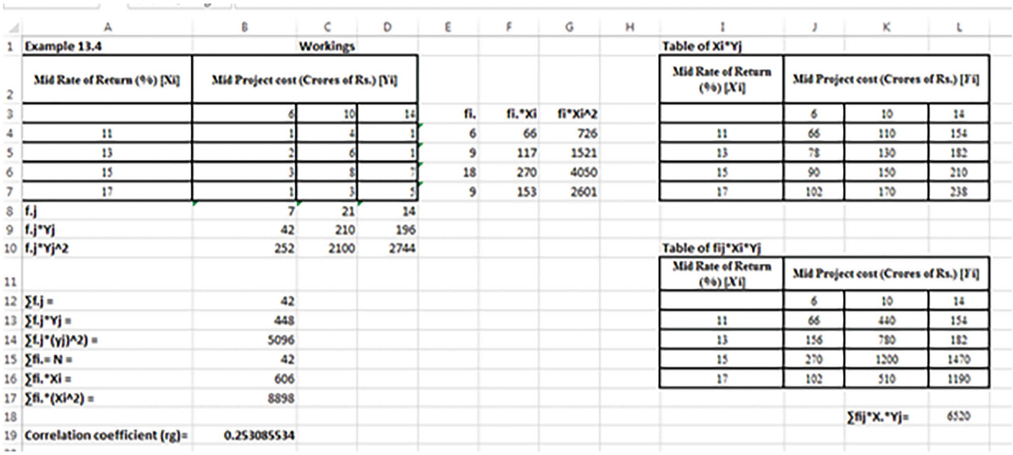
Since there is no function to compute r_g in Excel, a user-constructed Excel sheet is demonstrated for this example. The input of the data of Example 13.4 is shown in Figure 13.18. The screenshot of the working of Example 13.4 arise shown in Figure 13.19. The guidelines for the formulas of the working of Example 13.4 are shown in Figure 13.20. From Figure 13.19, one can see that the value of r_g is 0.253085534. This means that the association between the two variables, that is, project cost and rate of return, is very low.



A2: Mid Rate of Return (%) [Xi]

	A	B	C	D
1				
2	Mid Rate of Return (%) [Xi]	Mid Project cost (Crores of Rs.) [Yi]		
3		6	10	14
4	11	1	4	1
5	13	2	6	1
6	15	3	8	7
7	17	1	3	5
8				

Figure 13.18 Screenshot of input of Example 13.4



	A	B	C	D	E	F	G
1	Example 13.4		Workings				
2	Mid Rate of Return (%) [Xi]	Mid Project cost (Crores of Rs.) [Yi]					
3		6	10	14	fi.	fi.*Xi	fi*Xi^2
4	11	1	4	1	6	66	726
5	13	2	6	1	9	117	1521
6	15	3	8	7	18	270	4050
7	17	1	3	5	9	153	2601
8	f.j	7	21	14			
9	f.j*Yj	42	210	196			
10	f.j*Yj^2	252	2100	2744			
11							
12	∑f.j =	42					
13	∑f.j*Yj =	448					
14	∑f.j*(yj)^2) =	5096					
15	∑fi.= N =	42					
16	∑fi.*Xi =	606					
17	∑fi.*(Xi^2) =	8898					
18							
19	Correlation coefficient (rg)=	0.253085534					

Table of Xi*Yj

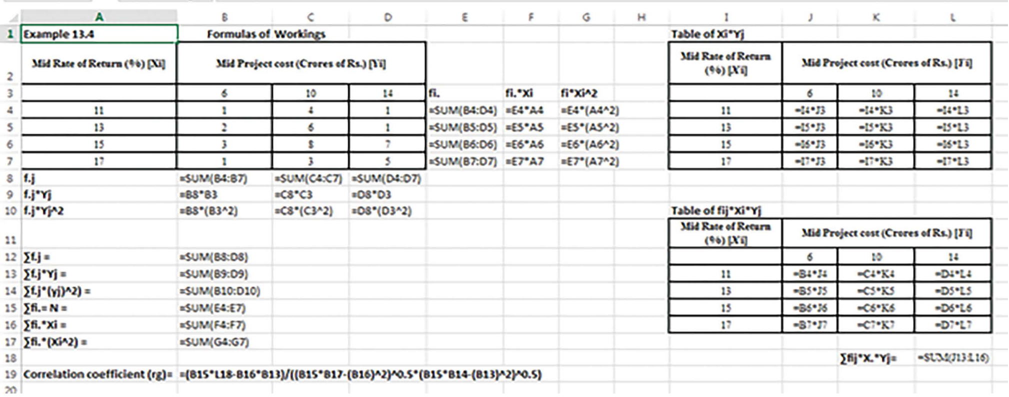
Mid Rate of Return (%) [Xi]	Mid Project cost (Crores of Rs.) [Yi]		
	6	10	14
11	66	110	154
13	78	130	182
15	90	150	210
17	102	170	238

Table of fij*Xi*Yj

Mid Rate of Return (%) [Xi]	Mid Project cost (Crores of Rs.) [Yi]		
	6	10	14
11	66	440	154
13	156	780	182
15	270	1200	1470
17	102	510	1190

∑fij*X.*Yj= 6520

Figure 13.19 Screenshot of working of r_g of Example 13.4



	A	B	C	D	E	F	G	H	I	J	K	L
1	Example 13.4	Formulas of Workings							Table of Xi*Yj			
2	Mid Rate of Return (%) [Xi]	Mid Project cost (Crores of Rs.) [Yi]							Mid Rate of Return (%) [Xi]	Mid Project cost (Crores of Rs.) [Yi]		
3		6	10	14	fi.	fi.*Xi	fi*Xi^2			6	10	14
4	11	1	4	1	=SUM(B4:D4)	=E4*A4	=E4*(A4^2)		11	=I4*J3	=I4*K3	=I4*L3
5	13	2	6	1	=SUM(B5:D5)	=E5*A5	=E5*(A5^2)		13	=I5*J3	=I5*K3	=I5*L3
6	15	3	8	7	=SUM(B6:D6)	=E6*A6	=E6*(A6^2)		15	=I6*J3	=I6*K3	=I6*L3
7	17	1	3	5	=SUM(B7:D7)	=E7*A7	=E7*(A7^2)		17	=I7*J3	=I7*K3	=I7*L3
8	f.j	=SUM(B4:B7)	=SUM(C4:C7)	=SUM(D4:D7)								
9	f.j*Yj	=B8*B3	=C8*C3	=D8*D3								
10	f.j*Yj^2	=B8*(B3^2)	=C8*(C3^2)	=D8*(D3^2)					Table of fij*Xi*Yj			
11									Mid Rate of Return (%) [Xi]	Mid Project cost (Crores of Rs.) [Yi]		
12	∑f.j =	=SUM(B8:D8)								6	10	14
13	∑f.j*Yj =	=SUM(B9:D9)							11	=B4*J4	=C4*K4	=D4*L4
14	∑f.j*(yj)^2) =	=SUM(B10:D10)							13	=B5*J5	=C5*K5	=D5*L5
15	∑fi.= N =	=SUM(E4:E7)							15	=B6*J6	=C6*K6	=D6*L6
16	∑fi.*Xi =	=SUM(F4:F7)							17	=B7*J7	=C7*K7	=D7*L7
17	∑fi.*(Xi^2) =	=SUM(G4:G7)										
18											∑fij*X.*Yj=	=SUM(J13:L16)
19	Correlation coefficient (rg)=	=(B15*L18-B16*B13)/((B15*B17-(B16)^2)^0.5*(B15*B14-(B13)^2)^0.5)										
20												

Figure 13.20 Screenshot of formulas of working of r_g of Example 13.4

13.5 Rank Correlation Using Excel Sheets and T.DIST.RT Function

A certain number of pairs of observations for two different variables are the subject of the rank correlation [2]. Take into account a situation where n units of an interesting product are ranked by two assessors. The objective now is to check whether these n pairs of ranks are connected. Spearman's correlation coefficient is another name for this correlation coefficient (r_s).

Another example can be the study of the EPS values for two consecutive years of a given number of companies in an industrial estate.

The formula to compute Spearman's rank correlation coefficient is as presented here.

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(X_i - Y_i)^2}{n(n^2 - 1)}$$

where

r_s is Spearman's rank correlation coefficient
n is the number of pairs of observations
X_i is the rank of the i^{th} observation of the variable 1, $i = 1, 2, 3, \ldots, n$
Y_i is the rank of the i^{th} observation of the variable 2, $i = 1, 2, 3, \ldots, n$

The null and alternate hypotheses to test the significance of the rank correlation coefficient are as stated.

H_0: $r_s = 0$ (The ranks of two variables (X_i and Y_i, $i = 1, 2, 3, \ldots, n$) are not correlated)
H_1: $r_s \neq 0$ (The ranks of two variables (X_i and Y_i, $i = 1, 2, 3, \ldots, n$) are correlated)

The significance of the rank correlation coefficient is tested using a two-tailed t test. The formula for the t statistic is as follows.

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} \quad \text{with } \alpha/2 \text{ significance level at right tail and } d.f. \text{ of } n-2$$

Example 13.5

The preference ratings of ten brands of washing machine by two judges are shown in Table 13.8.

Using Excel, determine the rank correlation between the judges' preference ratings for various washing machine brands. Also, test the rank correlation coefficient's significance at a 0.05 threshold of significance.

Solution

The data for this example are shown in Table 13.9.

The null and alternate hypotheses to test the significance of the rank correlation coefficient are as stated.

H_0: $r_s = 0$ (The ranks of two variables X_i and Y_i, $i = 1, 2, 3, \ldots, n$ are not correlated)
H_1: $r_s \neq 0$ (The ranks of two variables X_i and Y_i, $i = 1, 2, 3, \ldots, n$ are correlated)

The formula for the rank correlation coefficient is as follows.

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(X_i - Y_i)^2}{n(n^2 - 1)}$$

Table 13.8 Preference Ratings of Washing Machines

Brand	*Ratings by Judge 1*	*Ratings by Judge 2*
1	1	3
2	2	5
3	3	6
4	4	2
5	5	4
6	6	7
7	7	8
8	8	10
9	9	1
10	10	9

Table 13.9 Data for Example 13.5

Brand	*Ratings by Judge 1*	*Ratings by Judge 2*
1	1	3
2	2	5
3	3	6
4	4	2
5	5	4
6	6	7
7	7	8
8	8	10
9	9	1
10	10	9

where
r_s is Spearman's rank correlation coefficient
n is the number of pairs of observations
X_i is the rank of the i^{th} observation of the variable 1, $i = 1, 2, 3, \ldots, n$
Y_i is the rank of the i^{th} observation of the variable 2, $i = 1, 2, 3, \ldots, n$

The screenshot of the input of the data of Example 13.5 is shown in Figure 13.21. The screenshot of details of computations to obtain Spearman's correlation coefficient is shown in Figure 13.22. The formulas of details of computations to obtain Spearman's correlation coefficient are shown in Figure 13.23.

From Figure 13.22, it can be seen that Spearman's correlation coefficient is 0.406061.

The t value with respect to this correlation coefficient is 1.25679, which is less than the table t value (2.751524) when $\alpha/2$ is 0.25 and degree of freedom $(n - 2)$ is 8.

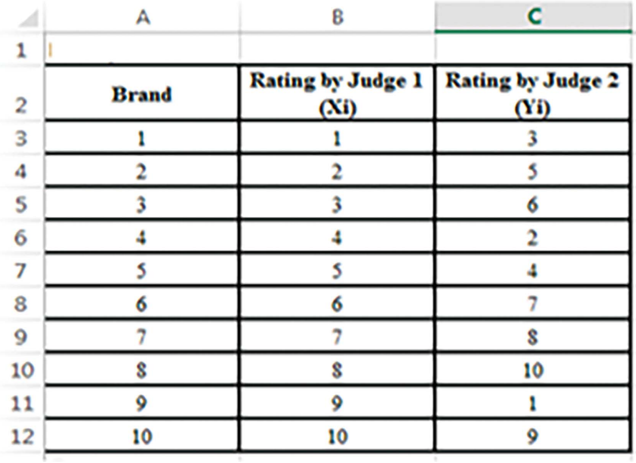
Hence, the null hypothesis is accepted, which means that the ranks of the brands given by the judges are not correlated.

13.6 Auto-Correlation Using Excel Sheets

The correlation between a set of data and another set of data that is derived from the first set of data with predetermined lag times is known as autocorrelation. Consider the values for a product's demand over n periods as stated in Table 13.10.

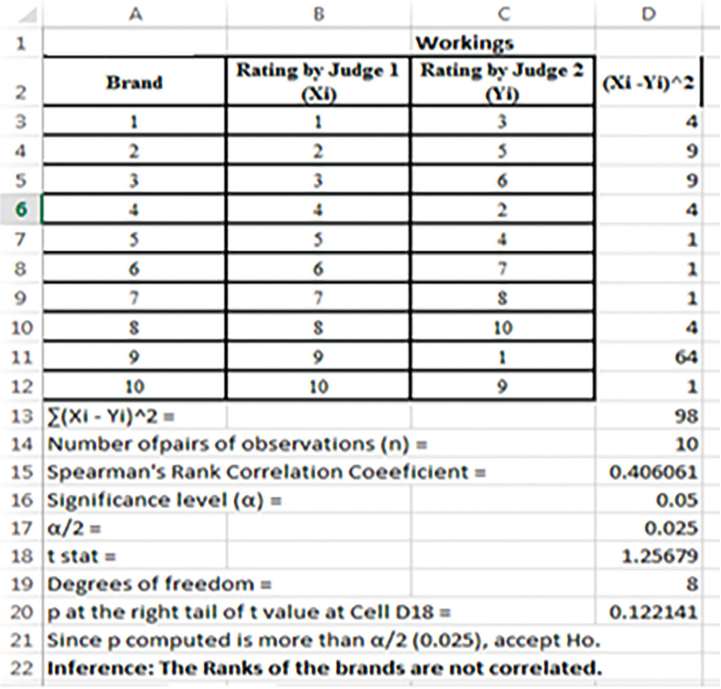
If the demand values shown in Table 13.10 are to be correlated with the same data with, say, one period of lag, then the second set of data are derived as shown in Table 13.11.

Now the correlation between the original data and the data with one period of lag is called the autocorrelation coefficient of the original data with one period of lag. If the number of periods of lag is k, then the correlation coefficient between the original data and the data with k periods of lag is called the auto correlation coefficient with k periods of lag.



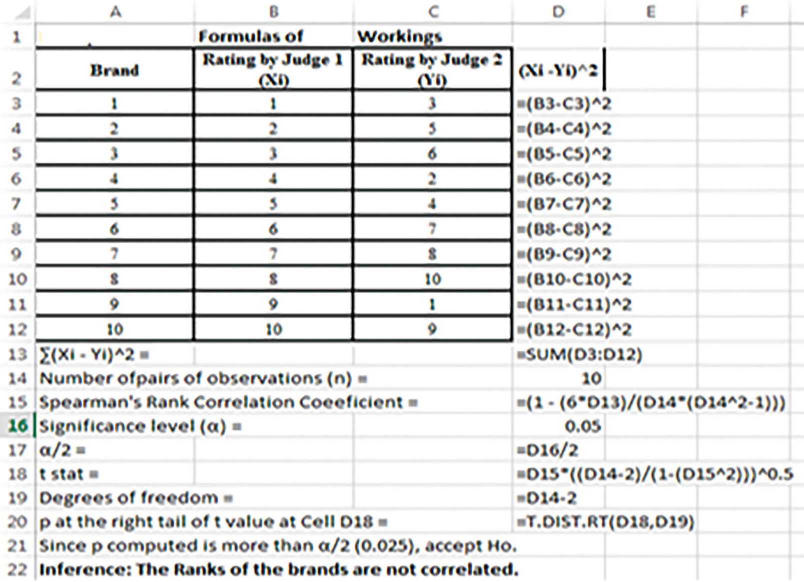
	A	B	C
1			
2	Brand	Rating by Judge 1 (Xi)	Rating by Judge 2 (Yi)
3	1	1	3
4	2	2	5
5	3	3	6
6	4	4	2
7	5	5	4
8	6	6	7
9	7	7	8
10	8	8	10
11	9	9	1
12	10	10	9

Figure 13.21 Screenshot of input of Example 13.5 in Excel sheet



	A	B	C	D
1			Workings	
2	Brand	Rating by Judge 1 (Xi)	Rating by Judge 2 (Yi)	(Xi -Yi)^2
3	1	1	3	4
4	2	2	5	9
5	3	3	6	9
6	4	4	2	4
7	5	5	4	1
8	6	6	7	1
9	7	7	8	1
10	8	8	10	4
11	9	9	1	64
12	10	10	9	1
13	∑(Xi - Yi)^2 =			98
14	Number ofpairs of observations (n) =			10
15	Spearman's Rank Correlation Coeeficient =			0.406061
16	Significance level (α) =			0.05
17	α/2 =			0.025
18	t stat =			1.25679
19	Degrees of freedom =			8
20	p at the right tail of t value at Cell D18 =			0.122141
21	Since p computed is more than α/2 (0.025), accept Ho.			
22	**Inference: The Ranks of the brands are not correlated.**			

Figure 13.22 Screenshot of computation of Spearman's correlation coefficient in Example 13.5



	A	B	C	D	E	F
1		Formulas of	Workings			
2	Brand	Rating by Judge 1 (Xi)	Rating by Judge 2 (Yi)	(Xi -Yi)^2		
3	1	1	3	=(B3-C3)^2		
4	2	2	5	=(B4-C4)^2		
5	3	3	6	=(B5-C5)^2		
6	4	4	2	=(B6-C6)^2		
7	5	5	4	=(B7-C7)^2		
8	6	6	7	=(B8-C8)^2		
9	7	7	8	=(B9-C9)^2		
10	8	8	10	=(B10-C10)^2		
11	9	9	1	=(B11-C11)^2		
12	10	10	9	=(B12-C12)^2		
13	∑(Xi - Yi)^2 =			=SUM(D3:D12)		
14	Number ofpairs of observations (n) =			10		
15	Spearman's Rank Correlation Coeeficient =			=(1 - (6*D13)/(D14*(D14^2-1)))		
16	Significance level (α) =			0.05		
17	α/2 =			=D16/2		
18	t stat =			=D15*((D14-2)/(1-(D15^2)))^0.5		
19	Degrees of freedom =			=D14-2		
20	p at the right tail of t value at Cell D18 =			=T.DIST.RT(D18,D19)		
21	Since p computed is more than α/2 (0.025), accept Ho.					
22	**Inference: The Ranks of the brands are not correlated.**					

Figure 13.23 Screenshot of formulas of computation of Spearman's correlation coefficient in Example 13.5

Table 13.10 Generalised Data for Demand

Period (t)	*Demand*
1	X_1
2	X_2
3	X_3
–	–
–	–
–	–
i	X_i
–	–
–	–
–	–
n	X_n

Table 13.11 Generalised Data for Demand With One Period of Lag of Data in Table 3.10

Period (t)	*Demand*	*Demand with One Period of Lag*
1	X_1	X_2
2	X_2	X_3
3	X_3	X_4
–	–	–
–	–	–
–	–	–
i	X_i	X_{i+1}
–	–	–
–	–	–
–	–	X_n
n	X_n	–

The formula to compute the auto-correlation coefficient with *k* periods of lag is presented as follows.

$$r_k = \frac{\sum_{i=1}^{n-k}\left(X_i - \bar{X}\right)\left(X_{i+k} - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

where

r_k is the autocorrelation with *k* periods of lag

X_{i} is i^{th} observation of the variable

$\bar{X}$ is the mean of the observations X_{i}, $i = 1, 2, 3, \ldots, n$

n is the number of observations of the variable *X*

Example 13.6

The demand values of a product over the past ten years are shown in Table 13.12. Find the auto-correlation coefficients with one-year lag (r_1) and two-year lag (r_2).

Table 13.12 Demand Values of Product

Year (i)	*Demand in '000 (D_i)*
1	40
2	45
3	52
4	60
5	68
6	75
7	87
8	90
9	95
10	110

Table 13.13 Data for Example 13.6 With One-Year Lag and Two-Year Lag

Year (i)	*Variable (X_i)*	*One-Year Lag Data (Y_1)*	*Two-Year Lag Data (Y_2)*
1	40	45	52
2	45	52	60
3	52	60	68
4	60	68	75
5	68	75	87
6	75	87	90
7	87	90	95
8	90	95	110
9	95	110	–
10	110	–	–

Solution

The data for Example 13.6 with one-year lag and two-year lag are shown in Table 13.13.

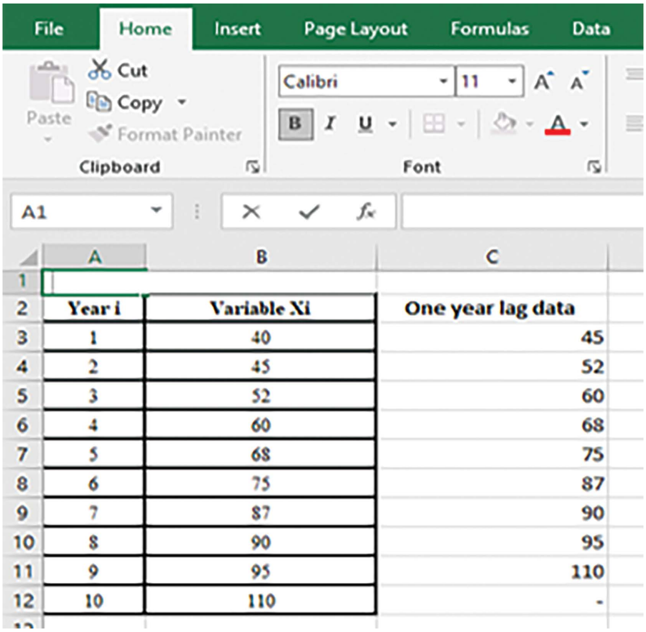
The formula to compute the auto-correlation is presented as follows.

$$r_k = \frac{\sum_{i=1}^{n-k}\left(X_i - \bar{X}\right)\left(X_{i+k} - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

where
k is the number of periods of lag
r_k is the autocorrelation with k periods of lag
X_i is i^{th} observation of the variable
$\bar{X}$ is the mean of the observations X_i, $i = 1, 2, 3, \ldots, n$
n is the number of observations of the variable X

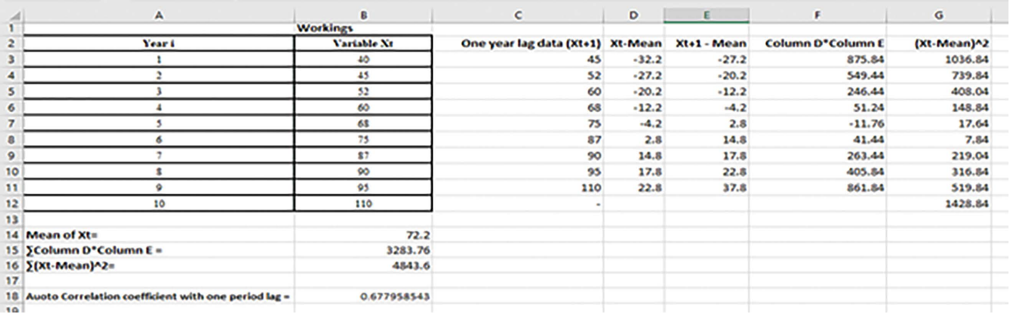
1. Computation of Auto-Correlation with One-Year Lag

   The data for computing the auto-correlation of the Example 13.6 data with a one-year lag are given as input in the Excel sheet shown in Figure 13.24 and are formed from the first two columns of Table 13.13. Figure 13.25 illustrates the steps involved



	A	B	C
1			
2	Year i	Variable Xi	One year lag data
3	1	40	45
4	2	45	52
5	3	52	60
6	4	60	68
7	5	68	75
8	6	75	87
9	7	87	90
10	8	90	95
11	9	95	110
12	10	110	-

Figure 13.24 Screenshot of input of data for autocorrelation coefficient with one-year lag



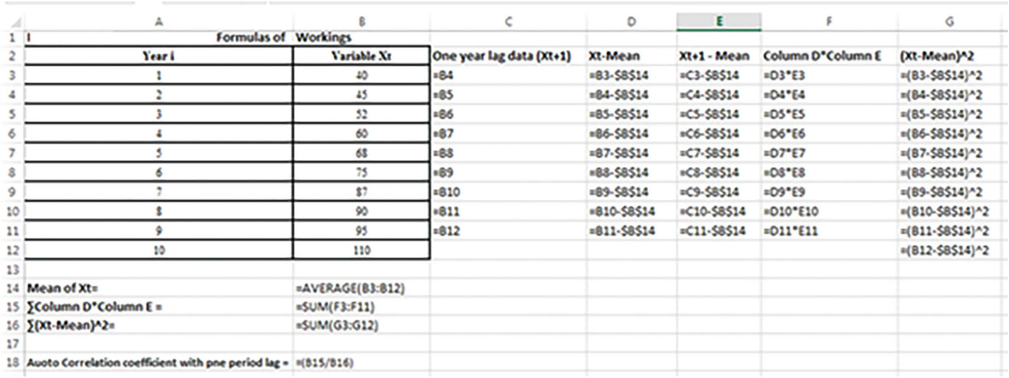
	A	B	C	D	E	F	G
1		Workings					
2	Year i	Variable Xt	One year lag data (Xt+1)	Xt-Mean	Xt+1 - Mean	Column D*Column E	(Xt-Mean)^2
3	1	40	45	-32.2	-27.2	875.84	1036.84
4	2	45	52	-27.2	-20.2	549.44	739.84
5	3	52	60	-20.2	-12.2	246.44	408.04
6	4	60	68	-12.2	-4.2	51.24	148.84
7	5	68	75	-4.2	2.8	-11.76	17.64
8	6	75	87	2.8	14.8	41.44	7.84
9	7	87	90	14.8	17.8	263.44	219.04
10	8	90	95	17.8	22.8	405.84	316.84
11	9	95	110	22.8	37.8	861.84	519.84
12	10	110	-				1428.84
13							
14	Mean of Xt=	72.2					
15	∑Column D*Column E =	3283.76					
16	∑(Xt-Mean)^2=	4843.6					
17							
18	Auoto Correlation coefficient with one period lag =	0.677958543					

Figure 13.25 Screenshot of working for autocorrelation coefficient of data with one-year lag

in calculating the autocorrelation coefficient for data with a one-year lag. Figure 13.26 displays the formulas for computing the autocorrelation coefficient of data with a one-year lag. The autocorrelation coefficient for the supplied data with a one-year lag is 0.677958543, as can be seen in Figure 13.25.

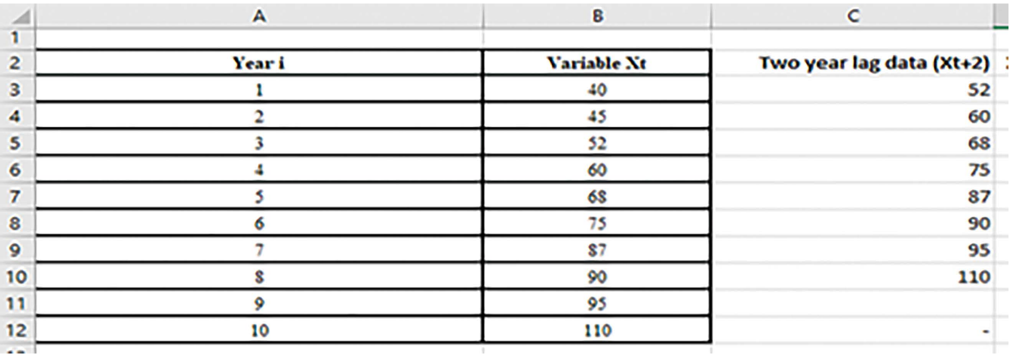
2. Auto-Correlation with Two-Year Lag

   The data to compute the auto-correlation of the data from Example 13.6 with a two-year lag are given as input in the Excel sheet shown in Figure 13.27, and they are formed by the first column and third column of Table 13.13. Figure 13.28 illustrates the steps involved in calculating the autocorrelation coefficient for data with a two-year lag. Figure 13.29 displays the formulas for computing the autocorrelation coefficient of the data with a two-year lag. The autocorrelation coefficient for the supplied data with a two-year lag is 0.419299694, as can be seen in Figure 13.28.



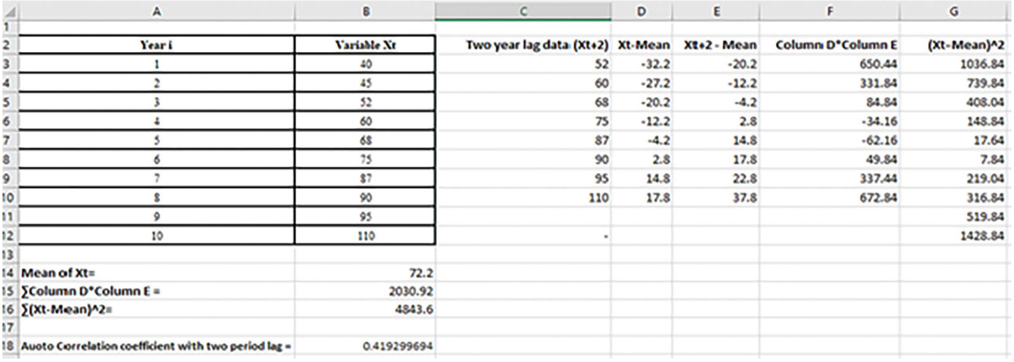
	A	B	C	D	E	F	G
1	I Formulas of	Workings					
2	Year i	Variable Xt	One year lag data (Xt+1)	Xt-Mean	Xt+1 - Mean	Column D*Column E	(Xt-Mean)^2
3	1	40	=B4	=B3-$B$14	=C3-$B$14	=D3*E3	=(B3-$B$14)^2
4	2	45	=B5	=B4-$B$14	=C4-$B$14	=D4*E4	=(B4-$B$14)^2
5	3	52	=B6	=B5-$B$14	=C5-$B$14	=D5*E5	=(B5-$B$14)^2
6	4	60	=B7	=B6-$B$14	=C6-$B$14	=D6*E6	=(B6-$B$14)^2
7	5	68	=B8	=B7-$B$14	=C7-$B$14	=D7*E7	=(B7-$B$14)^2
8	6	75	=B9	=B8-$B$14	=C8-$B$14	=D8*E8	=(B8-$B$14)^2
9	7	87	=B10	=B9-$B$14	=C9-$B$14	=D9*E9	=(B9-$B$14)^2
10	8	90	=B11	=B10-$B$14	=C10-$B$14	=D10*E10	=(B10-$B$14)^2
11	9	95	=B12	=B11-$B$14	=C11-$B$14	=D11*E11	=(B11-$B$14)^2
12	10	110					=(B12-$B$14)^2
13							
14	Mean of Xt=	=AVERAGE(B3:B12)					
15	∑Column D*Column E =	=SUM(F3:F11)					
16	∑(Xt-Mean)^2=	=SUM(G3:G12)					
17							
18	Auoto Correlation coefficient with pne period lag =	=(B15/B16)					

Figure 13.26 Screenshot of formulas for autocorrelation coefficient of data with one-year lag



	A	B	C
1			
2	Year i	Variable Xt	Two year lag data (Xt+2)
3	1	40	52
4	2	45	60
5	3	52	68
6	4	60	75
7	5	68	87
8	6	75	90
9	7	87	95
10	8	90	110
11	9	95	
12	10	110	-

Figure 13.27 Screenshot of input of data for autocorrelation coefficient with two-year lag

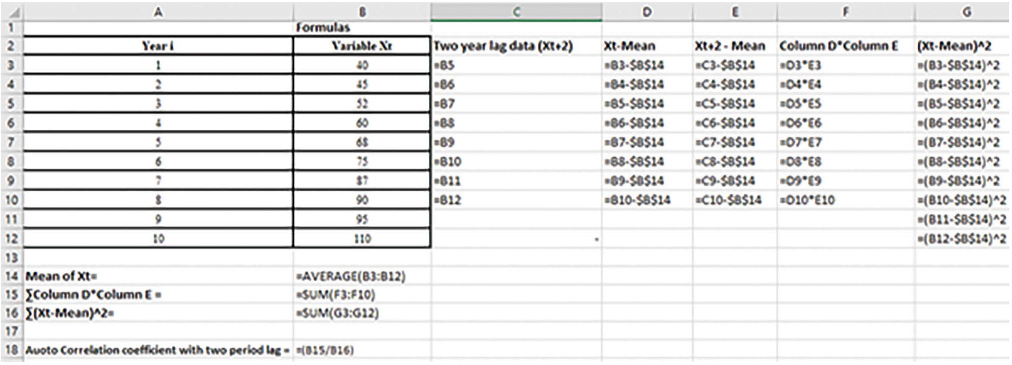


	A	B	C	D	E	F	G
1							
2	Year i	Variable Xt	Two year lag data (Xt+2)	Xt-Mean	Xt+2 - Mean	Column D*Column E	(Xt-Mean)^2
3	1	40	52	-32.2	-20.2	650.44	1036.84
4	2	45	60	-27.2	-12.2	331.84	739.84
5	3	52	68	-20.2	-4.2	84.84	408.04
6	4	60	75	-12.2	2.8	-34.16	148.84
7	5	68	87	-4.2	14.8	-62.16	17.64
8	6	75	90	2.8	17.8	49.84	7.84
9	7	87	95	14.8	22.8	337.44	219.04
10	8	90	110	17.8	37.8	672.84	316.84
11	9	95					519.84
12	10	110	-				1428.84
13							
14	Mean of Xt=	72.2					
15	∑Column D*Column E =	2030.92					
16	∑(Xt-Mean)^2=	4843.6					
17							
18	Auoto Correlation coefficient with two period lag =	0.419299694					

Figure 13.28 Screenshot of working for autocorrelation coefficient of data with two-year lag

13.7 Covariance Using Covariance.P/Covariance.S Functions

Another metric for examining the connection between two relevant variables is covariance. The two variables are positively correlated if the covariance value is positive; conversely, if it is negative, the two variables are negatively correlated. Think about the relationship between a country's agricultural output and rainfall. Since there is a



	A	B	C	D	E	F	G
1		Formulas					
2	Year i	Variable Xt	Two year lag data (Xt+2)	Xt-Mean	Xt+2 - Mean	Column D*Column E	(Xt-Mean)^2
3	1	40	=B5	=B3-$B$14	=C3-$B$14	=D3*E3	=(B3-$B$14)^2
4	2	45	=B6	=B4-$B$14	=C4-$B$14	=D4*E4	=(B4-$B$14)^2
5	3	52	=B7	=B5-$B$14	=C5-$B$14	=D5*E5	=(B5-$B$14)^2
6	4	60	=B8	=B6-$B$14	=C6-$B$14	=D6*E6	=(B6-$B$14)^2
7	5	68	=B9	=B7-$B$14	=C7-$B$14	=D7*E7	=(B7-$B$14)^2
8	6	75	=B10	=B8-$B$14	=C8-$B$14	=D8*E8	=(B8-$B$14)^2
9	7	87	=B11	=B9-$B$14	=C9-$B$14	=D9*E9	=(B9-$B$14)^2
10	8	90	=B12	=B10-$B$14	=C10-$B$14	=D10*E10	=(B10-$B$14)^2
11	9	95					=(B11-$B$14)^2
12	10	110					=(B12-$B$14)^2
13							
14	Mean of Xt=	=AVERAGE(B3:B12)					
15	∑Column D*Column E =	=SUM(F3:F10)					
16	∑(Xt-Mean)^2=	=SUM(G3:G12)					
17							
18	Auoto Correlation coefficient with two period lag =	=(B15/B16)					

Figure 13.29 Screenshot of formulas of working for autocorrelation coefficient of data with two-year lag

generally predicted positive association between these variables, it is anticipated that as rainfall increases, so will agricultural output. In rare situations, due to storm and heavy floods, the expected positive relationship between the rainfall and the agricultural output may not happen. But such an occurrence is very rare. Therefore, it is anticipated to be positive.

The formula for the covariance (COVAR) between two variables is as follows. It is the average of the products of the deviations of the means from their respective observations.

$$COVVAR(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

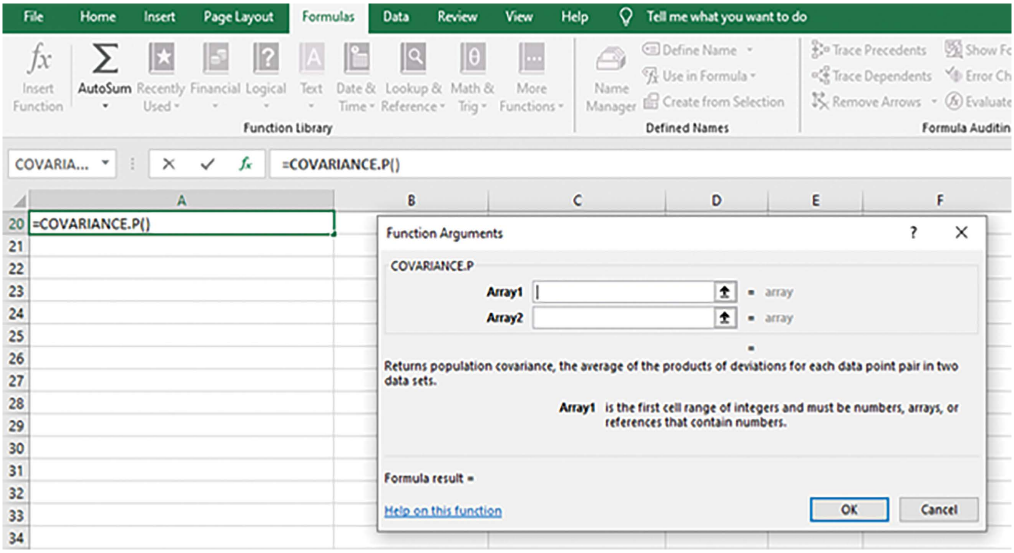
where
n is the number of pairs of observations of the variable *X* and the variable *Y*
X_i is the i^{th} observation of the variable *X*, *i* = 1, 2, 3, . . . , *n*
Y_i is the i^{th} observation of the variable *Y*, *i* = 1, 2, 3, . . . , *n*
$\bar{X}$ is the mean of the variable *X*
$\bar{Y}$ is the mean of the variable *Y*
COVARIANCE(X, Y) is the covariance between the variable *X* and the variable *Y*

In Excel, there are two functions relating to the covariance [3, 5, 6].

=COVARIANCE.P(Array 1, Array 2), which is shown in Figure 13.30, reached by clicking the sequence of buttons Formulas ⟹ More Functions ⟹ Statistical ⟹ COVARIANCE.P. This covariance is with population parameters.

=COVARIANCE.S(Array 1, Array 2), which is shown in Figure 13.31, reached by clicking the sequence of buttons Formulas ⟹ More Functions ⟹ Statistical ⟹ COVARIANCE.S. This covariance is with sample parameters.

The COVARIANCE function under the Data Analysis button under the Data button gives the covariance of two variables and further covariance of each pair of variables in a given set of variables in the form of a covariance matrix. The screenshots to use these functions are explained through an example at the end of this section.



COVARIA... =COVARIANCE.P()

20 =COVARIANCE.P()

Function Arguments

COVARIANCE.P

Array1 = array

Array2 = array

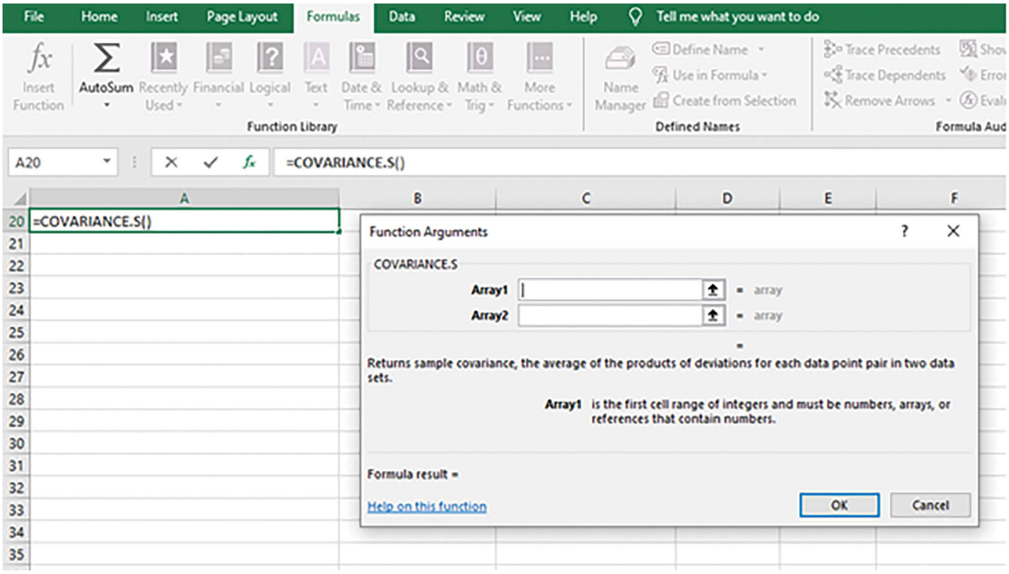
=

Returns population covariance, the average of the products of deviations for each data point pair in two data sets.

Array1 is the first cell range of integers and must be numbers, arrays, or references that contain numbers.

Formula result =

Figure 13.30 Screenshot after clicking the sequence of buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ COVARIANCE.P



A20 =COVARIANCE.S()

20 =COVARIANCE.S()

Function Arguments

COVARIANCE.S

Array1 = array

Array2 = array

=

Returns sample covariance, the average of the products of deviations for each data point pair in two data sets.

Array1 is the first cell range of integers and must be numbers, arrays, or references that contain numbers.

Formula result =

Figure 13.31 Screenshot after clicking the sequence of buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ COVARIANCE.S

In the dropdown menu of Figure 13.30/Figure 13.31, the range of cells containing the values (decimal or integer numbers) of the first variable should be entered in the cell against array 1, and the range of cells containing the values (decimal or integer numbers) of the second variable should be entered in the cell against array 2. Then the clicking the OK button in the dropdown menu of Figure 13.30/Figure 13.31 will give the result of the

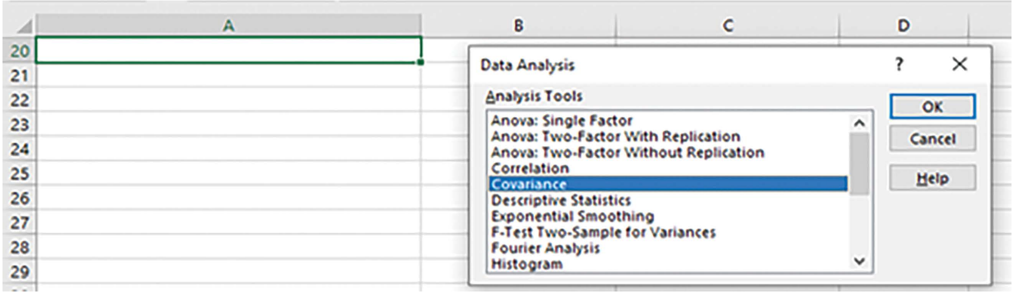
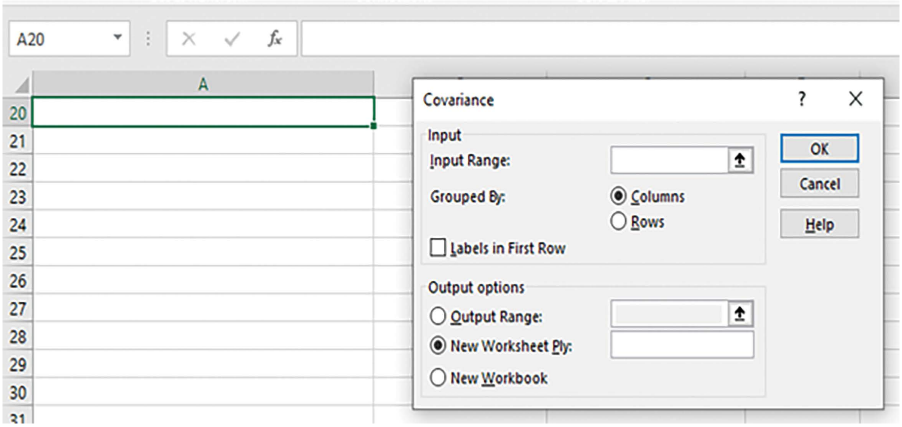


Figure 13.32 Screenshot for the sequence of button clicks, Home ⟹ Data ⟹ Data Analysis



Input

Input Range:

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 13.33 Screenshot after selecting Covariance in the dropdown menu of *Figure 13.32* and clicking the OK button

covariance between the two variables in the cell where the cursor has been located prior to clicking the sequence of buttons.

The covariance of each pair of variables in the given set of variables can be computed using the Covariance function under the Data Analysis button after clicking the Data button. A sample screenshot for this is shown in Figure 13.32. Clicking the OK button after selecting Covariance in the dropdown menu of Figure 13.32 gives the screenshot in Figure 13.33.

In Figure 13.33, the range of cells containing the data of all the variables including their column headings is to be entered in the box against Input Range. The grouped by columns button is also to be clicked, along with Labels in First Row. At the end, clicking the OK button in the dropdown menu of Figure 13.33 will give the covariance matrix in a new Excel sheet, from which the covariance between each pair of variables which are included in the range of the data can be seen.

Example 13.7

Infrastructure development spending in crores of rupees and growth in annual sales turnover in percentage of the industries in an industrial estate for the past five years are given

Table 13.14 Infrastructure Development Spending and Growth in Annual Sales Turnover of Industries

Year	*Infrastructure Development Spending (Lakhs of Rupees)*	*Growth in Annual Sales Turnover of Industries (Crores of Rupees)*
1	20	10
2	30	20
3	35	28
4	60	40
5	65	55

Table 13.15 Data for Example 13.7

Year	*Infrastructure Development Spending (Crores of Rupees)*	*Growth in Annual Sales Turnover of Industries (%)*
1	20	10
2	30	20
3	35	28
4	60	40
5	65	55

in Table 13.14. Find the covariance between the infrastructure development spending in crores of rupees and growth in annual sales turnover in percentage using Excel.

Solution

The data for Example 13.7 are shown in Table 13.15.

The formula for the covariance between two variables is as follows.

$$COVVAR(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

where

n is the number of pairs of observations of the variable X and the variable Y

X_i is the i^{th} observation of the variable X, $i = 1, 2, 3, \ldots, n$

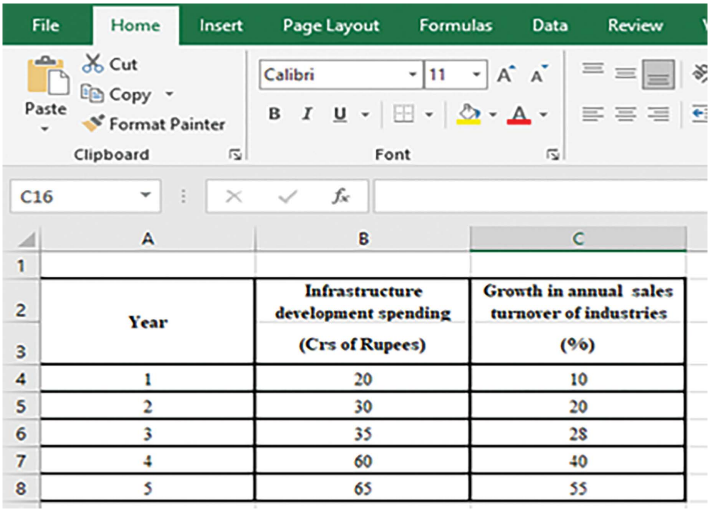
Y_i is the i^{th} observation of the variable Y, $i = 1, 2, 3, \ldots, n$

$\bar{X}$ is the mean of the variable X

$\bar{Y}$ is the mean of the variable Y

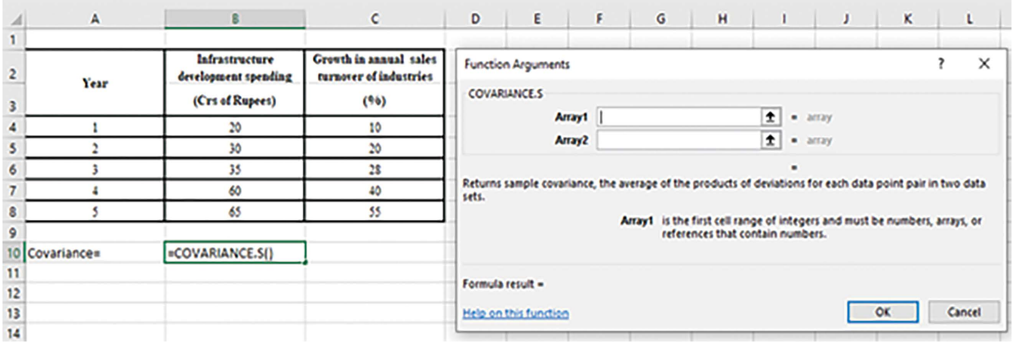
COVARIANCE(X, Y) is the covariance between the variable X and the variable Y

The input of this example in an Excel sheet is shown in Figure 13.34. The display for the sequence of button clicks Formulas ⟹ More Functions ⟹ Statistical ⟹ COVARIANCE.S is shown in Figure 13.35. The display after entering the range of cells containing data of the Infrastructure development spending in the cell against array 1 and the range of cells containing data of the growth in annual sales turnover of the industries in the cell against array 2 is shown in Figure 13.36. Clicking the OK button



	A	B	C
1			
2	Year	Infrastructure development spending	Growth in annual sales turnover of industries
3		(Crs of Rupees)	(%)
4	1	20	10
5	2	30	20
6	3	35	28
7	4	60	40
8	5	65	55

Figure 13.34 Screenshot of input of Example 13.7



	A	B	C
1			
2	Year	Infrastructure development spending	Growth in annual sales turnover of industries
3		(Crs of Rupees)	(%)
4	1	20	10
5	2	30	20
6	3	35	28
7	4	60	40
8	5	65	55
9			
10	Covariance=	=COVARIANCE.S()	

Function Arguments

COVARIANCE.S

Array1 = array

Array2 = array

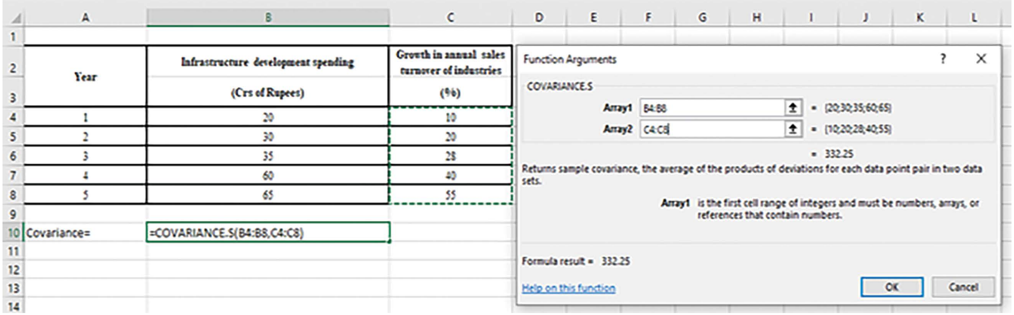
=

Returns sample covariance, the average of the products of deviations for each data point pair in two data sets.

Array1 is the first cell range of integers and must be numbers, arrays, or references that contain numbers.

Formula result =

Figure 13.35 Screenshot of display after clicking buttons, Formulas ⟹ More Functions ⟹ Statistical ⟹ COVARIANCE.S



	A	B	C
1			
2	Year	Infrastructure development spending	Growth in annual sales turnover of industries
3		(Crs of Rupees)	(%)
4	1	20	10
5	2	30	20
6	3	35	28
7	4	60	40
8	5	65	55
9			
10	Covariance=	=COVARIANCE.S(B4:B8,C4:C8)	

Function Arguments

COVARIANCE.S

Array1 B4:B8 = {20;30;35;60;65}

Array2 C4:C8 = {10;20;28;40;55}

= 332.25

Returns sample covariance, the average of the products of deviations for each data point pair in two data sets.

Array1 is the first cell range of integers and must be numbers, arrays, or references that contain numbers.

Formula result = 332.25

Figure 13.36 Screenshot after entering the ranges of cells containing data of two variables in the dropdown menu of Figure 13.35

in the dropdown menu of Figure 13.36 gives the result of the covariance between the two stated variables in cell B12 of Figure 13.37. The covariance between the two variables is 332.25.

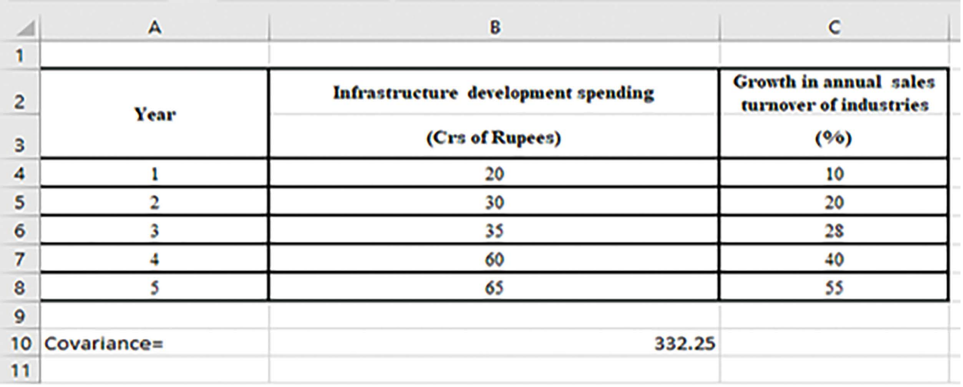
Example 13.8 (For Covariance Matrix Using Covariance Function Under Data Analysis Function)

Take a look at the data in Table 13.16 for the variables infrastructure development spending (lakhs of ₹), inflation (%), and growth in annual sales turnover of industries (crores of ₹). Using the Covariance function in Excel's Data Analysis feature, determine the covariance matrix of these variables.

Solution

The screenshot of the data for Example 13.8 is shown in Figure 13.38. The sequence of button clicks to obtain the covariance function under the Data Analysis button is as follows, which gives the screenshot in Figure 13.39.

Home ⟹ Data ⟹ Data Analysis

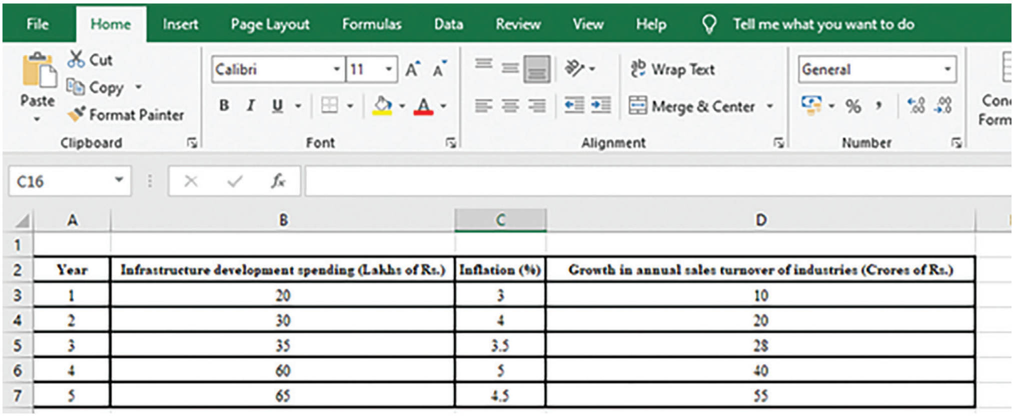


	A	B	C
1			
2	Year	Infrastructure development spending	Growth in annual sales turnover of industries
3		(Crs of Rupees)	(%)
4	1	20	10
5	2	30	20
6	3	35	28
7	4	60	40
8	5	65	55
9			
10	Covariance=	332.25	
11			

Figure 13.37 Screenshot after clicking the OK button in the dropdown menu of *Figure 13.36* to show result of COVARIANCE.S

Table 13.16 Data for Example 13.8

Year	*Infrastructure Development Spending (Lakhs of Rupees)*	*Inflation (%)*	*Growth in Annual Sales Turnover of Industries (Crores of Rupees)*
1	20	3	10
2	30	4	20
3	35	3.5	28
4	60	5	40
5	65	4.5	55



Year	Infrastructure development spending (Lakhs of Rs.)	Inflation (%)	Growth in annual sales turnover of industries (Crores of Rs.)
1	20	3	10
2	30	4	20
3	35	3.5	28
4	60	5	40
5	65	4.5	55

Figure 13.38 Screenshot of data for Example 13.8

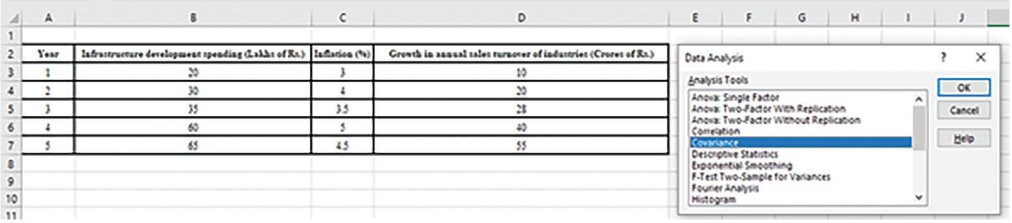
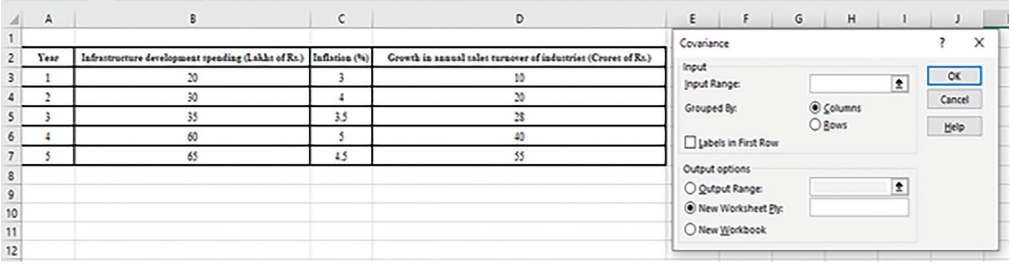

Year	Infrastructure development spending (Lakhs of Rs.)	Inflation (%)	Growth in annual sales turnover of industries (Crores of Rs.)
1	20	3	10
2	30	4	20
3	35	3.5	28
4	60	5	40
5	65	4.5	55

Figure 13.39 Screenshot for the sequence of button clicks, Home ⟹ Data<Insert Arrow> Data Analysis


Year	Infrastructure development spending (Lakhs of Rs.)	Inflation (%)	Growth in annual sales turnover of industries (Crores of Rs.)
1	20	3	10
2	30	4	20
3	35	3.5	28
4	60	5	40
5	65	4.5	55

Covariance

Input

Input Range:

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

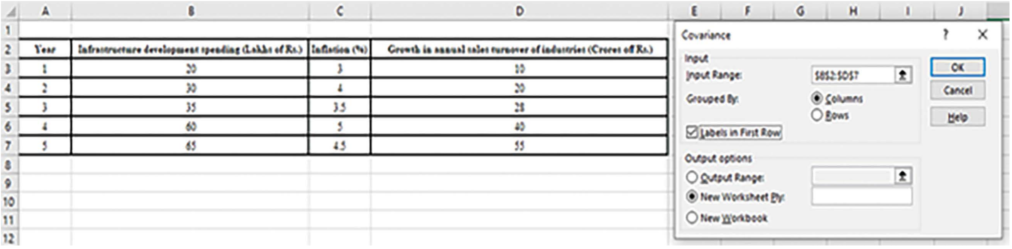
○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 13.40 Screenshot after clicking Covariance function and OK button in the dropdown menu of Figure 13.39

Selecting the Covariance button in the dropdown menu of Figure 13.39 and clicking OK gives the screenshot in Figure 13.40. The screenshot after filling the range of cells $B$2 to $D$7 in the box against Input Range, clicking Columns against Grouped By, and clicking Labels in First Row in the dropdown menu of Figure 13.40 is shown in Figure 13.41 Clicking the OK button in the dropdown menu of Figure 13.41 gives the covariance matrix among the variables, that is infrastructure development spending (lakhs of ₹), inflation (%) and growth in annual sales turnover of industries (crores of ₹) of a company, as shown in Figure 13.42.



Year	Infrastructure development spending (Lakhs of Rs.)	Inflation (%)	Growth in annual sales turnover of industries (Crores of Rs.)
1	20	3	10
2	30	4	20
3	35	3.5	28
4	60	5	40
5	65	4.5	55

Covariance

Input

Input Range: $B$2:$D$7

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

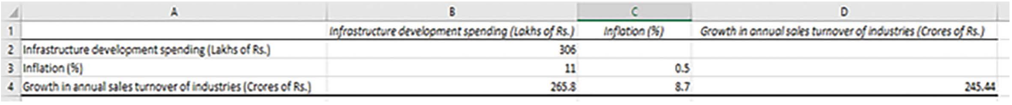
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 13.41 Screenshot after filling data in the dropdown menu of Figure 13.40



	Infrastructure development spending (Lakhs of Rs.)	*Inflation (%)*	*Growth in annual sales turnover of industries (Crores of Rs.)*
Infrastructure development spending (Lakhs of Rs.)	306		
Inflation (%)	11	0.5	
Growth in annual sales turnover of industries (Crores of Rs.)	265.8	8.7	245.44

Figure 13.42 Screenshot after clicking the OK button in the dropdown menu of Figure 13.41

Summary

- The degree of association between two variables is measured in terms of a correlation coefficient. The range of values of the correlation coefficient is from –1 to + 1.
- Consider a situation which deals with the ranks given by two different judges for *n* units of a product of interest. Now, the objective is to check whether these *n* pairs of ranks are correlated. This correlation coefficient is also called Spearman's correlation coefficient (r_s).
- Autocorrelation deals with the correlation of a given set of data with another set of data, which is derived from the first set of data with specified periods of lag.
- Covariance is another measure to study the relationship between two variables of interest. If the value of the covariance is positive, then the two variables are positively related, and if it is negative, then the two variables are negatively related.
- COVARIANCE(X, Y) is the covariance between the variable *X* and the variable *Y*.
- The Covariance function under Data Analysis gives covariance in the form of a correlation matrix for each pair of variables in a given set of variables.

Keywords

- Correlation coefficient is defined as the degree of association between two variables. The range of the correlation coefficient is from –1 to + 1.
- Spearman's correlation coefficient (r_s) is the degree of association between two different streams of ranks. An example of the ranks may be the ranks given by two different judges for *n* units of a product of interest.
- Autocorrelation deals with the correlation of a given set of data with another set of data, which is derived from the first set of data with a specified periods of lag.
- Covariance is another measure to study the relationship between two variables of interest. If the value of the covariance is positive, then the two variables are positively related, and if it is negative, then the two variables are negatively related.

- Covariance is the average of the products of the deviations of the means from their respective observations.

Review Questions

1. Define correlation coefficient and explain its range.
2. Give the formula for the correlation coefficient of ungrouped data and explain its variables and constants.
3. The annual R&D expenditure (lakhs of rupees) and the corresponding sales revenue (crores of rupees) of a company for the past ten years are shown in the following table.

   Find the correlation coefficient between the R&D expenditure and the annual revenue.

Year (i)	*Annual R&D (X_i)*	*Annual Revenue (Y_i)*
1	50	700
2	70	800
3	90	750
4	100	950
5	120	1050
6	150	1800
7	180	950
8	200	1600
9	225	1700
10	250	1850

   a. Using your own Excel sheet.
   b. Using an inbuilt function of Excel.

4. Using an appropriate Excel example, describe how to create a correlation matrix for each pair of variables in a given set of variables.
5. Give the formula for the correlation coefficient of grouped data and explain its variables and constants.
6. The following table shows data that a consultant gathered summarising the frequencies of employees based on the classification years of experience and skill index on a scale of 0 to 10. Using Excel, get the correlation coefficient for this set of grouped data.

Skill Index of Employee	*Experience of Employee in Years*			
	20–30	*30–40*	*40–50*	*50–60*
0–2	7	8	9	9
2–4	5	9	8	6
4–6	6	7	10	7
6–8	9	3	6	6
8–10	10	9	8	8

7. What is rank correlation? Give its equation and hypotheses.
8. The rankings of ten employees using two performance appraisal systems are shown in the following table. Compute the rank correlation coefficient of these data using Excel.

Employee	*Rank*	
	Performance Appraisal System 1	*Performance Appraisal System 2*
1	5	8
2	7	10
3	7	8
4	10	6
5	9	6
6	4	8
7	6	10
8	7	5
9	5	8
10	9	5

9. The yearly employee welfare fund spent in a company for the past ten years is given in the following table. Find the auto-correlation coefficients with one-year lag (r_1) and two-year lag (r_2) using Excel.

Year (i)	*Welfare fund in thousands of rupees* (X_1)
1	400
2	430
3	520
4	600
5	780
6	750
7	870
8	920
9	950
10	1100

10. What is covariance? Give its equation.
11. Give the formula for COVARIANCE in Excel and explain its arguments.
12. The following table shows the training expenditure in lakhs of rupees and the sales turnover in crores of rupees for a company during the previous five years. Using Excel, calculate the correlation between the training expenditure in lakhs of rupees and the sales turnover in crores of rupees.

Year	*Training Expenditure (Lakhs of Rupees)*	*Sales Turnover (Crores of Rupees)*
1	70	50
2	100	55
3	140	80
4	160	100
5	200	170

13. Illustrate the process of obtaining a covariance matrix between each pair of variables in a given set of variables using a suitable example in Excel.

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.

2. www.ablebits.com/office-addins-blog/2019/01/30/spearman-rank-correlation-Excel/ [June 25, 2020].
3. https://corporatefinanceinstitute.com/resources/Excel/functions/covariance-Excel-function/ [June 25, 2020].
4. www.Excel-easy.com/examples/correlation.html [June 27, 2020].
5. https://support.microsoft.com/en-us/office/covariance-p-function-6f0e1e6d-956d-4e4b-9943-cfef0bf9edfc [June 27, 2020].
6. https://support.microsoft.com/en-us/office/covariance-s-function-0a539b74-7371-42aa-a18f-1f5320314977 [June 27, 2020].

14 Forecasting

Learning Objectives

After reading this chapter, the readers will be able to

- Understand the concept of forecasting and its applications.
- Know the methods of forecasting.
- Design a model for linear regression using formulas in the conventional way.
- Design a model for linear regression using Excel.
- Analyse data using the moving average method of forecasting using Excel.
- Understand the working of the exponential method of forecasting.
- Develop a model for multiple linear regression.

14.1 Introduction

In actuality, making predictions about specific events is inevitable since doing so enables businesses to arrange their operations in accordance with societal needs, including those for food, clothes, housing, and other services. Take the example of transportation, which has integrated into everyone's daily lives. The number of automobile firms expanded rapidly as a result. The issue of market share is crucial in determining capacity and other associated activities like production and marketing because there are many companies in the automotive sector. Therefore, forecasting demand for a company's manufactured vehicle becomes a crucial task.

Such a prediction, which is made using historical data, is known as demand forecasting. Regression modelling is the process of creating a model from historical data by using the year as an independent variable and the demand as the dependent variable. The dependent variable, also known as the variable of prediction, and one or more independent variables, such as the year, gross domestic product, population size, inflation rate, and so on, are related by the regression model.

14.2 Methods of Forecasting

There are various methods of forecasting, which are classified into qualitative and quantitative forecasting methods.

The Delphi method is a qualitative forecasting technique that tries to forecast the future states of qualitative events, such as societal culture, computer technology advancement, personal lifestyle, and value system.

DOI: 10.4324/9781032671703-14

There are several quantitative forecasting techniques, including regression, exponential smoothing, and moving averages [1].

The moving average approach seeks to predict an item's demand in the near future, say, based on demand over the previous three or four weeks.

According to the demand for an item during the present period, the exponential smoothing method of forecasting attempts to predict demand for that item during the following period.

In order to create an equation for the dependent variable in terms of the presumed independent variables, a regression model is used. Linear and nonlinear regression are further categories for regression models. Each independent variable's power in a linear regression model is one. For a few examples of observations, this can be shown as a linear line with a specific slope. The independent variables' power in a nonlinear regression model can be any value. This can be seen by its plot, which takes the form of a nonlinear curve for some sample observations.

14.3 Linear Regression

As stated earlier, regression is defined as the dependence of a variable of interest (dependent variable) on one or more other variables (independent variables) [2].

A sample linear regression equation is as follows.

$$Y = a + bX$$

where
Y is the dependent variable
X is the independent variable
a is the intercept
b is the trend (slope)

One should note the fact that the number of independent variables in this equation is only one; hence, it is called a simple linear regression model. If the number of independent variables is more than one and each of them has a power of unity, then the corresponding regression equation is called a multiple linear regression model.

As stated earlier, X can be assumed as the time period in years and Y can be assumed as the demand for a product. A sample plot of the simple linear regression is shown in Figure 14.1.

In Figure 14.1, the X axis represents the independent variable X, and the Y axis represents the dependent variable Y. The inclined line in Figure 14.1 is the plot of the simple linear regression model, $Y = a + bX$, which intersects the Y axis at a distance of a units from the origin. The distance between the origin and the point of intersection of the regression line on the Y axis is called the intercept, which is a. This is the constant term of the regression model, which means that the value of Y is a when the value of X is 0. The slope of the regression line in the X-Y plane is b, which is the coefficient of the variable X in the regression model.

The formulas to compute the constants, that is, a and b of the model, are as follows.

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

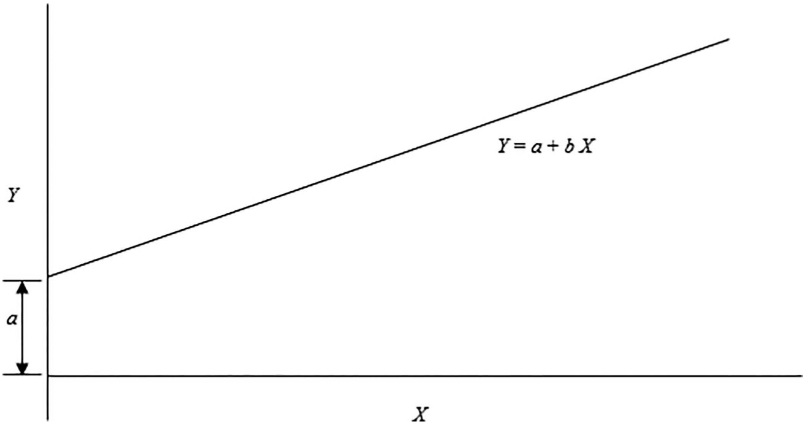


Figure 14.1 Sample plot of linear regression model

$$a = \bar{Y} - b\bar{X}$$

$$where, \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \; and \; \bar{Y} = \frac{\sum_{i=1}^{n} Y_I}{n}$$

and n is the number of pairs of observations made.

Note: When $\sum_{i=1}^{n} X_i = 0$, then the formulas for b and a are reduced as follows.

$$b = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \; and \, a = \bar{Y}$$

14.4 Linear Regression Using Excel Regression Function

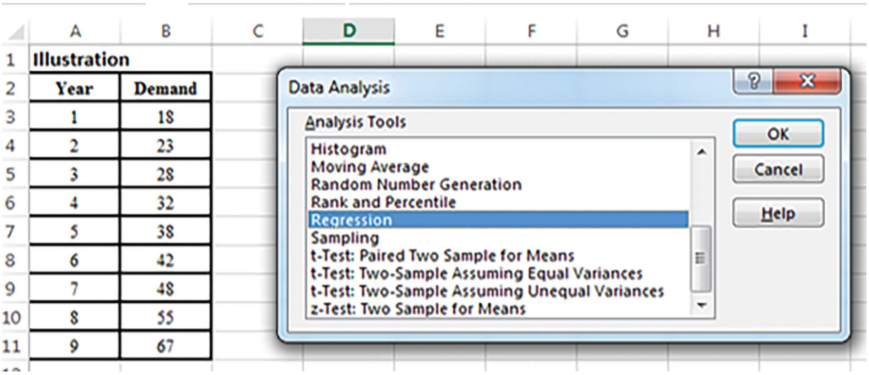
Consider the data shown in Table 14.1, which contains the year and sales of a product. Design a linear regression model to estimate the demand of the product for a given year.

The results of the linear regression can be obtained using Excel. The steps of solving it using Excel are as follows.

Step 1: Input the data in the Excel sheet.
Step 2: Select the Data button and Data Analysis button in the next level to show the screenshot in Figure 14.2.
Step 3: Select the option Regression from the dropdown menu of Figure 14.2, which gives the display shown in Figure 14.3 [3].
Step 4: Input the ranges of *Y* and *X*, level of significance, output range/New Worksheet, and other data in the cells of the dropdown menu of Figure 14.3.
Step 5: Click the OK button in the screenshot shown in Figure 14.3 after entering the data, which will give the output in the output range of the same worksheet or in a new worksheet.

Table 14.1 Sales of Product

Year	*Demand*
1	18
2	23
3	28
4	32
5	38
6	42
7	48
8	55
9	67

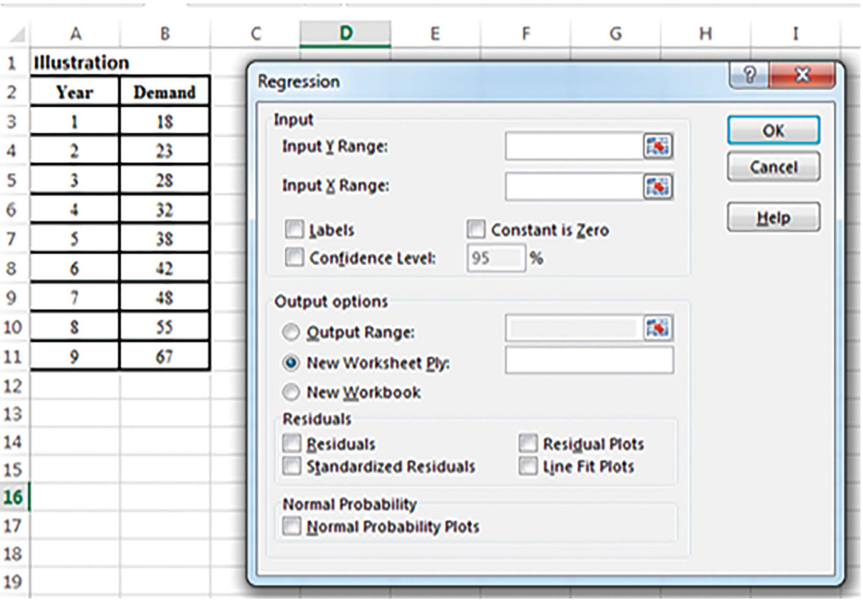


Illustration

Year	Demand
1	18
2	23
3	28
4	32
5	38
6	42
7	48
8	55
9	67

A
B
C
D
E
F
G
H
I
Data Analysis
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

Figure 14.2 Screenshot after clicking Data button and then clicking Data Analysis button



Illustration

Year	Demand
1	18
2	23
3	28
4	32
5	38
6	42
7	48
8	55
9	67

Regression

Input

Input Y Range:

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Figure 14.3 Screenshot after clicking regression option in the dropdown menu shown in Figure 14.2

Step 6: Check the value of p for the model in the ANOVA table (second table of the output). If it is less than 0.05, accept the regression model and then read the intercept and the slope of the independent variable from the second column of the third table of the output to design the regression model $Y = a + b\ X$; otherwise, state that the regression model is not adequate to predict the value of Y.

Example 14.1

The demand values of a product manufactured by Alpha Engineering Company for the past nine years are given in Table 14.2. Design a linear regression model to estimate the demand of the product using Excel.

Solution

The data for Example 14.1 are shown in Table 14.3.

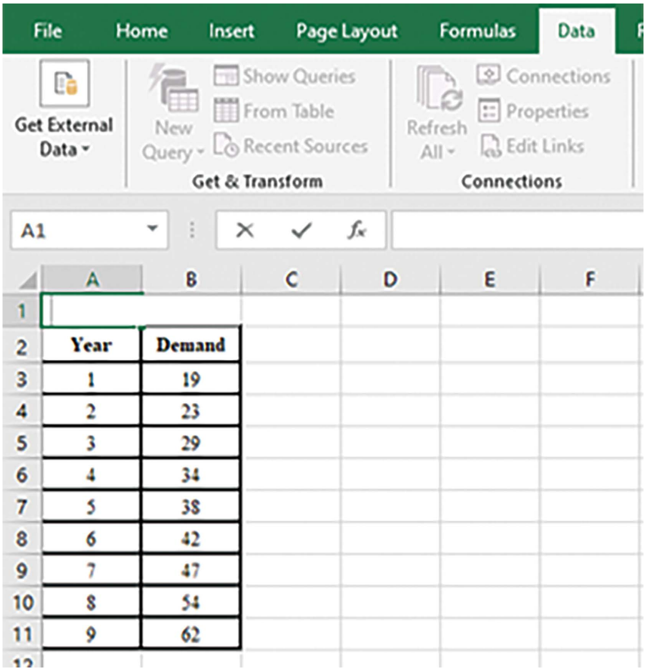
Step 1: Input the data in the Excel sheet as shown in the screenshot of Figure 14.4.
Step 2: Select the Data button and Data Analysis button in the next level to show the screenshot shown in Figure 14.5.
Step 3: Select the option Regression from the dropdown menu of Figure 14.5, which gives the display shown in Figure 14.6

Table 14.2 Data for Example 14.1

Year	*Demand*
1	19
2	23
3	29
4	34
5	38
6	42
7	47
8	54
9	62

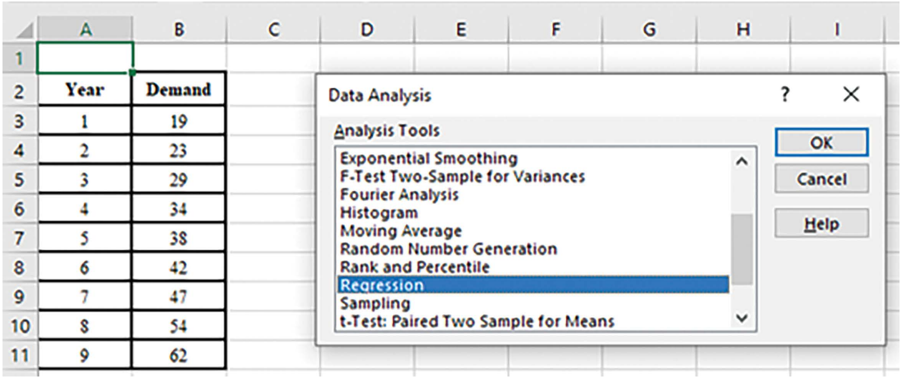
Table 14.3 Data for Example 14.1

Year	*Demand*
1	19
2	23
3	29
4	34
5	38
6	42
7	47
8	54
9	62



Year	Demand
1	19
2	23
3	29
4	34
5	38
6	42
7	47
8	54
9	62

Figure 14.4 Screenshot of input of data for Example 14.1



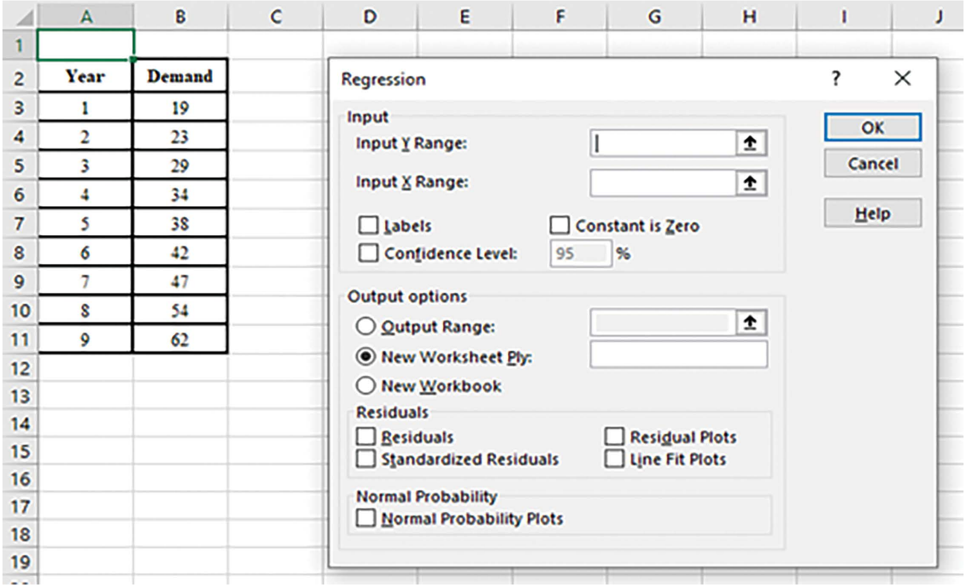
Year	Demand
1	19
2	23
3	29
4	34
5	38
6	42
7	47
8	54
9	62

Figure 14.5 Screenshot after clicking Data button and Data Analysis button

Step 4: Input the range of *Y* as B2:B11, range of *X* as A2:A11, click labels, retain the confidence level as 0.95 (1 – significance level of 0.05), and click New Worksheet in the dropdown menu of Figure 14.6, as shown in the screenshot shown in Figure 14.7.

Step 5: Click the OK button in the screenshot shown in Figure 14.7, which gives the output in a new worksheet as shown in Figure 14.8.

Step 6: Check the value of *p* for the model in the ANOVA table, which is in the second table of the output shown in Figure 14.8. Since it is less than 0.05, accept the regression model. Then read the intercept from the first row and the slope of the independent



Year	Demand
1	19
2	23
3	29
4	34
5	38
6	42
7	47
8	54
9	62

Regression

Input

Input Y Range:

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

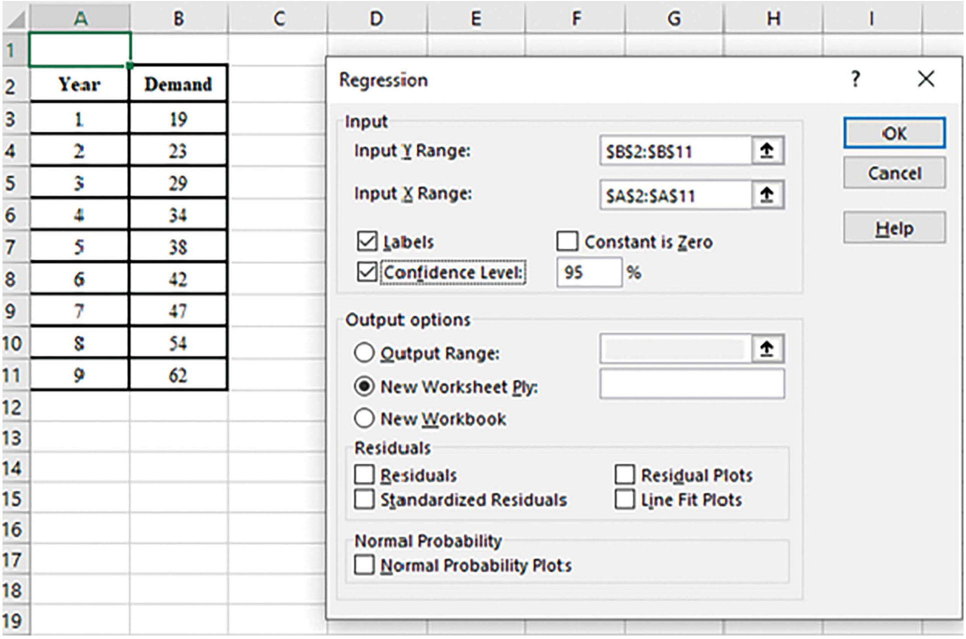
□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Figure 14.6 Screenshot in response to clicking Regression option in Figure 14.5



Year	Demand
1	19
2	23
3	29
4	34
5	38
6	42
7	47
8	54
9	62

Regression

Input

Input Y Range: $B$2:$B$11

Input X Range: $A$2:$A$11

☑ Labels □ Constant is Zero

☑ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

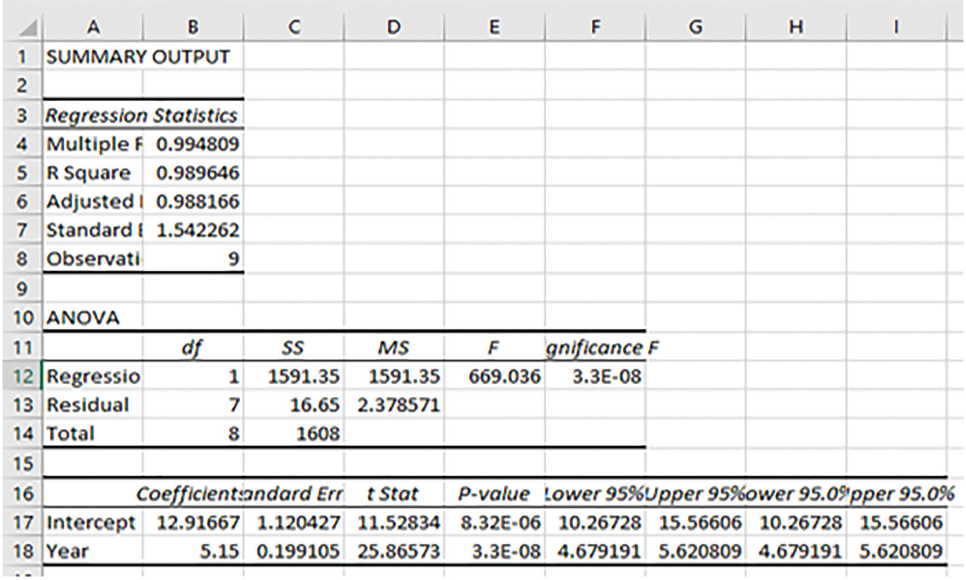
□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Figure 14.7 Screenshot after filling in necessary cells in the dropdown menu of Figure 14.6



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple F	0.994809							
5	R Square	0.989646							
6	Adjusted I	0.988166							
7	Standard I	1.542262							
8	Observati	9							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regressio	1	1591.35	1591.35	669.036	3.3E-08			
13	Residual	7	16.65	2.378571					
14	Total	8	1608						
15									
16		*Coefficient*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	12.91667	1.120427	11.52834	8.32E-06	10.26728	15.56606	10.26728	15.56606
18	Year	5.15	0.199105	25.86573	3.3E-08	4.679191	5.620809	4.679191	5.620809

Figure 14.8 Screenshot after clicking the OK button in the dropdown menu of Figure 14.7

variable Year, which is X from the second row of the third table in Figure 14.8 to design the regression model $Y = a + b\ X$, which is as follows.

$$Y = 12.91667 + 5.15X$$

14.5 Moving Average Method Using Moving Average Function in Excel

The value of a variable of interest is predicted using the moving average method of forecasting for the near term, such as a week or month. Using the moving average method of forecasting, it is possible to determine the hourly power consumption requirement in a mega-building.

Let Table 14.4's values for a product's demand be the case. The three-period moving average formula is as follows.

$$MA_i = \frac{(D_{i-2} + D_{i-1} + D_i)}{3}, i = 3, 4, 5, \ldots, , n$$

where

MA_{i} is the moving average for the period i, $i = 3, \ldots, n$

D_{i} is the demand of the period i, $i = 4, \ldots, n + 1$

n is the number of periods for which data are available.

The forecast of the period $i + 1$ is as follows.

$$F_{i+1} = MA_i \ for\, i = 3, 4, 5, , \ldots, n$$

Table 14.4 Demand Data, Moving Averages (MA_i), and Forecasts (F_{i+1})

Week	*Demand*	*Moving Average (MA_i)*	*Forecast (F_{i+1})*
1	D_1	–	–
2	D_2	–	–
3	D_3	$(D_1 + D_2 + D_3)/3$	–
4	D_4	$(D_2 + D_3 + D_4)/3$	$(D_1 + D_2 + D_3)/3$
5	D_5	$(D_3 + D_4 + D_5)/3$	$(D_2 + D_3 + D_4)/3$
6	D_6	$(D_4 + D_5 + D_6)/3$	$(D_3 + D_4 + D_5)/3$
7	D_7	$(D_5 + D_6 + D_7)/3$	$(D_4 + D_5 + D_6)/3$
8	D_8	$(D_6 + D_7 + D_8)/3$	$(D_5 + D_6 + D_7)/3$
9	D_9	$(D_7 + D_8 + D_9)/3$	$(D_6 + D_7 + D_8)/3$
10	D_{10}	$(D_8 + D_9 + D_{10})/3$	$(D_7 + D_8 + D_9)/3$
11	–	–	$(D_8 + D_9 + D_{10})/3$

A time interval of three or four periods will be used in the moving average method of forecasting. The average of the demand values from the first three periods is used as the moving average for period 3 and as the forecast value for period 4 in the three-period moving average. The forecast for period 5 is the moving average of periods 2 to 4, which is the average of the demand numbers for those periods. This reasoning keeps going until the third observation in the moving average is the last available data.

Example 14.2

Table 14.5 provides a summary of a product's monthly demand values. Using Excel, calculate the moving averages for periods 3 through 10, which will serve as the forecast values for periods 4 through 11, respectively.

Solution

The monthly demand data for Example 14.2 are shown in Table 14.6.

The computation of moving averages of the demand values using Excel is illustrated in the following steps.

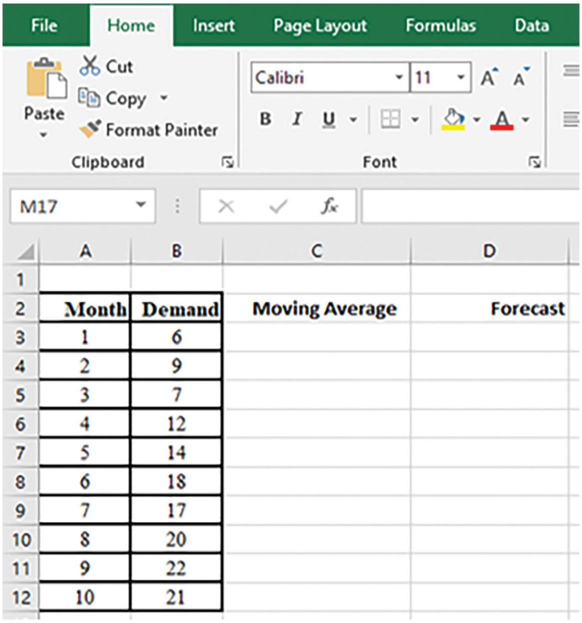
Step 1: Input the data of the given problem in an Excel sheet, as shown in Figure 14.9. In Figure 14.9, the column headings Moving Average and Forecast are inserted in cells C2 and D2, respectively, because the headings will not be present in the output of Excel. In the third column, cells C3 to C12 will be treated as the output range. Column D will have forecast values, which are to be copied manually at the end from cell D6 to D13.
Step 2: Click the Data button and Analyse button to show the screenshot in Figure 14.10.
Step 3: Select the option Moving Average in the dropdown menu of Figure 14.10 and then click OK to show the screenshot in Figure 14.11 [4].
Step 4: Enter the cells of the input range [B2:B12], click the checkbox Labels in First Row, enter the value of the interval as 3, and enter the range of cells of the output range as C3:C12 in the dropdown menu of Figure 14.11 as shown in Figure 14.12.
Step 5: Click the OK button in the dropdown menu of Figure 14.12 to show the output in column C from C3 to C12 as shown in Figure 14.13.
Step 6: In Figure 14.13, treat the moving average in column C for row *i*, MA_i, as the forecast in column D for row *i*+1, F_{i+1}, where row *i* = 5, 2, 3, . . . , 12 as shown in Figure 14.13.

Table 14.5 Monthly Demand Data

Month	*Demand*
1	6
2	9
3	7
4	12
5	14
6	18
7	17
8	20
9	22
10	21

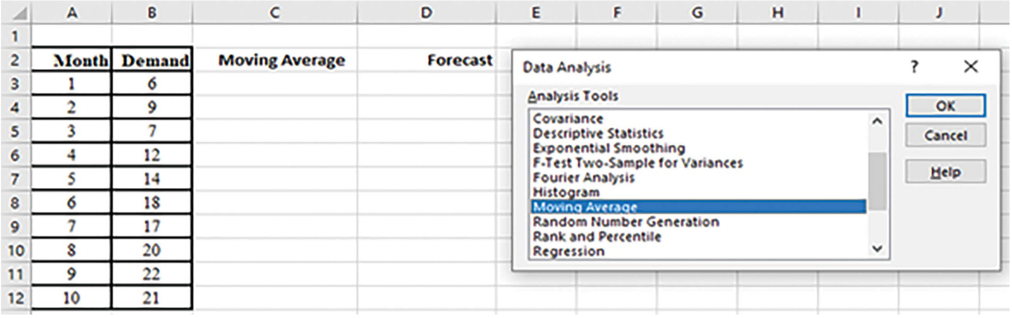
Table 14.6 Data for Example 14.2

Month	*Demand*
1	6
2	9
3	7
4	12
5	14
6	18
7	17
8	20
9	22
10	21



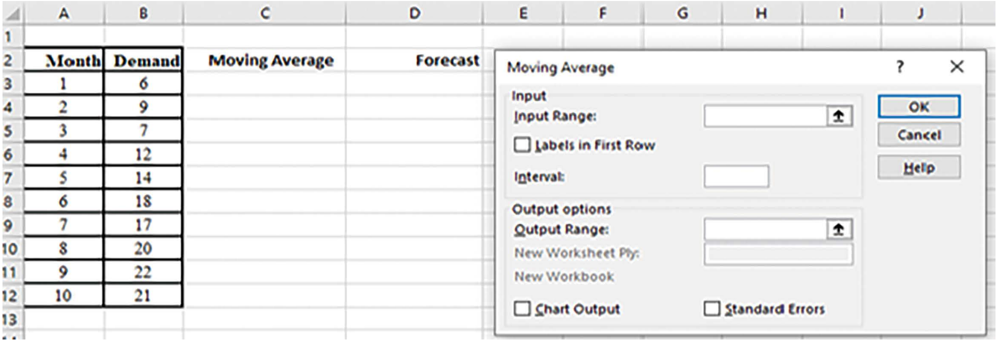
Month	Demand	Moving Average	Forecast
1	6		
2	9		
3	7		
4	12		
5	14		
6	18		
7	17		
8	20		
9	22		
10	21		

Figure 14.9 Screenshot of input of data for Example 14.2



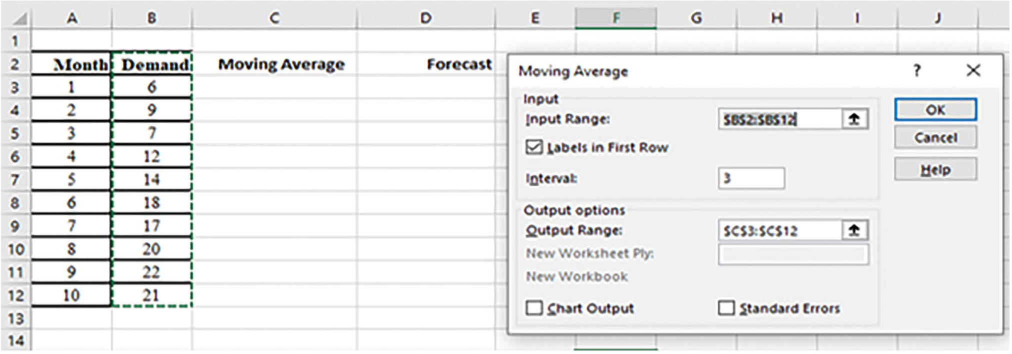
Month	Demand	Moving Average	Forecast
1	6		
2	9		
3	7		
4	12		
5	14		
6	18		
7	17		
8	20		
9	22		
10	21		

Figure 14.10 Screenshot after clicking Data button and then Data Analysis button



Month	Demand	Moving Average	Forecast
1	6		
2	9		
3	7		
4	12		
5	14		
6	18		
7	17		
8	20		
9	22		
10	21		

Figure 14.11 Screenshot after selecting Moving Average and clicking the OK button in Figure 14.10



Month	Demand	Moving Average	Forecast
1	6		
2	9		
3	7		
4	12		
5	14		
6	18		
7	17		
8	20		
9	22		
10	21		

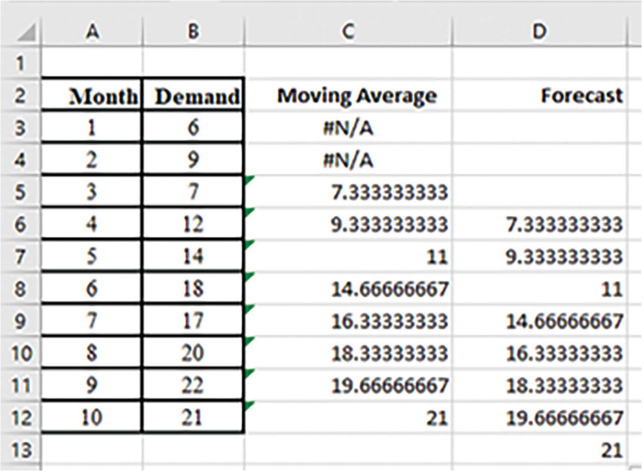
Input Range: $B$2:$B$12

☑ Labels in First Row

Interval: 3

Output Range: $C$3:$C$12

Figure 14.12 Screenshot after entering input range, clicking the checkbox Labels in Row, entering 3 as the value for the interval, and entering the cells of the output range in Figure 14.11



	A	B	C	D
1				
2	Month	Demand	Moving Average	Forecast
3	1	6	#N/A	
4	2	9	#N/A	
5	3	7	7.333333333	
6	4	12	9.333333333	7.333333333
7	5	14	11	9.333333333
8	6	18	14.66666667	11
9	7	17	16.33333333	14.66666667
10	8	20	18.33333333	16.33333333
11	9	22	19.66666667	18.33333333
12	10	21	21	19.66666667
13				21

Figure 14.13 Screenshot after clicking OK button in Figure 14.12 (result)

14.6 Exponential Smoothing Method of Forecasting Using Excel Sheets and Exponential Smoothing Function

Based on the demand and forecast from the immediate previous period, the exponential smoothing method of forecasting is used to estimate the demand for the current period. The exponential smoothing method of forecasting can be used to estimate the demand for liquor, hotel rooms, and so on. The time period of these realities will be in days/weeks. The formula for predicting using exponential smoothing is provided as follows.

$$F_t = F_{t-1} + \alpha\left(D_{t-1} - F_{t-1}\right)$$

where
F_{t} is the forecast of the period t
$F_{\mathrm{t-1}}$ is the forecast of the period $t-1$
$D_{\mathrm{t-1}}$ is the demand of the period $t-1$
α is a smoothing constant whose value will be more than 0 and less than 0.3

The exponential smoothing method can be carried out using the Exponential Smoothing function under Data Analysis, for which the sequence of button clicks is Home $\Longrightarrow$ Data $\Longrightarrow$ Data Analysis. The corresponding screenshot is shown in Figure 14.14. Selecting the Exponential Smoothing function in the dropdown menu of Figure 14.14 and then clicking the OK button will give a display as shown in Figure 14.15. For a given problem, fill in the linear range of cells plus one more empty successive cell containing the data for the independent variable *X* (actual demand values); damping factor, which is given by 1 – α, where α is the smoothing constant, a positive constant greater than 0 and less than or equal to 0.3; and the output range, which corresponds to the first address in the next row of the actual data, which specifies the beginning address for the forecast values.

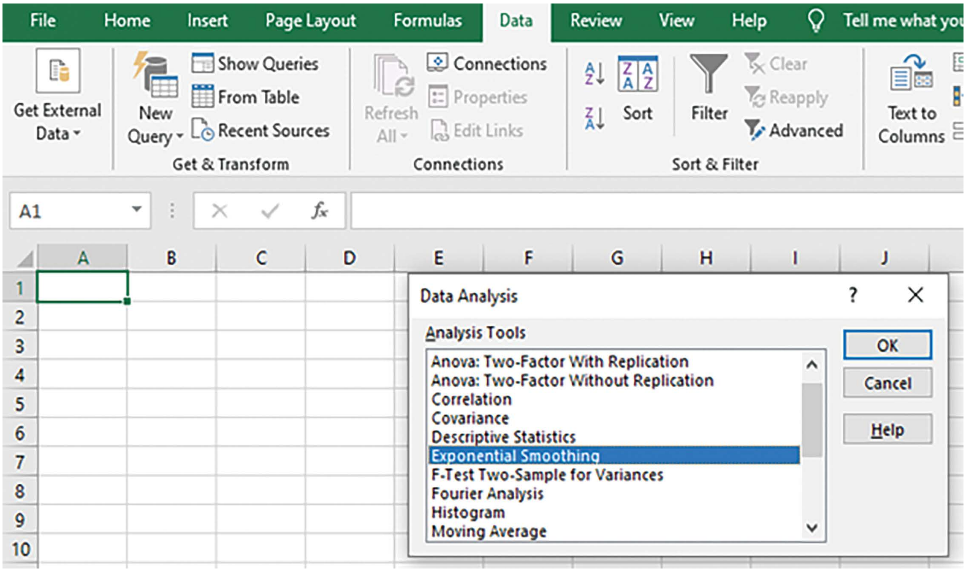


Figure 14.14 Screenshot after sequence of button clicks, Home ⟹ Data ⟹ Data Analysis

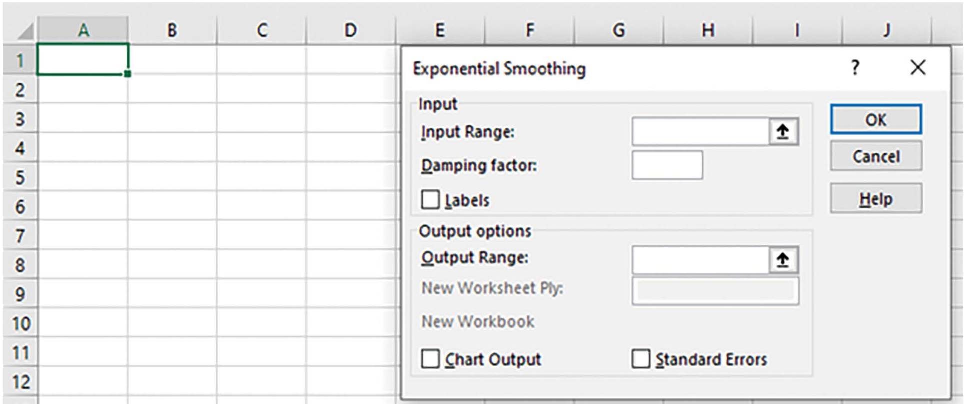


Figure 14.15 Screenshot after selecting Exponential Smoothing and clicking the OK button in the dropdown menu of Figure 14.14

Then clicking the OK button in the same dropdown menu will give the forecasted values starting from the address given in the output range.

Example 14.3

A firm uses simple exponential smoothing with $\alpha = 0.3$ to forecast demand. The forecast for the first week of January was 500 units, whereas actual demand turned out to be 450 units.

1. Forecast the demand for the second week of January.
2. Assume that the actual demand during the second week of January turned out to be 550 units. Forecast the demand up to the third week of February, assuming the subsequent demands as 475, 450, 470, 525, and 470 units.

Solution

Smoothing constant, $\alpha = 0.3$

A sample calculation to forecast the demand for week 2 is shown as follows.

Forecast for the first week of January, F_1 = 500 units

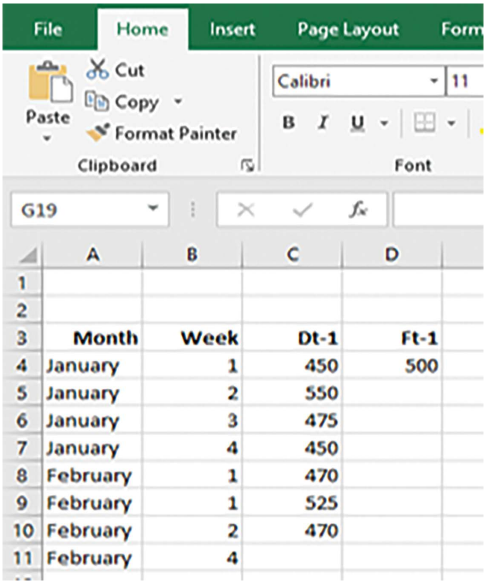
Demand for the first week of January, D_1 = 450 units

$$F_t = F_{t-1} + \alpha\left(D_{t-1} - F_{t-1}\right)$$

$$F_2 = F_1 + \alpha\left(D_1 - F_1\right) = 500 + 0.3 \times (450 - 500) = 485$$

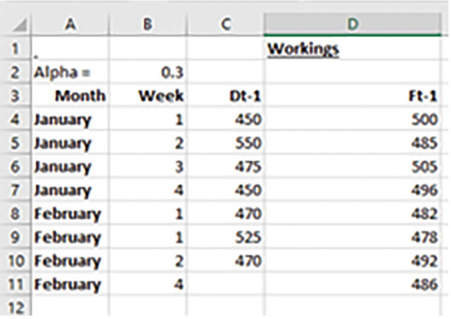
The given data are summarised in tabular format in Excel, and the corresponding screenshot is as shown in Figure 14.16.

The working of the exponential smoothing method of forecasting applied to the data in Figure 14.16 is shown in the screenshot shown in Figure 14.17 [5]. In this figure, the forecast values for week 2 of January to week 4 of February are shown from cells D5 to D11, respectively. The screenshot of the guidelines for the formulas of the Excel sheet shown in Figure 14.17 is shown in Figure 14.18.



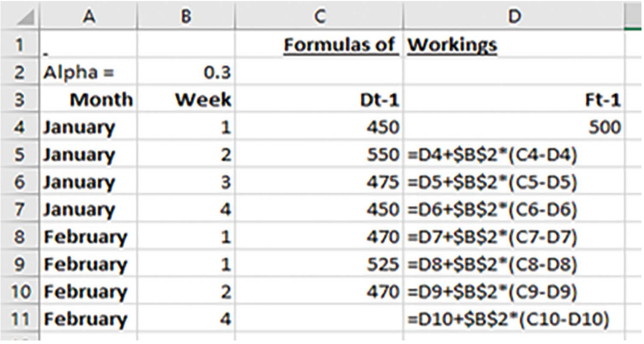
	A	B	C	D
1				
2				
3	Month	Week	Dt-1	Ft-1
4	January	1	450	500
5	January	2	550	
6	January	3	475	
7	January	4	450	
8	February	1	470	
9	February	1	525	
10	February	2	470	
11	February	4		

Figure 14.16 Screenshot of tabular format of given data



	A	B	C	D
1	-			Workings
2	Alpha =	0.3		
3	Month	Week	Dt-1	Ft-1
4	January	1	450	500
5	January	2	550	485
6	January	3	475	505
7	January	4	450	496
8	February	1	470	482
9	February	1	525	478
10	February	2	470	492
11	February	4		486
12				

Figure 14.17 Screenshot of working of Example 14.3



	A	B	C	D
1	-		Formulas of	Workings
2	Alpha =	0.3		
3	Month	Week	Dt-1	Ft-1
4	January	1	450	500
5	January	2	550	=D4+$B$2*(C4-D4)
6	January	3	475	=D5+$B$2*(C5-D5)
7	January	4	450	=D6+$B$2*(C6-D6)
8	February	1	470	=D7+$B$2*(C7-D7)
9	February	1	525	=D8+$B$2*(C8-D8)
10	February	2	470	=D9+$B$2*(C9-D9)
11	February	4		=D10+$B$2*(C10-D10)

Figure 14.18 Screenshot of formulas of working of Example 14.3

Example 14.4

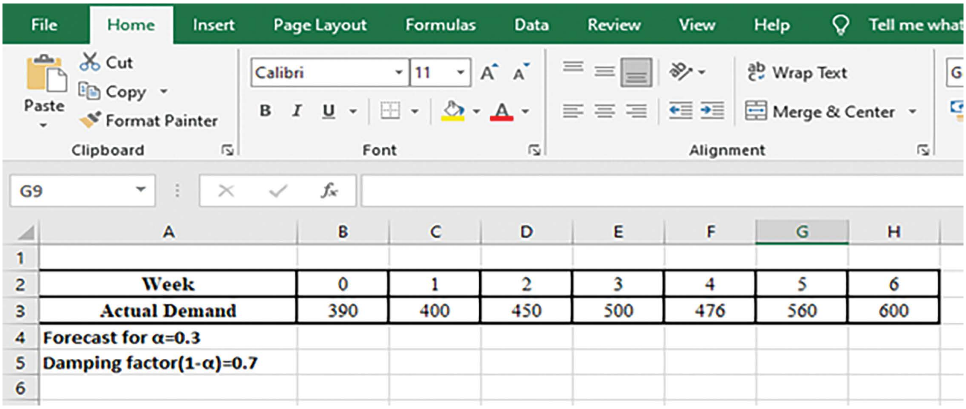
Consider the data given in Table 14.7, which gives the actual demand of a product from week 0 to week 6. Find the forecast values using the Exponential Smoothing function under the Data Analysis button by assuming a smoothing constant (α) as 0.3.

Solution

The screenshot of the data for Example 14.4 is shown in Figure 14.19. The exponential smoothing method applied to Example 14.4 is carried out using the Exponential Smoothing function under Data Analysis, for which the sequence of button clicks is Home $\Longrightarrow$ Data $\Longrightarrow$ Data Analysis. The corresponding screenshot is shown in Figure 14.20. Selecting the Exponential Smoothing function in the dropdown menu of Figure 14.20 and then clicking the OK button will give a display as shown in Figure 14.21. Fill in the linear range of cells plus one more empty successive cell containing the data for the independent variable X (actual demand values), that is, \$B\$3:\$I\$3; the damping factor as 0.7, which is 1 – 0.3; and Output Range, which corresponds to the first address in the next row of the actual data, which specifies the beginning address for the forecast values, which is \$B\$4, as shown in Figure 14.22. Then clicking the OK button in the dropdown menu of

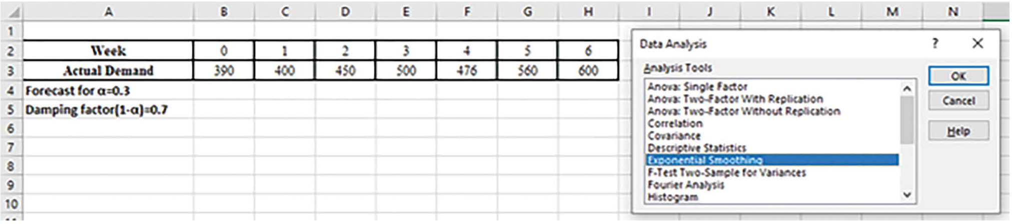
Table 14.7 Data for Example 14.4

Week	0	1	2	3	4	5	6
Actual Demand	390	400	450	500	476	560	600



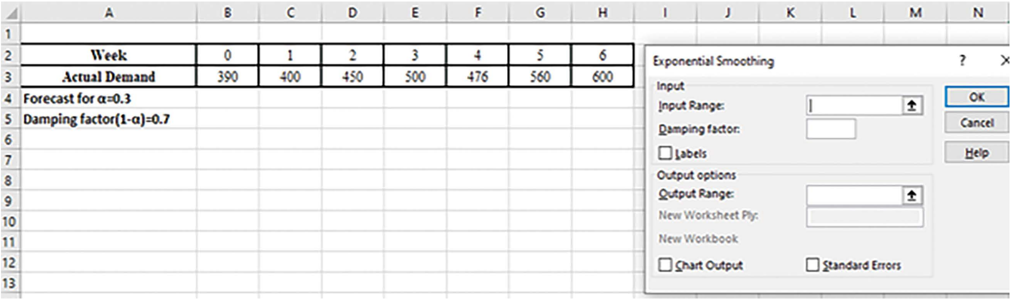
	A	B	C	D	E	F	G	H
1								
2	Week	0	1	2	3	4	5	6
3	Actual Demand	390	400	450	500	476	560	600
4	Forecast for α=0.3							
5	Damping factor(1-α)=0.7							
6								

Figure 14.19 Screenshot of data for Example 14.4



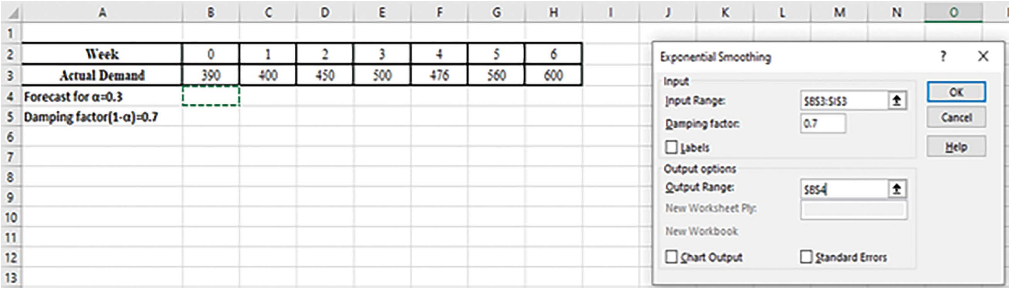
	A	B	C	D	E	F	G	H
1								
2	Week	0	1	2	3	4	5	6
3	Actual Demand	390	400	450	500	476	560	600
4	Forecast for α=0.3							
5	Damping factor(1-α)=0.7							
6								
7								
8								
9								
10								

Figure 14.20 Screenshot after clicking the sequence of buttons, Home ⟹ Data ⟹ Data Analysis



	A	B	C	D	E	F	G	H
1								
2	Week	0	1	2	3	4	5	6
3	Actual Demand	390	400	450	500	476	560	600
4	Forecast for α=0.3							
5	Damping factor(1-α)=0.7							
6								
7								
8								
9								
10								
11								
12								
13								

Figure 14.21 Screenshot after selecting Exponential Smoothing function in the dropdown menu of Figure 14.20 and clicking the OK button



	A	B	C	D	E	F	G	H
2	Week	0	1	2	3	4	5	6
3	Actual Demand	390	400	450	500	476	560	600
4	Forecast for α=0.3							
5	Damping factor(1-α)=0.7							

Exponential Smoothing

Input

Input Range: $B$3:$I$3

Damping factor: 0.7

□ Labels

Output options

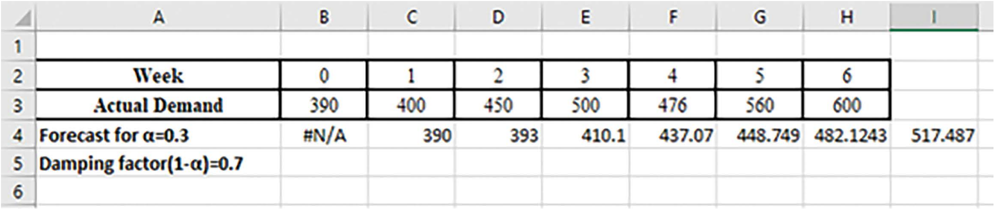
Output Range: $B$4

New Worksheet Ply:

New Workbook

□ Chart Output □ Standard Errors

Figure 14.22 Screenshot after filling data in the dropdown menu of Figure 14.21



	A	B	C	D	E	F	G	H	I
1									
2	Week	0	1	2	3	4	5	6	
3	Actual Demand	390	400	450	500	476	560	600	
4	Forecast for α=0.3	#N/A	390	393	410.1	437.07	448.749	482.1243	517.487
5	Damping factor(1-α)=0.7								
6									

Figure 14.23 Screenshot after clicking the OK button in the dropdown menu of Figure 14.22

Figure 14.22 will give the forecast values starting from the address given in the output range (cells B4:I4) as shown in Figure 14.23.

14.7 Multiple Linear Regression Using Excel Regression Function

As stated earlier, the multiple linear regression equation will contain more than one independent variable on its right-hand side [6, 7].

A generalised model for multiple linear regression equation is shown as follows.

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_iX_i + \ldots + a_nX_n$$

where
n is the number of independent variables
Y is the dependent variable
X_i is the i^{th} independent variable, for $i = 1, 2, 3, \ldots, n$
a_o is the intercept
a_i is the slope of the independent variable X_{i}, $i = 1, 2, 3, \ldots, n$

Multiple linear regression can be carried out using the Regression function under Data Analysis, for which the sequence of button clicks is Home $\Longrightarrow$ Data $\Longrightarrow$ Data Analysis. The corresponding screenshot is shown in Figure 14.24. Selecting the Regression function in the dropdown menu of Figure 14.24 and then clicking the OK button will give a display as shown in Figure 14.25. For a given problem, fill in the range of cells containing the data, including its heading (label) for the dependent variable *Y* in the box

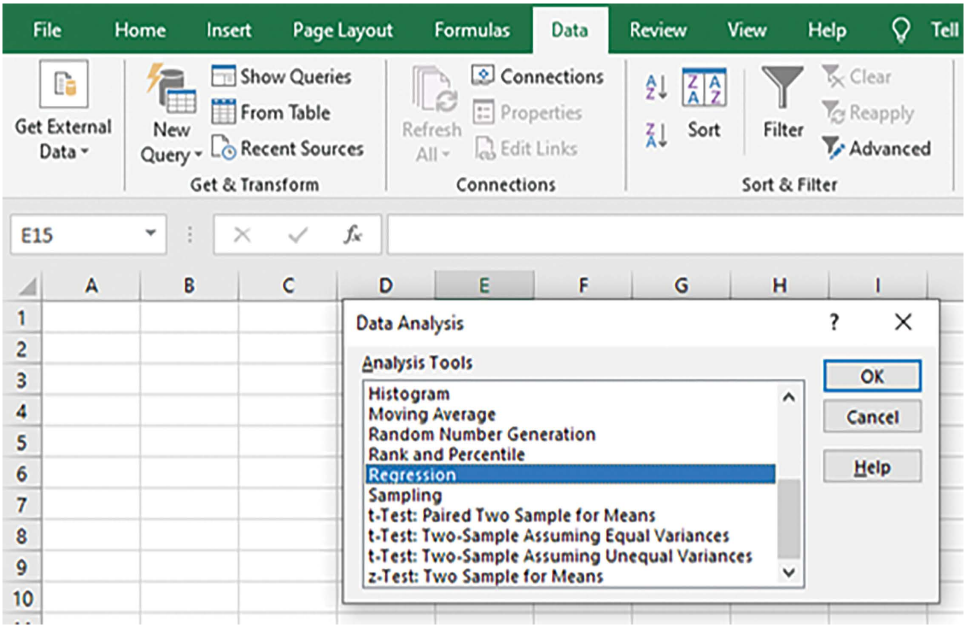
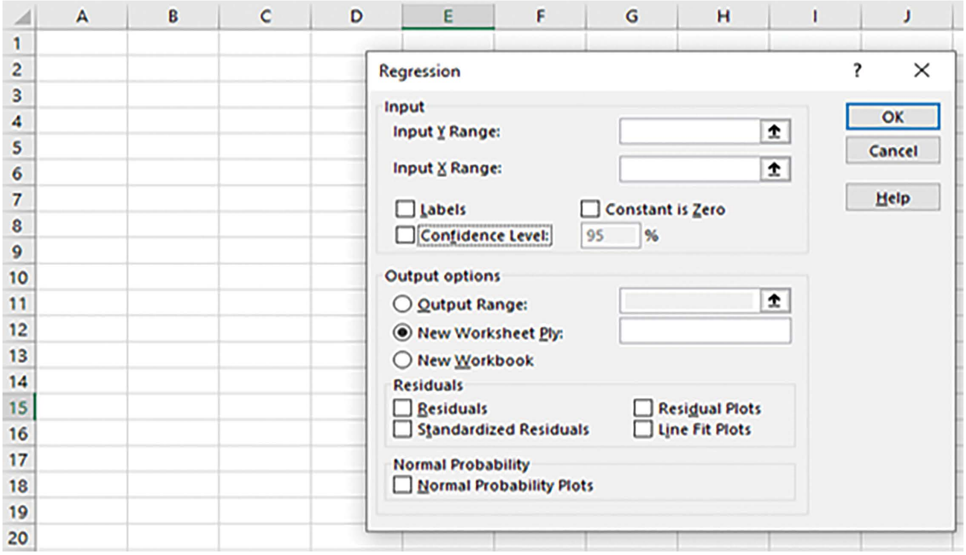


Figure 14.24 Screenshot after the sequence of button clicks, Home ⟹ Data ⟹ Data Analysis



Input

Input Y Range:

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Figure 14.25 Screenshot after selecting the Regression function from dropdown menu of Figure 14.24 and clicking the OK button

against Input Y Range; the range of cells containing the data of the independent variables, including their headings (labels), in the box against Input X Range; click labels; enter the confidence level; and click its checkbox. At the end, click the OK button in the dropdown menu of Figure 14.25 to show the result of the model of the multiple linear regression.

Example 14.5

Take a look at the information in Table 14.8, which includes data on the dependent variable Annual sales in crores of rupees during the previous six years as well as data on the independent variables Sales commission in percentage and Advertising Expenditure in lakhs of rupees. Utilising Excel's Regression feature under Data Analysis, fit a multiple linear regression model for this problem with a significance level of 0.05.

Solution

The generalised model of the multiple linear regression of Example 14.5 is as written.

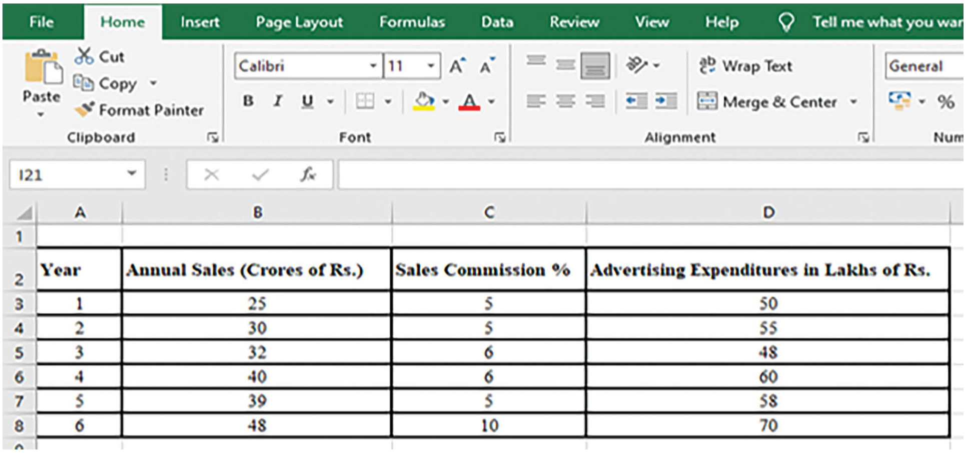
$$Y = a_0 + a_1 X_1 + a_2 X_2$$

Where
X_1 is the sales commission in percentage
X_2 is the advertising expenditure in lakhs of rupees
Y is the annual sales in crores of rupees
a_0 is the intercept
a_1 is the coefficient of X_1
a_2 is the coefficient of X_2

The screenshot of the data for Example 14.5 is shown in Figure 14.26. The multiple linear regression model applied to Example 14.5 is carried out using the Regression function under Data Analysis, for which the sequence of button clicks is Home $\Longrightarrow$ Data $\Longrightarrow$ Data Analysis. The corresponding screenshot is shown in Figure 14.27. Selecting the Regression function in the dropdown menu of Figure 14.27 and then the clicking the OK button will give a display as shown in Figure 14.28. Fill the data in the Input Y range as $B$2:$B$8 and the Input X Range as $C$2:$D$8, click Labels, and fill in Confidence Level as 95%, as shown in Figure 14.29. Then clicking the OK button in the dropdown menu of Figure 14.29 will give the results of the multiple linear regression model of Example 14.5, as shown in Figure 14.30. The third table in Figure 14.30 gives the coefficients, a_0, a_1, and a_2, of the model. The value of R square can be seen in the first table of Figure 14.30.

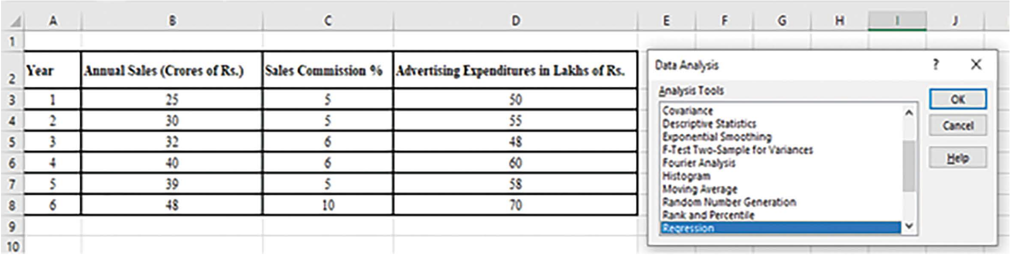
Table 14.8 Data for Example 14.5

Year	*Annual Sales (Crores of Rupees)*	*Sales Commission (%)*	*Advertising Expenditures (Lakhs of Rupees)*
1	25	5	50
2	30	5	55
3	32	6	48
4	40	6	60
5	39	5	58
6	48	10	70



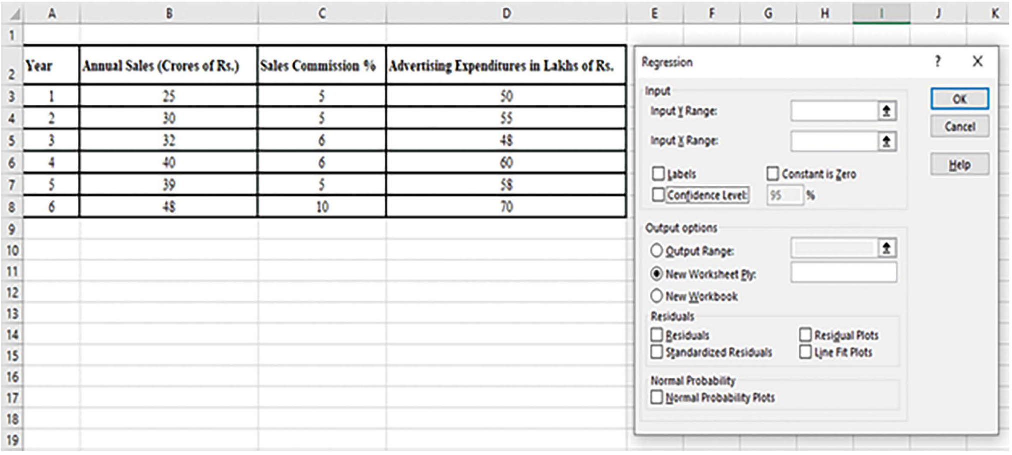
Year	Annual Sales (Crores of Rs.)	Sales Commission %	Advertising Expenditures in Lakhs of Rs.
1	25	5	50
2	30	5	55
3	32	6	48
4	40	6	60
5	39	5	58
6	48	10	70

Figure 14.26 Screenshot of data for Example 14.5



Year	Annual Sales (Crores of Rs.)	Sales Commission %	Advertising Expenditures in Lakhs of Rs.
1	25	5	50
2	30	5	55
3	32	6	48
4	40	6	60
5	39	5	58
6	48	10	70

Figure 14.27 Screenshot after the sequence of button clicks, Home ⟹ Data ⟹ Data Analysis



Year	Annual Sales (Crores of Rs.)	Sales Commission %	Advertising Expenditures in Lakhs of Rs.
1	25	5	50
2	30	5	55
3	32	6	48
4	40	6	60
5	39	5	58
6	48	10	70

Regression

Input

Input Y Range:

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

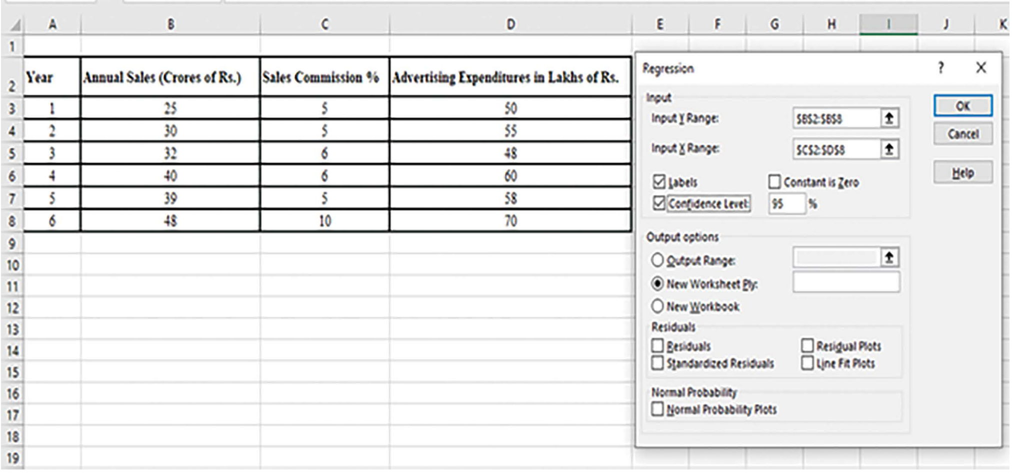
□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Figure 14.28 Screenshot after selecting Regression function in the dropdown menu of Figure 14.27 and clicking the OK button



	A	B	C	D
1				
2	Year	Annual Sales (Crores of Rs.)	Sales Commission %	Advertising Expenditures in Lakhs of Rs.
3	1	25	5	50
4	2	30	5	55
5	3	32	6	48
6	4	40	6	60
7	5	39	5	58
8	6	48	10	70

Regression

Input

Input Y Range: $B$2:$B$8

Input X Range: $C$2:$D$8

☑ Labels □ Constant is Zero

☑ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

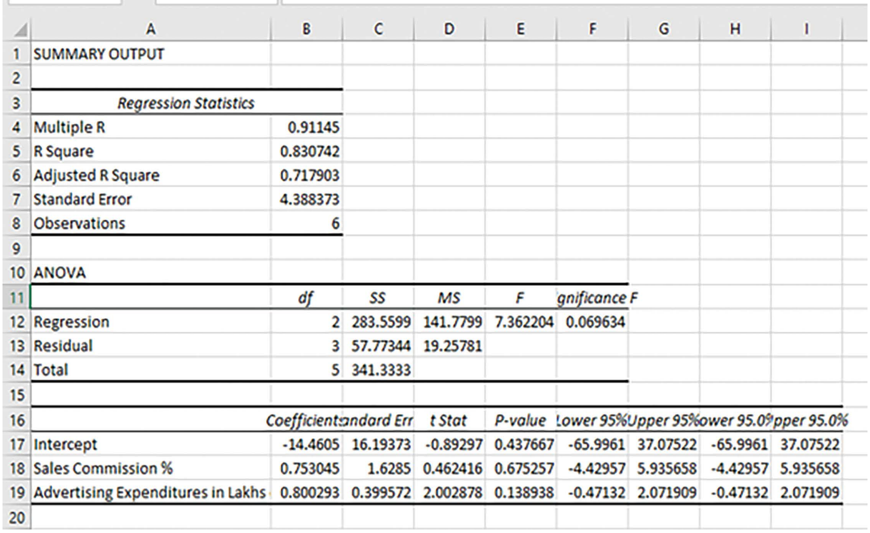
□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Figure 14.29 Screenshot after filling data in the dropdown menu of Figure 14.28



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.91145							
5	R Square	0.830742							
6	Adjusted R Square	0.717903							
7	Standard Error	4.388373							
8	Observations	6							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regression	2	283.5599	141.7799	7.362204	0.069634			
13	Residual	3	57.77344	19.25781					
14	Total	5	341.3333						
15									
16		*Coefficient*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	-14.4605	16.19373	-0.89297	0.437667	-65.9961	37.07522	-65.9961	37.07522
18	Sales Commission %	0.753045	1.6285	0.462416	0.675257	-4.42957	5.935658	-4.42957	5.935658
19	Advertising Expenditures in Lakhs	0.800293	0.399572	2.002878	0.138938	-0.47132	2.071909	-0.47132	2.071909
20									

Figure 14.30 Screenshot after clicking the OK button in the dropdown menu of Figure 14.29

From Figure 14.30, the model of the multiple linear regression of Example 14.5 is as written.

$$Y = -14.4605 + 0.753045X_1 + 0.800293X_2,$$

Where
X_1 is the sales commission in percentage
X_2 is the advertising expenditure in lakhs of rupees
Y is the annual sales in crores of rupees
The p value is 0.0696634. So, the model fits at a significance level of 0.05. Since R^2 is 0.830742, the model fit is good.

14.8 Time Series Analysis Using Excel Sheets and Regression Function

In the previous sections, forecasting techniques have been presented with the idea of predicting trends alone. In this section, time series analysis is presented. The time series data consist of trend (T), seasonal (S), cyclical (C), and random (R) components. These components are explained as follows.

The trend (T) component in the forecast deals with the increase or decrease of a dependent variable with an increase in the independent variable. The annual growth of the sales of a product due to its increasing popularity in the minds of customers sets an example of the trend component of the forecast.

The seasonal component (S) in the forecast delas with the regular and predictable changes that happen in a year. The pattern of demand that exists for the first year will be repeated in the following years too. The demand for a product will be low in some parts of the year, and it will be high during another part of the year. The demand for ice cream is the best example to explain the seasonality of demand. Its demand will be high during summer and low during winter. The demand for electricity will be more during the day and less during night, because many institutions and business establishments will be using electricity at a large scale during the day period.

The cyclical component (C) of the forecast is similar to seasonality, but its cycle time will be more than a year. The cycle time of cyclical components of the forecast will not be equal when it recurs. The business cycle of a nation is an example of the cyclical component of the forecast. The cycle time of this business cycle will not be the same in the next cycle because the individual phases, expansion, recession, contraction, and revival of the business cycle will have different timespans in the next cycle. This component will be present in the demand of a product too.

The random component (R) of the forecast will account for all explained factors in reality which will have impact on the demand of a product.

The forecast of a time series data is given by the following formula.

$$Y = TSCR$$

Where
Y is the forecast
T is the trend
S is the seasonal index
C is the cyclical index
R is random

The plots of a forecast with trend component, forecast with seasonality component, and forecast with cyclical component are shown in Figures 14.31, 14.32, and 14.33, respectively.

Just to have a feel for an integrated problem with all the components of the time series, an example is presented in the following, which will be solved using an Excel sheet after discussing the steps of the time series.

Example 14.6

Table 14.9 provides a summary of the demand for ice cream goods from a top ice cream manufacturer in lakhs of units. One can see from this table that the demand values are

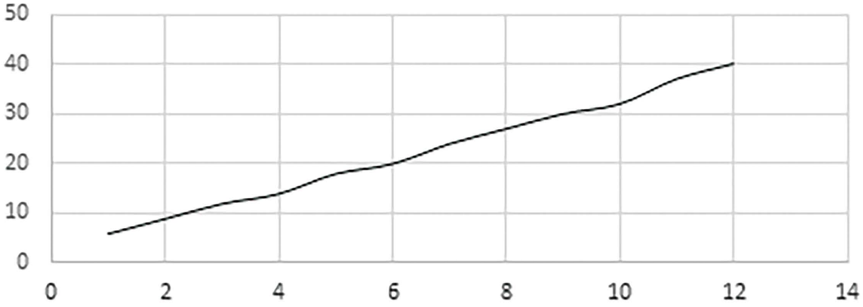


Figure 14.31 Forecast graph with trend component

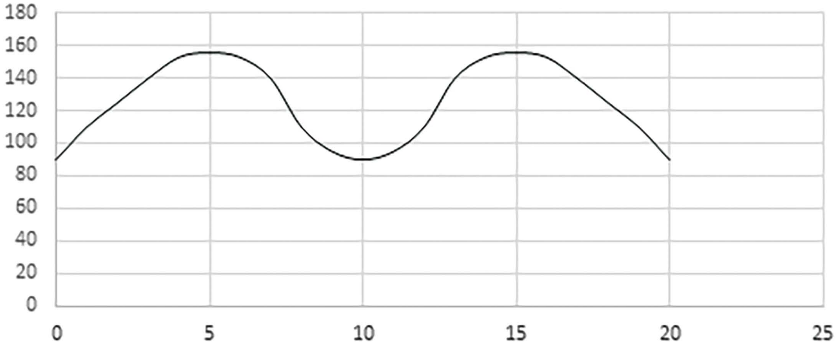


Figure 14.32 Forecast graph with seasonality component

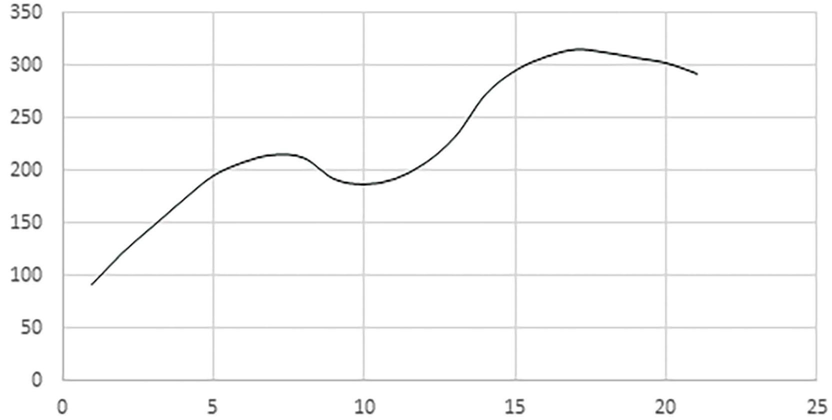


Figure 14.33 Forecast graph with cyclical component

higher in the second and third quarters when compared to the corresponding values in the first and fourth quarters. According to this observation, the first and third quarters correspond to the months of April through June and July through September, respectively, when higher demand values are anticipated, as shown in Table 14.9. The presence of seasonality in the provided data is confirmed by this finding. Further, the demand

Table 14.9 Time Series Data for Ice-Cream Demand

		Quarter			
		1	*2*	*3*	*4*
Year	1	80	106	162	126
	2	105	125	134	118
	3	100	138	154	142
	4	102	142	170	166

values have an increasing trend in each quarter over years, and one can verify that the total demand in each year also has an increasing trend over years. This ensures the presence of the trend component in the given data.

14.8.1 Steps of Time Series

The steps of the time series analysis are presented in the following.

Step 1: Arrange the data x_i (time period) and *Yi* (dependent variable) in serial order if given in matrix form, as shown in Table 14.9.

The forecast of the given data is given by the following formula.

$Y = TSCR$
Where
Y is the forecast
T is the trend
S is the seasonal index
C is the cyclical index
R is random

Step 2: Find the moving average (*MA*) of the given data for the assumed moving average period.

2.1. Find the moving total of the given data for the assumed moving average period.
2.2. Find the centred moving total of the given data by finding the average of successive two moving totals from top of the column. The mid-point of the previous two periods with respect to the first moving total is a decimal number. Hence, place the first average of the moving total in the previous row of the next column under the centred moving total with reference to the first moving total used to compute the centred moving total.
2.3. Divide the centred moving total obtained in Step 2.2 by the assumed moving average period to obtain the centred moving average of the given data.

Step 3: Find the seasonal indices from the centred moving averages (*MA*) using the following formula.

3.1. Find the seasonal index for the period *i* (*S*) using the following formula.
= Original time series data item of period *i* (Y_i)/centred moving average (*MA*) of the period *i*
3.2. Determination of smoothed seasonal indices.

3.2.1. Copy the seasonal indices obtained in Step 3.1 in a matrix form with, say, years on rows and seasons within year (quarters) on columns.

3.2.2. Find the average of the seasonal indices of each season in the matrix formed in Step 3.2.1.

3.2.3 Find the sum of the average of the seasonal indices of the seasons obtained in Step 3.2.2.

3.2.4 Find each smoothed seasonal index of each season using the following formula.

$$\mathit{Smoothed\,seasonal\,index\,of\,season\,j}$$

$$= \mathit{Average\,of\,sesonal\,index\,of\,season\,j} \times \frac{\mathit{Number\,of\,seasons}}{\mathit{Sum\,of\,average\,of\,seasonal\,indices\,of\,all\,seasons}}$$

Step 4: Find the deseasonalised value of y_i by using the following formula.

$$y_i = \frac{Y_i}{S}$$

Step 5: Fit a linear regression model for the data of x_i and y_i to determine the coefficients of the following model using the sequence of buttons Data $\Longrightarrow$ Data Analysis $\Longrightarrow$ Regression in Excel.

$$y = a + bx$$

Where
x is the time period
y is the deseasonalised forecast
a is the intercept of the regression model
b is the slope of the regression model associated with the variable X

Step 6: Find the estimate of the trend of y_i using the following formula.

$$\widehat{y_i} = a + bx_i$$

Step 7: Find the cyclical component from the deseasonalised value of y_i using the following formula.

$$C_i = \frac{y_i}{\widehat{y_i}}$$

Solution

Consider the data for Example 14.6 as reproduced in Table 14.10.

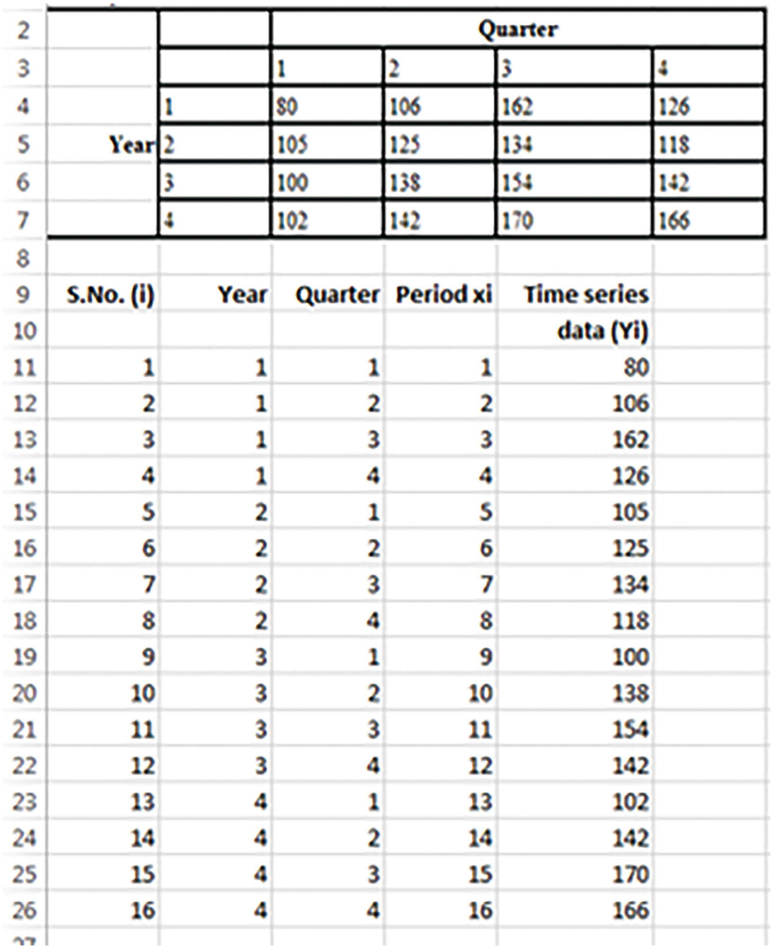
1. Find the deseasonalised forecast of the data.
2. Find the estimate of trend values of the data.
3. Find the cyclical component of the data.

Solution

Step 1: Arrange the data x_i (time period) and *Yi* (dependent variable), which are given in Table 14.9 in serial order if it is given in matrix form as shown in Figure 14.34. The plot of the data is shown in Figure 14.35.

Table 14.10 Time Series Data for Ice-Cream Demand

		Quarter			
		1	*2*	*3*	*4*
Year	1	80	106	162	126
	2	105	125	134	118
	3	100	138	154	142
	4	102	142	170	166



		Quarter			
		1	2	3	4
	1	80	106	162	126
Year	2	105	125	134	118
	3	100	138	154	142
	4	102	142	170	166

S.No. (i)	Year	Quarter	Period xi	Time series data (Yi)
1	1	1	1	80
2	1	2	2	106
3	1	3	3	162
4	1	4	4	126
5	2	1	5	105
6	2	2	6	125
7	2	3	7	134
8	2	4	8	118
9	3	1	9	100
10	3	2	10	138
11	3	3	11	154
12	3	4	12	142
13	4	1	13	102
14	4	2	14	142
15	4	3	15	170
16	4	4	16	166

Figure 14.34 Screenshot of rearranged data for Example 14.6

Step2: Find the moving average (*MA*) of the given data for a moving average period of four quarters.

2.1. Find the moving total of the given data with the moving average period of four quarters. The first moving total is to be placed in the fourth quarter (fourth period in the serial order) under the moving total column.

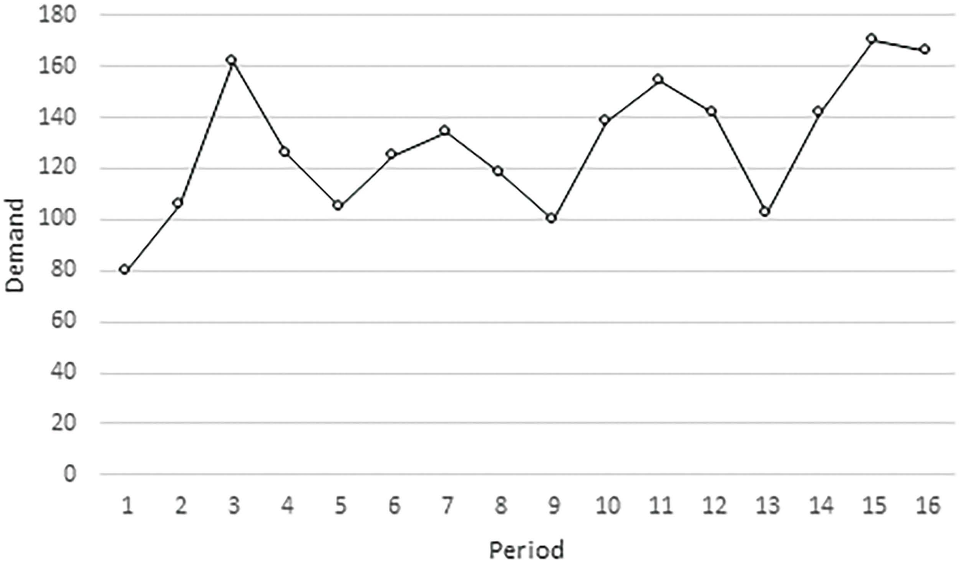
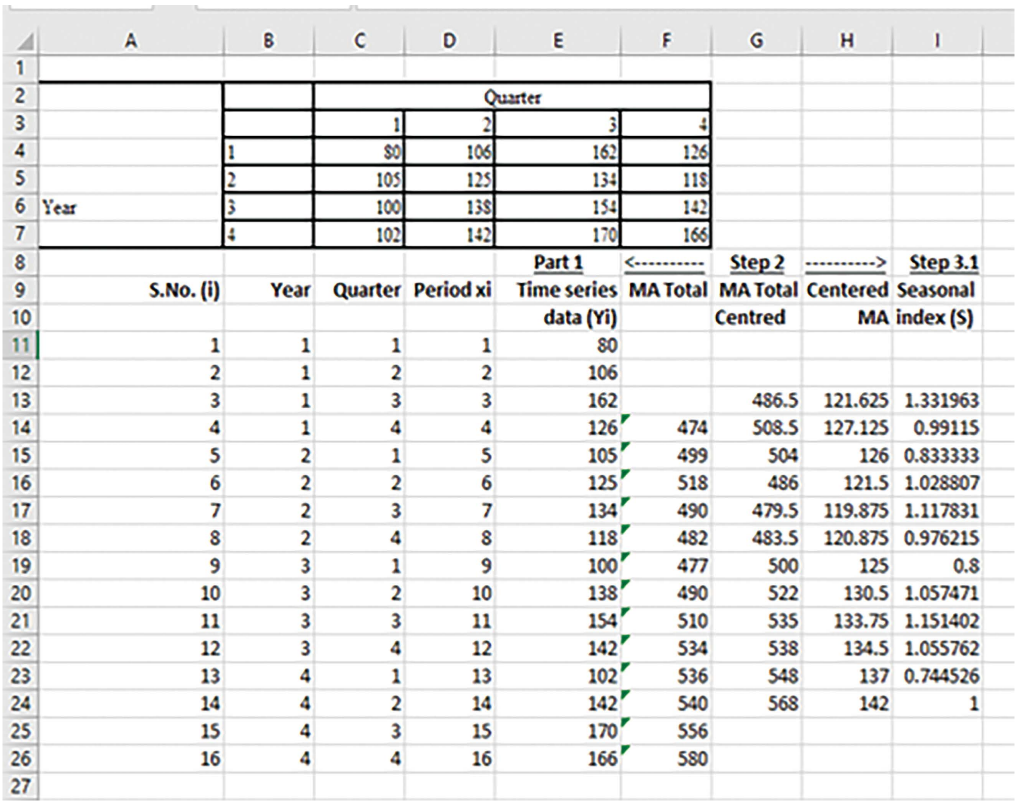


Figure 14.35 Plot of the data for Example 14.6



	A	B	C	D	E	F	G	H	I
1									
2			Quarter						
3			1	2	3	4			
4		1	80	106	162	126			
5		2	105	125	134	118			
6	Year	3	100	138	154	142			
7		4	102	142	170	166			
8					Part 1	<---------	Step 2	--------->	Step 3.1
9	S.No. (i)	Year	Quarter	Period xi	Time series	MA Total	MA Total	Centered	Seasonal
10					data (Yi)		Centred	MA	index (S)
11	1	1	1	1	80				
12	2	1	2	2	106				
13	3	1	3	3	162		486.5	121.625	1.331963
14	4	1	4	4	126	474	508.5	127.125	0.99115
15	5	2	1	5	105	499	504	126	0.833333
16	6	2	2	6	125	518	486	121.5	1.028807
17	7	2	3	7	134	490	479.5	119.875	1.117831
18	8	2	4	8	118	482	483.5	120.875	0.976215
19	9	3	1	9	100	477	500	125	0.8
20	10	3	2	10	138	490	522	130.5	1.057471
21	11	3	3	11	154	510	535	133.75	1.151402
22	12	3	4	12	142	534	538	134.5	1.055762
23	13	4	1	13	102	536	548	137	0.744526
24	14	4	2	14	142	540	568	142	1
25	15	4	3	15	170	556			
26	16	4	4	16	166	580			
27									

Figure 14.36 Screenshot of calculations from Step 2 to compute centred moving averages and Step 3 to compute seasonal index data

2.2. Find the centred moving total of the given data by finding the average of the successive two moving totals from the top of the column. The mid-point of the previous two periods with respect to the first moving total is a decimal number. Hence, place the first average of the moving total in the previous row of the next column under the centred moving total with reference to the first moving total used to compute the centred moving total.

2.3. Divide the centred moving total obtained in Step 2.2 by the assumed moving average period to obtain the centred moving average of the given data.

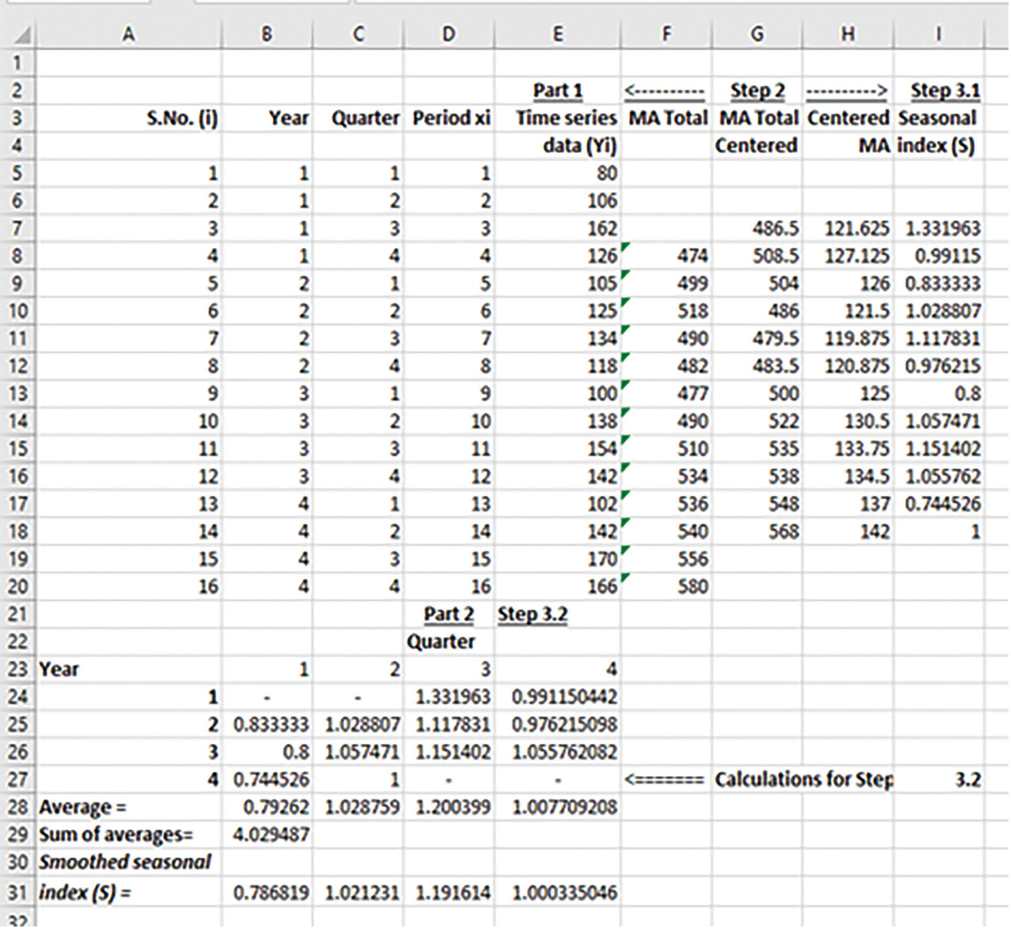
The results of these sub-steps are shown in columns F, G, and H of Figure 14.36.

Step 3: Find seasonal indices from the centred moving averages (*MA*) using the following formula.

3.1. Find the seasonal index for the period *i* (*S*) using the following formula as shown in column I of Figure 14.36.

= Original time series data item of period *i* (Y_i)/centred moving average (*MA*) of the period *i*

3.2. Determination of smoothed seasonal indices, as shown in Figure 14.37.



	A	B	C	D	E	F	G	H	I
1									
2					Part 1	<---------	Step 2	---------->	Step 3.1
3	S.No. (i)	Year	Quarter	Period xi	Time series	MA Total	MA Total	Centered	Seasonal
4					data (Yi)		Centered	MA	index (S)
5	1	1	1	1	80				
6	2	1	2	2	106				
7	3	1	3	3	162		486.5	121.625	1.331963
8	4	1	4	4	126	474	508.5	127.125	0.99115
9	5	2	1	5	105	499	504	126	0.833333
10	6	2	2	6	125	518	486	121.5	1.028807
11	7	2	3	7	134	490	479.5	119.875	1.117831
12	8	2	4	8	118	482	483.5	120.875	0.976215
13	9	3	1	9	100	477	500	125	0.8
14	10	3	2	10	138	490	522	130.5	1.057471
15	11	3	3	11	154	510	535	133.75	1.151402
16	12	3	4	12	142	534	538	134.5	1.055762
17	13	4	1	13	102	536	548	137	0.744526
18	14	4	2	14	142	540	568	142	1
19	15	4	3	15	170	556			
20	16	4	4	16	166	580			
21				Part 2	Step 3.2				
22				Quarter					
23	Year	1	2	3	4				
24	1	-	-	1.331963	0.991150442				
25	2	0.833333	1.028807	1.117831	0.976215098				
26	3	0.8	1.057471	1.151402	1.055762082				
27	4	0.744526	1	-	-	<=======	Calculations for Step		3.2
28	Average =	0.79262	1.028759	1.200399	1.007709208				
29	Sum of averages=	4.029487							
30	*Smoothed seasonal*								
31	*index (S) =*	0.786819	1.021231	1.191614	1.000335046				
32									

Figure 14.37 Screenshot of calculations of smoothed seasonal indices

3.2.1. Copy the seasonal indices obtained in Step 3.1 in a matrix form with, say, years on rows and seasons within year (quarters) on columns, as shown at the bottom of Figure 14.37.

3.2.2. Find the average of the seasonal indices of each season in the matrix formed in Step 3.2.1.

3.2.3. Find the sum of the average of the seasonal indices of the seasons obtained in Step 3.2.2.

3.2.4. Find each smoothed seasonal index of each season using the following formula. A plot of these smoothed seasonal indices is shown in Figure 14.38.

$$Smoothed\,seasonal\,index\,of\,season\,j$$

$$= Average\,of\,sesonal\,index\,of\,season\,j \times \frac{Number\,of\,seasons}{Sum\,of\,average\,of\,seasonal\,indices\,of\,all\,seasons}$$

Step 4: Find the deseasonalised value of y_i by using the following formula as shown in column G of Figure 14.39. This gives the answer to part (a) of the questions.

$$y_i = \frac{Y_i}{S}$$

Step 5: Fit a linear regression model for the data of x_i and y_i to determine the coefficients of the following model using the sequence of buttons Data ⟹ Data Analysis ⟹ Regression in Excel.

$$y = a + bx$$

Where
x is the time period
y is the deseasonalised forecast
a is the intercept of the regression model
b is the slope of the regression model associated with the variable X

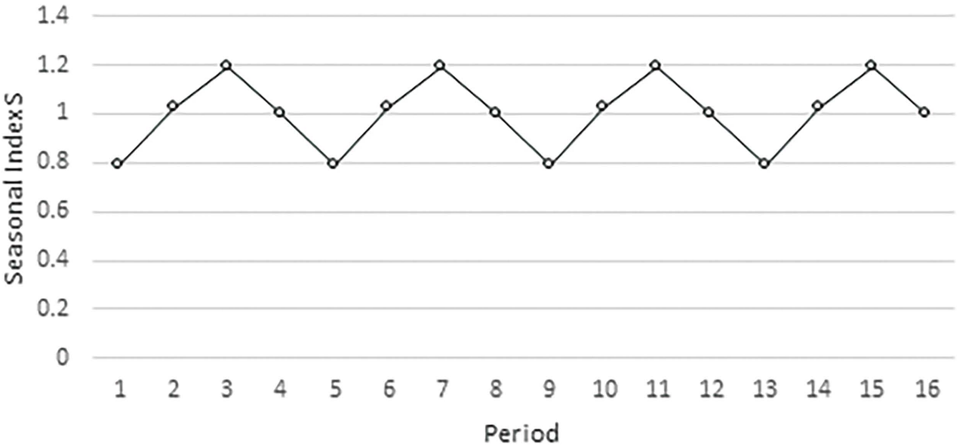
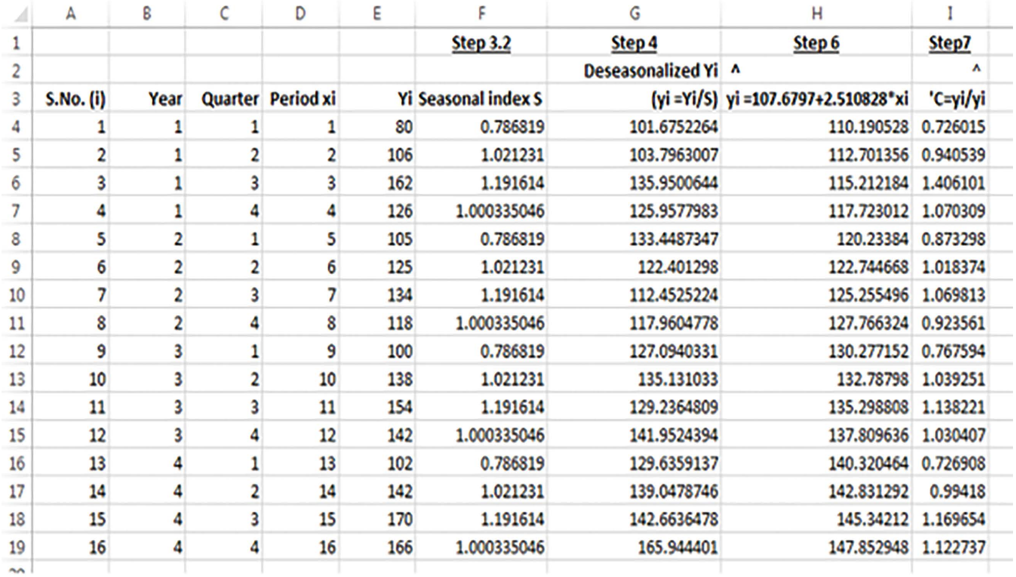


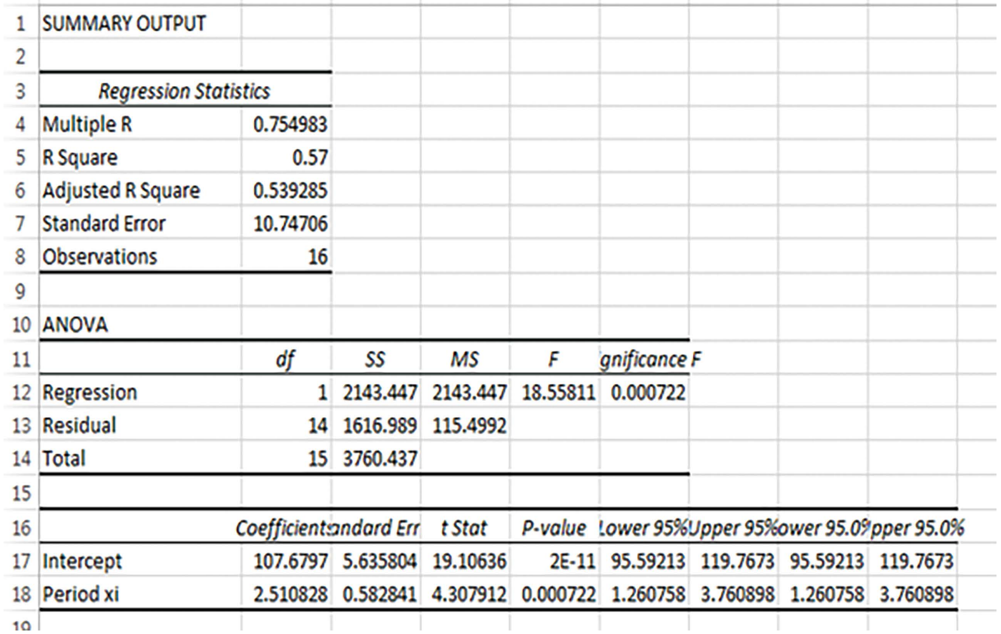
Figure 14.38 Plot of smoothed seasonal indices



	A	B	C	D	E	F	G	H	I
1						Step 3.2	Step 4	Step 6	Step7
2							Deseasonalized Yi	^	^
3	S.No. (i)	Year	Quarter	Period xi	Yi	Seasonal index S	(yi =Yi/S)	yi =107.6797+2.510828*xi	'C=yi/yi
4	1	1	1	1	80	0.786819	101.6752264	110.190528	0.726015
5	2	1	2	2	106	1.021231	103.7963007	112.701356	0.940539
6	3	1	3	3	162	1.191614	135.9500644	115.212184	1.406101
7	4	1	4	4	126	1.000335046	125.9577983	117.723012	1.070309
8	5	2	1	5	105	0.786819	133.4487347	120.23384	0.873298
9	6	2	2	6	125	1.021231	122.401298	122.744668	1.018374
10	7	2	3	7	134	1.191614	112.4525224	125.255496	1.069813
11	8	2	4	8	118	1.000335046	117.9604778	127.766324	0.923561
12	9	3	1	9	100	0.786819	127.0940331	130.277152	0.767594
13	10	3	2	10	138	1.021231	135.131033	132.78798	1.039251
14	11	3	3	11	154	1.191614	129.2364809	135.298808	1.138221
15	12	3	4	12	142	1.000335046	141.9524394	137.809636	1.030407
16	13	4	1	13	102	0.786819	129.6359137	140.320464	0.726908
17	14	4	2	14	142	1.021231	139.0478746	142.831292	0.99418
18	15	4	3	15	170	1.191614	142.6636478	145.34212	1.169654
19	16	4	4	16	166	1.000335046	165.944401	147.852948	1.122737

The regression model: $\hat{y}_i = 107.6797 + 2.510828$

Figure 14.39 Screenshot of copying result of Steps 3.2, 4, 6, and 7 of time series analysis.



1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.754983							
5	R Square	0.57							
6	Adjusted R Square	0.539285							
7	Standard Error	10.74706							
8	Observations	16							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regression	1	2143.447	2143.447	18.55811	0.000722			
13	Residual	14	1616.989	115.4992					
14	Total	15	3760.437						
15									
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	107.6797	5.635804	19.10636	2E-11	95.59213	119.7673	95.59213	119.7673
18	Period xi	2.510828	0.582841	4.307912	0.000722	1.260758	3.760898	1.260758	3.760898

Figure 14.40 Screenshot of Step 5 of time series analysis (regression model using Data ⟹ Data Analysis ⟹ Regression button based on data in column D (x_i) and column G (y_i) in Figure 14.39)

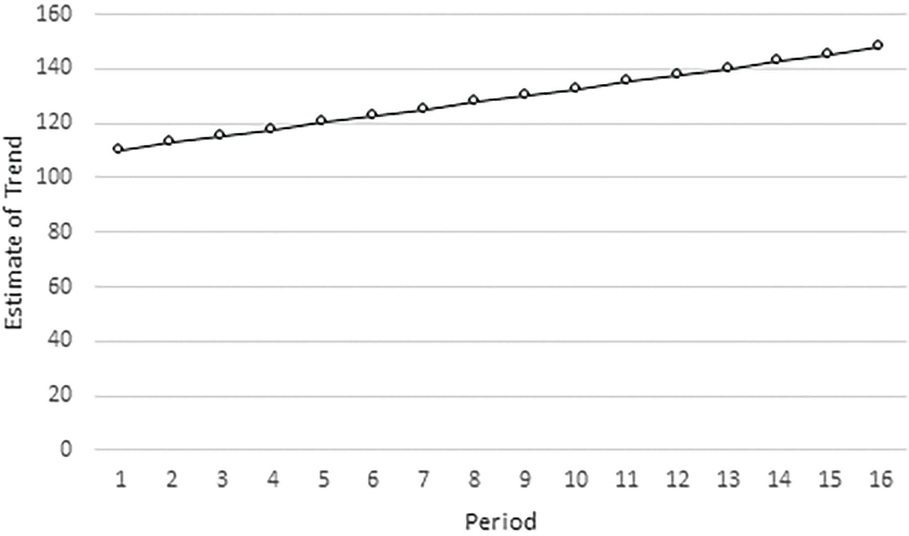


Figure 14.41 Trend pattern of data given in Example 14.6

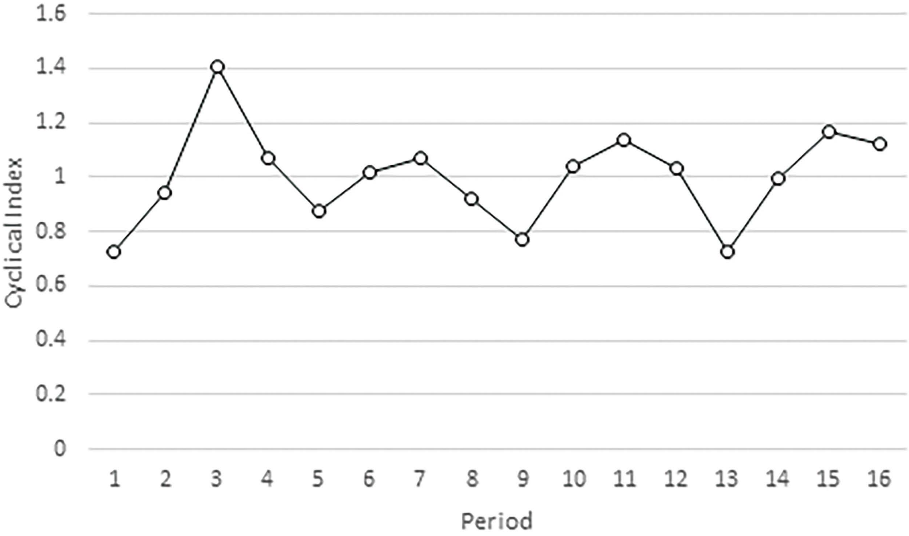


Figure 14.42 Plot of cyclical component of data

The result of this regression model is show in Figure 14.40, and the regression model is as follows.

$$The\,regression\,model : \widehat{y_i} = 107.6797 + 2.510828$$

Step 6: Find the estimate of the trend of y_i using the following formula, as shown in Figure 14.37. This gives the answer to part 2 of the questions.

$$\widehat{y_i} = a + bx_i$$

Step 7: Find the cyclical component from the value of y_i using the following formula as shown in column I of Figure 14.39. This gives the answer to part 3 of the questions. The plot of the cyclical indices is shown in Figure 14.42.

$$C_i = \frac{y_i}{\widehat{y_i}}$$

Summary

- The projection of an event, say, the demand for an item based on its past data, is called forecasting.
- The Delphi method is a qualitative forecasting method, which aims to predict the future states of qualitative events, that is, culture of a society, level of computer technology, lifestyle of people, value system, and so on.
- The moving average method aims to estimate the demand of an item in the short run, say, based on the demand of the past three or four weeks.
- The exponential smoothing method of forecasting aims to forecast the demand for an item for the next period based on the demand of the current period.
- The regression model aims to design an equation for the dependent variable in terms of the assumed independent variables.
- A regression is defined as the dependence of a variable of interest (dependent variable) on one or more other variables (independent variables).
- A sample linear regression equation is:

  $Y = a + bX$

  where

  Y is the dependent variable

  X is the independent variable

  a is the intercept

  b is the trend (slope)
- The sequence of button clicks for regression in Excel is Home ⟹ Data ⟹Data Analysis ⟹ Regression.
- The moving average method of forecasting predicts the value of a variable of interest in the short run, say, a week.
- In the moving average method of forecasting with three periods, the forecast of period i + 1 is

  $F_{i+1} = MA_i \; for\, i = 3,4,5, ,\ldots,n$
- The exponential smoothing method of forecasting is used to predict the demand of the current period based on the demand and forecast of the previous period.
- The multiple linear regression equation will contain more than one independent variable on its right-hand side.

- A generalised model for multiple linear regression equation is:

  $$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_iX_i + \ldots + a_nX_n$$

  where,

  n is the number of independent variables;
  Y is the dependent variable;
  X_i is the i^{th} independent variable, for $i = 1, 2, 3, \ldots, n$;
  a_o is the intercept;
  a_i is the slope of the independent variable X_{i}, for $i = 1, 2, 3, \ldots, n$.
- The Regression function under Data Analysis can be used to fit a model for multiple linear regression.
- Time series data consists of trend (*T*), seasonal (*S*), cyclical (*C*), and random (*R*) components.
- The trend (*T*) component in the forecast deals with the increase or decrease of a dependent variable with an increase in the independent variable.
- The seasonal component (*S*) in the forecast delas with the regular and predictable changes that happen in a year.
- The cyclical component (*C*) of the forecast is similar to seasonality, but its cycle time will be more than a year.
- The random component (*R*) of the forecast will account for all explained factors in reality that will have impact on the demand for a product.

Keywords

- Forecasting means the projection of an event, say, the demand for an item based on its past data.
- The Delphi method is a qualitative forecasting method, which aims to predict the future states of qualitative events, that is, culture of a society, level of computer technology, lifestyle of people, value system, and so on.
- The moving average method aims to estimate the demand of an item in the short run, say, based on the demand of the past three or four weeks.
- The exponential smoothing method of forecasting aims to forecast the demand for an item for the next period based on the demand of the current period.
- Regression is defined as the dependence of a variable of interest (dependent variable) on one or more other variables (independent variables).
- A multiple linear regression equation contains more than one independent variable on its right-hand side.
- R^2 is the value to check the fitness of the regression model. Any value beyond 0.75 is acceptable to justify the model fit of a regression model.

- Time series data consists of trend (*T*), seasonal (*S*), cyclical (*C*), and random (*R*) components.
- The trend (*T*) component in the forecast deals with the increase or decrease of a dependent variable with an increase in the independent variable.
- The seasonal component (*S*) in the forecast delas with the regular and predictable changes that happen in a year.
- The cyclical component (*C*) of the forecast is similar to seasonality, but its cycle time will be more than a year.

- The random component (R) of the forecast will account for all explained factors in reality that will have impact on the demand for a product.

Review Questions

1. Define forecast and explain its significance in business operations.
2. What is linear regression? Give the formulas to estimate the slope and constant of simple regression.
3. Explain the steps of estimating the coefficients of simple regression using Excel.
4. The demand values for a product manufactured by Beta Engineering Company for the past ten years are given in the following table. Design a linear regression model to estimate the demand of the product using Excel.

Year	*Demand*
1	20
2	25
3	32
4	36
5	40
6	45
7	52
8	56
9	65
10	70

5. The following table provides a summary of the monthly output of an export processing zone for the last 12 years, expressed in crores of rupees. Utilising Excel, design a linear regression model to predict the output of the export processing zone for a specific year in the future.

Year	1	2	3	4	5	6	7	8	9	10	11	12
Output (Crores of Rupees)	1,500	1,650	1,720	1,800	1,875	1,969	2,010	2,150	2,220	2,350	2,470	2,575

6. Explain the steps of the moving average method of forecasting using Excel.
7. The following table provides an overview of a product's weekly demand values. Utilising Excel, calculate the moving averages for periods 3 through 9, which will serve as the forecast values for periods 4 through 10.

Month	*Demand*
1	5
2	7
3	8
4	13
5	17
6	20
7	24
8	29
9	32

8. Discuss the applications of the exponential smoothing method of forecasting.

9. Exponential smoothing is used to assess a product's demand inside a corporation. To predict demand, a smoothing constant of 0.25 is utilised. The demand turned out to be 750 units as opposed to the 800 units forecast for the first week of February.

   a. Forecast the demand for the second week of February.
   b. Assume that the actual demand during the second week of February turned out to be 850 units. Forecast the demand up to March's third week, assuming the subsequent demands are 875, 900, 870, 925, and 940 units.

10. Give a generalised model of multiple linear regression and explain its components.
11. Take a look at the information in the following table, which includes data for two independent variables over the course of eight years, that is, R&D Expenditure in Lakhs of Rupees, X1 and Training Expenditure in Lakhs of Rupees, X2, and data for a dependent variable (Y), Profit in Crores of Rupees.

Year	*Profit (Crores of Rupees)*	*R&D Expenditure (Lakhs of Rupees)*	*Training Expenditure (Lakhs of Rupees)*
1	200	50	20
2	220	55	25
3	280	65	34
4	350	78	42
5	500	90	50
6	790	110	56
7	900	140	68
8	1200	180	80

   a. Fit a multiple linear regression model using the Regression function in Excel.
   b. Check the fitness of the model at a confidence level of 0.9.

12. What is a time series? Explain its components.
13. Give a sketch of the pattern of each of the following.

   a. Trend
   b. Seasonal
   c. Cyclical

14. List and explain the steps of time series analysis.
15. The following table shows the quarterly sales revenue for textile garments from a leading textile retailer for the previous four years, expressed in crores of rupees.

		Quarter			
		1	*2*	*3*	*4*
Year	1	30	50	80	70
	2	52	76	85	72
	3	55	90	105	95
	4	60	95	120	115

Time series sales data for textile garments:

a. Find the deseasonalised forecast of the data.
b. Find the estimate of trend values of the data.
c. Find the cyclical component of the data.

References

1. Panneerselvam, R., *Production and Operations Management* (3rd edition), Prentice-Hall of India, New Delhi, 2012.
2. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
3. www.statisticshowto.com/how-to-perform-Excel-2013-regression-analysis-Excel-2013/ [June 25, 2020].
4. www.Excel-easy.com/examples/moving-average.html [June 25, 2020].
5. www.Excel-easy.com/examples/exponential-smoothing.html [June 25, 2020].
6. https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775?gi=5c96ee7cd5e2 [July 9, 2020].
7. www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/ [July 9, 2020].

15 Analysis of Variance

Learning Objectives

After reading this chapter, the readers will be able to

- Understand the design and analysis of experiments.
- Distinguish between fixed factors and random factors.
- Analyse problems using analysis of variance (ANOVA) with a single factor (completely randomised design) through Excel.
- Solve problems with a randomised complete block design (ANOVA: two factor without replications) using Excel.
- Apply the Latin square design to solve problems with a single factor and two blocks without replications.
- Analyse problems with two factors including interactions among them through Excel.
- Understand the application of Yates' algorithm for a 2^n complete factorial experiment.

15.1 Introduction

Design and analysis of experiments focuses on different processes, machines, materials, employees, and other resources that are used to manufacture goods/provide services [1]. The concept of design and analysis of experiments is explained through an example.

Consider the manufacturing and selling of a costly washing machine. The washing machine is sold in five different states. The marketing chief wants to analyse whether the sales of the washing machine differ among these five states in terms of monthly sales revenue. The format of the monthly sales revenues of the last financial year in these states is shown in Table 15.1.

Specifying the experiment's hypotheses is the first stage in the design and analysis of experiments. The term "hypothesis" refers to a population's assumption, and it is divided into the null hypothesis (H_0) and alternate hypothesis categories (H_1). Null hypothesis refers to a population's favoured assumption. The complement of the null hypothesis is the alternate hypothesis. If the null hypothesis is not accepted, the alternative hypothesis is confirmed.

An illustrative null hypothesis and alternate hypothesis with respect to the example presented are:

H_0: There are no significant differences among the states in terms of the monthly sales revenue.

DOI: 10.4324/9781032671703-15

Table 15.1 Monthly Sales Revenues of Washing Machine in States

		State				
		1	*2*	*3*	*4*	*5*
Replication (Month)	1					
	2					
	3					
	4					
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					

H_1: There are significant differences among the states in terms of the monthly sales revenue.

The data for the example problem can be analysed in light of the hypotheses using analysis of variance.

The terminology used in the design of experiments is as follows [1, 2].

- Response (variable)
- Factor and level/treatment
- Replication

Response: An experiment's response is a measurement of an important dependent variable, which may be affected by one or more factors as well as their interactions. The monthly sales revenue of the example forms the response variable of the experiment.

Factor and Level/Treatment: A factor in an experiment is a parameter or entity which is suspected to have an effect on the response variable. The state (state 1/state 2/state 3/state 4/state 5) of the example forms the factor, which is suspected to have effect on the monthly sales revenue of the washing machine.

The different settings of the factor State are State 1, State 2, State 3, State 4, and State 5, which are called the treatments of that factor.

A treatment/level of a factor is a particular value, like 30 min out of three possible values, 30 min, 60 min, and 90 min, if the factor is time. It may be an option like male out of two possible options, male and female, if the factor is sex. For the example problem, the treatments/levels are State 1, State 2, State 3, State 4, and State 5.

The factors are classified into fixed factors and random factors, which are explained as follows.

Fixed Factor: There could actually be several levels to a factor. A factor is said to be a fixed factor if its inferences are limited to just the chosen subset of levels – a subset of all the levels that could be associated with it – and not to any other levels. Assume there are 50 machines in a factory if the factor of concern is machines. Only five machines are taken into consideration for the experiment. The factor machine is referred to as a fixed factor if the experiment's conclusions are limited to these five machines. Let the factor's name be Factor A and its level count be *a*. Then its effect is the ratio between the sum of squares of the ANOVA model component with respect to Factor A and the degrees of freedom of that factor as follows.

$$Effect\,of\;Fixed\,Factor\,A = \frac{\sum_{i=1}^{a} A_i^2}{a-1}$$

where
a is the number of levels of Factor A
(*a* – 1) is the degrees of freedom of Factor A
A_{i} is the effect of the i^{th} level of Factor A

Random Factor: A factor may have numerous levels, as previously mentioned. The factor is referred to as a random factor if the inferences of a set of levels chosen from the total number of levels for the purpose of performing the experiment are extended to all the levels of that factor. Assume there are 100 operators at a factory if the factor of concern is operator. Only five operators are taken into consideration for the experiment. The factor operator is referred to as a random factor if the experiment's conclusions are applied to all of its levels. Let Factor A be this random variable.

The effect of the treatments of the random Factor A on the response variable is given by the following formula, where *a* is the number of levels.

Variance component of the random Factor A = $\sigma_A{}^2$, where $\sigma_A{}^2$ is the variance.

Replication: Only a select few factors will be taken into account in an experiment. Unaccounted factors could have an impact and affect the experiment's results in some way. Repeated observations of the experiment for the response variable under the same experimental condition for each of the potential experimental conditions are used to estimate this error. Repeated observations under the same experimental condition are called replications.

Consider the example of analysing sales revenues of different states, as shown in Table 15.2.

Table 15.2 Sales Data for States

		State j								
		1	*2*	*3*	.	.	*j*	.	.	*a*
	1	Y_{11}	Y_{12}	Y_{13}	.	.	Y_{1j}	.	.	Y_{1a}
	2	Y_{21}	Y_{22}	Y_{23}	.	.	Y_{2j}	.	.	Y_{2a}
	3	Y_{31}	Y_{32}	Y_{33}	.	.	Y_{3j}	.	.	Y_{3a}
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
Replication i	*i*	Y_{i1}	Y_{i2}	Y_{i3}	.	.	Y_{ij}	.	.	Y_{ia}
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
	n	Y_{n1}	Y_{n2}	Y_{n3}	.	.	Y_{nj}	.	.	Y_{na}

The terminology in Table 15.2 is as follows.

- Y_{ij} is the *i*th sales data (*i*th replication) of the *j*th state.
- State is the factor which has an effect on the response variable Y_{ij}. Let it be Factor A.
- *a* is the number of states where the washing machine is sold. This is also known as the number of levels/treatments of Factor A.
- *n* is the number of sales data under each state. This is also known as the number of replications under each level of Factor A.

The ANOVA model of this situation is shown as follows.

$$Y_{ij} = \mu + A_j + e_{ij}$$

where
μ is the overall mean of the sales revenue
A_j is the effect of the *j*th treatment of Factor A (state) on the response
e_{ij} is the random error associated with the *i*th replication of the *j*th treatment/level of Factor A

The decomposition of total variability into its component parts is called analysis of variance. The partitioning of the total variability is presented as follows.

The total corrected sum of squares is as follows.

$$SS_{Total} = \sum_{i=1}^{n}\sum_{j=1}^{a}\left(Y_{ij} - \overline{Y_{..}}\right)^2$$

This equation is re-written by adding and subtracting the term $\overline{Y_{.j}}$ per the following presentation.

$$\sum_{i=1}^{n}\sum_{j=1}^{a}\left(Y_{ij} - \overline{Y_{..}}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{a}\left[\left(\overline{Y_{.j}} - \overline{Y_{..}}\right) + \left(Y_{ij} - \overline{Y_{.j}}\right)\right]^2$$

$$\sum_{i=1}^{n}\sum_{j=1}^{a}\left(Y_{ij} - \overline{Y_{..}}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{a}\left(\left(\overline{Y_{.j}} - \overline{Y_{..}}\right)^2\right) + \sum_{i=1}^{n}\sum_{j=1}^{a}\left(Y_{ij} - \overline{Y_{.j}}\right)^2 + 2\left(\sum_{i=1}^{n}\sum_{j=1}^{a}\left(\overline{Y_{.j}} - \overline{Y_{..}}\right)\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{a}\left(Y_{ij} - \overline{Y_{.j}}\right)\right)$$

The last term on the right-hand side of the equation is equal to 0 because of the following.

$$\sum_{i=1}^{n}\left(Y_{ij} - \overline{Y_{.j}}\right) = \left(Y_{.j} - n\overline{Y_{.j}}\right) = \left(Y_{.j} - n\left(\frac{Y_{.j}}{n}\right)\right) = 0$$

Therefore, the corrected total sum of squares is reduced to the following.

$$\sum_{i=1}^{n}\sum_{j=1}^{a}\left(Y_{ij} - \overline{Y_{..}}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{a}\left(\left(\overline{Y_{.j}} - \overline{Y_{..}}\right)^2\right) + \sum_{i=1}^{n}\sum_{j=1}^{a}\left(Y_{ij} - \overline{Y_{.j}}\right)^2$$

$$\sum_{i=1}^{n}\sum_{j=1}^{a}\left(Y_{ij}-\overline{Y}_{..}\right)^2 = n\sum_{j=1}^{a}\left(\overline{Y}_{.j}-\overline{Y}_{..}\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{a}\left(Y_{ij}-\overline{Y}_{.j}\right)^2$$

Total Sum of Squares = Sum of Squares of Treatment + Sum of Squares of Error.

$$SS_{Total} = SS_A + SS_{Error}$$

$$Mean\, sum\, of\, squares\, of\, Treatment = \frac{Sum\, of\, squares\, of\, Treatment}{Degrees\, of\, freedom\, of\, Treatment} = \frac{SS_A}{a-1}$$

$$Mean\, sum\, of\, squares\, of\, Error = \frac{Sum\, of\, squares\, of\, Error}{Degrees\, of\, freedom\, of\, Error}$$

$$= \frac{SS_{Error}}{a(n-1)} = \frac{SS_{Error}}{(N-a)}$$

where
N is the total number of observations in the experiment (*an*)

The hypotheses of the ANOVA model are as listed.

$$Null\, Hypothesis, H_0 : \overline{Y_1} = \overline{Y_2} = \overline{Y_3} = \cdots = \overline{Y_j} = \cdots = \overline{Y_a}$$

Alternate Hypothesis, H_1: Treatment means are not equal for at least one pair of the treatment means.

The generalised results are summarised in Table 15.3.

15.2 ANOVA With Single Factor (Completely Randomised Design) Using ANOVA: Single-Factor Function

The concept and application of a completely randomised design are demonstrated through an example [3].

The Beta Engineering Company investigates a quality problem in terms of surface finish. The investigator considers four different machines, 1, 2, 3, and 4, for this analysis. For each machine, four different observations are made. The four investigations of each machine are carried out at different points in time [early part of forenoon (Period 1), later part of forenoon (Period 2), early part of afternoon (Period 3), and later part of afternoon (Period 4)] randomly in the day shift of a day. In this experiment, four operator grades and four operators under each grade are considered for assignment to the machines. Here, O_{ij} is the j^{th} operator under the i^{th} operator grade. The objective of this study is to check whether there are significant differences among the machines in terms of surface finish of the component that is manufactured in those machines. Hence, it is called a completely randomised design.

A generalised experimental design of assigning the operators of the operator grades to different period and machine combinations is shown in Table 15.4. The values of the surface finish (response variable) as per the experimental design are shown in Table 15.5.

The ANOVA model of the completely randomised design with reference to the data shown in Table 15.5 is as follows.

$$Y_{ij} = \mu + T_j + e_{ij}$$

where
μ is the overall mean
Y_{ij} is the i^{th} observation under the j^{th} treatment of the factor Operator Grade
T_j is the effect of the j^{th} treatment of the factor Operator Grade
e_{ij} is the random error associated with the i^{th} observation under the j^{th} treatment of the factor Operator Grade

$Null\,Hypothesis, H_0 : T_1 = T_2 = T_3 = T_4$

Alternate Hypothesis, H_1: Treatment means are not equal for at least one pair of treatment means.

The relationship between different sums of squares of this ANOVA model is shown as follows.

Total sum of squares = Sum of squares of treatments + Sum of squares of errors.

that is, $SS_{Total} = SS_{Treatment} + SS_{Error}$

For the example problem, $SS_{Total} = SS_{Operator\,Grade} + SS_{Error}$

The generalised shortcut formulas to compute the sum of squares of different components of the ANOVA model are as follows.

$$SS_{Total} = \sum_{i=1}^{n}\sum_{j=1}^{a} Y_{ij}^2 - \frac{Y_{..}^2}{N}$$

$$SS_{Treatment} = \sum_{j=1}^{a} \frac{Y_{.j}^2}{n} - \frac{Y_{..}^2}{N}$$

$$SS_{Error} = SS_{Total} - SS_{Treatment}$$

where
a is the number of treatments (a = 4)
n is the number of replications under each treatment (n = 4 for each Operator Grade)
$Y_{..}$ is the sum of Y_{ij} over all values of i and j
$Y_{.j}$ is the sum of Y_{ij} over all values of i for a given j
N is the total number of observations in the experiments (4 × 4 = 16)

The distribution of the total sum of squares of this design is shown diagrammatically in Figure 15.1. The generalised results of the problem of this design are summarised in Table 15.6.

Table 15.3 Generalised Results of ANOVA With Single Factor

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Sum of Squares (MSS)*	*F Ratio*
Between Treatments	$a - 1$	$SS_{Treatment}$	$\frac{SS_{Treatment}}{a-1}$	$\frac{MSS_{Treatment}}{MSS_{Error}}$
Within Treatments	$a\ (n - 1)$	SS_{Error}	$\frac{SS_{Error}}{a(n-1)}$	
Total	$N - 1$	SS_{Total}		

Table 15.4 Data on Surface Finish

		Machine			
		1	*2*	*3*	*4*
Replications	Period 1	O_{11}	O_{21}	O_{31}	O_{12}
	Period 2	O_{41}	O_{13}	O_{22}	O_{32}
	Period 3	O_{14}	O_{23}	O_{33}	O_{42}
	Period 4	O_{24}	O_{34}	O_{43}	O_{44}

Table 15.5 Surface Finish Values of Completely Randomised Design

	Operator Grade j			
	1	*2*	*3*	*4*
Replication i	$Y_{11}\ (O_{11})$	$Y_{12}(O_{21})$	$Y_{13}(O_{31})$	$Y_{14}(O_{41})$
	$Y_{21}(O_{12})$	$Y_{22}(O_{22})$	$Y_{23}(O_{32})$	$Y_{24}(O_{42})$
	$Y_{31}(O_{13})$	$Y_{32}(O_{23})$	$Y_{33}(O_{33})$	$Y_{34}(O_{43})$
	$Y_{41}\ (O_{14})$	$Y_{42}(O_{24})$	$Y_{43}(O_{34})$	$Y_{43}(O_{44})$
Column Total $Y.j$	$Y_{\bullet 1}$	$Y_{\bullet 2}$	$Y_{\bullet 3}$	$Y_{\bullet 4}$

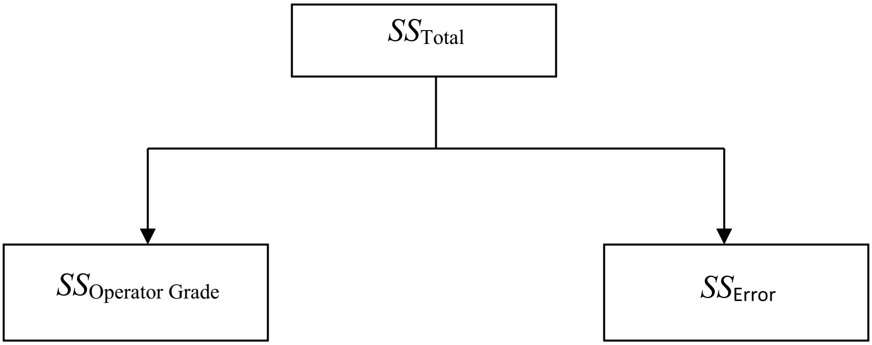


Figure 15.1 Distribution of total sum of squares of completely randomised design

Table 15.6 Results of Completely Randomised Design

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Sum of Squares*	*F Ratio*
Between Treatments	$a - 1$	$SS_{Operator\ Grade}$		$\frac{MSS_{Operator\ Grade}}{MSS_{Error}}$
Within Treatments	$a\ (n - 1)$	SS_{Error}	$\frac{SS_{Error}}{a(n-1)}$	
Total	$N - 1$	SS_{Total}		

Example 15.1

An agricultural officer wishes to investigate how four different fertiliser brands, A, B, C, and D, affect the yield (measured in tonnes) of a particular crop. For each fertiliser brand, four plots were used to create four replications of that fertiliser brand. Completely random design, often known as single-factor ANOVA, is the name of this design. Table 15.7 displays the experiment's data.

1. Write the corresponding ANOVA model.

2. Check whether each component of the ANOVA model has an effect on the yield of the crop at a significance level of 5%.

Solution

The data for the given problem are shown in Table 15.8.

1. The ANOVA model of this problem is as follows.

$$Y_{ij} = \mu + T_j + e_{ij}$$

where
Y_{ij} is the yield of the crop with respect to the i^{th} replication(plot) under the j^{th} treatment of fertiliser brand
μ is the overall mean
T_j is the effect of the j^{th} fertiliser brand on the yield of the crop
e_{ij} is the random error associated with the i^{th} replication of the yield of the crop with respect to the j^{th} fertiliser brand

The hypotheses of this experiment are as listed.

H_0: There are no significant differences among the yields of the crop under different levels of the factor Fertiliser brand.
H_1: There are significant differences among the yields of the crop under different levels of the factor Fertiliser brand.

Table 15.7 Yields for Different Fertilisers

		Fertiliser Brand			
		A	*B*	*C*	*D*
Replication (Plot)	1	100	150	120	70
	2	80	70	110	100
	3	68	90	85	78
	4	125	138	60	124

Table 15.8 Data for Example 15.1

		Fertiliser Brand			
		A	*B*	*C*	*D*
	1	100	150	120	70
	2	80	70	110	100
Replication (Plot)	3	68	90	85	78
	4	125	138	60	124

The Excel commands and screenshots for ANOVA with a single factor are presented in the following pages.

Step 1: Enter the data in four different columns, columns A, B, C, and D of the Excel sheet, as shown in Figure 15.2

Step 2: Click the Data button at the top of the ribbon, and then click the Data Analysis button, which is at the extreme right of the data button. This gives the screenshot in Figure 15.3.

Step 3: Click the option ANOVA: Single Factor from the dropdown menu shown in Figure 15.3. The response to this action is shown in Figure 15.4.

Step 4: Perform the following sub-steps in the dropdown menu of Figure 15.4 as shown in Figure 15.5.

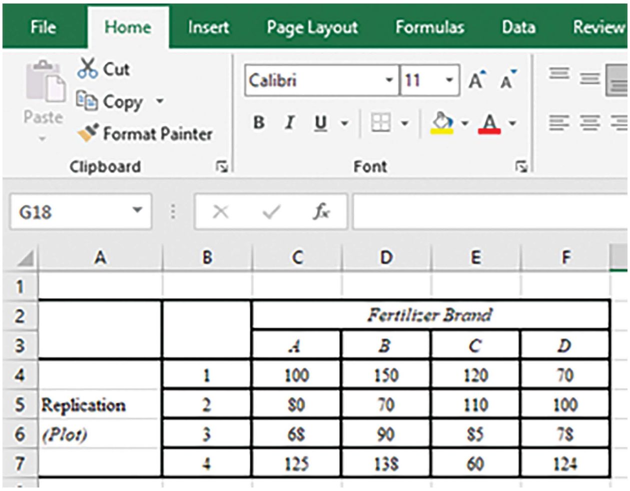
4.1: Copy the cell range $C$3:$F$7 in the Input Range of the dropdown menu shown in Figure 15.5.

4.2: Click the label in First Row box of the dropdown menu of Figure 15.5 to show the column labels, *A*, *B*, *C*, and *D*, which are the brands of the factor Fertiliser Brand.

4.3: In the dropdown menu, retain the value of α at 0.05. If it is different from 0.05, enter the corresponding value in that box to modify it.

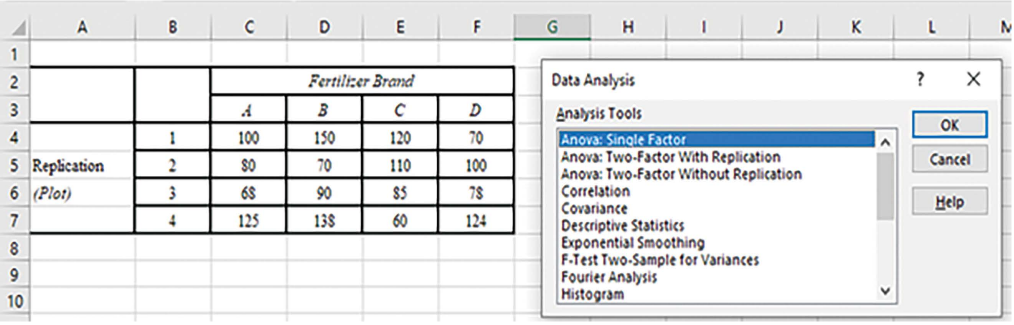
4.4: Click New Worksheet Ply in the dropdown menu to show the result in a new worksheet, as shown in Figure 15.6.

Step 5: Click the OK button in the dropdown menu shown in Figure 15.5, and the response to this action is shown in the screenshot in Figure 15.6, which gives the results of the ANOVA.



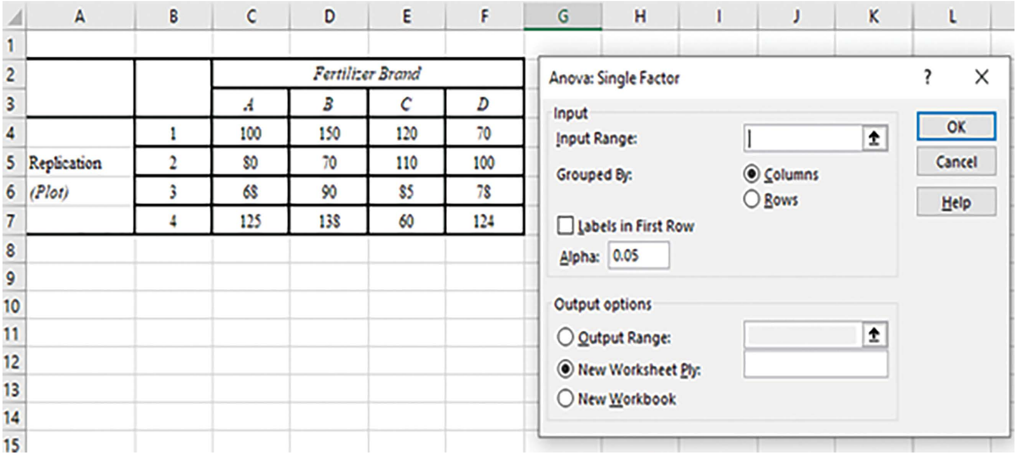
		Fertilizer Brand			
		A	B	C	D
	1	100	150	120	70
Replication	2	80	70	110	100
(Plot)	3	68	90	85	78
	4	125	138	60	124

Figure 15.2 Screenshot of input of Example 15.1



		Fertilizer Brand			
		A	B	C	D
	1	100	150	120	70
Replication	2	80	70	110	100
(Plot)	3	68	90	85	78
	4	125	138	60	124

Figure 15.3 Screenshot after clicking Data button and Data Analysis button in the ribbon



		Fertilizer Brand			
		A	B	C	D
	1	100	150	120	70
Replication	2	80	70	110	100
(Plot)	3	68	90	85	78
	4	125	138	60	124

Anova: Single Factor

Input

Input Range:

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Alpha: 0.05

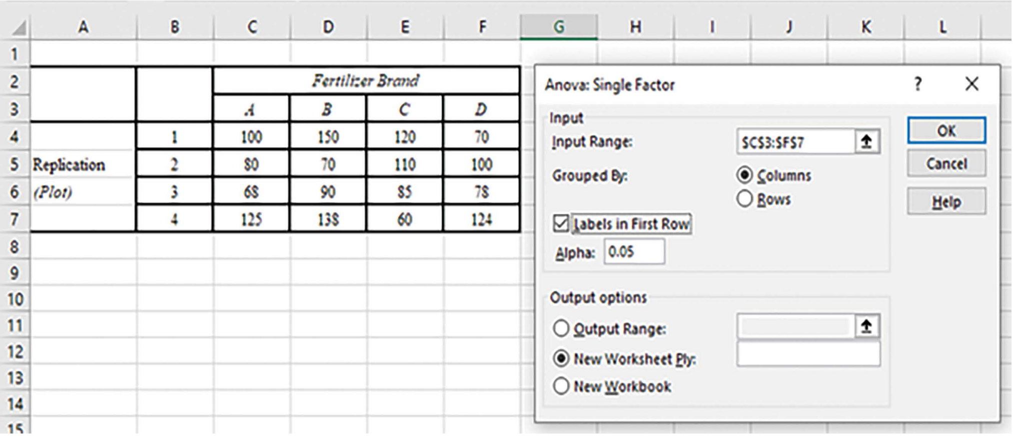
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 15.4 Screenshot after clicking ANOVA: Single Factor option from the dropdown menu in Figure 15.3



		Fertilizer Brand			
		A	*B*	*C*	*D*
	1	100	150	120	70
Replication	2	80	70	110	100
(Plot)	3	68	90	85	78
	4	125	138	60	124

Anova: Single Factor

Input

Input Range: $C$3:$F$7

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Alpha: 0.05

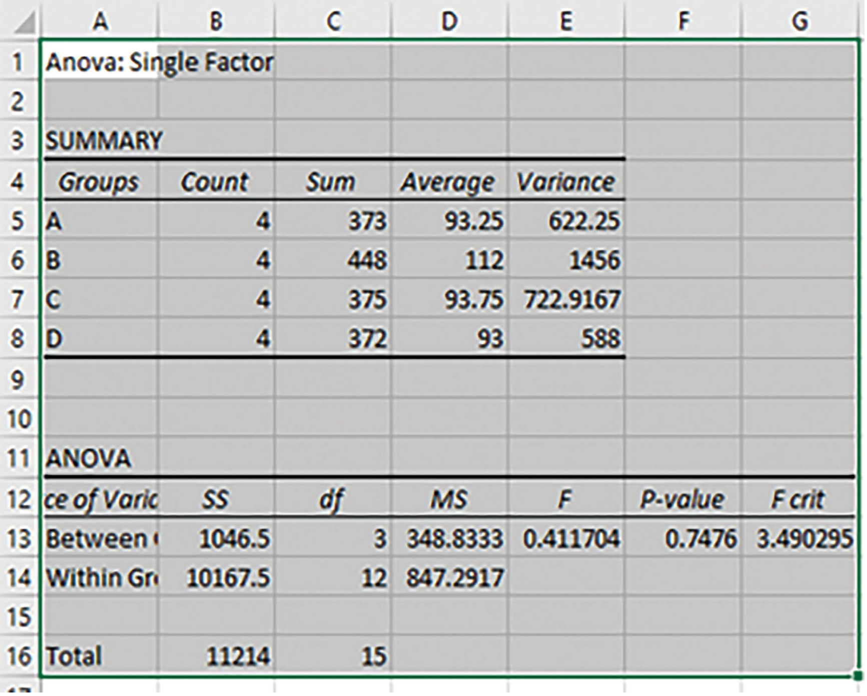
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 15.5 Screenshot of the actions of Step 4



Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
A	4	373	93.25	622.25
B	4	448	112	1456
C	4	375	93.75	722.9167
D	4	372	93	588

ANOVA

ce of Varic	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between (	1046.5	3	348.8333	0.411704	0.7476	3.490295
Within Gr	10167.5	12	847.2917			
Total	11214	15				

Figure 15.6 Screenshot of the response to clicking OK button in the dropdown menu in Figure 15.5 (ANOVA result)

Since the *p* value (0.7476) at the right tail of the F distribution is more than the given level of significance of 0.05, the null hypotheses is to be accepted.

Inference: There are no significant differences among the yields of crops of different levels of the factor Fertiliser brand.

Example 15.2

The R&D manager of a manufacturing company is unsure whether the sales region has an impact on the sales revenue (in crores of rupees). He opted to utilise a completely randomised design after gathering data for six periods under each of the six sales regions. The appropriate data are displayed in Table 15.9.

1. Write the corresponding ANOVA model.
2. Check whether the component of the ANOVA model has an effect on the sales revenue at a significance level of 5%.

Solution

The data for the given problem are shown in Table 15.10.

1. The ANOVA model of this problem is:

$$Y_{ij} = \mu + T_j + e_{ij}$$

where
Yij is the sales revenue with respect to the *i*th replication under the *j*th treatment of sales region
μ is the overall mean
T_j is the effect of the *j*th sales region on the sales revenue
e_{ij} is the random error associated with the sales revenue with regard to *i*th the replication under the *j*th sales region

Table 15.9 Sales of Sales Regions

		Sales Region					
		A	*B*	*C*	*D*	*E*	*F*
Replication	1	20	9	15	22	9	10
	2	25	7	14	8	12	13
	3	20	8	12	9	15	17
	4	15	13	30	11	20	15
	5	18	11	25	10	16	8
	6	28	20	17	10	20	20

Table 15.10 Data for Example 15.2

	Sales Region					
	A	*B*	*C*	*D*	*E*	*F*
Replication	20	9	15	22	9	10
	25	7	14	8	12	13
	20	8	12	9	15	17
	15	13	30	11	20	15
	18	11	25	10	16	8
	28	20	17	10	20	20

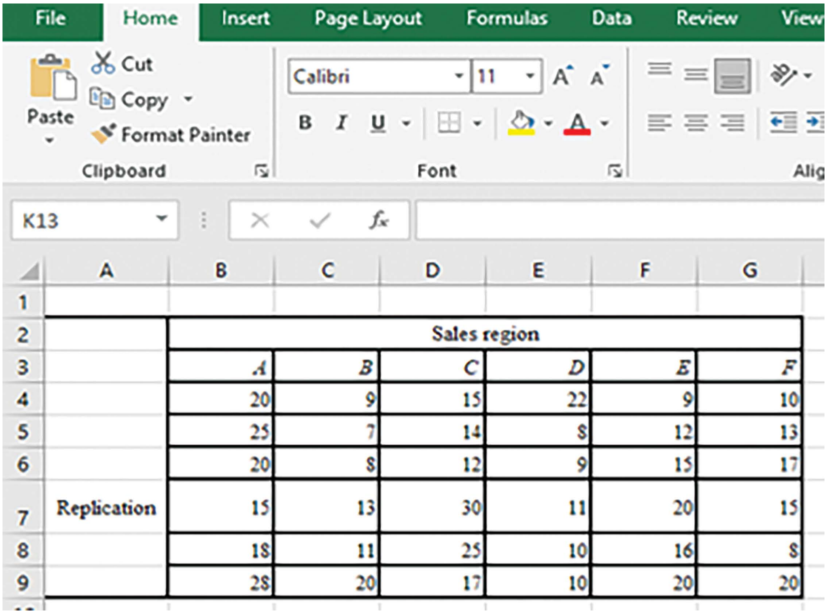
Hypotheses of Factor: Sales Region

H_0: There are no significant differences between the sales regions in terms of the sales revenue.
H_1: There are significant differences between the sales regions in terms of the sales revenue.

2. The steps of performing ANOVA for this problem in Excel are as follows.

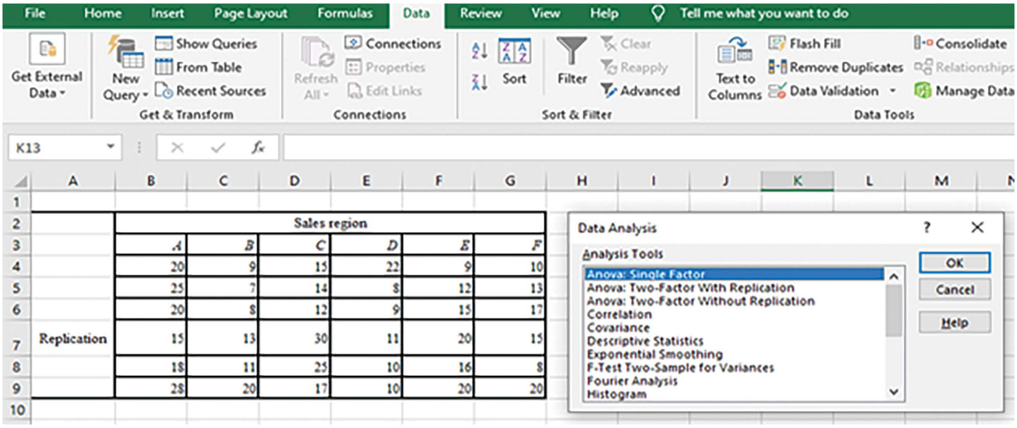
Step 1: Enter the data of the given problem in six different columns, that is, *A*, *B*, *C*, *D*, *E*, and *F* of the Excel sheet, as shown in the screenshot in Figure 15.7.
Step 2: Click the Data Analysis button, which is in the submenu of the Data button. The screenshot of the response to this action is shown in Figure 15.8.
Step 3: Click the option ANOVA: Single Factor from the dropdown menu shown in Figure 15.8. The response to this action is shown in Figure 15.9.
Step 4: Perform the following sub-steps in the dropdown menu of Figure 15.9, as shown in Figure 15.10.

4.1: Copy the cell range $C$3:$H$9 in the Input Range of the dropdown menu shown in Figure 15.10.
4.2: Click label in First Row box of the dropdown menu of Figure 15.10 to show the column labels, *A*, *B*, *C*, *D*, *E*, and *F*, which are the treatments of the factor Sales Region.
4.3: In the dropdown menu, retain the value of α at 0.05. If it is different from 0.05, enter the corresponding value in that box to modify it.
4.4: Click New Worksheet Ply in the dropdown menu to show the result in a new worksheet, as shown in Figure 15.11.



	Sales region					
	A	*B*	*C*	*D*	*E*	*F*
	20	9	15	22	9	10
	25	7	14	8	12	13
	20	8	12	9	15	17
Replication	15	13	30	11	20	15
	18	11	25	10	16	8
	28	20	17	10	20	20

Figure 15.7 Screenshot of data of given problem



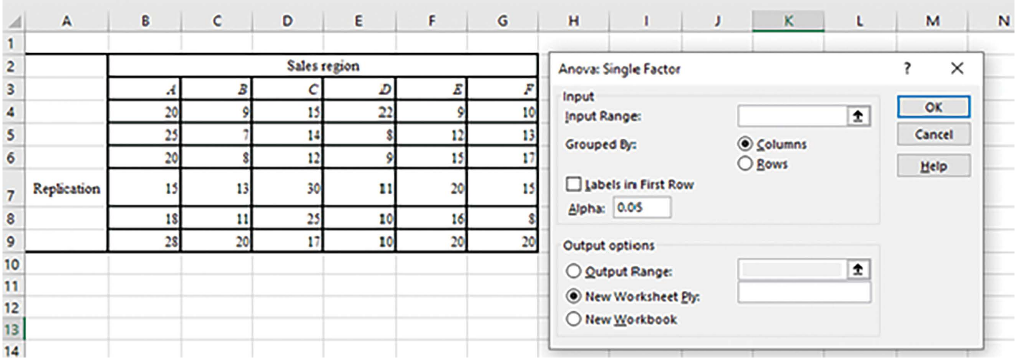
	A	B	C	D	E	F	G
2		Sales region					
3		*A*	*B*	*C*	*D*	*E*	*F*
4		20	9	15	22	9	10
5		25	7	14	8	12	13
6		20	8	12	9	15	17
7	Replication	15	13	30	11	20	15
8		18	11	25	10	16	8
9		28	20	17	10	20	20

Data Analysis

Analysis Tools

- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram

Figure 15.8 Screenshot of clicking Data button and then Data Analysis button



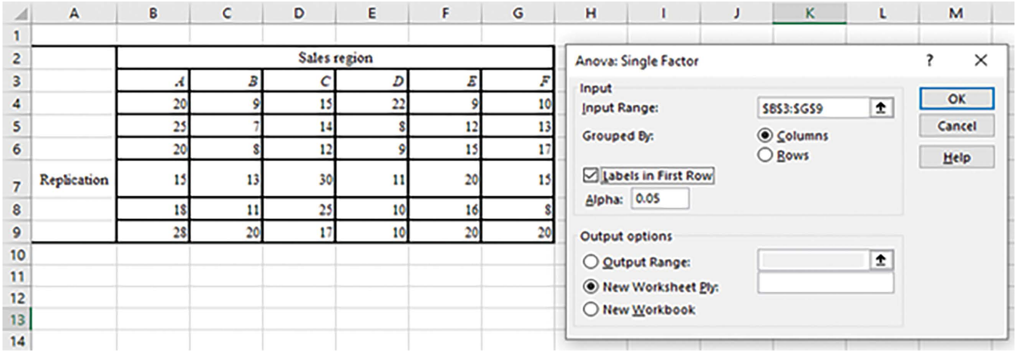
	A	B	C	D	E	F	G
2		Sales region					
3		*A*	*B*	*C*	*D*	*E*	*F*
4		20	9	15	22	9	10
5		25	7	14	8	12	13
6		20	8	12	9	15	17
7	Replication	15	13	30	11	20	15
8		18	11	25	10	16	8
9		28	20	17	10	20	20

Anova: Single Factor

Input
Input Range:
Grouped By: ● Columns ○ Rows
□ Labels in First Row
Alpha: 0.05

Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook

Figure 15.9 Screenshot of data and sequence of clicks of Data button and Data Analysis sub-button



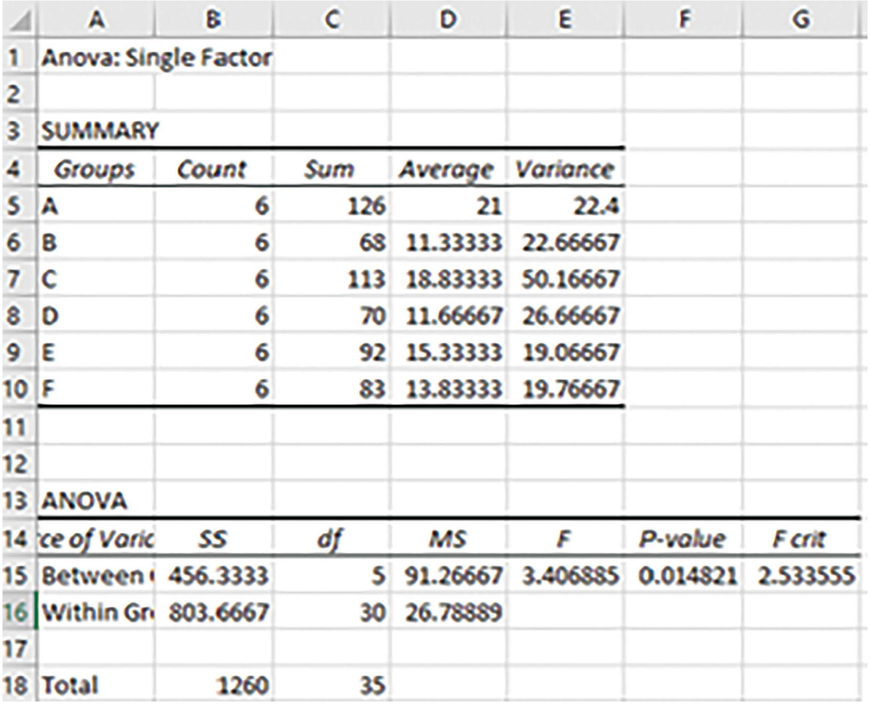
	A	B	C	D	E	F	G
2		Sales region					
3		*A*	*B*	*C*	*D*	*E*	*F*
4		20	9	15	22	9	10
5		25	7	14	8	12	13
6		20	8	12	9	15	17
7	Replication	15	13	30	11	20	15
8		18	11	25	10	16	8
9		28	20	17	10	20	20

Anova: Single Factor

Input
Input Range: $B$3:$G$9
Grouped By: ● Columns ○ Rows
☑ Labels in First Row
Alpha: 0.05

Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook

Figure 15.10 Screenshot after clicking ANOVA: Single Factor option from the dropdown menu in Figure 15.9 and implementing Step 4



	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	A	6	126	21	22.4		
6	B	6	68	11.33333	22.66667		
7	C	6	113	18.83333	50.16667		
8	D	6	70	11.66667	26.66667		
9	E	6	92	15.33333	19.06667		
10	F	6	83	13.83333	19.76667		
11							
12							
13	ANOVA						
14	*ce of Vari*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
15	Between	456.3333	5	91.26667	3.406885	0.014821	2.533555
16	Within Gr	803.6667	30	26.78889			
17							
18	Total	1260	35				

Figure 15.11 Screenshot of the results of clicking the OK button in the dropdown menu in Figure 15.10

Step 5: Click the OK button in the dropdown menu shown in Figure 15.10, and the response to this action is shown in the screenshot in Figure 15.11, which gives the results of the ANOVA.

Since the value of p (0.014821) at the right tail of the F distribution in Figure 15.11 is less than the given level of significance of 0.05, reject the null hypothesis.

Inference: There are significant differences between sales regions in terms of sales revenue.

15.3 Randomised Complete Block Design Using ANOVA: Two-Factor Without Replication Function

The concept of randomised complete block design [1, 2, 4] is demonstrated through the following example.

Consider the Beta Engineering Company's example given in Section 14.2 on the surface finish quality of the components produced by four different machines, 1, 2, 3, and 4, stated under a completely randomised design. In that example, four inspections are carried out at four different intervals, early part of forenoon (Period 1), later part of forenoon (Period 2), early part of afternoon (Period 3), and later part of afternoon (Period 4). In this experiment (Table 15.4), the objective is to test whether there are significant differences among the operator grades, which is considered a factor. In Table 15.4, one can check the fact that in Period 2 and Period 3, all the operator grades are present. But in Period 1, an operator from operator grade 4 is not used, and in Period 4, an operator from operator grade 1 is not used in that table. Hence, there is no homogeneity in terms of assignment of operator grades to the periods. So, a modified assignment of the operator grades to the "Machine and Period" combinations such that operators from all the operator grades are assigned to each period. There may be several possibilities of such an arrangement, but one such arrangement is shown in Table 15.11.

The rearrangement of the observations in Table 15.11 by keeping the operator grades in columns and periods in rows is shown in Table 15.12.

In Table 15.12:

$Y_i.$ is the total of the observations in the row i, where i = 1, 2, 3, 4

$Y._j$ is the total of the observations in the column j, where j = 1, 2, 3, 4

$Y..$ is the grand total of all the observations in Table 15.12

O_{jk} is the k^{th} operator under the j^{th} operator grade, where j = 1, 2, 3, and 4 and k = 1, 2, 3, and 4

The ANOVA model of the randomised complete block design with reference to the data shown in Table 15.12 is as follows.

$$Y_{ij} = \mu + B_J + T_j + e_{ij}$$

where

μ is the overall mean

Y_{ij} is the observation with respect to the j^{th} treatment of the factor (Operator Grade) and i^{th} block (Period)

B_i is the effect of the i^{th} block (Period)

Table 15.11 Experimental Combinations of Randomised Complete Block Design

		Machine			
		1	*2*	*3*	*4*
Period	1	O_{11}	O_{21}	O_{31}	O_{41}
	2	O_{42}	O_{12}	O_{22}	O_{32}
	3	O_{13}	O_{23}	O_{33}	O_{43}
	4	O_{24}	O_{34}	O_{44}	O_{14}

Table 15.12 Rearranged Data by Keeping Operator Grades in Columns and Periods in Rows

		Operator Grade j				*Row Total*
		1	*2*	*3*	*4*	$Y_{i.}$
	1	Y_{11} (O_{11})	$Y_{12}(O_{21})$	$Y_{13}(O_{31})$	$Y_{14}(O_{41})$	$Y_1.$
Period i	2	$Y_{21}(O_{12})$	$Y_{22}(O_{22})$	$Y_{23}(O_{32})$	$Y_{24}(O_{42})$	$Y_2.$
	3	$Y_{31}(O_{13})$	$Y_{32}(O_{23})$	$Y_{33}(O_{33})$	$Y_{34}(O_{43})$	$Y_3.$
	4	Y_{41} (O_{14})	$Y_{42}(O_{24})$	$Y_{43}(O_{34})$	$Y_{43}(O_{44})$	$Y_4.$
Column Total $Y._j$		$Y._1$	$Y._2$	$Y._3$	$Y._4$	$Y_{..}$

T_j is the effect of the j^{th} treatment of the factor (Operator Grade)
e_{ij} is the random error associated with the i^{th} block (Period) and the j^{th} treatment of the factor (Operator Grade)

Hypothesis with regard to Treatment (Operator Grade)

$: Null\,Hypothesis, H_0 : T_1 = T_2 = T_3 = T_4$

Alternate Hypothesis, H_1: Treatment means are not equal for at least one pair of treatment means.

Hypothesis with regard to Block (Period):

$Null\,Hypothesis, H_0 : B_1 = B_2 = B_3 = B_4$

Alternate Hypothesis, H_1: Block means are not equal for at least one pair of block means.

The relationship between different sums of squares of this model is shown as follows.

Total sum of squares = Sum of squares of blocks + Sum of squares of treatments + Sum of squares of errors.

that is, $SS_{Total} = SS_{Block} + SS_{Treatement} + SS_{Error}$

For the example problem, $SS_{Total} = SS_{Period} + SS_{Operator\,Grade} + SS_{Error}$

The generalised shortcut formulas to compute the sum of squares of different components of the model are given here.

$$SS_{Total} = \sum_{i=1}^{n}\sum_{j=1}^{a} Y_{ij}^2 - \frac{Y_{..}^2}{N}$$

$$SS_{Treatment} = \sum_{j=1}^{a} \frac{Y_{.j}^2}{b} - \frac{Y_{..}^2}{N}$$

$$SS_{Block} = \sum_{i=1}^{b} \frac{Y_{i.}^2}{a} - \frac{Y_{..}^2}{N}$$

$$SS_{Error} = SS_{Total} - SS_{Treatment} - SS_{Block}$$

where
a is the number of treatments, operator grades (a = 4)
b is the number of blocks, periods (b = 4)
$Y_{..}$ is the sum of Y_{ij} over all values of i and j
$Y_{,j}$ is the sum of Y_{ij} over all values of i for a given j
$Y_{i.}$ is the sum of Y_{ij} over all values of j for a given i
N is the total number of observations in the experiments (4 × 4 = 16)

The distribution of the total sum of squares of this design is shown diagrammatically in Figure 15.12.

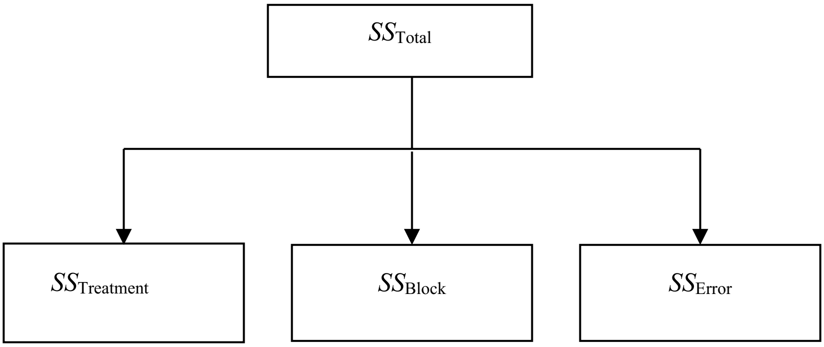


Figure 15.12 Distribution of total sum of squares of randomised complete block design

Table 15.13 Generalised Results of Completely Randomised Design

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Sum of squares*	*F Ratio*
Between Treatments	$a - 1$	$SS_{Treatment}$	$\frac{SS_{Ttreatment}}{a-1}$	$\frac{MSS_{Treatment}}{MSS_{Error}}$
Between Blocks	$b - 1$	SS_{Block}	$\frac{SS_{Block}}{b-1}$	$\frac{MSS_{Block}}{MSS_{Error}}$
Error	N – a – b +1	SS_{Error}	$\frac{SS_{Error}}{N-a-b+1}$	
Total	$N - 1$	SS_{Total}		

The generalised results of this problem per the randomised complete block design are summarised in Table 15.13.

Example 15.3

To make a product, there are four different technological options. The R&D manager believes that the kind of technology may have some bearing on the product's hourly output (measured in units). He chooses to adopt the randomised complete block design, since there can be variation from one plant to another. Table 15.14 has the appropriate data.

1. Write the ANOVA model of this situation.
2. Check whether each component of the ANOVA model has an effect on the output of the product at a significance level of 5%.

Solution

The data shown in Table 15.14 are shown in Table 15.15 in a special format, that is, Technology T_1, Technology T_2, Technology T_3, and Technology T_4 to represent the levels

Table 15.14 Hourly Output of Product

		Technology			
		T_1	T_2	T_3	T_4
Plant	P_1	73	68	74	71
	P_2	73	57	75	52
	P_3	45	38	68	40
	P_4	73	41	75	75

Table 15.15 Data for Example 15.3 in Special Format

		Technology			
		Technology 1	*Technology 2*	*Technology 3*	*Technology 4*
Plant	*Plant 1*	73	68	74	71
	Plant 2	73	57	75	52
	Plant 3	45	38	68	40
	Plant 4	73	41	75	75

of the factor Technology, and Plant P_1, Plant P_2, Plant P_3, and Plant P_4 to represent the levels of the block Plant.

1. The ANOVA model of this problem is:

$$Y_{ij} = \mu + B_i + T_j + e_{ij}$$

where
Y_{ij} is the hourly output of the product with regard to the i^{th} block (Plant) under the j^{th} treatment of Technology
μ is the overall mean
B_i is the effect of the i^{th} block (Plant) on the hourly output of the product
T_j is the effect of the j^{th} treatment of the factor (Technology)
e_{ij} is the random error associated with the hourly output of the product with regard to the i^{th} block (Plant) and j^{th} Technology

2. The different hypotheses of the components of the model are as follows.

Factor: Technology

H_0: There are no significant differences between technologies in terms of the hourly output of the product.
H_1: There are significant differences between technologies for at least one pair of technologies in terms of the hourly output of the product.

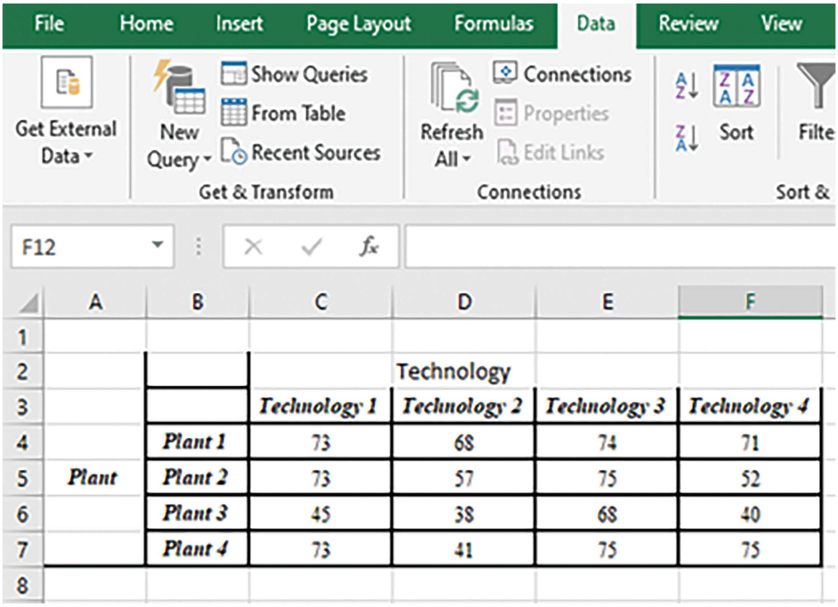
Block: Plant

H_0: There are no significant differences between plants in terms of the hourly output of the product.
H_1: There are significant differences between plants for at least one pair of plants in terms of the hourly output of the product.

The steps to carry out the ANOVA for this problem using Excel are as follows.

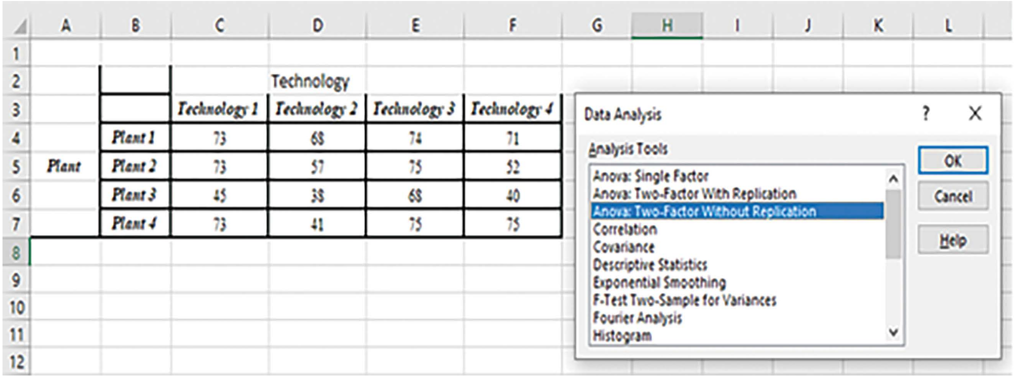
Step 1: Enter the data shown in Table 15.15 in an Excel sheet as shown in Figure 15.13.
Step 2: Click the Data Analysis button, which is in the submenu of the Data button, and the screenshot of the response to this action is shown in Figure 15.14.
Step 3: Click the option ANOVA: Two Factor without Replications from the dropdown menu shown in Figure 15.14. The response to this action is shown in Figure 15.15.
Step 4: Perform the following sub-steps in the dropdown menu of Figure 15.15, as shown in Figure 15.16.

4.1: Copy the cell range $B$3:$F$7 in the Input Range of the dropdown menu of Figure 15.16.
4.2 Click labels in the dropdown menu of Figure 15.16.



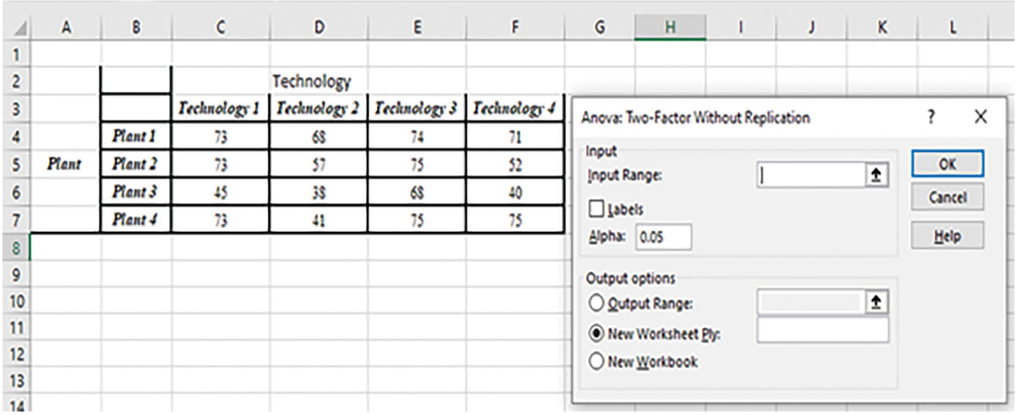
			Technology		
		Technology 1	*Technology 2*	*Technology 3*	*Technology 4*
	Plant 1	73	68	74	71
Plant	*Plant 2*	73	57	75	52
	Plant 3	45	38	68	40
	Plant 4	73	41	75	75

Figure 15.13 Screenshot of data for Example 15.3 in Excel sheet



			Technology		
		Technology 1	*Technology 2*	*Technology 3*	*Technology 4*
	Plant 1	73	68	74	71
Plant	*Plant 2*	73	57	75	52
	Plant 3	45	38	68	40
	Plant 4	73	41	75	75

Figure 15.14 Screenshot after clicking Data button and Data Analysis button for Example 15.3



		Technology			
		Technology 1	Technology 2	Technology 3	Technology 4
	Plant 1	73	68	74	71
Plant	Plant 2	73	57	75	52
	Plant 3	45	38	68	40
	Plant 4	73	41	75	75

Anova: Two-Factor Without Replication

Input

Input Range:

□ Labels

Alpha: 0.05

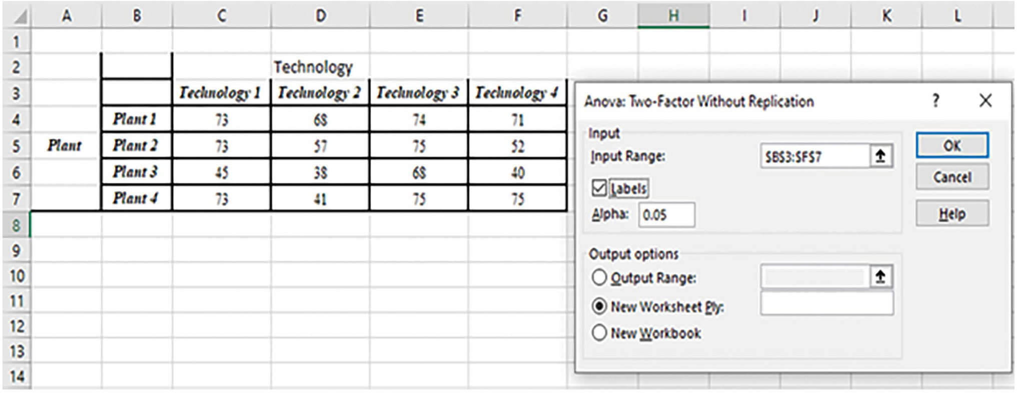
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 15.15 Screenshot of clicking ANOVA: Two Factor without Replications option from the dropdown menu of Figure 15.14



		Technology			
		Technology 1	Technology 2	Technology 3	Technology 4
	Plant 1	73	68	74	71
Plant	Plant 2	73	57	75	52
	Plant 3	45	38	68	40
	Plant 4	73	41	75	75

Anova: Two-Factor Without Replication

Input

Input Range: $B$3:$F$7

☑ Labels

Alpha: 0.05

Output options

○ Output Range:

● New Worksheet Ply:

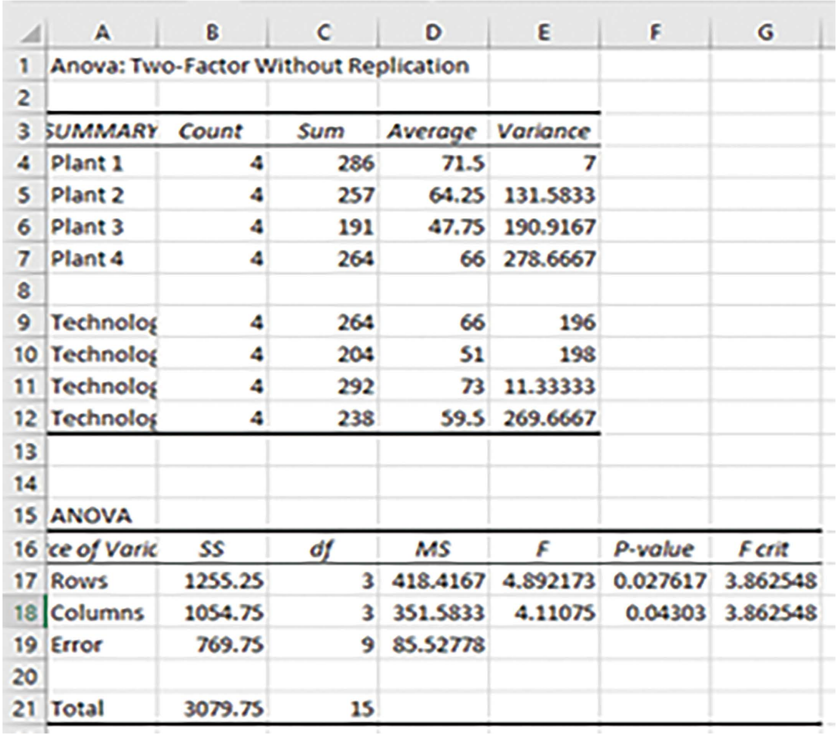
○ New Workbook

Figure 15.16 Screenshot after entering Input Range and clicking Labels in the dropdown menu of Figure 15.15

4.3: In the dropdown menu of Figure 15.16, the value of α is given as 0.05 by default. If it is different from 0.05, enter the corresponding value in that box to modify it.
4.4: In the dropdown menu of Figure 15.16, Click New Worksheet Ply to show the output in a new worksheet.
4.5: Click OK in the dropdown menu of Figure 15.16 to show the output in Figure 15.17.

From Figure 15.17, it can be seen that:

1. The p value at the right tail of the F distribution for the row (Period) is less than the given significance level of 0.05. Hence, reject its null hypothesis.
2. The p value at the right tail of the F distribution for the column (Technology) is less than the given significance level of 0.05. Hence, reject its null hypothesis.



	A	B	C	D	E	F	G
1	Anova: Two-Factor Without Replication						
2							
3	SUMMARY	Count	Sum	Average	Variance		
4	Plant 1	4	286	71.5	7		
5	Plant 2	4	257	64.25	131.5833		
6	Plant 3	4	191	47.75	190.9167		
7	Plant 4	4	264	66	278.6667		
8							
9	Technolog	4	264	66	196		
10	Technolog	4	204	51	198		
11	Technolog	4	292	73	11.33333		
12	Technolog	4	238	59.5	269.6667		
13							
14							
15	ANOVA						
16	ce of Varic	SS	df	MS	F	P-value	F crit
17	Rows	1255.25	3	418.4167	4.892173	0.027617	3.862548
18	Columns	1054.75	3	351.5833	4.11075	0.04303	3.862548
19	Error	769.75	9	85.52778			
20							
21	Total	3079.75	15				

Figure 15.17 Screenshot after clicking the OK button in the dropdown menu of Figure 15.16

Inference: The inferences with respect to the components of the ANOVA model are as follows.

1. There are significant differences between plants in terms of the hourly output of the product.
2. There are significant differences between technologies in terms of the hourly output of the product.

15.4 Latin Square Design Using Excel Sheets

The concept of Latin square design is demonstrated through an example, which is an extension of the example presented in the randomised complete block design.

Consider the template of the example introduced in the randomised complete block design shown in Table 15.11. In that design, homogeneity has been maintained in each period, which means that an operator from each operator grade is assigned to each period (Block). But such homogeneity is not seen in each of the machines. Hence, the operators from the operator grades should be assigned to different combinations of Period and Machine such that in each period, an operator from each of the operator grades is assigned to it, and in each machine, an operator from each of the operator grades is assigned to it. In this design, the period and the machine are treated as blocks, and the operator grade is treated as a factor. This modified design is called Latin square design, which is shown in Table 15.16. In this design, O_{kp} is the operator p from the operator grade k, where $k = 1, 2, 3$, and 4 and $p = 1, 2, 3$, and 4.

The experimental combinations associated with the data are shown in Table 15.17.

In Table 15.17:

Y_{ij} is the value of the surface finish with respect to the i^{th} period and j^{th} machine

$Y_{i}.$ is the total of the observations in row i

$Y._{j}$ is the total of the observations in column j

$Y_{..}$ is the grand total of the observations in Table 15.17.

The data in Table 15.17 are rearranged for different operator grades as shown in Table 15.18.

The ANOVA model of the Latin square design is as follows.

$$Y_{ijk} = \mu + B_i + M_j + T_k + e_{ijk}$$

Table 15.16 Experimental Combinations of Latin Square Design

		Machine			
		1	*2*	*3*	*4*
Period	1	O_{11}	O_{21}	O_{31}	O_{41}
	2	O_{22}	O_{32}	O_{42}	O_{12}
	3	O_{33}	O_{43}	O_{13}	O_{23}
	4	O_{44}	O_{14}	O_{24}	O_{34}

Table 15.17 Experimental Combinations Associated With Data for Latin Square Design

		Machine j				*Row Total ($Y_{i.}$)*
		1	*2*	*3*	*4*	
Period i	1	$(Y_{11})\ O_{11}$	$(Y_{12})\ O_{21}$	$(Y_{13})\ O_{31}$	$(Y_{14})\ O_{41}$	$Y_1.$
	2	$(Y_{21})\ O_{22}$	$(Y_{22})\ O_{32}$	$(Y_{23})\ O_{42}$	$(Y_{24})\ O_{12}$	$Y_2.$
	3	$(Y_{31})\ O_{33}$	$(Y_{32})\ O_{43}$	$(Y_{33})\ O_{13}$	$(Y_{34})\ O_{23}$	$Y_3.$
	4	$(Y_{41})\ O_{44}$	$(Y_{42})\ O_{14}$	$(Y_{43})\ O_{24}$	$(Y_{44})\ O_{34}$	$Y_4.$
Column Total ($Y._j$)		$Y._1$	$Y._2$	$Y._3$	$Y._4$	$Y_{..}$

Table 15.18 Data Latin Square Design With Respect to Operator Grade

	Operator Grade k			
	1	*2*	*3*	*4*
Column Total	$(Y_{11})\ O_{11}$	$(Y_{12})\ O_{21}$	$(Y_{13})\ O_{31}$	$(Y_{14})\ O_{41}$
	$(Y_{24})\ O_{12}$	$(Y_{21})\ O_{22}$	$(Y_{22})\ O_{32}$	$(Y_{23})\ O_{42}$
	$(Y_{33})\ O_{13}$	$(Y_{34})\ O_{23}$	$(Y_{31})\ O_{33}$	$(Y_{32})\ O_{43}$
	$(Y_{42})\ O_{14}$	$(Y_{43})\ O_{24}$	$(Y_{44})\ O_{34}$	$(Y_{41})\ O_{44}$
	Column 1 total	Column 2 total	Column 3 total	Column 4 total

where

μ is the overall mean

Y_{ijk} is the observation in the i^{th} row (Period) and j^{th} column (Machine) for the k^{th} treatment of the factor, *Operator Grade*

B_{i} is the effect of the i^{th} block representing Period

M_{j} is the effect of the j^{th} block representing Machine

T_{k} is the effect of the k^{th} treatment of the factor Operator Grade

e_{ijk} is the random error associated with the i^{th} row (Period) and the j^{th} column(Machine) for the k^{th} treatment of the factor (Operator Grade)

Hypothesis With Regard to Treatment (Operator Grade)

$$Null\,Hypothesis, H_0 : T_1 = T_2 = T_3 = T_4$$

Alternate Hypothesis, H_1: Treatment means are not equal for at least one pair of treatment means.

Hypothesis With Regard to Rows (Period):

$$Null\,Hypothesis, H_0 : B_1 = B_2 = B_3 = B_4$$

Alternate Hypothesis, H_1: Row means are not equal for at least one pair of row means.

Hypothesis With Regard to Columns (Machine):

$$Null\,Hypothesis, H_0 : M_1 = M_2 = M_3 = M_4$$

Alternate Hypothesis, H_1: Column means are not equal for at least one pair of column means.

The relationship between different sums of squares of this ANOVA model is shown as follows.

$$\begin{aligned} Total\,sum\,of\,squares &= Sum\,of\,squares\,of\,treatment\,(Operator) \\ &+ Sum\,of\,squares\,of\,rows\,(Period) \\ &+ Sum\,of\,squares\,of\,columns\,(Machine) \\ &+ Sum\,of\,squares\,of\,error \end{aligned}$$

$$SS_{Total} = SS_{Row} + SS_{Column} + SS_{Treatment} + SS_{Error}$$

For the example problem,

$$SS_{Total} = SS_{Period} + SS_{Machine} + SS_{Operator\,Grade} + SS_{Error}$$

The generalised shortcut formulas to compute the sum of squares of different components of the model are as follows.

$$SS_{Total} = \sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij.}^2 - \frac{Y_{...}^2}{N}\ (Subscript\ k\ is\ dummy\ in\ this\ formula)$$

$$SS_{Treatment} = \sum_{k=1}^{c} \frac{Y_{..k}^2}{a} - \frac{Y_{...}^2}{N}$$

$$SS_{Block(Row)} = \sum_{i=1}^{a} \frac{Y_{i..}^2}{b} - \frac{Y_{...}^2}{N} \quad (Subscript\,k\,is\,dummy\,in\,this\,formula)$$

$$SS_{Block(Column)} = \sum_{j=1}^{b} \frac{Y_{.j.}^2}{a} - \frac{Y_{...}^2}{N} \quad (Subscript\,k\,is\,dummy\,in\,this\,formula)$$

$$SS_{Error} = SS_{Total} - SS_{Treatment} - SS_{Block(Row)} - SS_{Block(Column)}$$

where
a is the number of blocks, Periods ($a = 4$)
b is the number of blocks, Machines ($b = 4$)
c is the number of treatments, operator grades ($c = 4$)
$Y_{...}$ is the sum of Y_{ijk} over all values of *i*, *j*, and k
$Y_{.j.}$ is the sum of Y_{ijk} over all values of *i* and k for a given *j*
$Y_{i..}$ is the sum of Y_{ijk} over all values of *j* and k for a given *i*
$Y_{..k}$ is the sum of Y_{ijk} over all values of *i* and *j* for a given *k*
N is the total number of observations in the experiments (4 × 4 = 16)

The distribution of sum of squares of this design is shown diagrammatically in Figure 15.18.

The generalised results for the example problem of this design are summarised in Table 15.19.

Example 15.4

The director of the Alpha School of Management is interested in examining how faculty influences trainees' average performance on a scale of 0 to 10. Because there may be variation from one subject to another as well as from one batch to another, Latin square design is the relevant design. In this design, five distinct faculty members (A, B, C, D, and E) were allocated to five different subjects of the training programme conducted for five different batches. Table 15.20 contains the data according to this design.

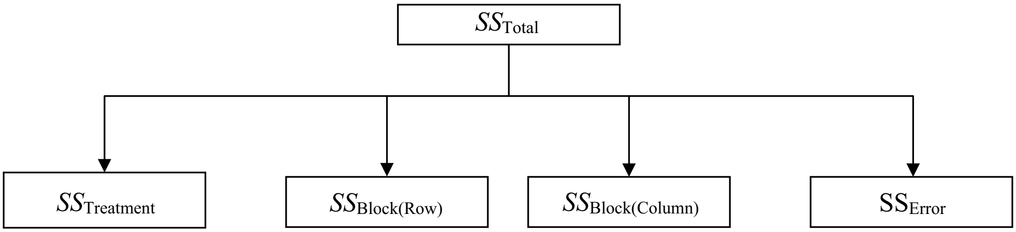


Figure 15.18 Distribution of sum of squares of Latin square design

Table 15.19 Generalised Results of Completely Randomised Design

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Sum of Squares (MSS)*	*F ratio*
Between Treatments	$c - 1$	$SS_{Treatment}$	$\frac{SS_{Treatment}}{c-1}$	$\frac{MSS_{Treatment}}{MSS_{Error}}$
Between Blocks(Row)	$a - 1$	$SS_{Block(Row)}$	$\frac{SS_{Block(Row)}}{a-1}$	$\frac{MSS_{Block(Row)}}{MSS_{Error}}$
Between Blocks(Column)	$b - 1$	$SS_{Block(Column)}$	$\frac{SS_{Block(Column)}}{b-1}$	$\frac{MSS_{Block(Column)}}{MSS_{Error}}$
Error	$N– a– b– c+2$	SS_{Error}	$\frac{SS_{Error}}{N-a-b-c+2}$	
Total	$N - 1$	SS_{Total}		

Table 15.20 Data for Latin Square Design

		Batch				
		1	*2*	*3*	*4*	*5*
	1	$A = 10$	$B = 6$	$C = 6$	$D = 6$	$E = 8$
	2	$B = 7$	$C = 6$	$D = 5$	$E = 1$	$A = 4$
Subject	3	$C = 5$	$D = 3$	$E = 3$	$A = 2$	$B = 1$
	4	$D = 6$	$E = 4$	$A = 1$	$B = 2$	$C = 5$
	5	$E = 4$	$A = 2$	$B = 3$	$C = 8$	D = 9

1. Write the corresponding ANOVA model.
2. Check whether each component of the ANOVA model has an effect on the performance of the participants at a significance level of 5%.

Solution

The data for Example 15.4 are shown in Table 15.21, and the same data are again shown in Table 15.22 by removing the tag of the faculty codes, that is, *A*, *B*, *C*, *D,* and *E*. To compute the sum of squares of the faculty, the data shown in Table 15.21 are rearranged as shown in Table 15.23.

1. The ANOVA model of this problem is:

$$Y_{ijk} = \mu + S_i + B_j + F_k + e_{ijk}$$

Table 15.21 Data for Example 15.4

		Batch				
		1	*2*	*3*	*4*	*5*
Subject	1	*A* = 10	*B* = 6	C = 6	*D* = 6	*E* = 8
	2	*B* = 7	*C* = 6	*D* = 5	*E* = 1	*A* = 4
	3	*C* = 5	*D* = 3	*E* = 3	*A* = 2	*B* = 1
	4	*D* = 6	*E* = 4	*A* = 1	*B* = 2	*C* = 5
	5	*E* = 4	*A* = 2	*B* = 3	C = 8	D = 9

Table 15.22 Rearranged Data for the Data Shown in Table 15.21

		Batch				
		1	2	3	4	5
Subject	1	10	6	6	6	8
	2	7	6	5	1	4
	3	5	3	3	2	1
	4	6	4	1	2	5
	5	4	2	3	8	9

Table 15.23 Rearranged Data for Example 15.4 to Compute Sum of Squares of Faculty

	Faculty				
	A	B	C	D	E
Replication	10	7	5	6	4
	2	6	6	3	4
	1	3	6	5	3
	2	2	8	6	1
	4	1	5	9	8
Column total	19	19	30	29	20

where

Y_{ijk} is the performance of the participants with respect to the kth faculty teaching the ith subject for the jth batch

μ is the overall mean of the performance of the participants

S_i is the effect of the ith subject on the performance of the participants

B_j is the effect of the jth batch on the performance of the participants

F_k is the effect of the kth faculty on the performance of the participants

e_{ijk} is the random error associated with the performance of the participants with respect to the kth faculty teaching the ith subject for the jth batch

The hypotheses of the components of the ANOVA model are as follows.

Factor: Faculty

H_0: There are no significant differences between the faculties in terms of the performance of the participants.

H_1: There are significant differences between the faculties for at least one pair of faculties in terms of the performance of the participants.

Block 1: Subject

H_0: There are no significant differences between the subjects in terms of the performance of the participants.
H_1: There are significant differences between the subjects for at least one pair of subjects in terms of the performance of the participants

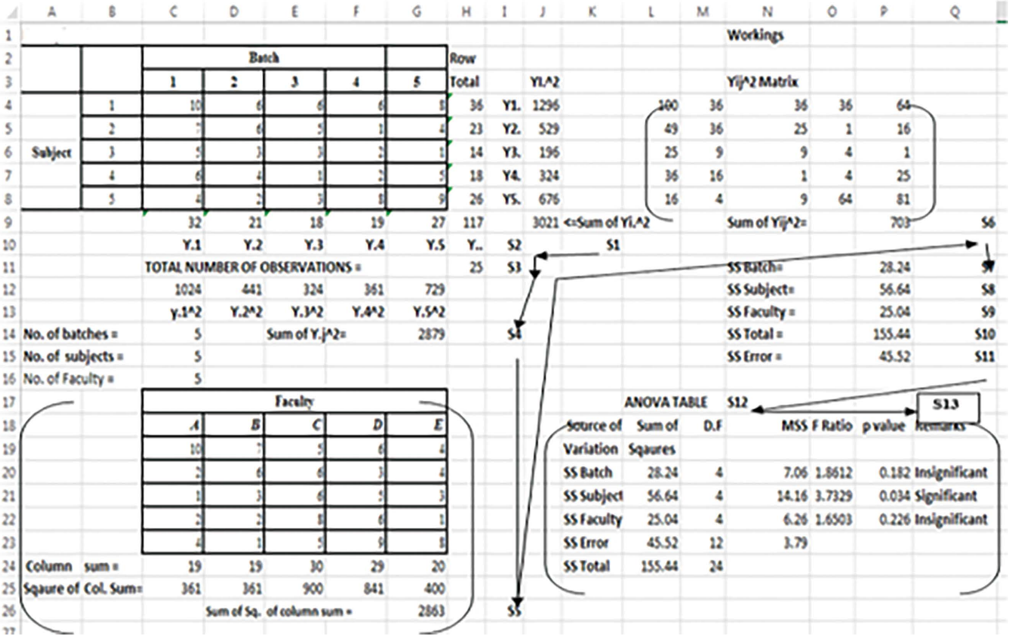
Block 2: Batch

H_0: There are no significant differences between the batches in terms of the performance of the participants.
H_1: There are significant differences between the batches for at least one pair of batches in terms of the performance of the participants.

The data shown in Table 15.22 are copied to an Excel sheet, and the working of the Latin square Design is shown in Figure 15.19. The guidelines for the formulas in Excel of Latin Square Design applied to the example are shown in Figure 15.20.

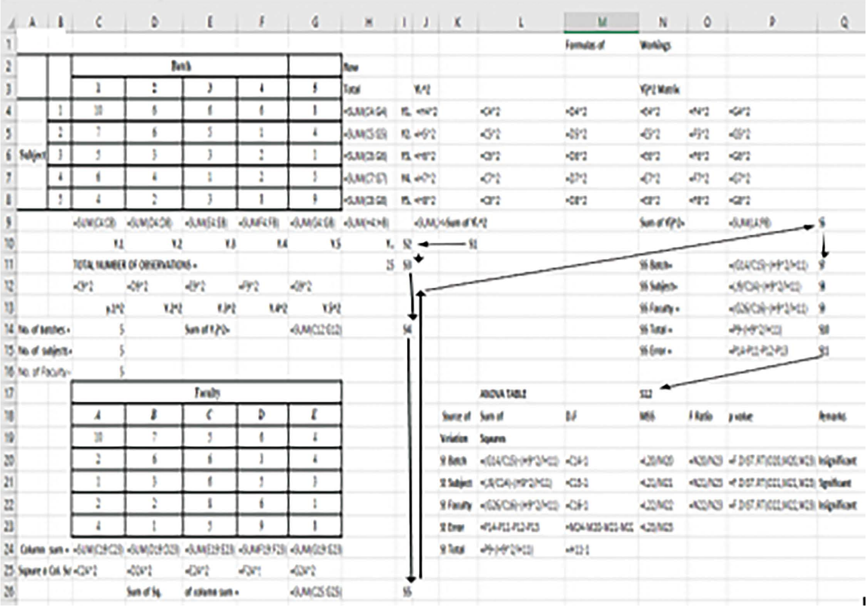
The ANOVA uses the F test, for which the value of p is to be computed for given F value, numerator degrees of freedom, and denominator degrees of freedom. The Excel formula to compute the p value is shown as follows.

$$= \text{F.DIST.RT}\left(\text{X, Deg\_freedom1, Deg\_freedom2}\right)$$



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1														Workings			
2					Batch			Row									
3			1	2	3	4	5	Total		YI.^2				Yij^2 Matrix			
4		1	10	6	6	6	8	36	Y1.	1296		100	36	36	36	64	
5		2	7	6	5	1	4	23	Y2.	529		49	36	25	1	16	
6	Subject	3	5	3	3	2	1	14	Y3.	196		25	9	9	4	1	
7		4	6	4	1	2	5	18	Y4.	324		36	16	1	4	25	
8		5	4	2	3	8	9	26	Y5.	676		16	4	9	64	81	
9			32	21	18	19	27	117		3021	<=Sum of Yi.^2			Sum of Yij^2=		703	S6
10			Y.1	Y.2	Y.3	Y.4	Y.5	Y..	S2		S1						
11			TOTAL NUMBER OF OBSERVATIONS =					25	S3					SS Batch=		28.24	S7
12			1024	441	324	361	729							SS Subject=		56.64	S8
13			y.1^2	Y.2^2	Y.3^2	Y.4^2	Y.5^2							SS Faculty =		25.04	S9
14	No. of batches =		5		Sum of Y.j^2=		2879		S4					SS Total =		155.44	S10
15	No. of subjects =		5											SS Error =		45.52	S11
16	No. of Faculty =		5														
17					Faculty							ANOVA TABLE		S12		S13	
18			A	B	C	D	E				Source of	Sum of	D.F	MSS	F Ratio	p value	Remarks
19			10	7	5	6	4				Variation	Sqaures					
20			2	6	6	3	4				SS Batch	28.24	4	7.06	1.8612	0.182	Insignificant
21			1	3	6	5	3				SS Subject	56.64	4	14.16	3.7329	0.034	Significant
22			2	2	8	6	1				SS Faculty	25.04	4	6.26	1.6503	0.226	Insignificant
23			4	1	5	9	8				SS Error	45.52	12	3.79			
24	Column sum =		19	19	30	29	20				SS Total	155.44	24				
25	Sqaure of Col. Sum=		361	361	900	841	400										
26				Sum of Sq. of column sum =			2863		S5								
27																	

Figure 15.19 Excel working of Latin square design applied in Example 15.4



	Batch					Row Total	Y_i.^2
Subject	1	2	3	4	5		
1	10	6	6	6	1	=SUM(C4:G4)	Y1. =H4^2
2	7	6	5	1	4	=SUM(C5:G5)	Y2. =H5^2
3	5	3	3	2	1	=SUM(C6:G6)	Y3. =H6^2
4	6	4	1	2	5	=SUM(C7:G7)	Y4. =H7^2
5	4	2	3	1	9	=SUM(C8:G8)	Y5. =H8^2

TOTAL NUMBER OF OBSERVATIONS = 25

No. of batches = 5

No. of subjects = 5

No. of Faculty = 5

Sum of Y.j^2 = =SUM(C12:G12)

Faculty				
A	B	C	D	E
10	7	5	6	4
2	6	6	1	4
1	3	6	5	3
2	2	8	6	1
4	1	5	9	8

Column sum =

Square a Col. Su =

Sum of Sq. of column sum =

Formulas of Workings

Yij^2 Matrix

Sum of Yij^2 = S5

SS Batch = S7

SS Subject = S8

SS Faculty = S9

SS Total = S10

SS Error = S11

ANOVA TABLE S12

Source of Variation	Sum of Squares	D.F	MSS	F Ratio	p value	Remarks
SS Batch	[illegible]	[illegible]	[illegible]	[illegible]	=F.DIST.RT([illegible])	Insignificant
SS Subject	[illegible]	[illegible]	[illegible]	[illegible]	=F.DIST.RT([illegible])	Significant
SS Faculty	[illegible]	[illegible]	[illegible]	[illegible]	=F.DIST.RT([illegible])	Insignificant
SS Error	[illegible]	[illegible]	[illegible]			
SS Total	[illegible]	[illegible]				

Figure 15.20 Guidelines for formulas of Excel working of Latin square design applied in Example 15.4

The steps of the working of the Latin square design shown in Figure 15.19 are as follows. The steps are indicated in Figure 15.19 to clarify the sequence of calculations.

Step 1: Find the sum of the squares of row (Subject) totals.
Step 2: Find the grand total of all observations.
Step 3: Compute the total number of observations.
Step 4: Compute the sum of the squares of the column (Batch) totals.
Step 5: Compute the sum of the squares of the column totals of faculty.
Step 6: Find the sum of the squares of the observations.
Step 7: Find the sum of squares of row (subject).
Step 8: Find the sum of squares of column (batch).
Step 9: Compute the sum of squares of the factor Faculty.
Step 10: Compute the sum of squares of the total.
Step 11: Find the sum of squares of errors.
Step 12: Compute the values of ANOVA table.
Step 13: State inferences.

Results:

From Figure 15.19, it can be seen that:

1. The *p* value (0.182151) at the right tail of the F distribution of the column (Batch) is more than the given significance level of 0.05. Hence, accept its null hypothesis.

Inference: There are no significant differences between the levels of the batch in terms of the performance of the participants.

2. The p value (0.033859) at the right tail of the F distribution of the row (Subject) is less than the given significance level of 0.05. Hence, reject its null hypothesis.

Inference: There are significant differences between the levels of the subject in terms of the performance of the participants.

3. The p value (0.225581) at the right tail of the F distribution of the treatment (Faculty) is more than the given significance level of 0.05. Hence, accept its null hypothesis.

Inference: There are no significant differences between the levels of the faculty in terms of the performance of the participants.

15.5 Complete Factorial Experiment With Two Factors Using ANOVA: Two-Factor With Replication Function

Consider an experiment with two factors. Let the first factor be A with a levels/treatments and the other factor be B with b levels/treatments. For each experimental combination of factor A and factor B, n replications are carried out to better represent the error of the experiment. A generalised experimental design of the two-factor complete factorial experiment is shown in Table 15.24.

Table 15.24 Generalised Data Format of Two-Factor Complete Factorial Experiment

		Factor B							
		1	*2*	.	.	*j*	.	.	*b*
	1	Y_{111}	Y_{121}	.	.	Y_{1j1}	.	.	Y_{1b1}
		Y_{112}	Y_{122}	.	.	Y_{1j2}	.	.	Y_{1b2}
		.	.	.	.	.	.	.	.
		.	.	.	.	.	.	.	.
		Y_{11n}	Y_{12n}	.	.	Y_{1jn}	.	.	Y_{1bn}
	2	Y_{211}	Y_{221}	.	.	Y_{2j1}	.	.	Y_{2b1}
		Y_{212}	Y_{222}	.	.	Y_{2j2}	.	.	Y_{2b2}
		.	.	.	.	.	.	.	.
		.	.	.	.	.	.	.	.
		Y_{21n}	Y_{22n}	.	.	Y_{2jn}	.	.	Y_{2bn}
	.			.	.		.	.	
	.			.	.		.	.	
Factor A	*i*	Y_{i11}	Y_{i21}	.	.	Y_{ij1}	.	.	Y_{ib1}
		Y_{i12}	Y_{i22}	.	.	Y_{ij2}	.	.	Y_{ib2}
		.	.	.	.	.	.	.	.
		.	.	.	.	.	.	.	.
		Y_{i1n}	Y_{i2n}	.	.	Y_{ijn}	.	.	Y_{ibn}
	.			.	.		.	.	
	.			.	.		.	.	
	a	Y_{a11}	Y_{a21}	.	.	Y_{aj1}	.	.	Y_{ab1}
		Y_{a12}	Y_{a22}	.	.	Y_{aj2}	.	.	Y_{ab2}
		.	.	.	.	.	.	.	.
		.	.	.	.	.	.	.	.
		Y_{a1n}	Y_{a2n}	.	.	Y_{ajn}	.	.	Y_{abn}

The generalised ANOVA model of the complete factorial design with two factors is as follows.

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + e_{ijk}$$

where

μ is the overall mean

Y_{ijk} is the k^{th} replication for the i^{th} treatment of factor A and the j^{th} treatment of factor B

A_i is the effect of the i^{th} treatment of factor A on the response

B_j is the effect of the j^{th} treatment of factor B on the response

AB_{ij} is the effect of the i^{th} treatment of factor A and the j^{th} treatment of factor B (two-way interaction effect) on the response

e_{ijk} is the random error associated with the k^{th} replication under the i^{th} treatment of factor A and the j^{th} treatment of factor B

Hypothesis With Regard To Treatment of Factor A:

$$Null\,Hypothesis, H_0 : A_1 = A_2 = A_3 = \cdots = A_a$$

Alternate Hypothesis, H_1: Treatment means are not equal for at least one pair of the treatment means of factor A.

Hypothesis With Regard To Treatment of Factor B:

$$Null\,Hypothesis, H_0 : B_1 = B_2 = B_3 = \cdots = B_b$$

Alternate Hypothesis, H_1: Treatment means are not equal for at least one pair of the treatment means of factor B.

Hypothesis With Regard to Interaction Component, AB:

$$Null\,Hypothesis, H_0 : A_1B_1 = A_1B_2 = \cdots\cdot = A_aB_b$$

Alternate Hypothesis, H_1: Interaction means are not equal for at least one pair of interaction means.

The relationship between different sums of squares of this ANOVA model is shown as follows.

$$Total\,sum\,of\,squares = Sum\,of\,squares\,of\,Factor\,A + Sum\,of\,squares\,of\,Factor\,B \\ + Sum\,of\,squares\,of\,Interaction\,AB + Sum\,of\,squares\,of\,error$$

that is, $SS_{Total} = SS_A + SS_B + SS_{AB} + SS_{Error}$

The generalised shortcut formulas to compute the sum of squares of different components of the model are as follows.

$$SS_{Total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}^2 - \frac{Y_{\dots}^2}{N}$$

$$SS_{Row} = \sum_{i=1}^{a} \frac{Y_{i..}^2}{bn} - \frac{Y_{...}^2}{N}$$

$$SS_{Column} = \sum_{j=1}^{b} \frac{Y_{.j.}^2}{an} - \frac{Y_{...}^2}{N}$$

$$SS_{Subtotal} = \sum\nolimits_{i=1}^{a} \sum\nolimits_{j=1}^{b} \frac{Y_{ij.}^2}{n} - \frac{Y_{...}^2}{N}$$

$$SS_{Interaction} = SS_{Subtotal} - SS_{Row} - SS_{Column}$$

$$SS_{Error} = SS_{Total} - SS_{Row} - SS_{Column} - SS_{Interaction}$$

$Y_{...}$ is the sum of Y_{ijk} over all values of i, j, and k
$Y_{.j.}$ is the sum of Y_{ijk} over all values of i and k for a given j
$Y_{i..}$ is the sum of Y_{ijk} over all values of j and k for a given i
$Y_{..k}$ is the sum of Y_{ijk} over all values of i and j for a given value of k
N is the total number of observations in the experiments (abn)

The generalised results of this design are summarised in Table 15.25.

Example 15.5

The sales manager of a renowned textile showroom in Chennai is interested in learning how clients rate the quality of their customer service on a scale of 0 to 10. Both the clients' monthly income level and their nature of profession are treated as fixed factors

Table 15.25 Generalised Results of Complete Factorial Experiment

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Sum of Squares (MSS)*	*F ratio*
Between Rows(A)	a – 1	SS_{Row}	$\frac{SS_{Row}}{a-1}$	$\frac{MSS_A}{MSS_{Error}}$
Between Columns(B)	b – 1	SS_{Column}	$\frac{SS_{Column}}{b-1}$	$\frac{MSS_B}{MSS_{Error}}$
Interaction between Rows and Columns (AB)	(a-1)(b-1)	$SS_{Interaction}$	$\frac{SS_{Interaction}}{(a-1)(b-1)}$	$\frac{MSS_{AB}}{MSS_{Error}}$
Error	ab(n-1)	SS_{Error}	$\frac{SS_{Error}}{ab(n-1)}$	
Total	N – 1	SS_{Total}		

in this experiment. Under each experimental combination, two separate consumers were sampled, and the related ratings are displayed in Table 15.26.

1. Write the ANOVA model of this factorial experiment.
2. Check the significance of each of the components of the ANOVA model at a significance level of 0.05.

Solution

The data for Example 15.5 are shown in Table 15.27.

1. The ANOVA model of this problem is:

$$Y_{ijk} = \mu + I_i + P_j + IP_{ij} + e_{ijk}$$

where

Y_{ijk} is the service quality with respect to the k^{th} replication under i^{th} level of Income and j^{th} level of Profession

μ is the overall mean of the service quality

I_i is the effect of the i^{th} level of Income on the service quality

P_j is the effect of the j^{th} level of Profession on the service quality

IP_{ij} is the interaction effect of the i^{th} level of Income and j^{th} level of Profession on the service quality

e_{ijk} is the random error associated with the k^{th} replication under i^{th} level of Income and j^{th} level of Profession

Table 15.26 Service Quality Ratings Given by Customers

		Nature of Profession			
		Engineer	*Doctor*	*Lawyer*	*Others*
Income Level	Less than ₹10,000	3	3	8	10
		1	8	2	9
	More than ₹10,000	3	10	9	2
		7	4	7	8

Table 15.27 Data for Example 15.5

		Nature of Profession (P)			
		Engineer	*Doctor*	*Lawyer*	*Others*
Income Level (I)	Less than ₹10,000	3	3	8	10
		1	8	2	9
	More than ₹10,000	3	10	9	2
		7	4	7	8

2. The hypotheses of the two factors and their interaction term are as follows.

Factor: Income level

H_0: There are no significant differences between the income levels in terms of the service quality.
H_1: There are significant differences between the income levels in terms of the service quality.

Factor: Profession

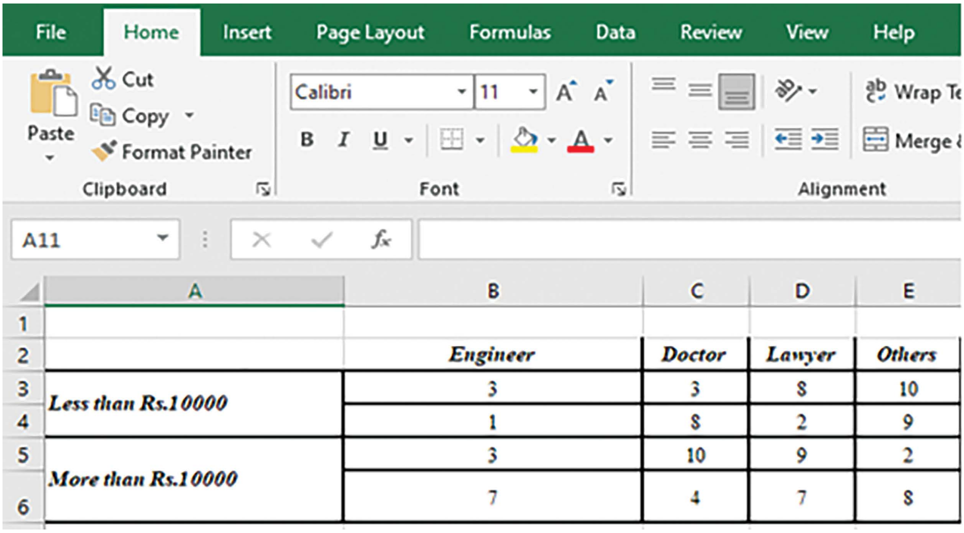
H_0: There are no significant differences between the levels of profession in terms of the service quality.
H_1: There are significant differences between the levels of profession in terms of the service quality.

Interaction: Income Level X Profession

H_0: There are no significant differences between different pairs of interaction terms of income and profession in terms of the service quality.
H_1: There are significant differences between different pairs of interaction terms of income and profession in terms of the service quality.

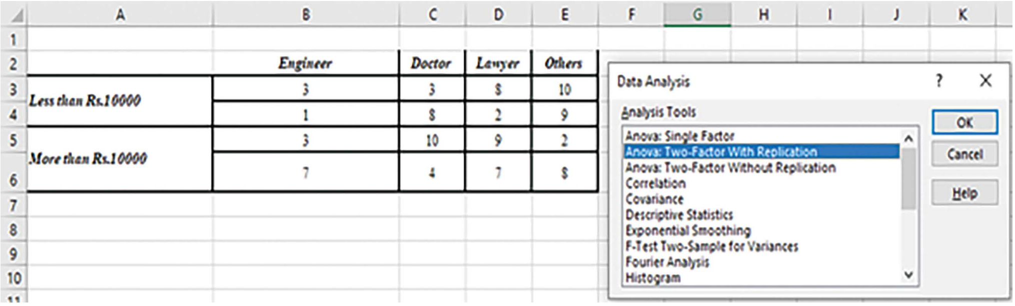
The steps of solving this problem per complete factorial experiment using Excel are as follows.

Step 1: Input the data in an Excel sheet as shown in Figure 15.21.
Step 2 Click the Data Analysis button, which is in the submenu of the Data button; the screenshot of the response to this action is shown in Figure 15.22.



	A	B	C	D	E
1					
2		*Engineer*	*Doctor*	*Lawyer*	*Others*
3	*Less than Rs.10000*	3	3	8	10
4		1	8	2	9
5	*More than Rs.10000*	3	10	9	2
6		7	4	7	8

Figure 15.21 Data for Example 15.5 edited in Excel sheet



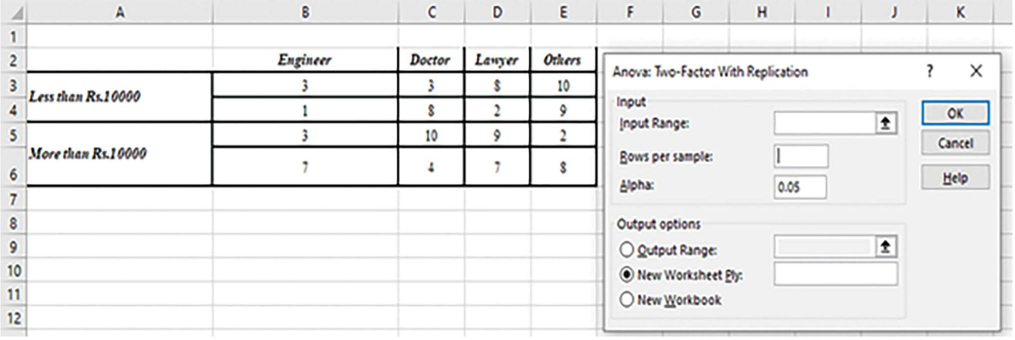
	Engineer	Doctor	Lawyer	Others
Less than Rs.10000	3	3	8	10
	1	8	2	9
More than Rs.10000	3	10	9	2
	7	4	7	8

Data Analysis

Analysis Tools

- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram

Figure 15.22 Screenshot of data and sequence of clicks of Data button and Data Analysis sub-button



	Engineer	Doctor	Lawyer	Others
Less than Rs.10000	3	3	8	10
	1	8	2	9
More than Rs.10000	3	10	9	2
	7	4	7	8

Anova: Two-Factor With Replication

Input

Input Range:

Rows per sample:

Alpha: 0.05

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 15.23 Screenshot after clicking ANOVA: Two Factor with replication in the dropdown menu of Figure 15.22

Step 3: Click the option ANOVA: Two Factor with Replications from the dropdown menu shown in Figure 15.22. The response to this action is shown in Figure 15.23.

Step 4: Perform the following sub-steps in the dropdown menu shown in Figure 15.23 to show the filled version in Figure 15.24.

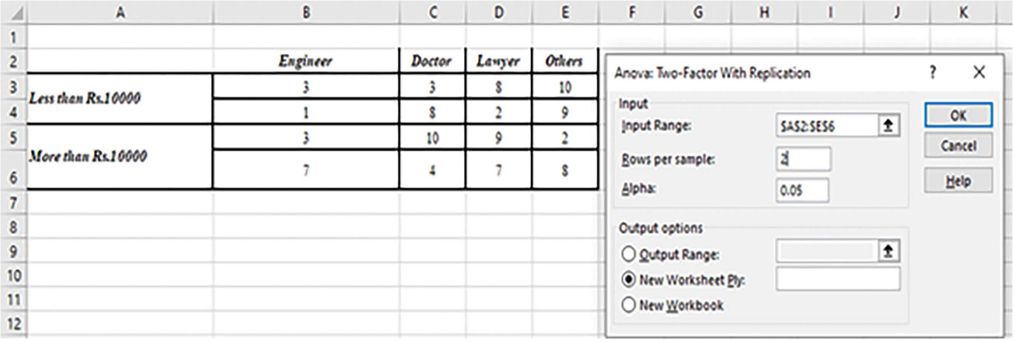
4.1: Copy the cell range $A$2:$E$6 in the Input Range of the dropdown menu shown in Figure 15.24.

4.2: In the dropdown menu of Figure 15.24, enter Rows per Sample as 2, which is the number of replications under each experimental combination.

4.3: In the dropdown menu of Figure 15.24, the value of Alpha is given as 0.05 by default. If it is different from 0.05, enter the corresponding value in that box to modify it.

4.4: In the dropdown menu of Figure 15.24, Click New Worksheet Ply to show the output in a new worksheet.

4.5 Click OK in the dropdown menu of Figure 15.24 to give the results of ANOVA as shown in Figure 15.25.



	A	B	C	D	E
1					
2		*Engineer*	*Doctor*	*Lawyer*	*Others*
3	*Less than Rs.10000*	3	3	8	10
4		1	8	2	9
5	*More than Rs.10000*	3	10	9	2
6		7	4	7	8

Anova: Two-Factor With Replication

Input

Input Range: $A$2:$E$6

Rows per sample: 2

Alpha: 0.05

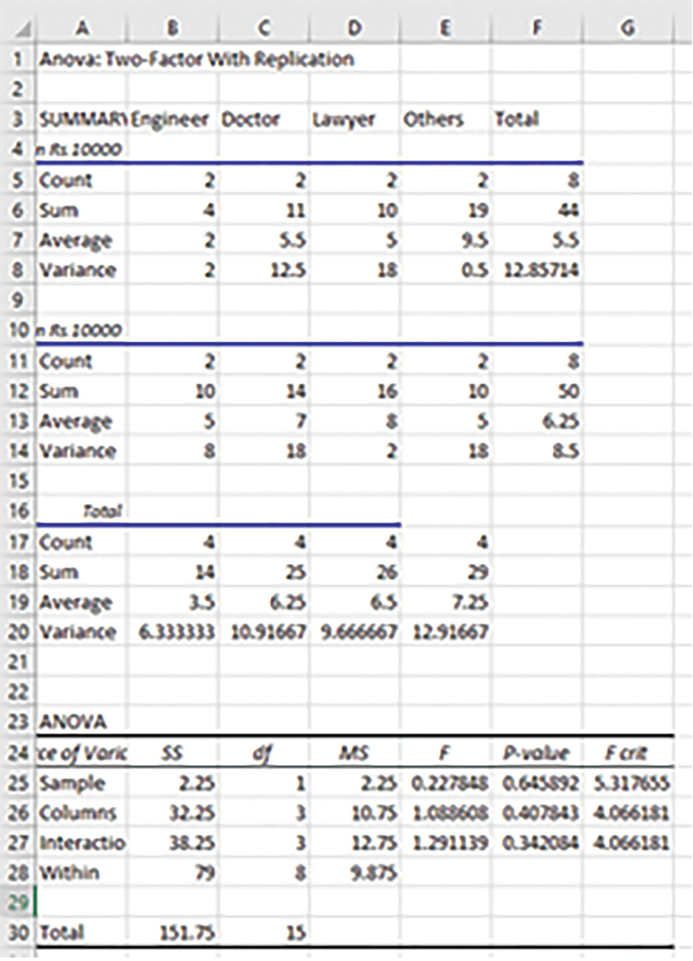
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 15.24 Screenshot after entering required fields using sub-steps of Step 4



	A	B	C	D	E	F	G
1	Anova: Two-Factor With Replication						
2							
3	SUMMARY	Engineer	Doctor	Lawyer	Others	Total	
4	n Rs 10000						
5	Count	2	2	2	2	8	
6	Sum	4	11	10	19	44	
7	Average	2	5.5	5	9.5	5.5	
8	Variance	2	12.5	18	0.5	12.85714	
9							
10	n Rs 10000						
11	Count	2	2	2	2	8	
12	Sum	10	14	16	10	50	
13	Average	5	7	8	5	6.25	
14	Variance	8	18	2	18	8.5	
15							
16	Total						
17	Count	4	4	4	4		
18	Sum	14	25	26	29		
19	Average	3.5	6.25	6.5	7.25		
20	Variance	6.333333	10.91667	9.666667	12.91667		
21							
22							
23	ANOVA						
24	ce of Varic	SS	df	MS	F	P-value	F crit
25	Sample	2.25	1	2.25	0.227848	0.645892	5.317655
26	Columns	32.25	3	10.75	1.088608	0.407843	4.066181
27	Interactio	38.25	3	12.75	1.291139	0.342084	4.066181
28	Within	79	8	9.875			
29							
30	Total	151.75	15				

Figure 15.25 Output in response after clicking the OK button in the dropdown of Figure 15.24 (ANOVA results)

Results:

From Figure 15.25, it can be seen that:

1. The p value (0.645892) of the sample (row: Income) is more than the given significance level of 0.05. Hence, accept its null hypothesis.

   Inference: There are no significant differences between the levels of the income in terms of the service quality.

2. The p value (0.407843) of the columns (Column: Nature of Profession) is more than the given significance level of 0.05. Hence, accept its null hypothesis.

   Inference: There are no significant differences between the levels of the nature of profession in terms of the service quality.

3. The p value (0.342084) of the interaction between the income and nature of profession is more than the given significance level of 0.05. Hence, accept its null hypothesis.

   Inference: There are no significant differences between the interaction terms of the income and the nature of profession in terms of the service quality.

15.6 Yates' Algorithm for 2^n Factorial Experiment Using Excel Sheets and F.DIST.RT Function

Yates' algorithm is a generalised algorithm that calculates the sum of squares for each component of the 2^n factorial experiment model, where n is the total number of factors and each factor has just two levels.

The steps of this algorithm are as follows.

Step 1: Arrange the standard order of the model components column wise. Let it be column x.

The standard orders of a few designs are as follows.
2^2 Design: 1, *a*, *b*, *ab*
2^3 Design: 1, *a, b, ab, c, ac, bc, abc*
2^4 Design: 1, *a, b, ab, c, ac, bc, abc, d, ad, bd, abd, cd, acd, bcd, abcd*

Step 2: Write the response totals of the corresponding model components in the next column. Let it be column y.

Step 3: Obtain the entries of Column 1 using the following steps.

3.1: Obtain each of the first half of the entries from top in Column 1 by adding the consecutive pair of entries from the top of the Column y.

3.2: Obtain each of the second half of the entries from the $(2^n/2)+1$ position in Column 1 by adding the consecutive pair of entries from the top of Column y by subtracting the first entry from the second entry in that pair.

Step 4*:* Obtain the entries of Column 2 using the results of Column 1 and by following the steps as followed for Column 1.

Step 5*:* Obtain the entries of Column 3 using the results of Column 2 and by following the steps as followed for Column 2.

Step 6*:* Obtain the entries of the remaining columns up to Column n in the same manner, where n is the total number of factors.

Step 7: Find the sum of squares of each component of the ANOVA model using the following formula.

$$SS = \frac{(Corresponding\,entry\,in\,Column\,n)^2}{k \times 2^n}$$

where

n is the total number of factors

k is the number of replications under each treatment combination of the factorial table

Step 8: Find the total sum of squares in the usual way.

Step 9: The error sum of squares is obtained using the following formula.

$$Sum\,of\,squares\,of\,error = Total\,sum\,of\,squares -$$

$$(Sum\,of\,squares\,of\,model\,components\,on\,the\,right\,hand\,side\,of\,the\,ANOVA\,model)$$

Step 10: Perform ANOVA calculations and draw conclusions.

Example 15.6

A company is interested in evaluating the value addition made by its personnel to its business operations on a scale of 0 to 10. The employees' U.G. qualifications, sex, and work experience are taken into consideration as factors in this regard. Table 15.28 displays the corresponding values that employees contributed to the functioning of the business.

1. Write the ANOVA model of this situation.
2. Perform the relevant ANOVA using Yates' algorithm and state the inferences at the significance level of 5%.

Solution

The data for Example 15.6 are shown in Table 15.29.

1. The ANOVA model of this problem is:

$$Y_{ijk} = \mu + E_i + D_j + ED_{ij} + S_k + ES_{ik} + DS_{jk} + EDS_{ijk} + e_{ijkl}$$

Table 15.28 Value Additions of Employees to Business Operations

Work Experience	*U.G Degree*			
	Eng.		*Commerce*	
	Sex		*Sex*	
	Male	*Female*	*Male*	*Female*
Less than 5 years	9	3	5	3
	8	7	9	5
5 years and above	10	5	8	6
	10	10	9	7

Table 15.29 Data for Example 15.6

Work Experience	*U.G Degree*			
	Eng.		*Commerce*	
	Sex		*Sex*	
	Male	*Female*	*Male*	*Female*
Less than 5 years	9	3	5	3
	8	7	9	5
5 years & above	10	5	8	6
	10	10	9	7

where

Y_{ijkl} is the contribution of the employee with regard to lth replication under i^{th} work experience, j^{th} degree, and k^{th} sex

μ is the overall mean of the contribution of the employee

E_i is the effect of the i^{th} work experience on the contribution of the employee

D_j is the effect of the j^{th} degree on the contribution of the employee

ED_{ij} is the interaction effect of the i^{th} work experience and j^{th} degree on the contribution of the employee

S_k is the effect of k^{th} sex on the contribution of the employee

ES_{ik} is the interaction effect of the i^{th} work experience and k^{th} sex on the contribution of the employee

DS_{jk} is the interaction effect of the j^{th} degree and k^{th} sex on the contribution of the employee

EDS_{ijk} is the interaction effect of the i^{th} work experience, j^{th} degree, and k^{th} sex on the contribution of the employee

e_{ijkl} is the random error associated with the lth replication under i^{th} work experience, j^{th} degree and k^{th} sex

2. The hypotheses of the ANOVA model are stated as follows.

Factor: Work Experience

H_0: There are no significant differences between work experiences in terms of the contribution of the employee.

H_1: There are significant differences between work experiences in terms of the contribution of the employee.

Factor: Degree

H_0: There are no significant differences between degrees in terms of the contribution of the employee.

H_1: There are significant differences between degrees in terms of the contribution of the employee.

Interaction: Work Experience X Degree

H_0: There are no significant differences between different pairs of interaction terms of work experience and degree in terms of the contribution of the employee.

H_1: There are significant differences between different pairs of interaction terms of work experience and degree in terms of the contribution of the employee.

Factor: Sex

H_0: There are no significant differences between sexes in terms of the contribution of the employee.

H_1: There are significant differences between sexes in terms of the contribution of the employee.

Interaction: Work Experience X Sex

H_0: There are no significant differences between different pairs of interaction terms of work experience and sex in terms of the contribution of the employee.

H_1: There are significant differences between different pairs of interaction terms of work experience and sex in terms of the contribution of the employee.

Interaction: Degree X Sex

H_0: There are no significant differences between different pairs of interaction terms of degree and sex in terms of the contribution of the employee.

H_1: There are significant differences between different pairs of interaction terms of degree and sex in terms of the contribution of the employee.

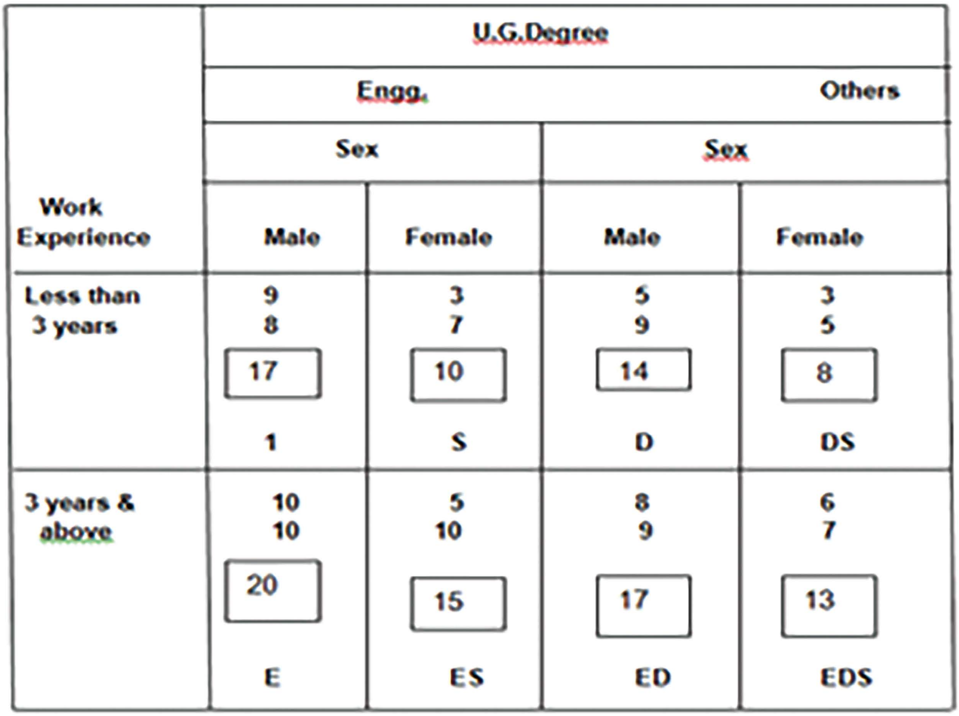
Interaction: Work Experience X Degree X Sex

H_0: There are no significant differences between different combinations of interaction terms of work experience, degree and sex in terms of the contribution of the employee.

H_1: There are significant differences between different combinations of work experience, degree and sex in terms of the contribution of the employee.

The subtotals for each of the components of the model are shown in Table 15.30.

Table 15.30 Subtotals of the Components of the Model



	U.G.Degree			
	Engg.		Others	
	Sex		Sex	
Work Experience	Male	Female	Male	Female
Less than 3 years	9 8 17 1	3 7 10 S	5 9 14 D	3 5 8 DS
3 years & above	10 10 20 E	5 10 15 ES	8 9 17 ED	6 7 13 EDS

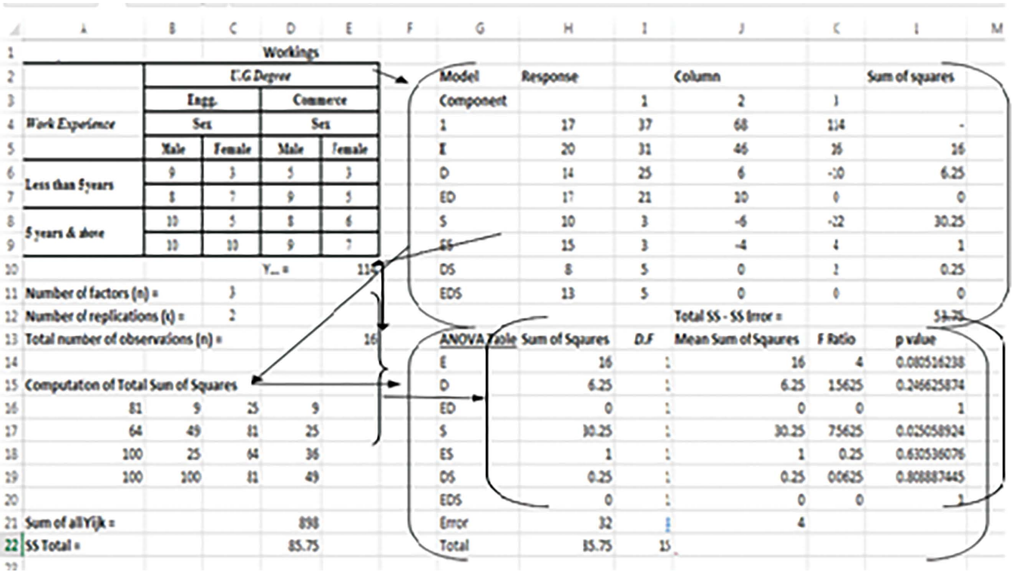
The ANOVA uses the F test for which the value of *p* is to be computed for the given F value, numerator degrees of freedom, and denominator degrees of freedom. The Excel formula to compute the *p* value is shown as follows.

$$= F.DIST.RT\left(X, \deg\_freedom1, deg\_freedom2\right)$$

The screenshot of the working of this problem is shown in Figure 15.26. The screenshot for the guidelines for the formulas of the working of the problem is shown in Figure 15.27.

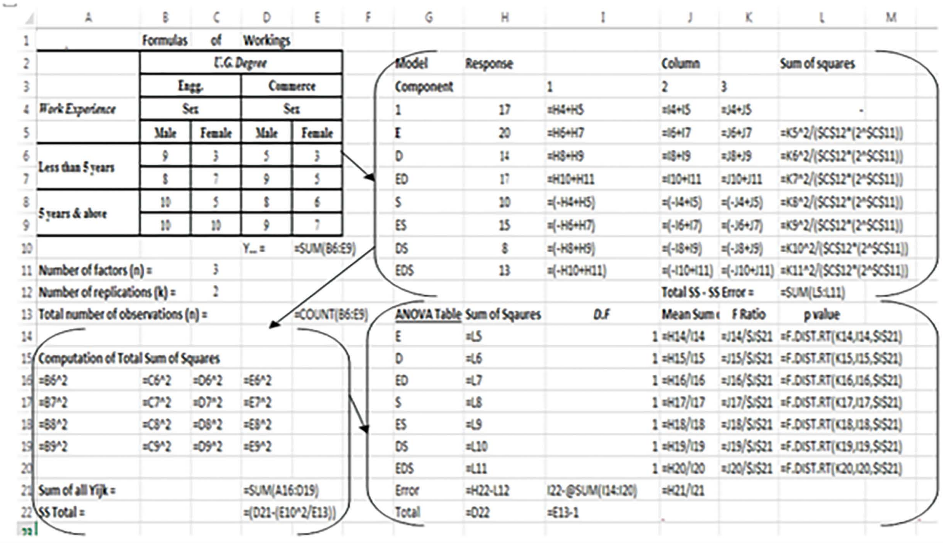
Based on the *p* values shown in Figure 12.26, the inferences for the components of the *ANOVA* model are presented here.

1. There are no significant differences between the work experiences in terms of contribution of employee.
2. There are no significant differences between the degrees in terms of contribution of employee.
3. There are no significant differences between the interaction terms of work experience and degree in terms of contribution of employee.
4. There are significant differences between the sexes in terms of contribution of employee.
5. There are no significant differences between the interaction terms of work experience and sex in terms of contribution of employee.
6. There are no significant differences between the interaction terms of degree and sex in terms of contribution of employee.
7. There are no significant differences between the interaction terms of work experience, degree and sex in terms of contribution of employee.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1				Workings									
2		*U.G Degree*					Model	Response		Column		Sum of squares	
3		Engg.		Commerce			Component		1	2	3		
4	*Work Experience*	Sex		Sex			1	17	37	68	114	-	
5		Male	Female	Male	Female		E	20	31	46	16	16	
6	Less than 5years	9	3	5	3		D	14	25	6	-10	6.25	
7		8	7	9	5		ED	17	21	10	0	0	
8	5 years & above	10	5	8	6		S	10	3	-6	-22	30.25	
9		10	10	9	7		ES	15	3	-4	4	1	
10				Y... =	114		DS	8	5	0	2	0.25	
11	Number of factors (n) =		3				EDS	13	5	0	0	0	
12	Number of replications (k) =		2							Total SS - SS Error =		53.75	
13	Total number of observations (n) =				16		ANOVA Table	Sum of Squares	D.F	Mean Sum of Sqaures	F Ratio	p value	
14							E	16	1	16	4	0.080516238	
15	Computation of Total Sum of Squares						D	6.25	1	6.25	1.5625	0.246625874	
16	81	9	25	9			ED	0	1	0	0	1	
17	64	49	81	25			S	30.25	1	30.25	7.5625	0.025058924	
18	100	25	64	36			ES	1	1	1	0.25	0.630536076	
19	100	100	81	49			DS	0.25	1	0.25	0.0625	0.808887445	
20							EDS	0	1	0	0	1	
21	Sum of allYijk =			898			Error	32	8	4			
22	SS Total =			85.75			Total	85.75	15				

Figure 15.26 Sequence of calculations of Yates' algorithm applied to Example 15.5



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Formulas	of	Workings									
2		*U.G. Degree*					Model	Response		Column		Sum of squares	
3		Engg.		Commerce			Component		1	2	3		
4	*Work Experience*	Sex		Sex			1	17	=H4+H5	=I4+I5	=J4+J5	-	
5		Male	Female	Male	Female		E	20	=H6+H7	=I6+I7	=J6+J7	=K5^2/($C$12*(2^$C$11))	
6	Less than 5 years	9	3	5	3		D	14	=H8+H9	=I8+I9	=J8+J9	=K6^2/($C$12*(2^$C$11))	
7		8	7	9	5		ED	17	=H10+H11	=I10+I11	=J10+J11	=K7^2/($C$12*(2^$C$11))	
8	5 years & above	10	5	8	6		S	10	=(-H4+H5)	=(-I4+I5)	=(-J4+J5)	=K8^2/($C$12*(2^$C$11))	
9		10	10	9	7		ES	15	=(-H6+H7)	=(-I6+I7)	=(-J6+J7)	=K9^2/($C$12*(2^$C$11))	
10				Y... =	=SUM(B6:E9)		DS	8	=(-H8+H9)	=(-I8+I9)	=(-J8+J9)	=K10^2/($C$12*(2^$C$11))	
11	Number of factors (n) =		3				EDS	13	=(-H10+H11)	=(-I10+I11)	=(-J10+J11)	=K11^2/($C$12*(2^$C$11))	
12	Number of replications (k) =		2							Total SS - SS Error =		=SUM(L5:L11)	
13	Total number of observations (n) =				=COUNT(B6:E9)		ANOVA Table	Sum of Sqaures	D.F	Mean Sum	F Ratio	p value	
14							E	=L5	1	=H14/I14	=J14/$J$21	=F.DIST.RT(K14,I14,$I$21)	
15	Computation of Total Sum of Squares						D	=L6	1	=H15/I15	=J15/$J$21	=F.DIST.RT(K15,I15,$I$21)	
16	=B6^2	=C6^2	=D6^2	=E6^2			ED	=L7	1	=H16/I16	=J16/$J$21	=F.DIST.RT(K16,I16,$I$21)	
17	=B7^2	=C7^2	=D7^2	=E7^2			S	=L8	1	=H17/I17	=J17/$J$21	=F.DIST.RT(K17,I17,$I$21)	
18	=B8^2	=C8^2	=D8^2	=E8^2			ES	=L9	1	=H18/I18	=J18/$J$21	=F.DIST.RT(K18,I18,$I$21)	
19	=B9^2	=C9^2	=D9^2	=E9^2			DS	=L10	1	=H19/I19	=J19/$J$21	=F.DIST.RT(K19,I19,$I$21)	
20							EDS	=L11	1	=H20/I20	=J20/$J$21	=F.DIST.RT(K20,I20,$I$21)	
21	Sum of all Yijk =			=SUM(A16:D19)			Error	=H22-L12	I22-@SUM(I14:I20)	=H21/I21			
22	SS Total =			=(D21-(E10^2/E13))			Total	=D22	=E13-1				

Figure 15.27 Guidelines for formulas of Yates' algorithm applied to Example 15.6

Summary

- The response in an experiment is a measurement of a dependent variable of interest, which may be influenced by the effects of one or more factors and their interactions.
- A factor in an experiment is a parameter or entity which is suspected to have an effect on the response variable.

- The different settings of a factor are called the treatments of that factor.
- The repeated observations under the same experimental condition are called replications.
- The model of the completely randomised design is:

  $Y_{ij} = \mu + A_j + e_{ij}$, where μ is the overall mean of the sales revenue, A_j is the effect of the j^{th} treatment of factor A (state) on the response, and e_{ij} is the random error associated with the i^{th} replication of the j^{th} treatment of factor A.

- ANOVA with a single factor and one block is called a completely randomised block design.
- A block in ANOVA brings homogeneity in the rows or columns of the block.
- The total sum of squares of the Latin square design is the sum of squares of treatment, sum of squares of block A, sum of squares of block B, and sum of squares of error.
- The complete factorial experiment contains two factors with replications for each experimental combination of the treatments of the factors.
- Yates' algorithm is a generalised algorithm which gives the sum of squares of different components of the model of 2^n factorial experiment, where n is the total number of factors and each factor is with two levels.

Keywords

- Response in an experiment is a measurement of a dependent variable of interest, which may be influenced by the effects of one or more factors and their interactions.
- Factor in an experiment is a parameter or entity which is suspected to have an effect on the response variable.
- Treatment refers to different settings of a factor.
- Fixed factor means that the inferences of a selected set of levels of a factor, which is a subset of the total possible levels of that factor, are restricted to only to the selected set of levels of that factor.
- Random factor means that the inferences of a set of levels selected from the total number of levels for the purpose of conducting an experiment are extended to all the levels of that factor.
- Replications are repeated observations under the same experimental condition.
- A randomised block design is ANOVA with a single factor and one block.
- A block in ANOVA brings homogeneity in its rows or columns.
- The Latin square design has a single factor with two blocks without replications.
- A complete factorial experiment contains two factors with replications for each experimental combination of the treatments of the factors.
- Yates' algorithm is a generalised algorithm which gives the sum of squares of different components of the model of 2^n factorial experiment, where n is the total number of factors and each factor is with two levels.

Review Questions

1. Define ANOVA and explain the terminology using a suitable example.
2. Distinguish between fixed factor and random factor.

3. a. Give a suitable example for a completely randomised design.
   b. Give the model of ANOVA of a completely randomised design and explain its components.
4. Give the generalised format of an ANOVA table of a completely randomised design.
5. The following table provides a summary of the monthly sales revenue produced by four salespeople, A, B, C, and D, in lakhs of rupees during the past six months.

		Salesperson			
		A	*B*	*C*	*D*
Replication	1	30	27	40	45
	2	34	40	45	34
	3	28	38	30	28
	4	40	50	35	56
	5	55	36	28	40
	6	40	45	46	56

   a. Write the appropriate ANOVA model.
   b. Using Excel, determine whether each ANOVA model component has an impact on the monthly sales revenue at a significance level of 10%.

6. A company has four distinct salespeople who generate revenue in lakhs of rupees over the course of four different quarters. The marketing manager thinks the salesperson might have an effect on sales revenue. He chooses the randomised complete block design as a result. The following table displays the pertinent data.

		Salesperson			
		1	*2*	*3*	*4*
Quarter	1	30	80	40	10
	2	30	70	50	52
	3	50	80	80	40
	4	30	10	50	50

   a. Write the ANOVA model of this situation.
   b. Check whether each component of the ANOVA model has an effect on the sales revenue at a significance level of 0.01 using Excel.

7. Give a suitable example of Latin square design.
8. Give the model of the Latin square design and explain its components.
9. Give the generalised format of the ANOVA of a Latin square design.
10. A company's marketing manager is interested in researching how salespeople affect the average sales revenue realised. For each of the company's five different products, five separate salespeople (A, B, C, D, and E) were assigned to five different regions. He chose the Latin square design, as indicated in the following table, and received the average sales revenue for the salespeople for various design settings, as shown in the same table in crores of rupees.

		Region				
		1	*2*	*3*	*4*	*5*
Product	1	*A* = 12	*B* = 8	*C* = 5	*D* = 7	*E* = 8
	2	*B* = 9	*C* = 7	*D* = 7	*E* = 6	*A* = 4
	3	*C* = 7	*D* = 6	*E* = 4	*A* = 4	*B* = 5
	4	*D* = 8	*E* = 6	*A* = 3	*B* = 3	*C* = 5
	5	*E* = 6	*A* = 4	*B* = 2	*C* = 8	D = 4

a. Write the ANOVA model of this situation.
b. Check whether each component of the ANOVA model has an effect on the sales revenue at a significance level of 0.05 using Excel.

11. Give a suitable example of a complete factorial experiment.
12. Give the model of a complete factorial experiment and explain its components.
13. Give the generalised format of an ANOVA of the complete factorial experiment.
14. The director of the Alpha School of Management is interested in examining how faculty influences trainees' average performance on a scale of 0 to 10. Three separate subjects of the training programme were assigned to three different faculty members (A, B, and C). According to the assumed complete factorial experiment with two factors, faculty and subject, he has taken three replications under each experimental combination. The following table contains the data according to this design.

Subject	*Faculty*		
	A	*B*	*C*
1	10	6	6
	6	8	7
	6	5	2
2	4	5	4
	3	3	2
	4	6	4
3	3	2	5
	5	4	2
	9	8	9

a. Write the corresponding ANOVA model.
b. Check whether each component of the ANOVA model has an effect on the performance of the participants at a significance level of 0.10 using Excel.

15. Explain the steps of Yates' algorithm applied to a 2^n complete factorial experiment using a suitable example.
16. The shop floor manager of a manufacturing company believes that three factors, machine (machine 1 and machine 2), operator (operator 1 and operator 2), and shift (shift 1 and shift 2, will affect the production volume in units of a product manufactured in milling machines. The following table shows the design of a 2^3 complete factorial experiment with two replications along with the data that were gathered regarding the production volume per shift of the product.

Shift	*Milling Machine*			
	1		*2*	
	Operator		*Operator*	
	1	*2*	*1*	*2*
1	100	80	120	105
	90	78	110	100
2	95	110	115	90
	115	105	112	80

a. Write the corresponding ANOVA model.

b. Check whether each component of the ANOVA model has an effect on the production volume per shift of the product at a significance level of 0.05 using Yates' algorithm through Excel.

References

1. Panneerselvam, R., *Design and Analysis of Experiments*, PHI Learning Private Limited, New Delhi, 2012.
2. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
3. www.real-statistics.com/design-of-experiments/completely-randomized-design/ [June 25, 2020].
4. www.real-statistics.com/two-way-anova/ [June 25, 2020].

16 Charts

Learning Objectives

After reading this chapter, you will be able to

- Recognise the importance of charts to analyse a given set of data.
- Construct a pie chart foe a given set of data to visualise the relative size of each item in the data.
- Analyse the given set of data using a bar chart, which will clearly show the relative difference among the data items proportional to their frequencies.
- Implement a stacked bar chart for a given set of data if there are multiple instances for each item in the data set.
- Construct a line chart for the given data set to study the frequencies through a piecewise linear graph.
- Implement a multiple-line chart for a given set of data if there are instances for many variables for a given instance of another variable/item in that data set.

16.1 Introduction

Although frequency tables can be used to describe the data, it is impossible to quickly understand the pattern or trend in the data using these tables. Thus, a different method of expressing the data becomes necessary, leading to the creation of charts and graphs. A chart merely displays the relative positioning of the bars' heights or the relative distribution of the chart's regions for various values or instances of the variable of interest.

Business managers will gain from using these charts in the majority of real-world scenarios by having their decision-making process facilitated. The decision maker utilising such charts should pay particular attention to differentiating the frequencies around the maximum or minimum based on the purpose of the study, as the charts simply show the relative difference among the frequencies for different values of an interest variable.

Charts are classified into the following types [1, 2].

- Pie chart
- Bar chart
- Stacked bar chart
- Line chart
- Multiple-line chart

DOI: 10.4324/9781032671703-16

16.2 Pie Charts

The shape of a pie chart is round. 360 degrees are included in the circle's circumference. In proportion to the frequency of the relevant variable, this 360° will be partitioned.

Example 16.1

Consider the M.B.A. course at a prestigious business school as an example. As everyone is aware, any undergraduate degree with a specific minimum mark in that qualifying degree qualifies for entrance to the M.B.A. programme. According to the business school's admissions data, undergraduate degrees in the humanities, sciences, engineering, medicine, and law are available. Table 16.1 lists the number of students in a class of 60 students who hold each of these degrees.

Construct a pie chart for the data given in Table 16.1.

Solution

The data for Example 16.1 are shown in Table 16.2.

The steps to construct the pie chart are as follows.

Step 1: Copy the data in Table 16.2 in the top left corner of an Excel sheet, as shown the screenshot of Figure 16.1.

Step 2: Select the data on undergraduate degree and frequency of student, including heading, from the range of cells B2:C7.

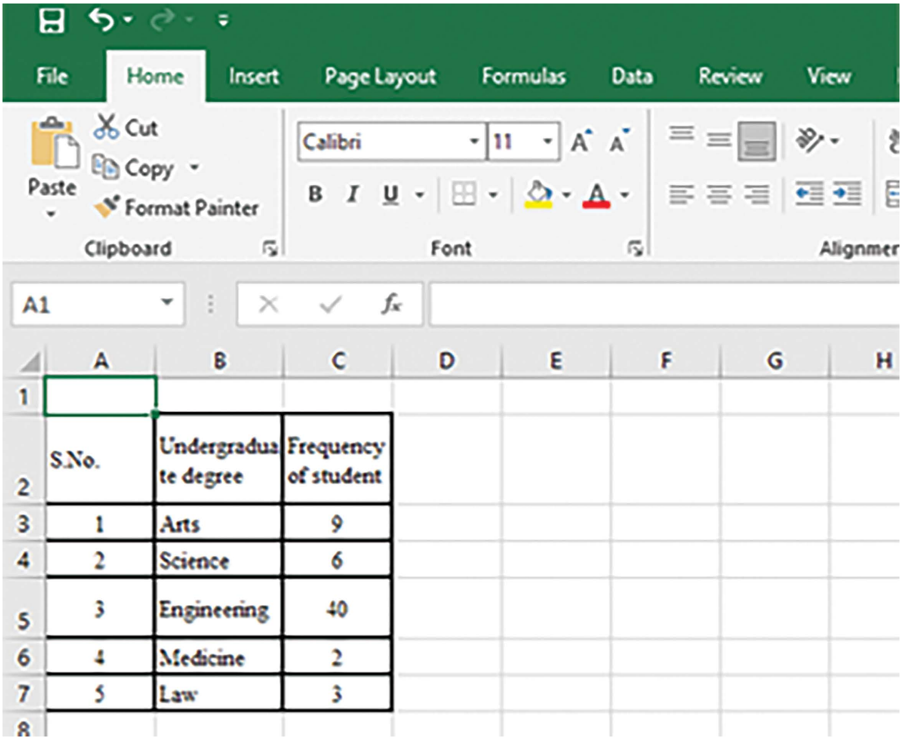
Step 3: Click the Insert button in the ribbon, which will give a cluster of charts with the bottom heading Charts, as shown in Figure 16.2.

Table 16.1 Data for Example 16.1

S. No.	*Undergraduate Degree*	*Frequency of Student*
1	Arts	9
2	Science	6
3	Engineering	40
4	Medicine	2
5	Law	3

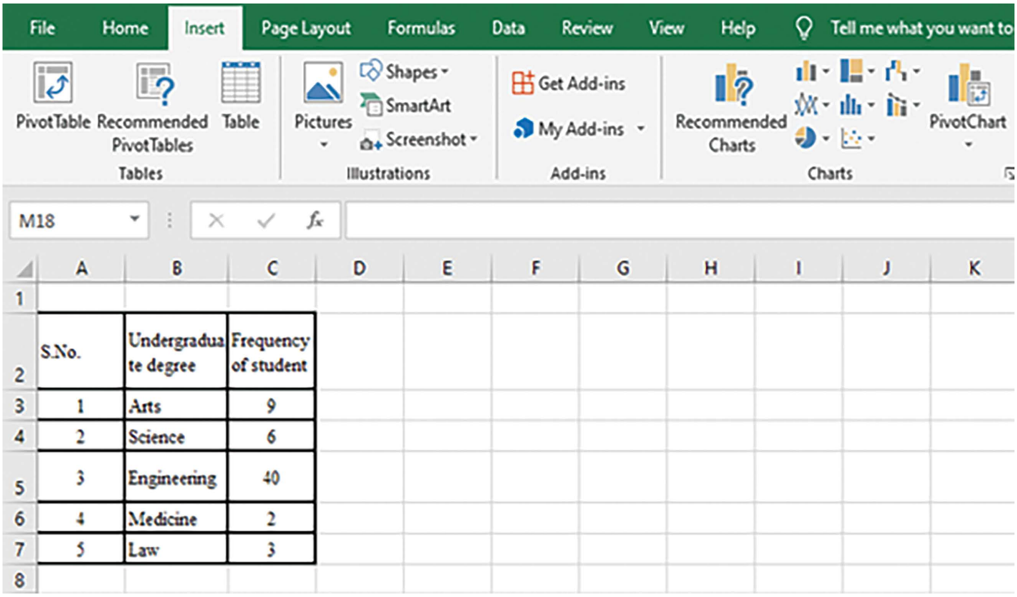
Table 16.2 Frequencies of Students With Different Undergraduate Degrees in MBA Class

S. No.	*Undergraduate Degree*	*Frequency of Student*
1	Arts	9
2	Science	6
3	Engineering	40
4	Medicine	2
5	Law	3



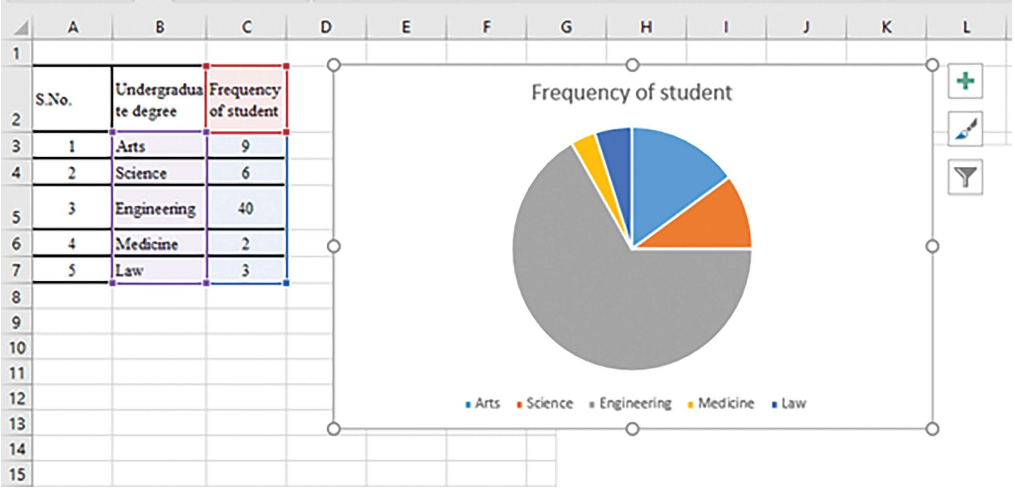
S.No.	Undergraduate degree	Frequency of student
1	Arts	9
2	Science	6
3	Engineering	40
4	Medicine	2
5	Law	3

Figure 16.1 Screenshot of data for Example 16.1 in Excel sheet



S.No.	Undergraduate degree	Frequency of student
1	Arts	9
2	Science	6
3	Engineering	40
4	Medicine	2
5	Law	3

Figure 16.2 Screenshot after clicking the Insert button in the ribbon



S.No.	Undergraduate degree	Frequency of student
1	Arts	9
2	Science	6
3	Engineering	40
4	Medicine	2
5	Law	3

Figure 16.3 Screenshot after clicking 2-D Pie option under Pie chart option under Charts in Figure 16.2

Step 4: Select the options under Pie chart (in circular form), which will display the following types of pie chart [4].

- 2-D pie
- 3-D pie
- Doughnut
- More pie charts

Step 5: Select the 2-D pie option, which will give a pie chart, and then click in the chart area to show a view, as shown in Figure 16.3.

To the right of the pie chart, there are three flags, the + symbol (chart elements), brush symbol (style), and funnel symbol (values).

5.a. Clicking the + symbol will give the following options.

i. Chart title
ii. Data labels
iii. Legend

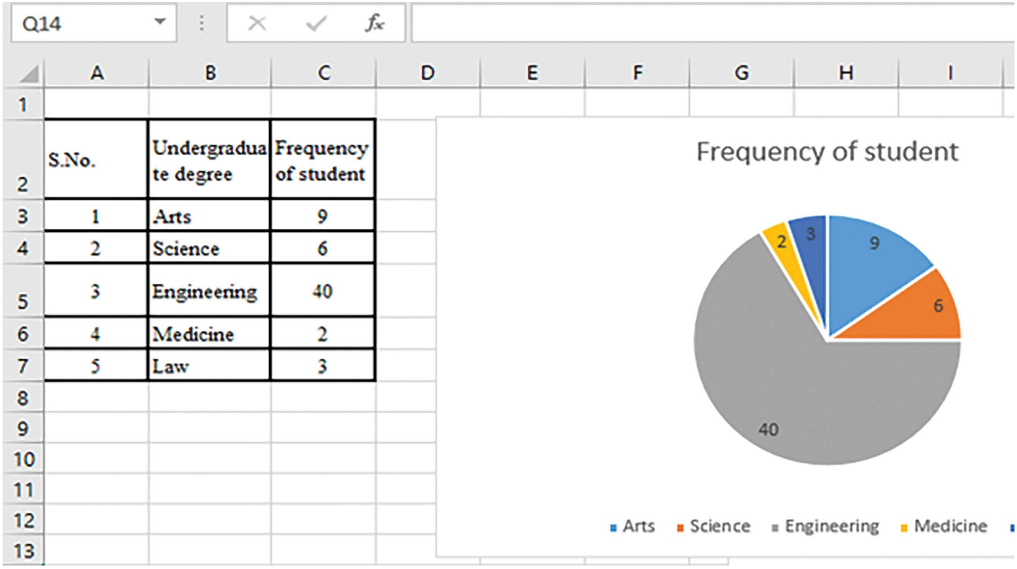
In these options, the first and the third options are already selected. Clicking the data labelled will add the frequencies in the respective regions of the pie chart. After clicking this option, the pie chart is shown in Figure 16.4.

5.b. Clicking the brush symbol will give three different styles.

Style 1: As shown in Figure 16.4.

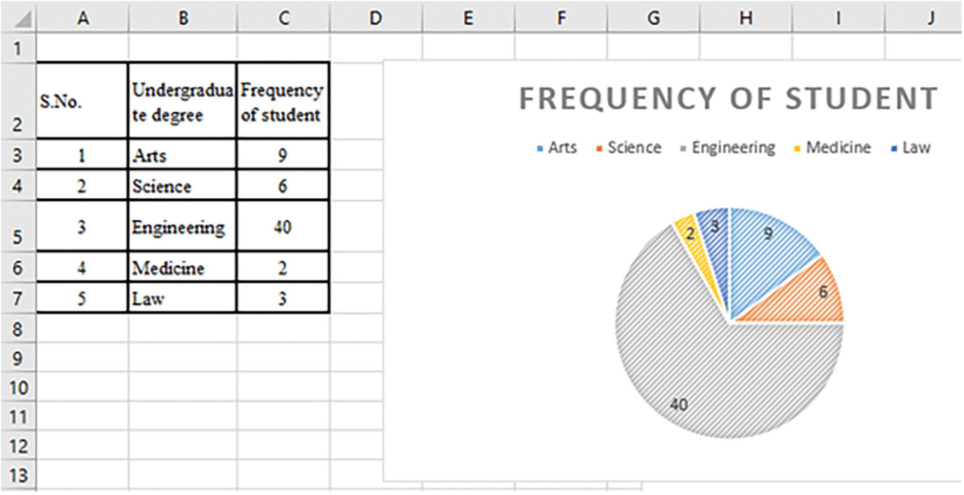
Style 2: Hatching in different slices of the pie chart shown in Figure 16.4, as shown in Figure 16.5.

Style 3: Percentage of frequencies marked in different slices of Figure 16.4, as shown in Figure 16.6.



S.No.	Undergraduate degree	Frequency of student
1	Arts	9
2	Science	6
3	Engineering	40
4	Medicine	2
5	Law	3

Figure 16.4 Pie chart of Example 16.1 after executing Step 5.a

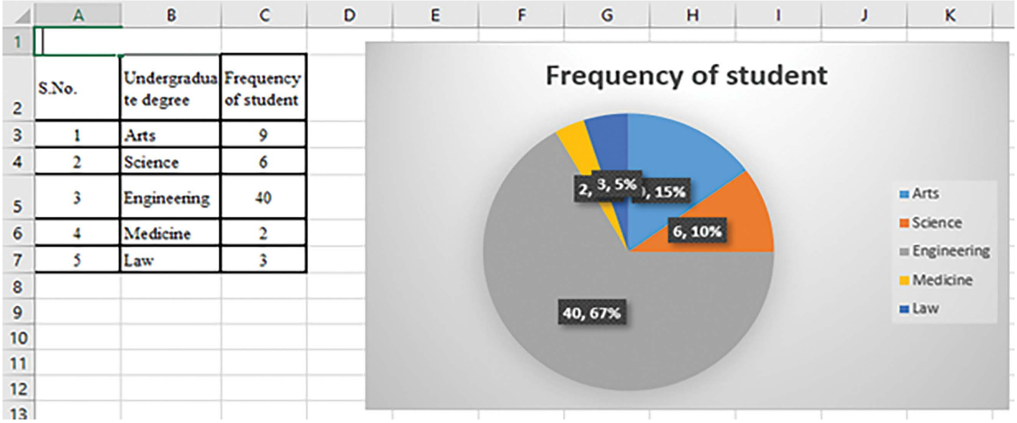


S.No.	Undergraduate degree	Frequency of student
1	Arts	9
2	Science	6
3	Engineering	40
4	Medicine	2
5	Law	3

Figure 16.5 Pie chart with hatching of slices of Example 16.1 after performing Step 5.b

16.3 Bar (Column) Charts

Vertical bars set up against several values or instances of a variable of interest make up a bar (column) chart. The *X* axis displays the values of a quantitative variable, such as the average age of employees, or the instances of a qualitative variable, such as the gender of employees. The *Y* axis displays the frequencies for various values of the variable.



S.No.	Undergraduate degree	Frequency of student
1	Arts	9
2	Science	6
3	Engineering	40
4	Medicine	2
5	Law	3

Figure 16.6 Pie chart with percentages of frequencies of Example 16.1 after performing Step 5.b

Table 16.3 Age Distribution of Employees

Age Interval Number	*Age Interval*	*Frequency of Age Interval*
1	21–25	75
2	25–30	100
3	30–35	140
4	35–40	200
5	40–45	170
6	45–50	150
7	50–55	100
8	55–60	65

Example 16.2

Consider the age of 1000 employees working in a company. The distribution of their ages is shown in Table 16.3.

Construct a bar chart for the data given in Table 16.3.

Solution

The data for Example 16.2 are shown in Table 16.4.

The steps to construct the bar chart are as follows.

Step 1: Copy the data in Table 16.4 in the top left corner of an Excel sheet, as shown in Figure 16.7.

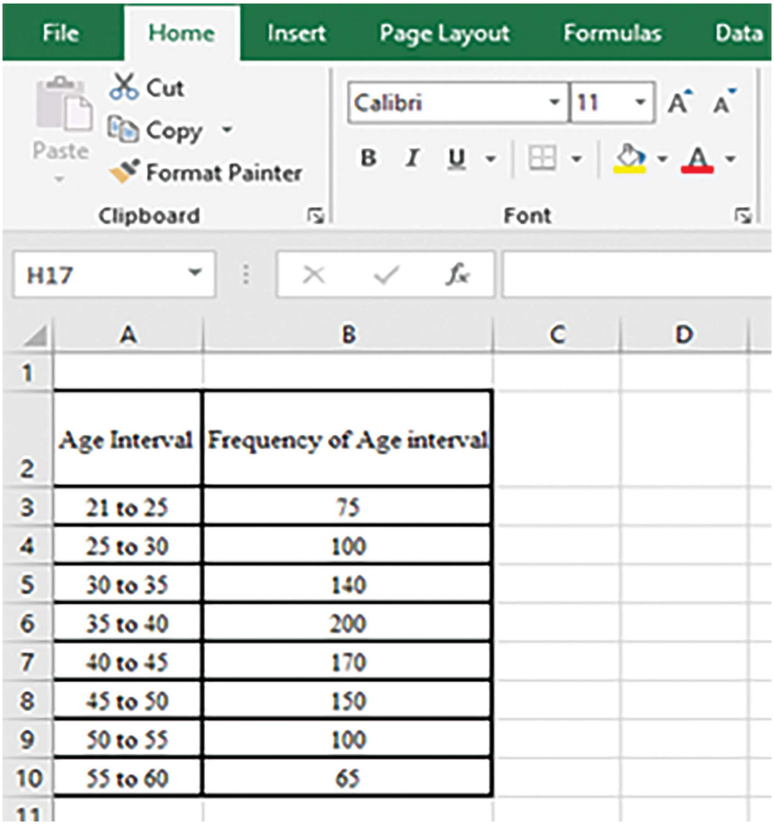
Step 2: Select the data on age interval and frequency of age interval from the range of cells A3:B10, excluding headings.

Step 3: Click the Insert button in the ribbon, which will give a cluster of charts with the bottom heading Charts, as already shown in Figure 16.2 of Example 16.1.

Step 4: Select the option Bar Chart (Chart with vertical strips) [3], which will display all possible column/bar charts, as shown in Figure 16.8.

Table 16.4 Age Distribution of Employees

Age Interval	Frequency of Age Interval
21–25	75
25–30	100
30–35	140
35–40	200
40–45	170
45–50	150
50–55	100
55–60	65



	A	B
1		
2	Age Interval	Frequency of Age interval
3	21 to 25	75
4	25 to 30	100
5	30 to 35	140
6	35 to 40	200
7	40 to 45	170
8	45 to 50	150
9	50 to 55	100
10	55 to 60	65

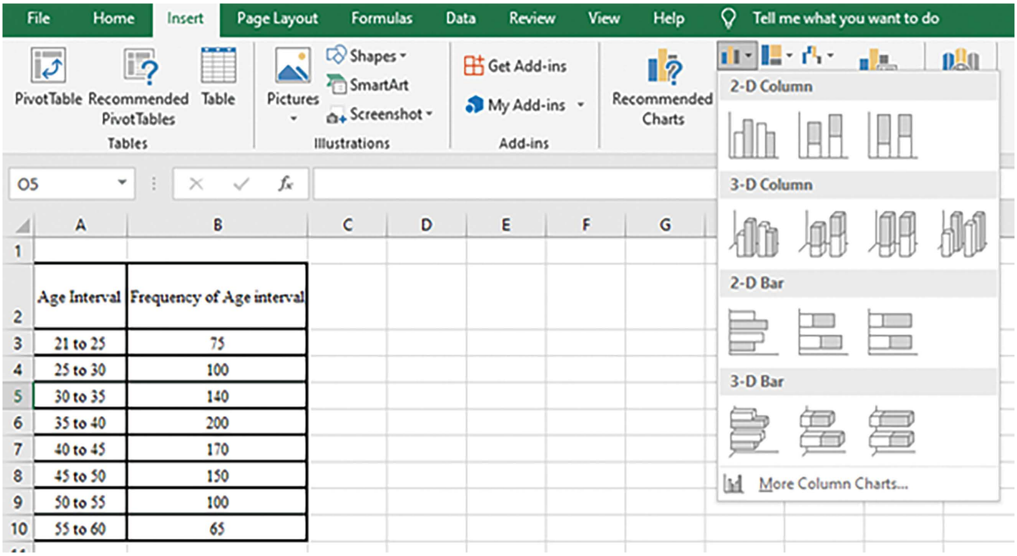
Figure 16.7 Screenshot of data for Example 16.2 in Excel sheet

Step 5: Select the 2-D column chart option, which gives the first-level bar chart of Example 16.2, as shown in Figure 16.9.

To the right of the bar chart, there are three flags, the + symbol (chart elements), brush symbol (style), and funnel symbol (values).

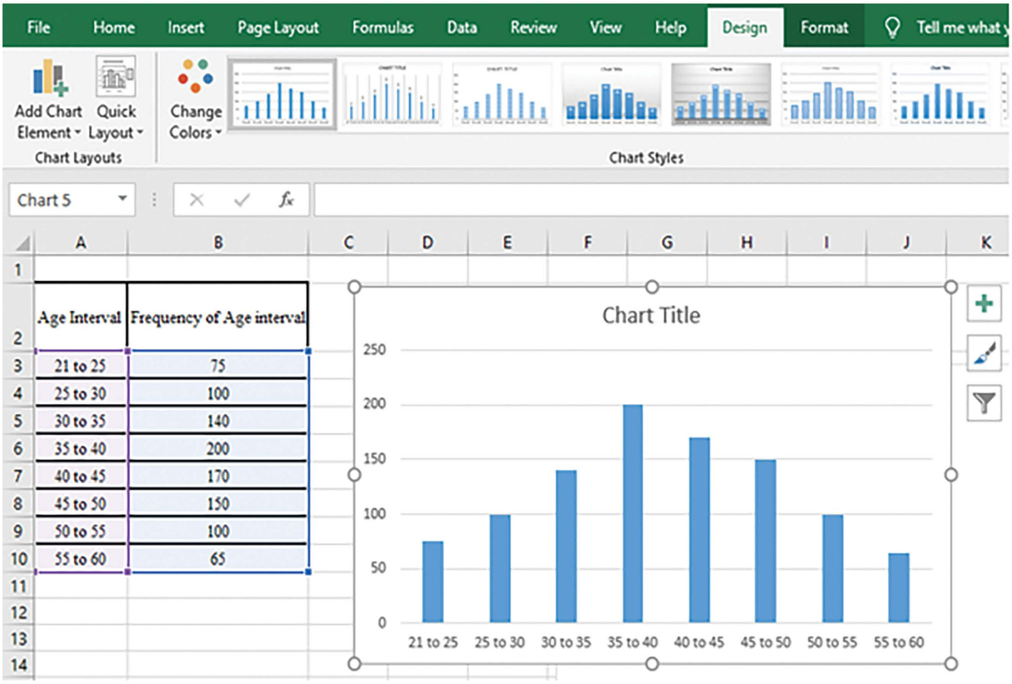
Clicking the + symbol will give the following options.

1. Axes
2. Axis titles



Age Interval	Frequency of Age interval
21 to 25	75
25 to 30	100
30 to 35	140
35 to 40	200
40 to 45	170
45 to 50	150
50 to 55	100
55 to 60	65

Figure 16.8 Screenshot of display for bar chart (Column Chart) options



Age Interval	Frequency of Age interval
21 to 25	75
25 to 30	100
30 to 35	140
35 to 40	200
40 to 45	170
45 to 50	150
50 to 55	100
55 to 60	65

Figure 16.9 Screenshot after clicking the 2-D column chart option at the top left corner of options in Figure 16.8 and then clicking in the chart region

3. Chart title
4. Data labels
5. Data table
6. Error bars
7. Gridlines
8. Legend
9. Trend line

In these options, the first and third options are already selected. Clicking the second option, Axis Titles, will enable us to type the title of each axis. Here, Age Interval is entered for the *X* axis and Frequency is entered for the *Y* axis. Clicking option 4, Data Labels, will add the frequencies at the top of the columns in the chart. Since by default the chart title appears at the top of the chart, unclick option 3 under the + symbol to remove it.

The options under the brush symbol (style) and funnel symbol (values) need not be changed.

Now the final bar (column) chart from the Excel sheet is copied and pasted as in Figure 16.10.

16.4 Multi-Bar (Columns) Charts

A multi-bar (columns) chart represents a variable of interest in the form of vertical bars for various values or instances. Either a qualitative or a quantitative variable may be the one on the *X* axis. For illustration, the values of the qualitative variable salesperson are shown on the *X* axis. The *Y* axis plots the values of various quarterly sales against each *X* axis value.

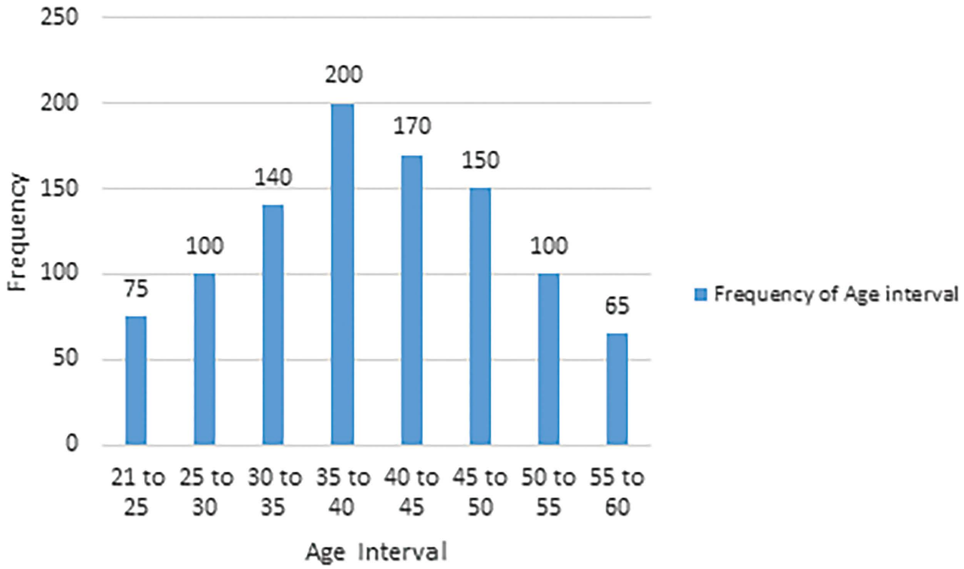


Figure 16.10 Final bar (column) chart of Example 16.2

Example 16.3

Consider the quarterly sales of six salespeople in a year, as shown in Table 16.5.

Construct a multi-bar chart for the data given in Table 16.5.

Solution

The data for Example 16.3 are shown in Table 16.6.

The steps to construct the multi-bar chart are as follows.

Step 1: Copy the data in Table 16.6 in the top left corner of an Excel sheet, as shown in Figure 16.11.

Step 2: Select the data on quarterly sales of salespeople, including the subheading quarter, from the range of cells B3:E9.

Step 3: Click the Insert button in the ribbon, which will give a cluster of charts with the bottom heading Charts, as already shown in Figure 16.2 for Example 16.1.

Step 4: Select the option of bar chart (chart with vertical strips), which will display all possible column/bar charts, as shown in Figure 16.12.

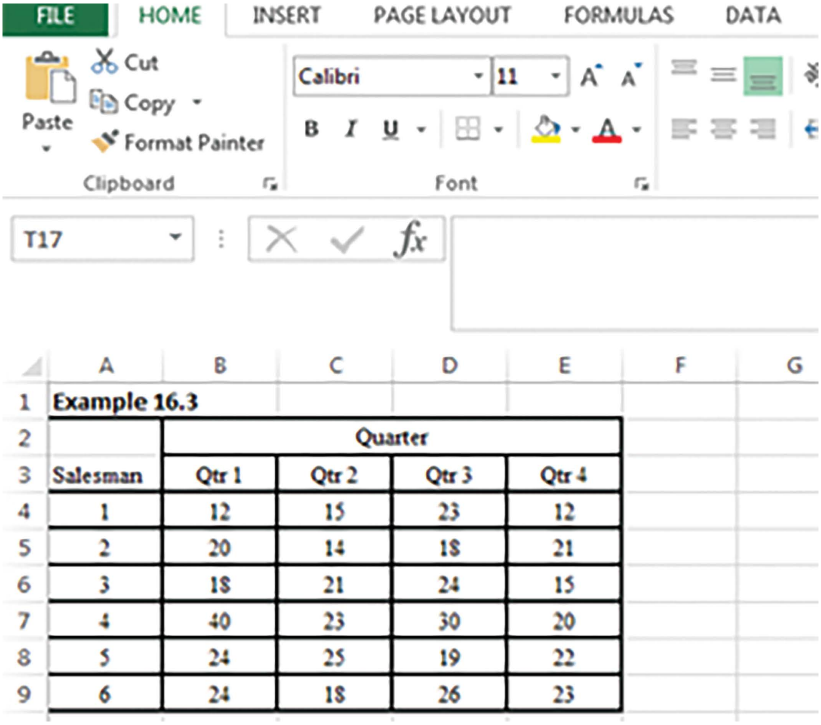
Step 5: Select the first option from the left under the 2-D column chart option in the drop-down menu of Figure 16.12, which gives the first level bar chart of Example 16.3, as shown in Figure 16.13.

Table 16.5 Quarterly Sales of Salespeople

Salesperson	*Quarter*			
	1	*2*	*3*	*4*
1	12	15	23	12
2	20	14	18	21
3	18	21	24	15
4	40	23	30	20
5	24	25	19	22
6	24	18	26	23

Table 16.6 Quarterly Sales of Salespeople

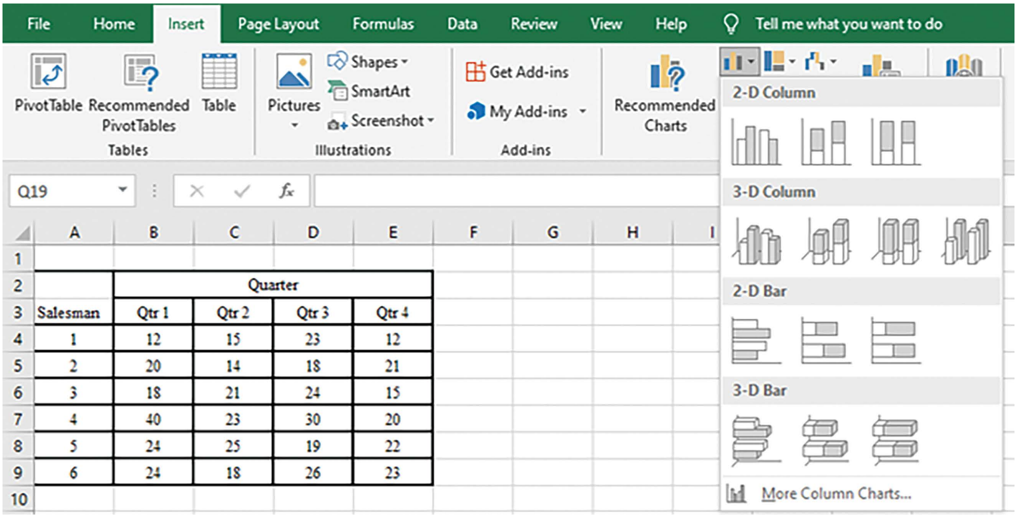
Salesperson	*Quarter*			
	Qtr 1	*Qtr 2*	*Qtr 3*	*Qtr 4*
1	12	15	23	12
2	20	14	18	21
3	18	21	24	15
4	40	23	30	20
5	24	25	19	22
6	24	18	26	23



Example 16.3

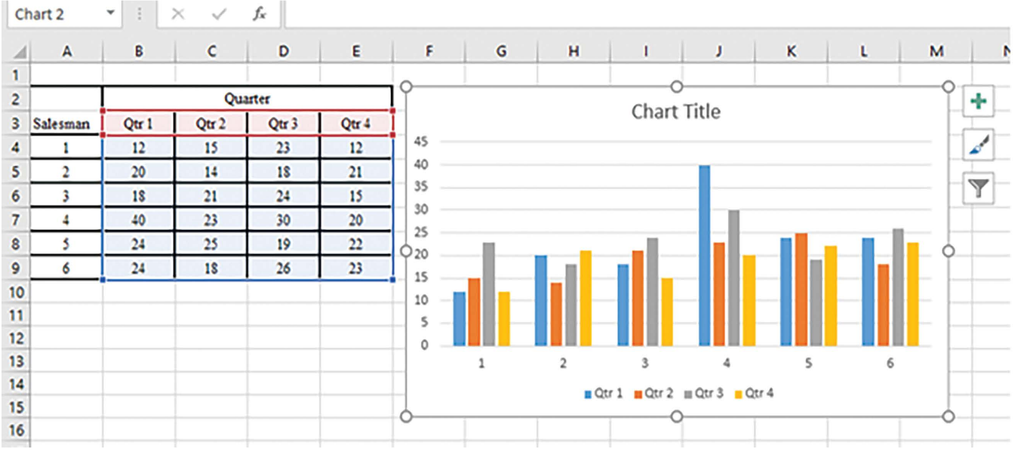
	Quarter			
Salesman	Qtr 1	Qtr 2	Qtr 3	Qtr 4
1	12	15	23	12
2	20	14	18	21
3	18	21	24	15
4	40	23	30	20
5	24	25	19	22
6	24	18	26	23

Figure 16.11 Screenshot of data for Example 16.3 in Excel sheet



	Quarter			
Salesman	Qtr 1	Qtr 2	Qtr 3	Qtr 4
1	12	15	23	12
2	20	14	18	21
3	18	21	24	15
4	40	23	30	20
5	24	25	19	22
6	24	18	26	23

Figure 16.12 Screenshot of display for bar chart (Column Chart) options



	Quarter			
Salesman	Qtr 1	Qtr 2	Qtr 3	Qtr 4
1	12	15	23	12
2	20	14	18	21
3	18	21	24	15
4	40	23	30	20
5	24	25	19	22
6	24	18	26	23

Figure 16.13 Screenshot after clicking the 2-D column chart option at the top left corner of options in Figure 16.12 and then clicking in the chart region

To the right of the bar chart in Figure 16.13, there are three flags, the + symbol (chart elements), brush symbol (style), and funnel symbol (values).

Clicking the + symbol will give the following options.

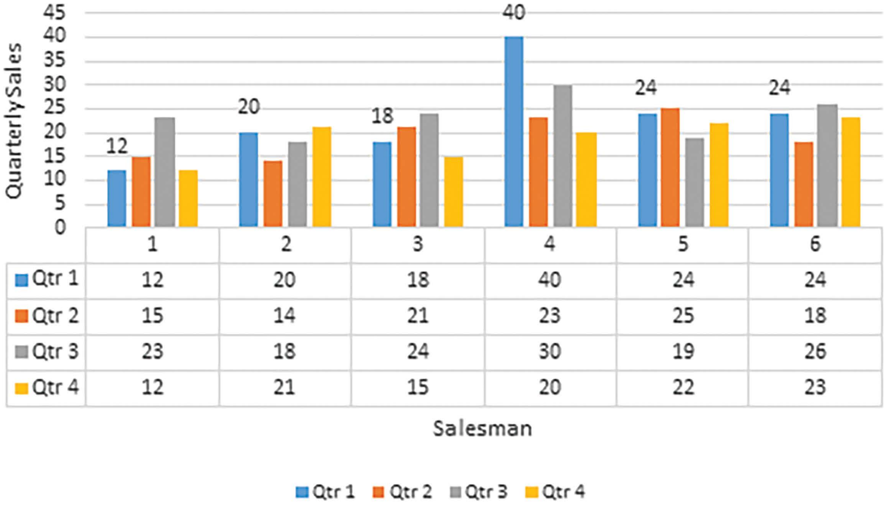
1. Axes
2. Axis titles
3. Chart title
4. Data labels
5. Data table
6. Error bars
7. Gridlines
8. Legend
9. Trend line

In these options, the first, third, seventh, and eight options are already selected. Since there is no chart title in the data, the chart title will not appear, except the appearance of “Chart Title” at the top in Figure 16.13. “Chart Title” at the top of the figure can be removed by unclicking option 3 under + symbol. Clicking the second option, Axis Titles, will enable us to enter the title of each axis. Here, Salesperson is entered for the *X* axis and Quarterly Sales is entered for the *Y* axis. Clicking option 4, Data Labels, will add the frequencies at the top of the columns in the chart. Clicking option 5, Data Table, will present the entire data of the problem below the chart.

The options under the brush symbol (style) and funnel symbol (values) need not be changed.

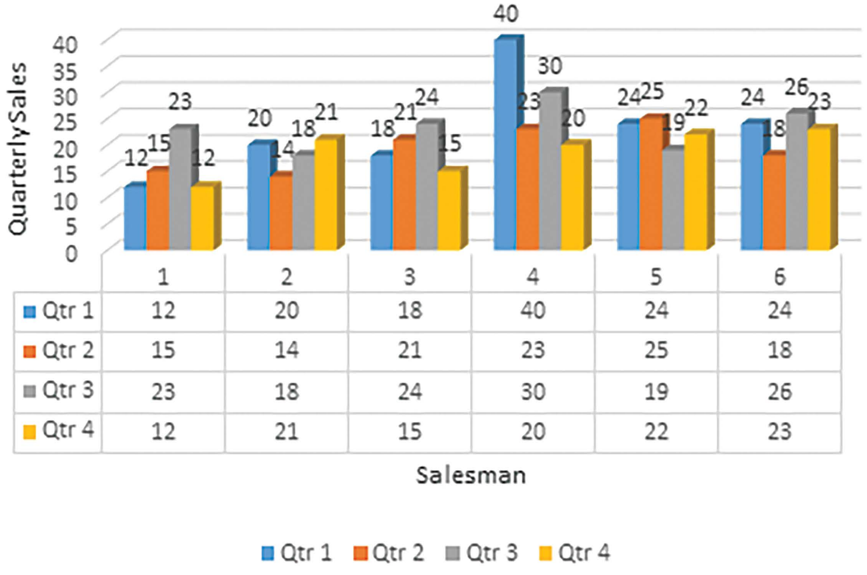
Now the final multi-bar (columns) chart from the Excel sheet is copied and pasted as in Figure 16.14.

Just to view the 3-D multi-bar (columns) chart, the entire process is repeated with the selection of the 3-D Column chart option in Step 5, and the corresponding final multi-bar (columns) chart is shown in Figure 16.15.



	1	2	3	4	5	6
Qtr 1	12	20	18	40	24	24
Qtr 2	15	14	21	23	25	18
Qtr 3	23	18	24	30	19	26
Qtr 4	12	21	15	20	22	23

Figure 16.14 Final 2-D multi-bar (columns) chart of Example 16.3



	1	2	3	4	5	6
Qtr 1	12	20	18	40	24	24
Qtr 2	15	14	21	23	25	18
Qtr 3	23	18	24	30	19	26
Qtr 4	12	21	15	20	22	23

Figure 16.15 Final 3-D multi-bar (columns) chart of Example 16.3

16.5 Stacked Bar Charts

The vertical bars of a stacked bar chart are divided into smaller rectangles in accordance with the instances of the various values of the variable on the *X* axis.

A qualitative variable, such as year, can have its values assumed on the *X* axis as an example. The values of several quarterly sales can be assumed on the *Y* axis for each value on the *X* axis.

Example 16.4

Consider the quarterly sales of a company during the past six years, as shown in Table 16.7.

Construct a stacked bar chart for the data given in Table 16.7.

Solution

The data for Example 16.4 are shown in Table 16.8.

The steps to construct the stacked bar chart are as follows.

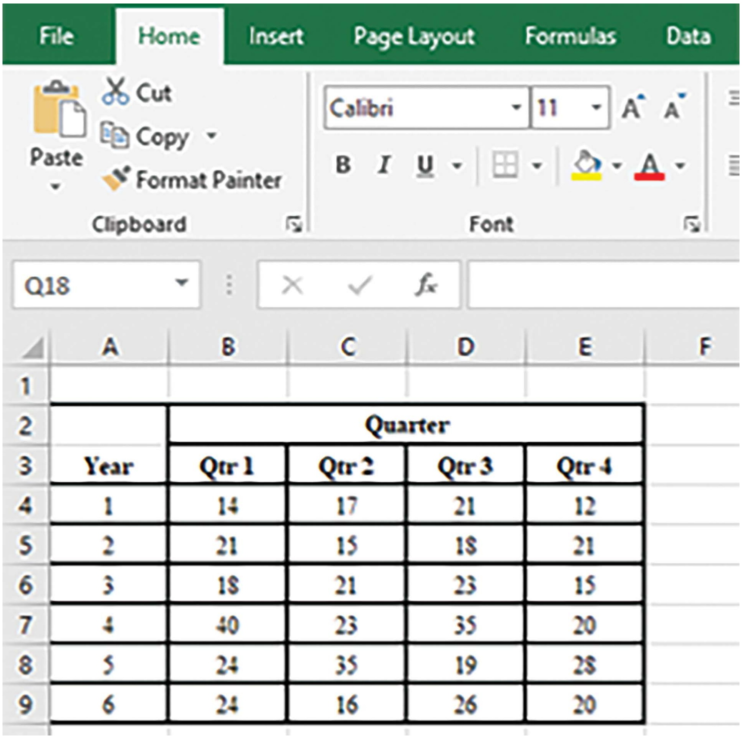
Step 1: Copy the data in Table 16.8 in the top left corner of an Excel sheet, as shown in Figure 16.16.
Step 2: Select the data on quarterly sales of the years, including the subheading of quarter, from the range of cells B3:E9.
Step 3: Click the Insert button in the ribbon, which will give a cluster of charts with the bottom heading Charts, as already shown in Figure 16.2 of Example 16.1.
Step 4: Select the option of bar chart (chart with vertical strips), which will display all possible column/bar charts, as shown in Figure 16.17.
Step 5: Select the 2-D stacked bar (column) chart option 2 from left at the top of Figure 16.17 and click in the chart region to show the first-level stacked bar chart of Example 16.4, as shown in Figure 16.18.
To the right of the bar chart, there are three flags, the + symbol (chart elements), brush symbol (style), and funnel symbol (values).

Table 16.7 Quarterly Sales of Past Six Years

Year	*Quarter*			
	Qtr 1	*Qtr 2*	*Qtr 3*	*Qtr 4*
1	14	17	21	12
2	21	15	18	21
3	18	21	23	15
4	40	23	35	20
5	24	35	19	28
6	24	16	26	20

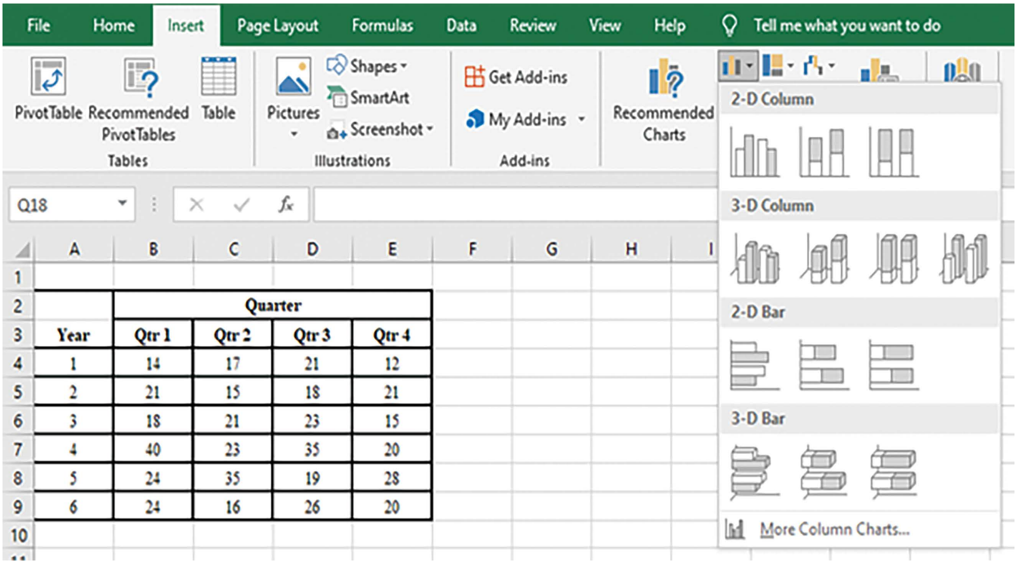
Table 16.8 Data for Example 16.4

Year	*Quarter*			
	Qtr 1	*Qtr 2*	*Qtr 3*	*Qtr 4*
1	14	17	21	12
2	21	15	18	21
3	18	21	23	15
4	40	23	35	20
5	24	35	19	28
6	24	16	26	20



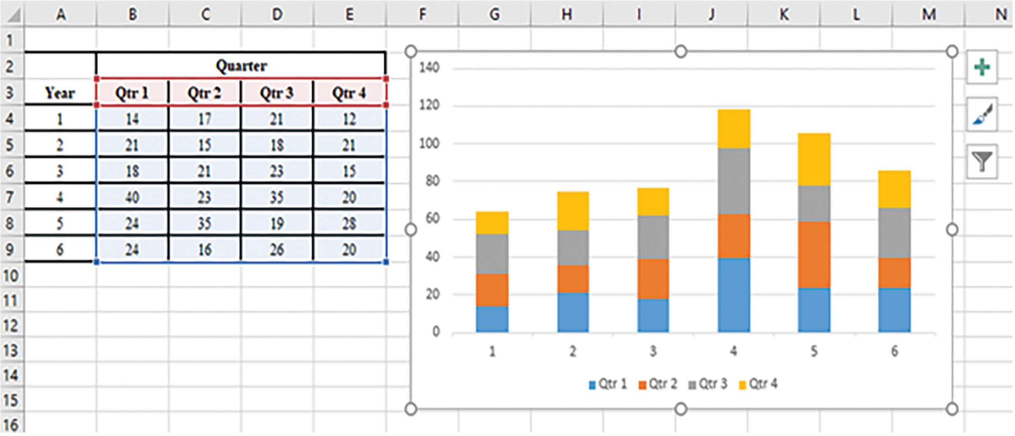
	Quarter			
Year	Qtr 1	Qtr 2	Qtr 3	Qtr 4
1	14	17	21	12
2	21	15	18	21
3	18	21	23	15
4	40	23	35	20
5	24	35	19	28
6	24	16	26	20

Figure 16.16 Screenshot of data for Example 16.4



	Quarter			
Year	Qtr 1	Qtr 2	Qtr 3	Qtr 4
1	14	17	21	12
2	21	15	18	21
3	18	21	23	15
4	40	23	35	20
5	24	35	19	28
6	24	16	26	20

Figure 16.17 Screenshot of display for stacked bar chart (Column Chart) options



	Quarter			
Year	Qtr 1	Qtr 2	Qtr 3	Qtr 4
1	14	17	21	12
2	21	15	18	21
3	18	21	23	15
4	40	23	35	20
5	24	35	19	28
6	24	16	26	20

Figure 16.18 Screenshot after clicking 2-D stacked bar (column) chart option 2 from left at the top of Figure 16.17 and then clicking in the chart region

Clicking the + symbol will give the following options.

1. Axes
2. Axis titles
3. Chart title
4. Data labels
5. Data table
6. Error bars
7. Gridlines
8. Legend
9. Trend-line

In these options, the first, third, seventh, and eight options are already selected. Since there is no chart title in the data, the chart title will not appear, except for the appearance of Chart Title at the top in Figure 16.18. Chart Title at the top of the figure can be removed by unclicking option 3. Clicking the second option, Axis Titles, will enable us to type the title of each axis. Here, Year is typed for the *X* axis and Quarterly Sales is typed for the *Y* axis. Clicking option 4, Data Labels, will add the frequencies within the sub-boxes of each vertical bar in the chart.

The options under the brush symbol (style) and funnel symbol (values) need not be changed.

Now the final stacked bar (column) chart from the Excel sheet is copied and pasted, as in Figure 16.19.

Just to view the 3-D multi-bar (columns) chart, the entire process is repeated with the selection of the 3-D Column chart option in Step 5, and the corresponding final stacked multi-bar (columns) chart is shown in Figure 16.20.

16.6 Line Charts

A line chart is a piecewise linear graph that is built on the *X*-*Y* plane. Typically, an independent variable is placed on the *X* axis, while the dependent variable or variables are

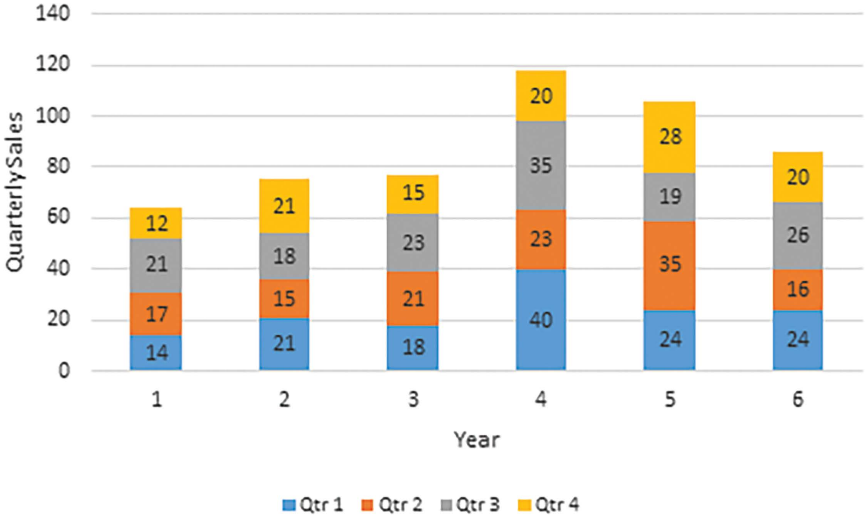


Figure 16.19 Final 2-D stacked-bar (column) chart of Example 16.4

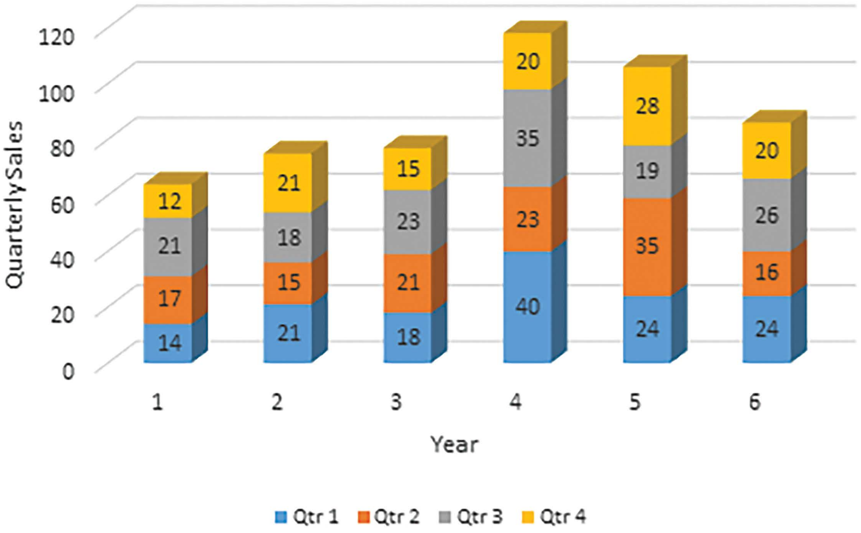


Figure 16.20 Final 3-D stacked-bar (column) chart of Example 16.4

placed on the *Y* axis. The graph's trend for the variable *Y* with regard to the independent variable *X* can be shown by plotting the line connecting the two variables *X* and *Y*. A line graph can be used to examine the trend in a product's demand values over time.

Example 16.5

Consider the annual sales of a product during the past six years, as shown in Table 16.9.

Construct a line chart for the data given in Table 16.9.

Table 16.9 Annual Sales (Lakhs of Rupees) of Product

Year	*Annual Sales*
1	90
2	110
3	140
4	130
5	150
6	165

Table 16.10 Annual Sales (Lakhs of Rupees) of Product

Year	*Annual Sales*
1	90
2	110
3	140
4	130
5	150
6	165

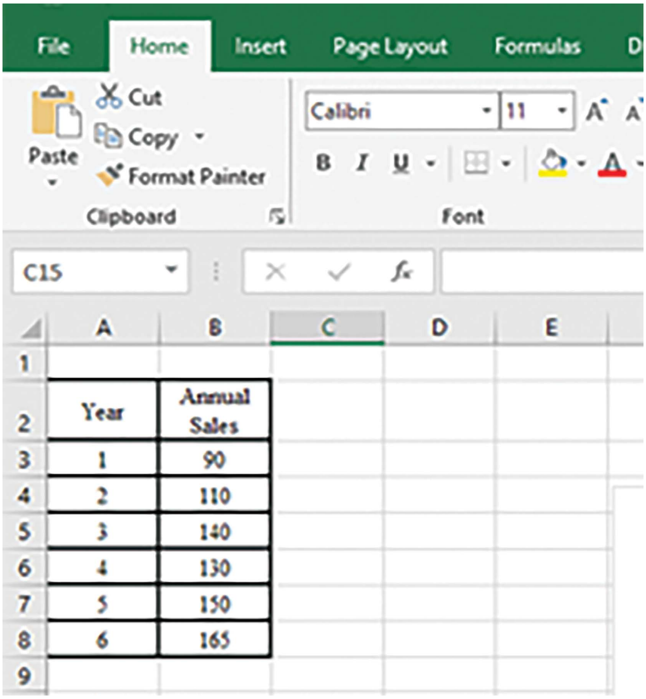
Solution

The data for Example 16.5 are shown in Table 16.10.

The steps to construct the line chart with labels are as follows.

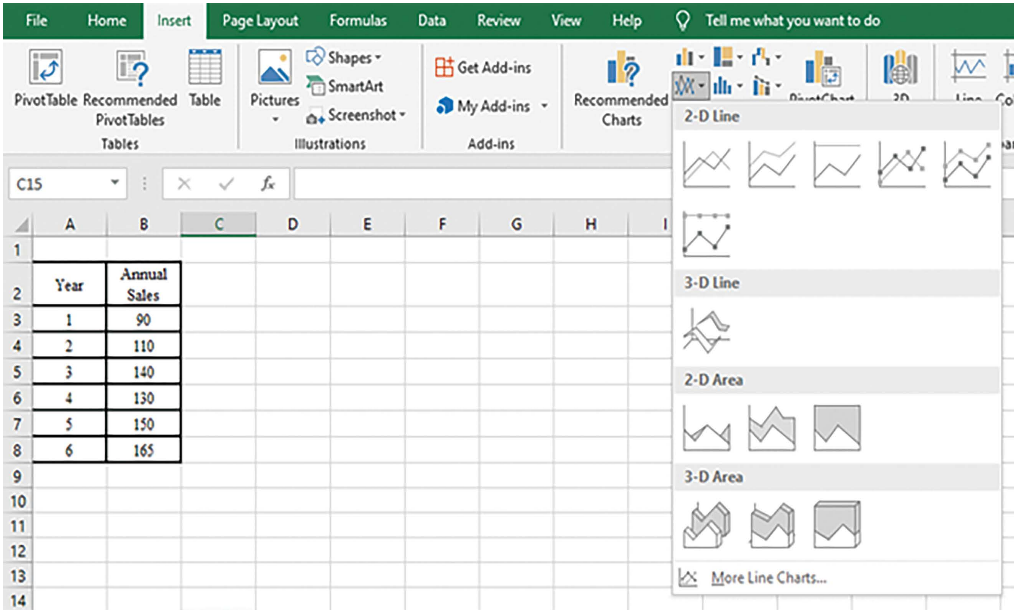
Step 1: Copy the data in Table 16.10 in the top left corner of an Excel sheet, as shown in Figure 16.21.
Step 2: Select the data on annual sales of the years in the range of cells B3:B8.
Step 3: Click the Insert button in the ribbon, which will give a cluster of charts with the bottom heading Charts, as already shown in Figure 16.2 of Example 16.1.
Step 4: Select the option of line chart [5], which will display all possible line charts, as shown in Figure 16.22.
Step 5: Select 2-D line chart with markings (first from left in the second row) from Figure 16.22, and click in the chart region to show a first-level line chart with the markings for Example 16.5, as shown in Figure 16.23.
To the right of the bar chart, there are three flags, the + symbol (chart elements), brush symbol (style), and funnel symbol (values).
Clicking the + symbol will give the following options.

1. Axes
2. Axis titles
3. Chart title
4. Data labels
5. Data table
6. Error bars
7. Gridlines
8. Legend
9. Trend-line



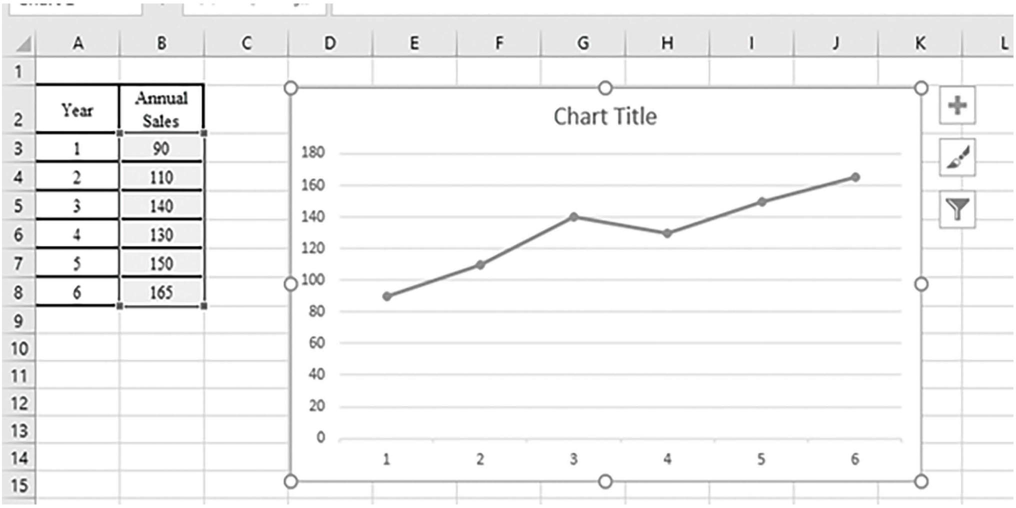
Year	Annual Sales
1	90
2	110
3	140
4	130
5	150
6	165

Figure 16.21 Screenshot of data for Example 16.5 in Excel sheet



Year	Annual Sales
1	90
2	110
3	140
4	130
5	150
6	165

Figure 16.22 Screenshot of display for line chart (Column Chart) options



Year	Annual Sales
1	90
2	110
3	140
4	130
5	150
6	165

Figure 16.23 Screenshot after clicking 2-D line chart option 1 (Line with markers) from the left in the second row of Figure 16.22 and then clicking in the chart region

In these options, the first, third, and seventh are already selected. Since there is no chart title in the data, the chart title will not appear, except for the appearance of Chart Title at the top in Figure 16.23. Chart Title at the top of the figure can be removed by unclicking option 3, because the chart title will be typed later at the bottom of the figure. Clicking the second option, Axis Titles, will enable us to type the title of each axis. Here, Year is typed for the *X* axis and Annual Sales is typed for the *Y* axis. Clicking option 4, Data Labels, will add the annual sales against the respective years marked in the *X* axis of the chart.

The options under the brush symbol (style) and funnel symbol (values) need not be changed.

Now the final line chart from the Excel sheet is copied and pasted, as in Figure 16.24.

16.7 Multiple-Line Charts

The structure of a multiple-line chart is made up of numerous piecewise linear lines built on an *X*-*Y* plane. A group of dependent variables will typically be taken on the *Y* axis and an independent variable will typically be taken on the *X* axis. The trend of each dependent variable (*Y*) with regard to the independent variable (*X*) is shown in the graph by the line that connects them. This multiple-line chart can be used to study the trend of a firm's R&D expenditure values and the demand values of a product made by that company with respect to year.

Example 16.6

Consider the R&D expenditures and annual sales of a product during the past six years, as shown in Table 16.11.

Construct a multi-line chart with markings for the data given in Table 16.11.

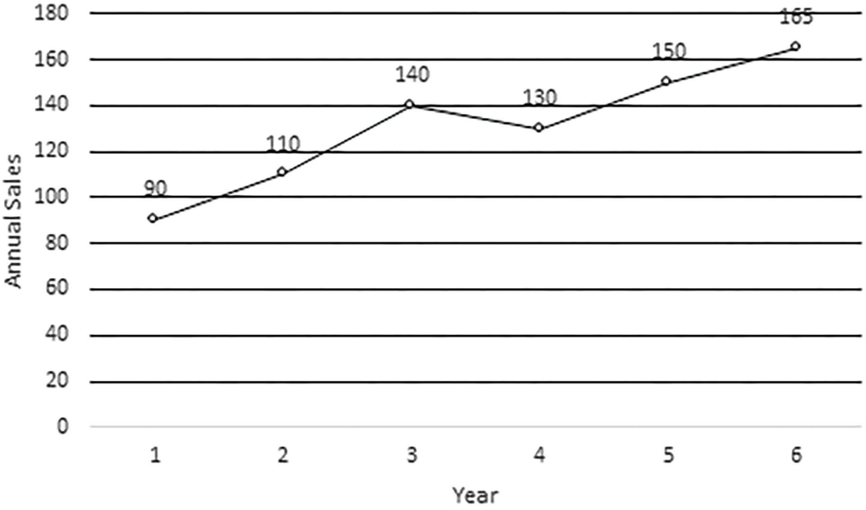


Figure 16.24 Final 2-D line chart of Example 16.5

Table 16.11 Data for Example 16.6

Year	*R&D Expenditure (Lakhs of Rupees)*	*Annual Sales (Crores of Rupees)*
1	4	10
2	6	14
3	5	16
4	7	20
5	6	19
6	8	24

Solution

The data for Example 16.6 are shown in Table 16.12.

The steps to construct the multiple-line chart are as follows.

Step 1: Copy the data in Table 16.12 in the top left corner of an Excel sheet, as shown in Figure 16.25.

Step 2: Select the data on R&D Expenditures and annual sales of the years in the range of cells B4:C9.

Step 3: Click the Insert button in the ribbon, which will give a cluster of charts with a bottom heading Charts, as already shown in Figure 16.2 of Example16.1.

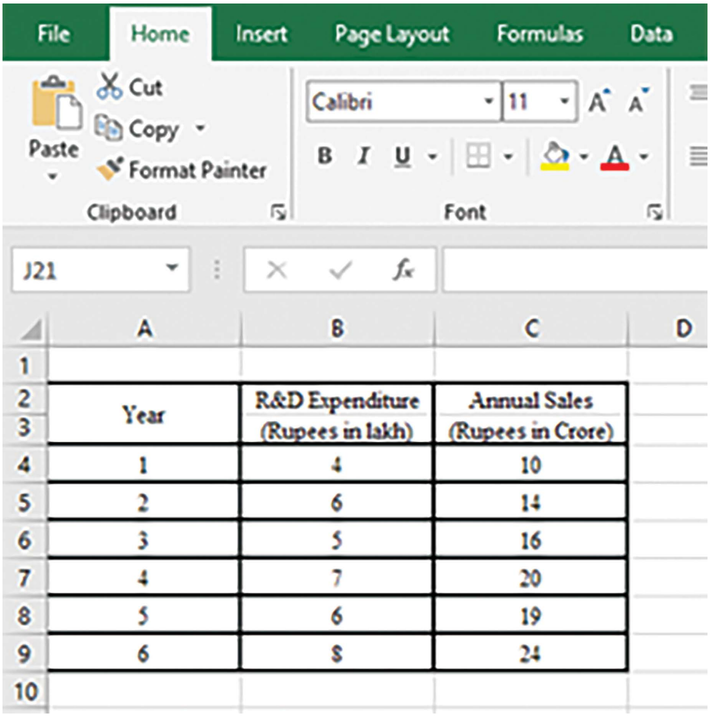
Step 4: Select the option of line chart [5], which will display all possible line charts, as shown in Figure 16.26.

Step 5: Select 2-D line chart with markings (first from left in the second row) from the dropdown menu of Figure 16.26 and click in the chart region to show a first-level line chart with markings for Example 16.6, as shown in Figure 16.27.

To the right of the bar chart, there are three flags, the + symbol (chart elements), brush symbol (style), and funnel symbol (values).

Table 16.12 Data for Example 16.6

Year	*R&D Expenditure (Rupees in Lakh)*	*Annual Sales (Rupees in Crore)*
1	4	10
2	6	14
3	5	16
4	7	20
5	6	19
6	8	24

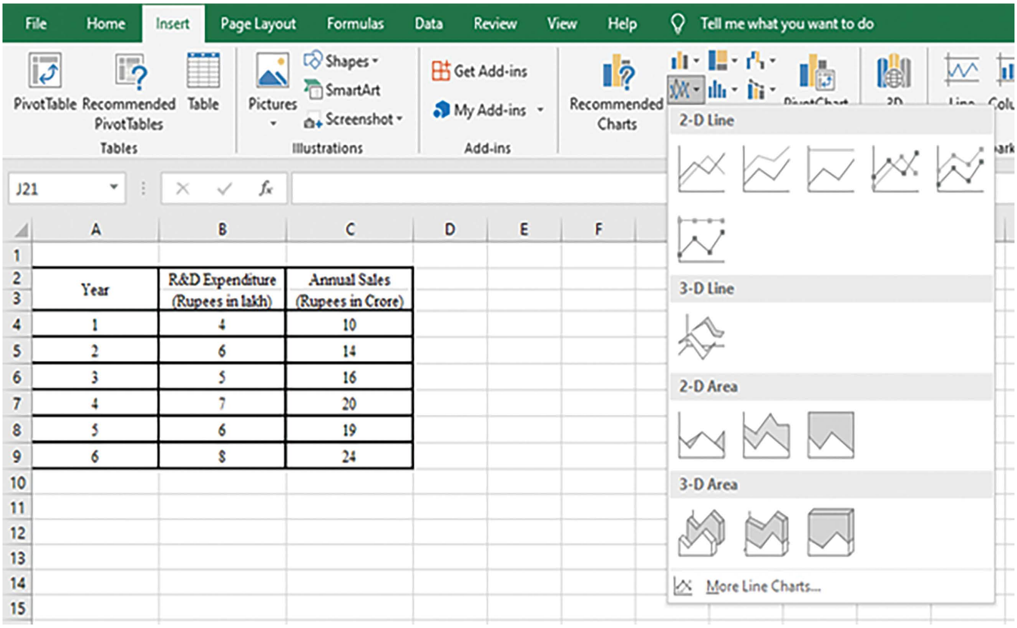


Year	R&D Expenditure (Rupees in lakh)	Annual Sales (Rupees in Crore)
1	4	10
2	6	14
3	5	16
4	7	20
5	6	19
6	8	24

Figure 16.25 Screenshot of data for Example 16.6 in Excel sheet

Clicking the + symbol will give the following options.

1. Axes
2. Axis titles
3. Chart title
4. Data labels
5. Data table
6. Error bars
7. Gridlines
8. Legend
9. Trend-line



Year	R&D Expenditure (Rupees in lakh)	Annual Sales (Rupees in Crore)
1	4	10
2	6	14
3	5	16
4	7	20
5	6	19
6	8	24

Figure 16.26 Screenshot of display for multi-line chart (Column Chart) options

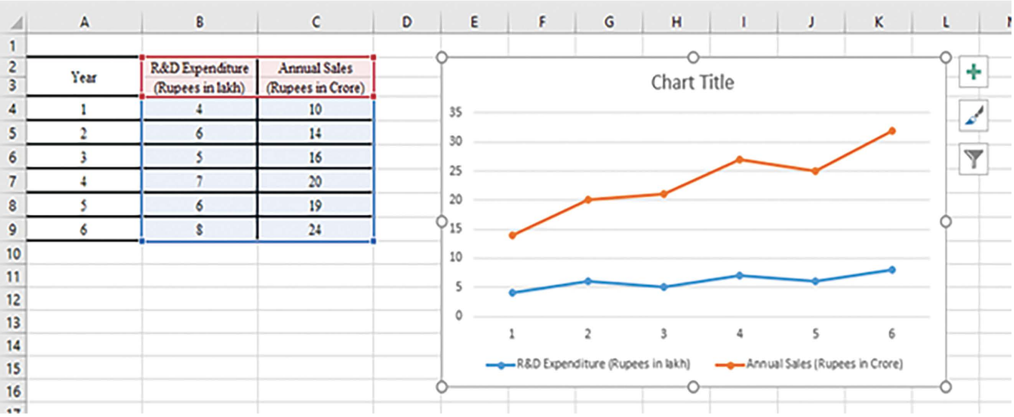


Chart Title
35
30
25
20
15
10
5
0
1
2
3
4
5
6
R&D Expenditure (Rupees in lakh)
Annual Sales (Rupees in Crore)

Year	R&D Expenditure (Rupees in lakh)	Annual Sales (Rupees in Crore)
1	4	10
2	6	14
3	5	16
4	7	20
5	6	19
6	8	24

Figure 16.27 Screenshot after clicking 2-D multi-line chart option 1 (Line with markers) from left in the second row of Figure 16.26 and then clicking in the chart region

In these options, the first, third, and seventh are already selected. Since there is no chart title in the data, the chart title will not appear, except for the appearance of Chart Title at the top in Figure 16.27. Chart Title at the top of the figure can be removed by unclicking option 3, because the chart title will be typed later at the bottom of the figure. Clicking the second option, Axis Titles, will enable us to type the title of each axis. Here, Year is typed for the *X* axis and Rupee unit as the case may be is typed for the *Y* axis. Clicking option 4, Data Labels, will add the R&D Expenditure and annual sales against respective years marked on the *X* axis of the chart.

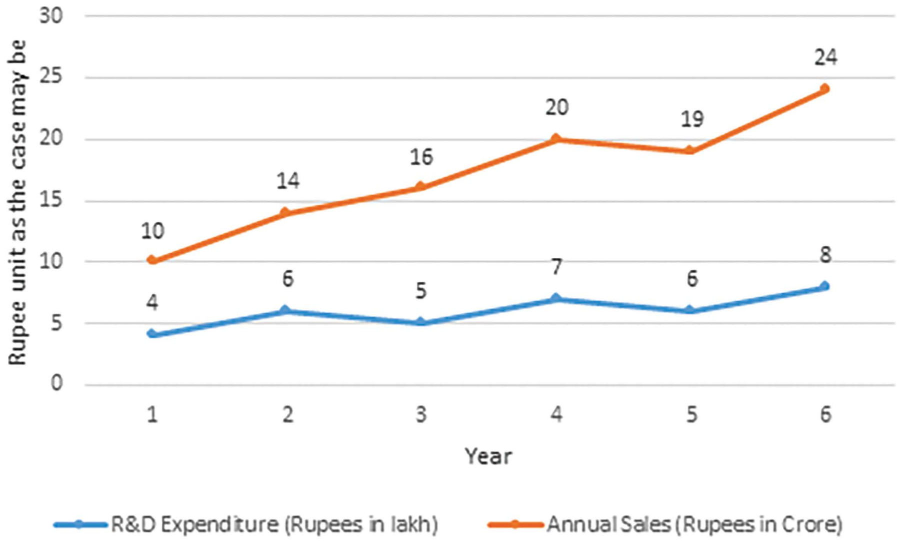


Figure 16.28 Final 2-D multi-line chart of Example 16.6

The options under the brush symbol (style) and funnel symbol (values) need not be changed.

Now the final multiple-line chart from the Excel sheet is copied and pasted, as in Figure 16.28.

Summary

- Charts/graphs form an alternative way of representing data when compared to tabular form.
- A pie chart is in the form of a circle, in which 360° will be divided proportionately according to the frequencies of the variable of concern.
- A bar (column) chart is in the form of vertical bars erected against different values/instances of a variable of interest on the *X* axis.
- A multi-bar (columns) chart is in the form of vertical bars for multiple instances against different values/instances of a variable of interest on the *X* axis.
- A stacked bar chart is in the form of vertical bars such that each vertical bar is subdivided into smaller rectangles according to the instances of the respective value of the variable on the *X* axis.
- A line chart is in the form of a piecewise linear graph constructed on the *X*-*Y* plane.
- Normally, in a line chart, an independent variable will be on the *X* axis and dependent variable(s) will be on the *Y* axis.
- A multiple-line chart is in the form of several piecewise linear lines constructed on the *X*-*Y* plane. Normally, an independent variable will be on the *X* axis and a set of dependent variables will be on the *Y* axis.

Keywords

- Charts/graphs form an alternative way of representing data when compared to tabular form.

- A pie chart is in the form of a circle, in which 360° will be divided proportionately according to the frequencies of the variable of concern.
- A bar (column) chart is in the form of vertical bars erected against different values/instances of a variable of interest on the *X* axis.
- A multi-bar (columns) chart is in the form of vertical bars for multiple instances against different values/instances of a variable of interest on the *X* axis.
- A stacked bar chart is in the form of vertical bars such that each vertical bar is subdivided into smaller rectangles according to the instances of the respective value of the variable on the *X* axis.
- A line chart is in the form of a piecewise linear graph constructed on the *X*-*Y* plane.
- A multiple-line chart is in the form of several piecewise linear lines constructed on the *X*-*Y* plane. Normally, an independent variable will be on the *X* axis and a set of dependent variables will be on the *Y* axis.

Review Questions

1. What is a pie chart? Give a sample pie chart.
2. Give the steps to construct a pie chart in Excel.
3. Take the B.Tech course at a prestigious engineering college as an example. Students from all over the nation submit applications for admission to that college. The following table displays the frequency distribution of students in terms of the percentage of various geographic areas of the nation to which they are belong. Create a pie chart in Excel using the data from the table.

S. No.	*Region of the Country*	*Frequency of Student (%)*
1	North	25
2	South	15
3	West	12
4	East	10
5	North East	3

4. What is a bar chart/column chart? Give a sample of such a chart.
5. List and explain the steps of constructing a bar/column chart in Excel.
6. Think about a company's six sales representatives' annual sales revenues. The following table displays the sales revenue distribution. Create a bar chart in Excel using the data from the table.

Salesperson	*Annual Sales (Crores of Rupees)*
1	100
2	130
3	110
4	160
5	140
6	135

7. What is a multi-bar (column) chart?
8. List and explain the steps of constructing a multi-bar (column) chart in Excel.

9. Take a look at the following table, which displays the grade point average data for four semesters of ten students in an M.Tech programme. Use Excel to create a multi-bar chart using the data from the table.

Student	*Grade Point Average*			
	Semester			
	1	*2*	*3*	*4*
1	2.5	2.8	2.4	2.9
2	2.2	2.1	2.5	2.6
3	2.3	1.7	2.0	1.5
4	2.9	2.8	2.7	2.9
5	1.5	1.9	2.2	2.3
6	2.3	2.7	2.8	2.8
7	2.3	2.5	1.7	1.9
8	2.9	2.8	2.6	.29
9	2.3	2.7	2.0	1.5
10	2.3	1.75	2.5	2.5

10. What is a stacked bar chart? Give a sample chart.
11. List and explain the steps of constructing a stacked bar chart in Excel.
12. Consider the manufacturing line of a company's weekly production in units during the previous six months, as indicated in the accompanying table.

Month	*Production Quantity*			
	Week			
	Week 1	*Week 2*	*Week 3*	*Week 4*
1	1,000	1,050	950	1,100
2	1,100	1,200	1,150	1,175
3	1,190	1,150	1,000	1,200
4	1,000	1,075	1,150	1,225
5	1,225	1,100	1,150	1,200
6	1,250	1,200	1,900	1,750

Construct a stacked multi-bar chart for the data given in the table using Excel.

13. What is a line chart? Give a sample line chart.
14. List and explain the steps of constructing a line chart in Excel.
15. Think about a company's productivity indices over the previous eight years, which are displayed in the accompanying table on a scale of 0 to 10. Using Excel, create a line chart with markings for the data in the table.

Year	*Productivity Index*
1	7.00
2	7.50
3	8.50
4	9.00
5	8.50
6	9.75
7	9.90
8	9.50

16. What is a multi-line chart? Give a sample line chart.
17. Think about how many direct and indirect labours there are in a company, as shown in the accompanying table. Using Excel, create a multi-line chart with labels for the data in the table.

Year	*Number of Direct Labours*	*Number of Indirect Labours*
1	600	240
2	700	260
3	760	180
4	900	340
5	700	410
6	1000	270

References

1. Panneerselvam, R., *Research Methodology* (2nd edition), PHI Learning Private Limited, New Delhi, 2014.
2. https://edu.gcfglobal.org/en/Excel2016/charts/1/ [June 25, 2020].
3. www.Excel-easy.com/examples/pie-chart.html [July 8, 2020].
4. www.techonthenet.com/Excel/charts/bar_chart2016.php [June 27, 2020].
5. www.quora.com/How-can-I-create-a-multiple-line-graph-in-Excel.

17 Linear Programming

Learning Objectives

After gong through this chapter, you will be able to

- Understand linear programming and its applications.
- Recognise the assumptions of linear programming.
- Analyse the components of linear programming.
- Use graphical method for linear programming.
- Apply the simplex method for linear programming problems.
- Apply Excel to solve linear programming problems.

17.1 Introduction

A mathematical model for a decision-making problem in the real world is called linear programming. It is a method of mathematical programming that seeks to maximise a system's performance metric while adhering to a certain set of management-imposed restrictions. The profit, cost, sales revenue, benefits, and so on may be used as a system's performance indicators. A number of resources will be necessary for the system. Each of those resources has a finite amount of availability. A collection of constraints is created when such system limitations are included in the linear programming model. The resources could be capital, manpower, machine hours, raw materials, and so on.

A sample set of applications of linear programming is as follows.

1. Product mix problem
2. Manpower scheduling problem
3. Diet problem
4. Capital budgeting problem
5. Cargo loading problem
6. Transportation problem
7. Assignment problem
8. Project selection problem

DOI: 10.4324/9781032671703-17

17.2 Assumptions of Linear Programming

The mathematical model of a system is developed with certain assumptions. The assumptions of linear programming are as follows [1].

1. Linearity
2. Divisibility
3. Non-negativity
4. Additivity

17.2.1 Linearity

The linearity assumption deals with the linear relationship between the quantity of a resource used and the level of an activity which uses that resource.

If the quantity of material required to manufacture one unit of a product is 10 kg, then the quantity of the material required to manufacture 5 units of that product is 50 kg. This sets an example of linearity in linear programming.

17.2.2 Divisibility

The nature of values allowed for decision variables is related to the linear programming model's divisibility assumption. The linear programming model's decision variables are allowed to have fractional values in accordance with the divisibility assumption.

17.2.3 Non-Negativity

The sign of the values allowed for decision variables is related to the non-negativity assumption of a linear programming model. The linear programming approach allows decision variables to have any value which is 0 and above 0. Negative decision variable values are not allowed in the model.

17.2.4 Additivity

A linear programming model's additivity assumption denotes the addition of individual activity output levels to produce the overall output for a particular combination of activity levels.

17.3 Components of Linear Programming Model

A linear programming model has four components, which are as follows.

1. Decision variables
2. Objective function
3. Set of constraints
4. Non-negative variables

A company manufactures two products, P_1 and P_2, and the profit per unit of product P_1 and that of product P_2 are ₹ 90 and ₹ 60, respectively. The products require wood and plastic beading to manufacture them. One unit of product P_1 requires 8 kg of wood and 6 kg of plastic beading. One unit of product P_2 requires 5 units of wood and 3 kg of plastic beading. The total quantity of wood available per day is 800 kg and the plastic beading available per day is 700 kg.

Formula a linear programming model to find the production volumes of the products such that the total profit is maximised subject to the material availability constraints.

17.3.1 Component 1: Decision Variables

The decision variables of a linear programming model are the levels of activities of the decision environment.

Let

X_1 be the production volume of product P_1 per day

X_2 be the production volume of product P_2 per day

Based on the definition of the decision variables, a linear programming model of the problem is presented as follows.

$$Maximise\,Z = 90X_1 + 60X_2\,[Objective\,Function]$$

Subject to

$$8X_1 + 5X_1 \leq 800 \qquad [Constraint\,for\,wood\,availability]$$

$$6X_1 + 3X_2 \leq 700 \qquad [Constraint\,for\,plastic\,beading\,availability]$$

Where

$$X_1 \geq 0\,and\,X_1 \geq 0 \qquad [Non-negativity\,constraints]$$

17.3.2 Component 2: Objective Function

Component 2 of the linear programming model is the objective function, which represents the total profit of manufacturing X_1 units of product $P_1\,and\,X_2$ units of product P_2.

For the given problem, the objective function is as presented below, which is the sum of the multiples of the profit/unit of the products with the respective production volumes of the products.

$$Maximise\,Z = 90X_1 + 60X_2$$

Note: The objective function may be either the maximisation type or minimisation type.

17.3.3 Component 3: Constraints

For the given problem, there are two constraints. The first constraint limits the total amount of wood consumed to manufacture X_1 units of product P_1 and X_2 units of product P_2 to a maximum of 800 kg per day

The second constraint limits the total amount of plastic beading consumed to manufacture X_1 units of product P_1 and X_2 units of product P_2 to a maximum of 700 kg per day. These constraints are as follows.

$8X_1 + 5X_1 \leq 800$ $\quad$ $[\textit{Constraint for wood availability}]$

$6X_1 + 3X_2 \leq 700$ $\quad$ $[\textit{Constraint for plastic beading availability}]$

Note: The type of the constraint may be " $\leq$ constraint " *or* " $\geq$ constraint" *or* " = *constraint* "

17.3.4 Component 4: Non-Negativity Constraints

The non-negativity constraints ensure that the value of each decision variable is greater than or equal to zero. For the given problem, the non-negativity constraints are as follows.

$X_1 \geq 0 \, and \, X_1 \geq 0$ $\;[Non-negativity\ constraints]$

17.4 Examples of Mathematical Models

This section illustrates linear programming models used in various industrial contexts after briefly introducing the concept of linear programming models. The readers' ability to use mathematical modelling would be enhanced by understanding these models.

Example 17.1

A machine shop employs 8 skilled operators and 12 semi-skilled operators for making a product in two designs, Model 1 and Model 2. The production of a unit of Model 1 requires four hours work by a skilled operator and two hours work by a semi-skilled operator. Model 2 requires three hours work by a skilled operator and four hours work by a semi-skilled operator. According to the employees' union rule, no operator can work more than eight hours per day. The profit for Model 1 is ₹ 2,000 per unit and that for Model 2 is ₹ 1,200 per unit. Formulate a linear programming model for this manufacturing situation to determine the production volume of each model such that the total profit is maximised.

Solution

Number of hours per day per operator = 8 hours
Number of skilled operators = 8
Number of semiskilled operators = 12
Number of hours of skilled operators available per day = 64 hours
Number of hours of semi-skilled operators available per day = 96 hours

The details of the processing time of the models by different operators are summarised in Table 17.1.

Let
X_1 be the production volume per day of Model 1
X_2 be the production volume per day of Model 2

Table 17.1 Details of Processing Times of Models by Operators

Resource	*Product*		*Resource Availability*
	Model 1	*Model 2*	
Hours of skilled operator	4	3	64 hours
Hours of semi-skilled operator	2	4	96 hours
Profit per unit (₹)	2,000	1,200	

Table 17.2 Processing Times (in Minutes) of the Operations of the Four Products on the Three Machines

Machine	*Product*				*Number of Minutes Available per Day*
	A	*B*	*C*	*D*	
Lathe	5	4	6	4	4,800
Milling machine	2	4	5	5	4,320
Drilling machine	1	3	2	4	2,160

The linear programming model of this problem is shown as follows.

$$Maximize\, Z = 2000X_1 + 1200X_2$$

Subject to

$$4X_1 + 3X_2 \leq 64$$

$$2X_1 + 4X_2 \leq 96$$

Where

$X_1 \geq 0$ and $X_1 \geq 0$

Example 17.2

Four products, A, B, C, and D, are produced by a company, and their relative profits per unit are ₹ 3000, ₹ 2000, ₹ 4000, and ₹ 2500, respectively. These products need to be processed on a lathe, a milling machine, and a drilling machine. Table 17.2 displays the processing times for the four products on the three machines. The last column of Table 17.2 displays the machine capabilities in minutes.

Waste from the raw material of product C serves as the raw material for product D. Every time two pieces of product C are produced, raw materials are produced for one unit of product D. The range of the daily demand for product B is between 20 and 80 units. The market will see complete sales of the other products.

Develop a linear programming model to determine the production volumes of the products each day that will maximise the company's overall profit.

Table 17.3 Data for Example 17.2

Machine	*Product*				*Number of Minutes Available per Day*
	A	*B*	*C*	*D*	
Lathe	5	4	6	4	4,800
Milling machine	2	4	5	5	4,320
Drilling machine	1	3	2	4	2,160
Profit per unit (₹)	3,000	2,000	4,000	2,500	

Solution

The data for Example 17.2 are summarised in Table 17.3.

Let

X_1 be the production volume per day of product A

X_2 be the production volume per day of product B

X_3 be the production volume per day of product C

X_4 be the production volume per day of product D

The linear programming model for the given problem is shown as follows.

$$Maximize\, Z = 3000X_1 + 2000X_2 + 4000X_3 + 2500X_4$$

$$Subject\, to$$

$$5X_1 + 4X_2 + 6X_3 + 4X_4 \leq 4800$$

$$2X_1 + 4X_2 + 5X_3 + 5X_4 \leq 4320$$

$$X_1 + 3X_2 + 2X_3 + 4X_4 \leq 2160$$

$$X_3 - 2X_4 = 0$$

$$\mathrm{X}_2 \geq 20$$

$$X_2 \leq 80$$

$$\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \mathrm{X}_4 \geq 0$$

17.5 Graphical Method for Linear Programming

The models presented in the earlier section need to be solved to obtain the values of the decision variables stated in them. If the number of decision variables in the linear programming model is only two, one can use graphical method to solve that model.

The steps of solving the linear programming model using the graphical method are as follows.

Step 1: Assign the variable, say, X_1 to the X axis and another variable, say, X_2, to the Y axis of the X-Y plane.

Step 2: Introduce non-negative constraints, that is, $\mathrm{X}_1 \geq 0$ and $\mathrm{X}_2 \geq 0$ to eliminate the solution spaces in the second quadrant, third quadrant, and fourth quadrant of the X-Y plane.

Step 3: For each constraint of the model, perform the following.

3.1: Equate the left-hand side of the constraint to the right-hand side of the constraint irrespective of the type of the constraint ($\leq$ *type conatraint*, $\geq$ *type constraint and* $=$ *type constrain*.

3.2: Set the value of $X_1 = 0$ in the constraint and solve for the value of X_2.

3.3: Set the value of $X_2 = 0$ in the constraint and solve for the value of X_1.

Step 4: Take an appropriate scale for the values of X_1 and X_2, mainly to accommodate all the coordinates of the points of the constraints in the graph.

Step 5: For each constraint of the model, perform he following.

5.1: Mark the two points of the constraint in the first quadrant of the *X*-*Y* plane and connect those points.

5.2: Substitute the coordinates of the origin of the graph in the left-hand side of the constraint and check whether it satisfies the nature of the relation ($\leq$ *or* $\geq$) of that constraint.

If yes, the side containing the origin with respect to the line of the constraint is the feasible side of that constraint; otherwise, the other side of the line of the constraint is the feasible side of that constraint.

Special case: If the constraint is an equal to type constraint, the solution space for that constraint is on the line, which represents that constraint.

Step 6: Find the intersection of the feasible regions of the solution space of all the constraints, which represents the solution space that contains the optimal solution of the given model.

Step 7: Determination of the coordinates of the optimal solution of the model.

7.1: Evaluate the objective function value by substituting the coordinates of X_1 and X_2 of each of the corner points of the final feasible solution space.

7.2: Find the corner point of the feasible solution space, which has the least objective function value in the case of minimisation problem and maximum objective function value in the case of maximisation problem.

Step 8: Treat the coordinates of the corner point identified in Step 7.2 as the coordinates of the optimal solution and the corresponding objective function value as the optimal value of the objective function of the model.

Example 17.3

Solve the following LP problem graphically.

$$Maximise\, Z = 20X_1 + 80X_2$$

$$Subject\, to$$

$$4X_1 + 6X_2 \leq 90$$

$$8X_1 + 6X_2 \leq 100$$

$$5X_1 + 4X_2 \leq 80$$

$$X_1, X_2 \geq 0$$

Solution

The given linear programming model is shown as follows.

$$Maximise\, Z = 20X_1 + 80X_2$$

Subject to

$$4X_1 + 6X_2 \leq 90$$

$$8X_1 + 6X_2 \leq 100$$

$$5X_1 + 4X_2 \leq 80$$

$$X_1, X_2 \geq 0$$

Coordinates for constraint 1

$$4X_1 + 6X_2 \leq 90$$

Convert this constraint into an equation as follows.

$$4X_1 + 6X_2 = 90$$

When $X_1 = 0, X_2 = 15$
$X_2 = 0, X_1 = 22.5$

Coordinates for constraint 2

$$8X_1 + 6X_2 \leq 100$$

Convert this constraint into an equation as follows.

$$8X_1 + 6X_2 = 100$$

When $X_1 = 0, X_2 = 16.67$
$X_2 = 0, X_1 = 12.5$

Coordinates for constraint 3

$$5X_1 + 4X_2 \leq 80$$

Convert this constraint into an equation as follows.

$$5X_1 + 4X_2 = 80$$

When $X_1 = 0, X_2 = 20$
$X_2 = 0, X_1 = 16$

Since the values of the coordinates of the two points in each of the three constraints are not very high values, there is no need to take scale.

Plot constraint 1 in the *X-Y* plane as shown in Figure 17.1. Then substituting the coordinates (0,0) of the origin in the left-hand side of the constraint 1 gives 0, which is less than its right-hand side. Hence, the side of constraint 1 which contains the origin is the feasible side.

Plot constraint 2 in the *X-Y* plane as shown in Figure 17.1. Then substituting the coordinates (0,0) of the origin in the left-hand side of constraint 2 gives 0, which is less than its right-hand side. Hence, the side of constraint 2 which contains the origin is the feasible side.

Plot constraint 3 in the *X-Y* plane as shown in Figure 17.1. Then substituting the coordinates (0,0) of the origin in the left-hand side of constraint 3 gives 0, which is less than its right-hand side. Hence, the side of constraint 3 which contains the origin is the feasible side.

The closed polygon representing the feasible region of the given model is A-B-C-D.

The objective function value for the coordinates of each of the corner points of the closed feasible polygon is computed as follows.

$Z(A)$ = 20 x 0 + 80 x 0 = 0
$Z(B)$ = 20 x 12.5 + 80 x 0 = 250
$Z(C)$ = 20 x 2.5 + 80 x (40/3) = 1116.67
$Z(D)$ = 20 x 0 + 80 x 15 = 1200

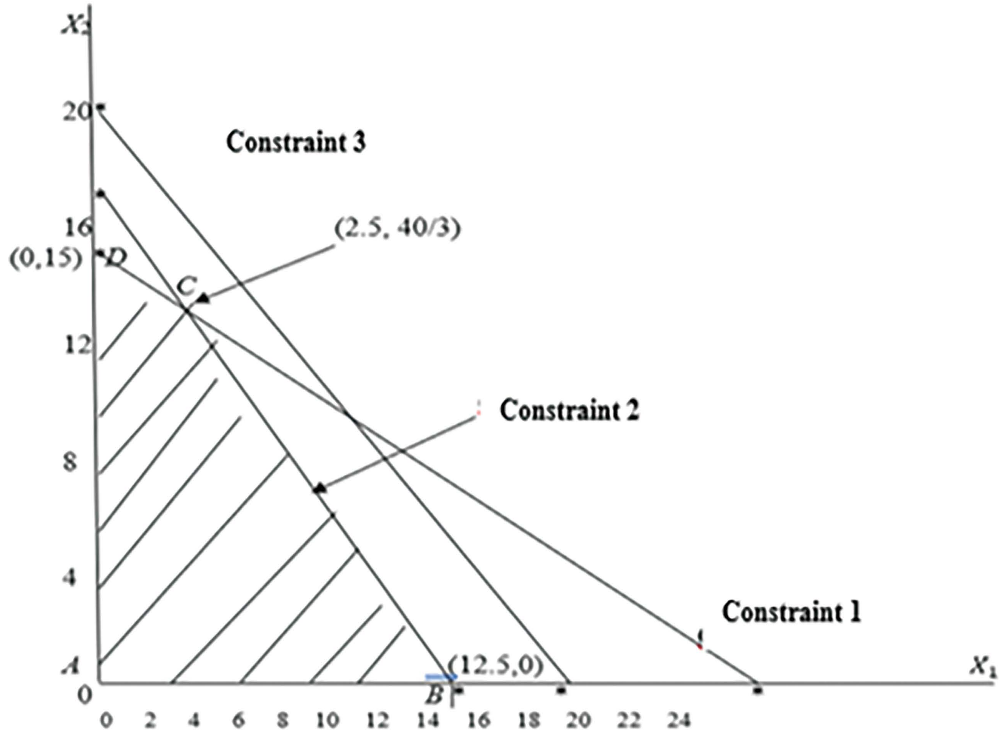


Figure 17.1 Plot of constraints of Example 17.3

The objective function value is the maximum for the corner point *D* of the closed feasible polygon. The corresponding solution is:

$Z_{\text{opt}} = 1200$, $X_1{}^* = 0$ & $X_2{}^* = 15$

Example 17.4

Solve the following LP problem graphically.

$$\begin{aligned}&Maximise\, Z = 60X_1 + 90X_2\\&Subject\, to\\&X_1 + 2X_2 \le 40\\&2X_1 + 3X_2 \le 90\\&X_1 - X_2 \ge 10\\&X_1 \text{ and } X_2 \ge 0\end{aligned}$$

Solution

The given linear programming model is shown as follows.

$$\begin{aligned}&Maximise\, Z = 60X_1 + 90X_2\\&Subject\, to\\&X_1 + 2X_2 \le 40\\&2X_1 + 3X_2 \le 90\\&X_1 - X_2 \ge 10\\&X_1 \text{ and } X_2 \ge 0\end{aligned}$$

Coordinates for constraint 1:

$$X_1 + 2X_2 \le 40$$

Convert this constraint into an equation as follows.

$$X_1 + 2X_2 = 40$$

When $X_1 = 0$, $X_2 = 20$ and when $X_2 = 0$, $X_1 = 40$

Coordinates for constraint 2:

$$2X_1 + 3X_2 \le 90$$

Convert this constraint into an equation as follows.

$$2X_1 + 3X_2 = 90$$

When $X_1 = 0$, $X_2 = 30$ and when $X_2 = 0$, $X_1 = 45$

Coordinates for constraint 3:

$$X_1 - X_2 \geq 10$$

Convert this constraint into an equation as follows.

$$X_1 - X_2 = 10$$

When $X_1 = 0$, $X_2 = -10$ and when $X_2 = 0$, $X_1 = 10$

Since the values of the coordinates of the two points in each of the three constraints are not very high values, there is no need to take scale.

Plot constraint 1 in the *X*-*Y* plane as shown in Figure 17.2. Then substituting the coordinates (0,0) of the origin in the left-hand side of the constraint 1 gives 0, which is less than its right-hand side. Hence, the side of constraint 1 which contains the origin is the feasible side.

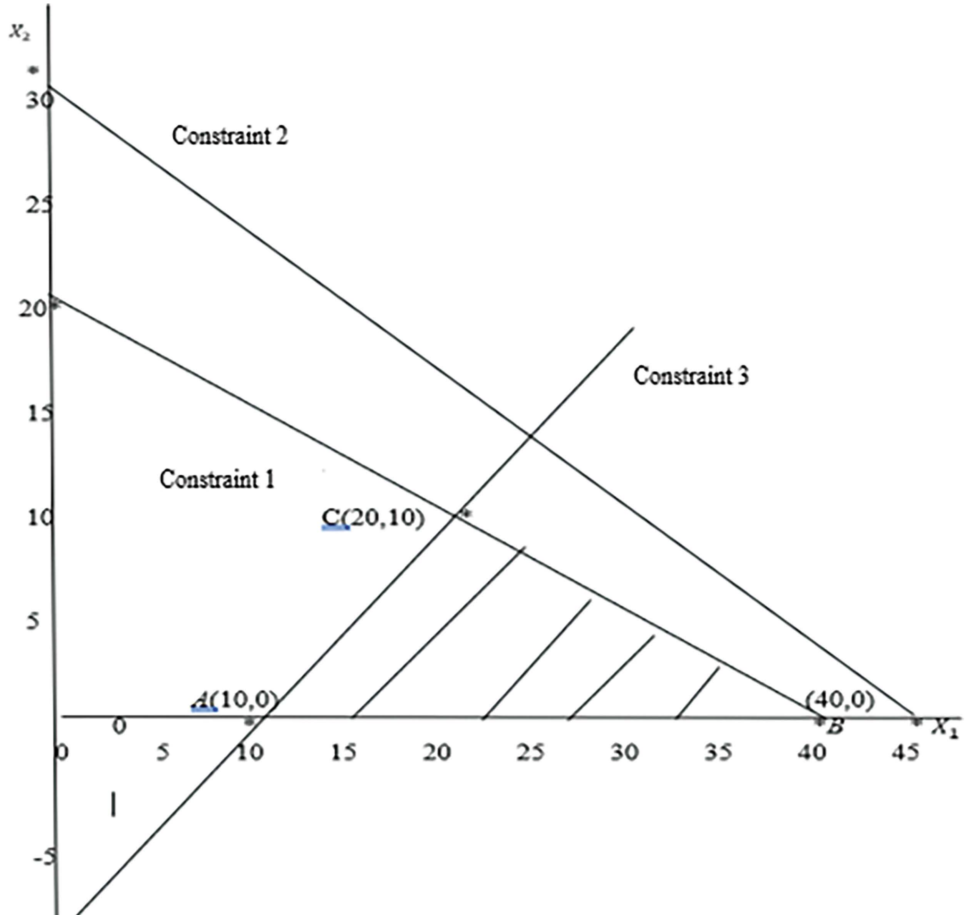


Figure 17.2 Plot of constraints of Example 17.4

Plot constraint 2 in the *X*-*Y* plane as shown in Figure 17.2. Then substituting the coordinates (0,0) of the origin in the left-hand side of the constraint 2 gives 0, which is less than its right-hand side. Hence, the side of constraint 2 which contains the origin is the feasible side.

Plot constraint 3 in the *X*-*Y* plane as shown in Figure 17.2. Then substituting the coordinates (0,0) of the origin in the left-hand side of constraint 3 gives 0, which is not greater than its right-hand side. Hence, the side of the constraint 3 which does not contain the origin is the feasible side.

The closed polygon representing the feasible region of the given model is A-B-C.

The objective function value for the coordinates of each of the corner points of the closed feasible polygon is computed as follows.

$Z(A) = 60 \text{ x } 10 + 90 \text{ x } 0 = 600$
$Z(B) = 60 \text{ x } 40 + 90 \text{ x } 0 = 2400$
$Z(C) = 60 \text{ x } (20) + 90 \text{ x } (10) = 2100$

The objective function value is the maximum for the corner point *B* of the closed feasible polygon.

$Z_{opt} = 2400$, $X_1^* = 40$ & $X_2^* = 0$

17.6 Simplex Method

George Dantzig developed the simplex approach in 1947 to address the challenge of linear programming. This is an iterative process that starts with a feasible starting solution and keeps going until an optimality condition is met. The solution's feasibility is maintained during each iteration. By adding a slack variable to the less than or equal to constraint and a surplus variable to the greater than or equal to constraint, the original linear programming model is transformed into a standard form.

Consider the following linear programming problem.

$$Maximise\, Z = 10X_1 + 8X_2$$

$$Subject\, to$$

$$2X_1 + 5X_2 \le 100$$

$$4X_1 + 5X_2 \ge 200$$

Where X_1 and $X_2 \ge 0$

A slack variable is a variable that is introduced in a less than or equal to constraint, which equates the left-hand side of that constraint to the right-hand side of that constraint.

Consider first constraint as follows.

$$2X_1 + 5X_2 \le 100$$

Since the left-hand side of the constraint is less than or equal to its right-hand side, a variable called a slack variable, say, S_1, is introduced with a plus sign on the left-hand side of the constraint to balance both sides with zero as its coefficient in the objective function as follows.

$$2X_1 + 5X_2 + S_1 = 100$$

A surplus variable is a variable that is introduced in a greater than or equal to constraint, which equates the left-hand side of that constraint to the right-hand side of that constraint.

Consider the second constraint as follows.

$$4X_1 + 5X_2 \geq 200$$

Since the left-hand side of the constraint is greater than or equal to its right-hand side, a variable called surplus variable, say, S_2 , is introduced with a minus sign on the left-hand side of the constraint to balance both sides with zero as its coefficient in the objective function as follows.

$$4X_1 + 5X_2 - S_2 = 200$$

Then a check for the canonical form of the model is carried out. A canonical form of the linear programming model contains a basic variable in each of the constraints.

Basic Variable: A basic variable is a variable in a constraint such that it has a unit coefficient in that constraint and zero coefficient in the remaining constraints.

Consider the less than or equal to constraint which was balanced using a slack variable S_1 along with another constraint with a greater than or equal to constraint which was balanced using a surplus variable as follows.

$$2X_1 + 5X_2 + S_1 = 100$$

$$4X_1 + 5X_2 - S_2 = 200$$

Since the model contains only these two constraints, then the variable S_1 can be treated as the basic variable of the first constraint, because it has a unit coefficient in that constraint and a zero coefficient in the other constraint, which is the requirement of a basic variable.

The surplus variable S_2 in the second constraint cannot act as the basic variable of that constraint, because it has –1 as its coefficient. Hence, another variable, say, R_1, is introduced on the left-hand side of the second constraint with a coefficient of +1 to serve as the basic variable of this constraint and $-M$ as its coefficient in the objective function, because the objective function is the maximisation type, where M is a very high value (in the case of a minimisation problem, an artificial variable is to be included in the objective function with $+M$ as its coefficient). The variable R_1 is introduced in this constraint just to serve as the basic variable; hence, in the final solution, it should not be present. So the variable R_1 is called an artificial variable.

Based on these guidelines, the canonical form of the given model is as shown.

$$\textit{Maximise}\, Z = 10X_1 + 8X_2 - M \times R_1$$

Subject to

$$2X_1 + 5X_2 + S_1 = 100$$

$$4X_1 + 5X_2 - S_2 + R_1 = 200$$

Where X_1, X_2, S_1, S_2, $R_1 \geq 0$

17.6.1 Steps of Simplex Method

The steps of the simplex method are presented in this section. The given canonical form of the model will be presented in the form of initial basic feasible table. The initial simplex table is shown in Table 17.4 for the canonical form of the model of the previous section.

In Table 17.4, the coefficients of the left-hand side variables of the constraints as shown in bold font put together are called technological coefficients $\left[a_{ij}\right]$. The basic variables are shown in the basic variable column. The maximum number of variables, which will have non-zero values, is equal to the number of constraints in the model. The objective function coefficients (CB_i) of the basic variables are shown under the CB_i column.

Let

m be the number of constraints

n be the total number of variables, including slack variables, surplus variables, and artificial variables

In Table 17.4, Z_j in column *j* is the sum of the multiples of the technological coefficients in that column with the corresponding value in the CB_i column.

$$Z_j = \sum_{i=1}^{m} \left(a_{ij} \times CB_i\right), \, for \; j = 1, 2, 3, \ldots, n+1,$$

including the solution column along with the total number of variables.

The criterion row values are given by the following formula.

$$Criterion\,Row\,value\,of\,the\,column\,j = C_j - Z_j$$

Table 17.4 Initial Simplex Table

CB_i	*Column j*	*1*	*2*	*3*	*4*	*5*	*Solution Column (*SC_i*)* *(Right-Hand Side of Constraint)*	*Ratio*
	C_j	*10*	*8*	*0*	*0*	*−M*		
	Basic Variable	X_1	X_2	S_1	S_2	R_1		
0	S_1	**2**	**5**	**1**	**0**	**0**	100	11*
−*M*	R_1	**4**	**5**	**0**	**−1**	**1**	200	26
Z_j		−4*M*	−5*M*	0	*M*	−*M*	−200*M*	
Criterion row : $C_j - Z_j$		10 + 4*M*	8 + 5*M*	0	−*M*	0		

17.6.1.1 Conditions for Optimality

The optimality of the solution obtained in the current iteration is governed by the following conditions.

For maximisation-type problems:

If all the values of $C_j - Z_j$ are 0 and negative, then optimality is reached and the values in the solution column represent the values of the corresponding basic variables in the second column of the table.

For minimisation-type problems:

If all the values of $C_j - Z_j$ are 0 and positive, then optimality is reached and the values in the solution column represent the values of the corresponding basic variables in the second column of the table.

17.6.1.2 Selection of Entering Variable/Key Column

In the current iteration, a promising non-basic variable will be selected as the basic variable to enter the set of basic variables in the next iteration.

If optimality is not reached in the current iteration, then a non-basic variable is to be identified as the entering variable using the following steps.

Select a non-basic variable which has the maximum positive criterion row value as the entering variable/key column. In case of a tie, break the tie randomly.

17.6.1.3 Selection of Leaving Variable/Key Row

In this iteration, a non-promising basic variable will be selected as the leaving variable, which will leave the set of basic variables in the next iteration. In effect, the entering variable will replace this leaving variable in the next iteration.

Find the ratio of the solution column values $\left[SC_I\right]$ and the corresponding technological coefficients $\left[a_{ij}\right]$ of the key column values only for positive denominators (technological coefficients). Then select the row with the minimum ratio as the key row/leaving variable to maintain feasibility.

The steps of the simplex method are as follows.

Step 1: Convert the given linear programming model into its canonical form.
Step 2: Set Iteration (I) = 1.
Step 3: Represent the canonical form of the given model in initial simplex table.
Step 4: Find Z_j for j = 1, 2, 3, . . . , n+1, including the solution column using the following formula.

$$Z_j = \sum_{i=1}^{m}\left(a_{ij} \times CB_i\right), \textit{for } j = 1, 2, 3, \ldots, n+1,$$

Step 5: Find the values of the criterion row for j = 1, 2, 3, . . . , n.

$$\textit{Criterion Row value of the column } j = C_j - Z_j, j = 1,2,3,\ldots,n+1$$

Step 6: Optimality check.

For maximisation-type problems:

If all the values of $C_j - Z_j$ are 0 and negative, then optimality is reached and the values in the solution column represent the values of the corresponding basic variables in the second column of the table. Then go to Step 12. Otherwise, go to Step 7.

For minimisation-type problems:

If all the values of $C_j - Z_j$ are 0 and positive, then optimality is reached and the values in the solution column represent the values of the corresponding basic variables in the second column of the table. Then go to Step 12. Otherwise, go to Step 7.

Step 7: Selection of key column/entering variable

For maximisation-type problems:

Find the non-basic variable column with the maximum positive criterion row value and treat it as the key column/entering variable.

For minimisation-type problems:

Find the non-basic variable column with the highest negative criterion row value and treat it as the key column/entering variable.

Step 8: Selection of key row/leaving variable
Find the ratio of the solution column values and the correcting technological coefficients of the key column values only for positive denominators (technological coefficients). Then select the row with the minimum ratio as the key row/leaving variable to maintain feasibility.
Step 9: Increment the iteration number by 1 ($I = I + 1$).
Step 10: Obtain a simplex table for the I^{th} iteration from the $I - 1$th iteration using the following formula.

10.1: Find the key value or pivot value using the following formula.
Key value or pivot value = The value at the intersection of the key row and the key column in the technological coefficient matrix of the iteration $I - 1$.
10.2:

$$New\,value\,of\,the\,technological\,coefficient\,a_{ij}\,in\,the\,current\,table\left(Iteration\,I\right)$$
$$= \frac{Key\,row\,value\,with\,respect\,to\,a_{ij} \times Key\,column\,value\,with\,respect\,to\,a_{ij}}{Key\,value}$$

Step 11: Go to Step 4.
Step 12: Print the values of the objective function value at the Z_{n+1} location of the current table and basic variables.

Example 17.5

Solve the following LP problem using the simplex method.

$$\begin{aligned}&\textit{Maximise}\, Z = 5X_1 + 3X_2 + 7X_3\\&\textit{Subject to}\\&X_1 + X_2 + 2X_3 \le 22\\&3X_1 + 2X_2 + X_3 \le 26\\&X_1 + X_2 + X_3 \le 18\end{aligned}$$

Where X_1, X_2 *and* $X_3 \geq 0$

Solution

The given model is as follows.

$$\begin{aligned}&\textit{Maximise}\, Z = 5X_1 + 3X_2 + 7X_3\\&\textit{Subject to}\\&X_1 + X_2 + 2X_3 \le 22\\&3X_1 + 2X_2 + X_3 \le 26\\&X_1 + X_2 + X_3 \le 18\end{aligned}$$

Where X_1, X_2 *and* $X_3 \geq 0$

The canonical form of the given problem is:

$$\begin{aligned}&\textit{Maximise}\, Z = 5X_1 + 3X_2 + 7X_3\\&\textit{Subject to}\\&X_1 + X_2 + 2X_3 + S_1 = 22\\&3X_1 + 2X_2 + X_3 + S_2 = 26\\&X_1 + X_2 + X_3 + S_3 = 18\end{aligned}$$

Where X_1, X_2, X_3, S_1, S_2, *and* $S_3 \geq 0$

17.6.1.4 Application of Simplex Method

The application of the simplex method to this problem is presented as follows.

Iteration 1

The initial simplex table is presented in Table 17.5.

Table 17.5 Initial Simplex Table

CB_i	*C_j*	*5*	*3*	*7*	*0*	*0*	*0*	*Solution*	*Ratio*
	Basic Variable	X_1	X_2	X_3	S_1	S_2	S_3		
0	S_1	**1**	**1**	**2**	**1**	**0**	**0**	**22**	**11***
0	S_2	3	2	**1**	0	1	0	26	26
0	S_3	1	1	**1**	0	0	1	18	18
Z_j		0	0	**0**	0	0	0	0	
Criterion Row $:C_j - Z_j$		5	3	7*	0	0	0		

In Table 17.5, there are positive criterion row values; hence the solution of the current iteration is not optimal.

The maximum positive criterion row value is for the non-basic variableX_3. Treat it as the entering variable. The minimum ratio with respect to the column of this variable is 11, and the corresponding basic variable is S_1, which is the leaving variable So this variable is replaced by the entering variable X_3 in the next iteration.

Iteration 2

The corresponding next simplex table is shown in Table 17.6 using the following formula.

In Table 17.5, the key value = 2.

$$New\,value\,of\,the\,technological\,coefficient\,a_{ij}\,in\,the\,current\,table$$
$$= \frac{Key\,row\,value\,with\,respect\,to\,a_{ij} \times Key\,column\,value\,with\,respect\,to\,a_{ij}}{Key\,value}$$

In Table 17.6, there is a positive criterion row value for the non-basic variable X_1 ; hence the solution of the current iteration is not optimal.

The only positive criterion row value is for the non-basic variableX_1. Treat it as the entering variable. The minimum ratio with respect to the column of this variable is 6, and the corresponding basic variable is S_2, which is the leaving variable. So this variable is replaced by the entering variable X_1.

Iteration 3

The corresponding next simplex table is shown in Table 17.7 using the following formula.

In Table 17.6, the key value $=\frac{5}{2}$.

$$New\,value\,of\,the\,technological\,coefficient\,a_{ij}\,in\,the\,current\,table$$
$$= \frac{Key\,row\,value\,with\,respect\,to\,a_{ij} \times Key\,column\,value\,with\,respect\,to\,a_{ij}}{Key\,value}$$

In Table 17.7, all the values of the criterion row are 0 and negative. Hence, the optimality condition is reached.

The optimum solution: $X_1 = 6$, $X_2 = 0$, $X_3 = 8$, and $Z_{\text{opt}} = 86$

Table 17.6 Simplex Table of Iteration 2

CB_i	C_j	5	3	7	0	0	0	*Solution*	*Ratio*
	Basic Variable	X_1	X_2	X_3	S_1	S_2	S_3		
7	X_3	**1/2**	1/2	1	1/2	0	0	11	22
0	S_2	**5/2**	**3/2**	**0**	**–1/2**	**1**	**0**	**15**	**6***
0	S_3	**1/2**	1/2	0	–1/2	0	1	7	14
Z_j		**7/2**	7/2	7	7/2	0	0	77	
$C_j - Z_j$		**3/2***	–1/2	0	–7/2	0	0		

Table 17.7 Simplex Table of Iteration 3

CB_i	C_j	5	3	7	0	0	0	*Solution*
	Basic Variable	X_1	X_2	X_3	S_1	S_2	S_3	
7	X_3	0	1/5	1	3/5	–1/5	0	8
5	X_1	1	3/5	0	–1/5	2/5	0	6
0	S_3	0	1/5	0	–2/5	–1/5	1	4
Z_j		5	22/5	7	4	3/5	0	86
$C_j - Z_j$		0	–7/5	0	–4	–3/5	0	

17.7 Solving Linear Programming Problems Using Excel

Excel has an add-in feature called Excel Add-ins. Excel Add-ins in turn has a Solver Add-in option, which will include Solver under Analysis, which appears at top right corner of the Excel sheet. The Solver function is used to solve linear programming models, which come under the decision-making environment of operations research [2].

If this option is not available in the Excel sheet, use the following steps to include it.

Step 1: Click File in the Excel sheet.

Step 2: Click Options in the menu that appears at the leftmost side of the screen in response to clicking File.

Step 3: Under Excel options, which appears after clicking Options, as mentioned in Step 2, click Add-ins in the dropdown menu, which is present in the leftmost column.

Step 4: In the response screen after clicking Add-ins, click Go . . . , which is present at the bottom of the screen.

Step 5: Click the checkbox against Solver Add-in in the response screen after clicking Go . . ., and then click the OK button, which will add the Solver function under the Data button of the Excel sheet.

Example 17.6

Two products, product 1 and product 2, are produced by a corporation. Two distinct machines, machine 1 and machine 2, are used to machine each product. The profit made from producing and selling each unit of product 1 is ₹ 10, while the profit made from producing and selling each unit of product 2 is ₹. Table 17.8 displays the amount of time

Table 17.8 Machining Times of Products

Machine	*Product 1*	*Product 2*	*Maximum Number of Hours Available*
Machine 1	2	5	100
Machine 2	4	5	200

needed to machine each product on each of the two machines as well as the maximum hours that can be used on each machine.

a) Create a linear programming model to find the production volume of each product such that the total profit is maximised.
b) Solve the linear programming model using the Solver function.

Solution

The data for Example 17.6 are shown in Table 17.9.

Let

X_1 be the production volume of product 1

X_2 be the production volume of product 2

A linear programming model of this problem is presented in the following.

$$\textit{Maximise}\, Z = 10X_1 + 8X_2$$

Subject to

$$2X_1 + 5X_2 \leq 100$$

$$4X_1 + 5X_2 \leq 200$$

$$\textit{where}\ \mathrm{X}_1 \geq 0\ \textit{and}\ X_1 \geq 0$$

An Excel template to suit to this data is shown in Figure 17.3.

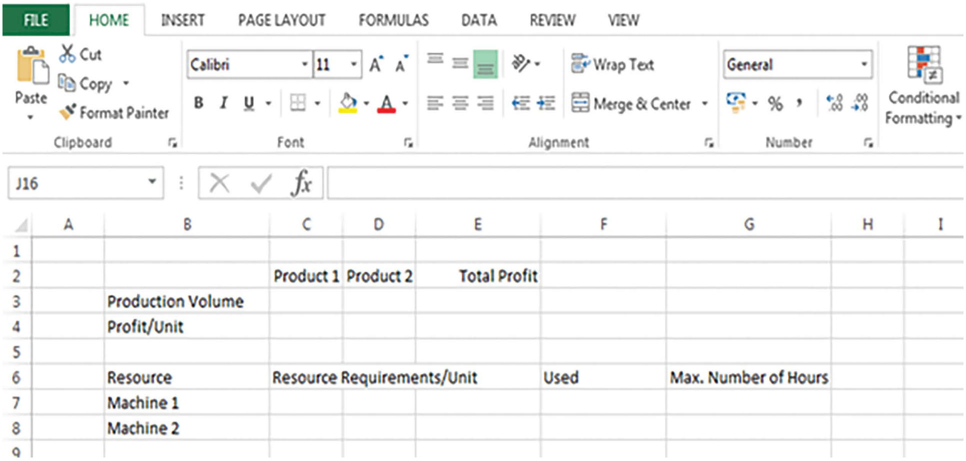
The cell definitions in Figure 17.3 are presented in Table 17.10. One should note the = symbol used in each item in the last column of Table 17.10, which signifies formula used in the respective cell. The data like profit/unit of product 1 in cell C4 is given in formula form =10. The same is applicable for all other data in addition to the usage of formulas to compute the total profit in cell E4, total number of hours utilised in machine 1 in cell F7, total number of hours utilised in machine 2 in cell G7, maximum number of hours available in machine 1 in cell G7, and the maximum number of hours available in machine 2 in Cell G8.

The steps of solving the linear programming model using the Solver Add-in in Excel are as follows.

- Enter the data of the model in an Excel sheet, as shown in Figure 17.4.
- Click the Data button, which gives a display as shown in Figure 17.5.
- Click the Solver Add-in in Figure 17.5, which gives a display as shown in Figure 17.6.

Table 17.9 Data of Example 17.6

Profit/Unit (₹)	*10*	*8*	
Machine	*Product 1*	*Product 2*	*Maximum Number of Hours Available*
Machine 1	2	5	100
Machine 2	4	5	200

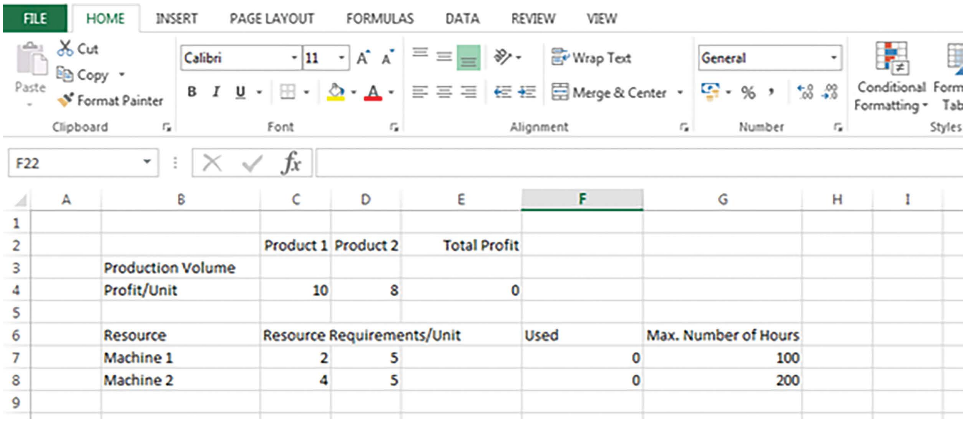


	A	B	C	D	E	F	G	H	I
1									
2			Product 1	Product 2	Total Profit				
3		Production Volume							
4		Profit/Unit							
5									
6		Resource	Resource Requirements/Unit			Used	Max. Number of Hours		
7		Machine 1							
8		Machine 2							
9									

Figure 17.3 Excel template to suit to linear programming model of Example 17.6

Table 17.10 Cell Definitions of Fig 17.3

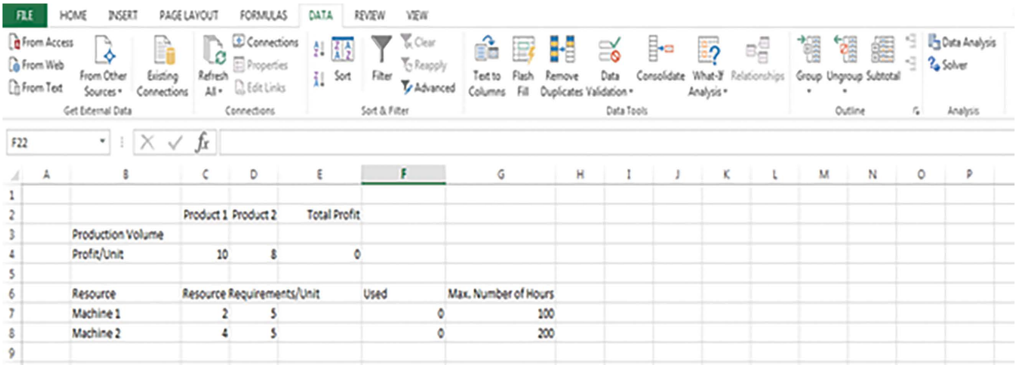
S. No.	*Cell Address*	*Definition*	*Remark*
1-2	C3:D3	Production volume of Product 1 and Product 2, respectively	Initially, these cells will have any value
3-4	C4:D4	Profit/unit of Product 1 and Product 2, respectively	C4: =10 D4: = 8
5	E4	Formula of objective function	=C4*C3+D4*D3
6-7	C7:C8	Machining times of the Product 1 on Machine 1 and Machine 2, respectively.	C7: =2 C8: =4
8-9	D7:D8	Machining times of the Product 2 on Machine 1 and Machine 2, respectively.	D7: =5 D8: =5
10-11	F7:F8	Total amount of hours of Machine 1 and that of Machine 2 utilized, respectively	F7: =C7*C3+D7*D3 F8: =C8*C3+D8*D3
12-13	G7:G8	Maximum number of hours available in Machine 1 and Machine 2, respectively.	G7: =100 G8: =200



F22

	A	B	C	D	E	F	G	H	I
1									
2			Product 1	Product 2	Total Profit				
3		Production Volume							
4		Profit/Unit	10	8	0				
5									
6		Resource	Resource Requirements/Unit			Used	Max. Number of Hours		
7		Machine 1	2	5		0	100		
8		Machine 2	4	5		0	200		
9									

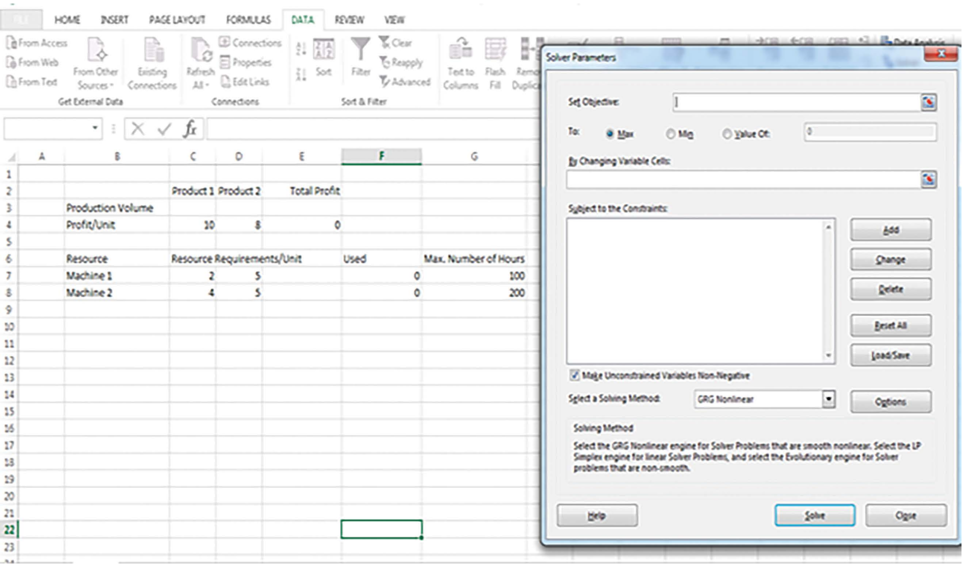
Figure 17.4 Data for model filled in Excel sheet



F22

	A	B	C	D	E	F	G
1							
2			Product 1	Product 2	Total Profit		
3		Production Volume					
4		Profit/Unit	10	8	0		
5							
6		Resource	Resource Requirements/Unit			Used	Max. Number of Hours
7		Machine 1	2	5		0	100
8		Machine 2	4	5		0	200
9							

Figure 17.5 Screenshot after clicking Data button



	A	B	C	D	E	F	G
1							
2			Product 1	Product 2	Total Profit		
3		Production Volume					
4		Profit/Unit	10	8	0		
5							
6		Resource	Resource Requirements/Unit			Used	Max. Number of Hours
7		Machine 1	2	5		0	100
8		Machine 2	4	5		0	200

Set Objective:

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells:

Subject to the Constraints:

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 17.6 Screenshot after clicking Solver Add-in in Figure 17.5

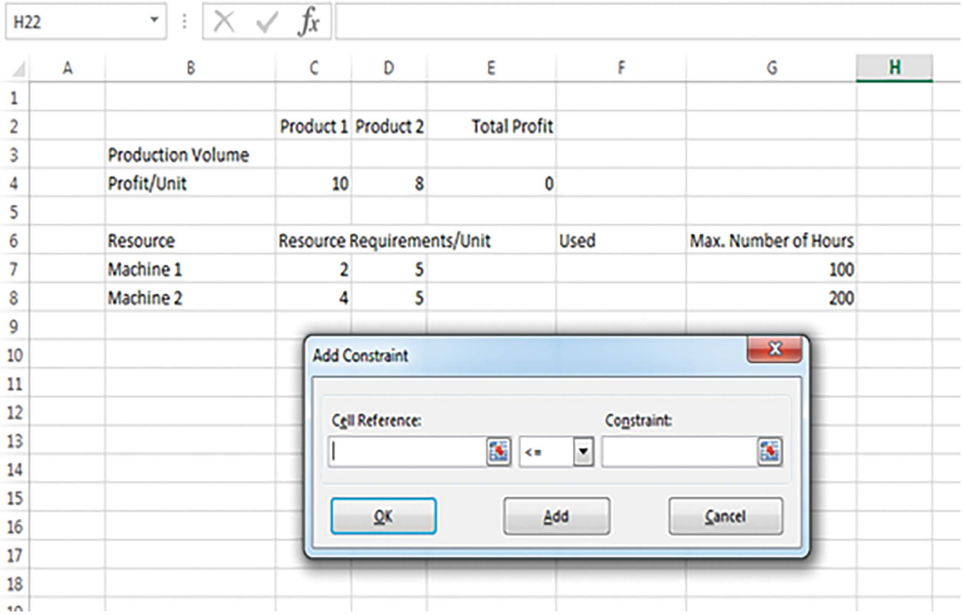
- Click the following addresses in the dropdown menu of Figure 17.6 and then click OK to show the display as shown in Figure 17.7.

1. Copy the formula for total profit in cell E4 in the box against Set Objective in the dropdown menu.
2. Select the type of objective function that is the max type for this problem.
3. Copy the cell address C3:D3 of production volumes of the products in the box against By Changing Variable Cells in the dropdown menu.
4. Click the Add button in the dropdown menu to add the first constraint, which gives a display as in Figure 17.8.

   4.1: Copy the cell formula in cell F7 in the box against the cell reference in the dropdown menu of Figure 17.8.
   4.2: Select the " ≤ " symbol from the next-level dropdown menu that is present at the middle of the dropdown menu of Figure 17.8.
   4.3: Copy cell G7 in the box against Constraint in the dropdown menu of Figure 17.8.
   4.4: Click the Add button in the dropdown menu of Figure 17.8 if there are more constraints to enter; otherwise, click the OK button to indicate the end of adding constraints.

   After clicking the Add button to add the second constraint, follow steps 4.1 to 4.4 to add that constraint. Now, click the OK button in the dropdown menu of adding constraints, which gives a display as in Figure 17.9.

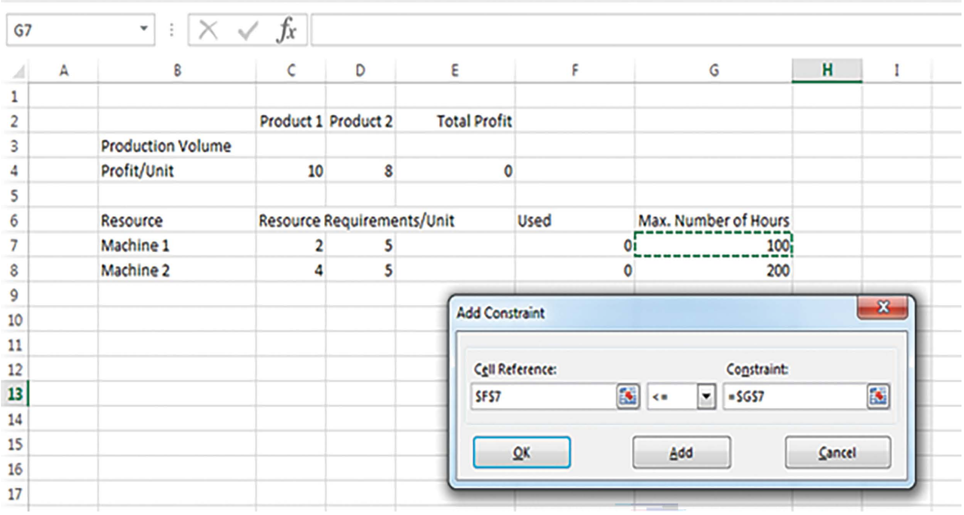
5. Click the tick mark in the checkbox for Make unconstraint variables non-negative in Figure 17.9.



H22

	A	B	C	D	E	F	G	H
1								
2			Product 1	Product 2	Total Profit			
3		Production Volume						
4		Profit/Unit	10	8	0			
5								
6		Resource	Resource Requirements/Unit			Used	Max. Number of Hours	
7		Machine 1	2	5			100	
8		Machine 2	4	5			200	

Figure 17.7 Screenshot after clicking Add button to add a constraint



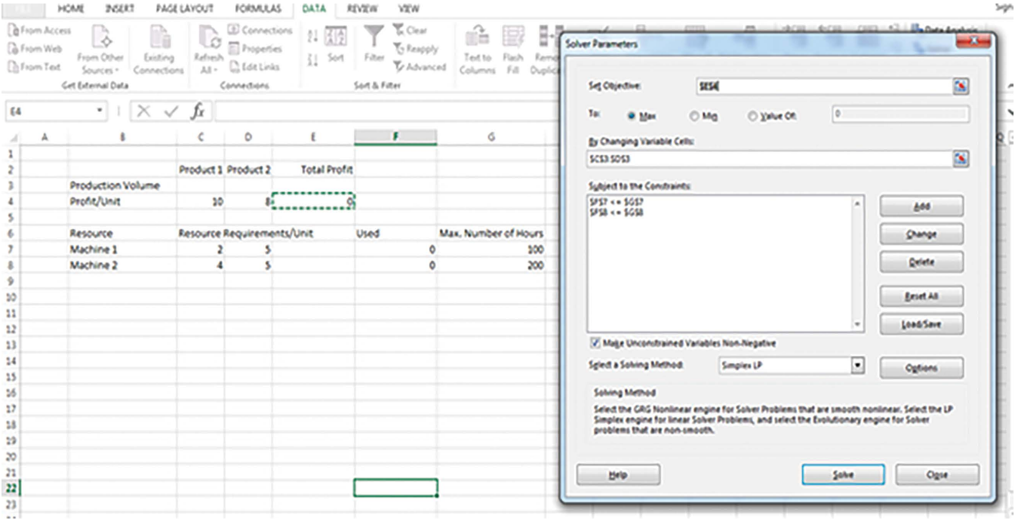
G7

	A	B	C	D	E	F	G	H	I
1									
2			Product 1	Product 2	Total Profit				
3		Production Volume							
4		Profit/Unit	10	8	0				
5									
6		Resource	Resource Requirements/Unit			Used	Max. Number of Hours		
7		Machine 1	2	5		0	100		
8		Machine 2	4	5		0	200		

Add Constraint

Cell Reference: $F$7 &lt;= Constraint: =$G$7

Figure 17.8 Screenshot after entering the left-hand side formula and right-hand side constant of first constant in the dropdown menu of Figure 17.7



E4

	A	B	C	D	E	F	G
2			Product 1	Product 2	Total Profit		
3		Production Volume					
4		Profit/Unit	10	8	0		
6		Resource	Resource Requirements/Unit			Used	Max. Number of Hours
7		Machine 1	2	5		0	100
8		Machine 2	4	5		0	200

Solver Parameters

Set Objective: $E$4

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells: $C$3:$D$3

Subject to the Constraints:
$F$7 <= $G$7
$F$8 <= $G$8

☑ Make Unconstrained Variables Non-Negative

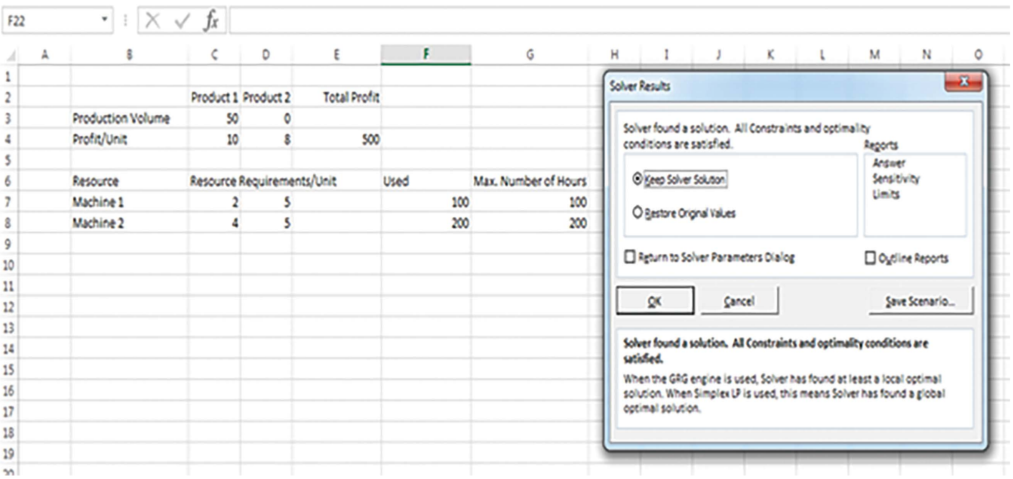
Select a Solving Method: Simplex LP

Solving Method
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 17.9 Screenshot after entering all the cells in the dropdown menu of Figure 17.6

6. Click the Simplex LP option in the dropdown menu against Select Solving Method in Figure 17.9.
7. Click the Solve button at the bottom of Figure 17.9, which gives a display as shown in Figure 17.10.

- Click Answer under Report in Figure 17.10 and click the OK button, which gives the results of the given linear programming model as shown in Figure 17.11.



F22

	B	C	D	E	F	G
2		Product 1	Product 2	Total Profit		
3	Production Volume	50	0			
4	Profit/Unit	10	8	500		
6	Resource	Resource Requirements/Unit			Used	Max. Number of Hours
7	Machine 1	2	5		100	100
8	Machine 2	4	5		200	200

Solver Results

Solver found a solution. All Constraints and optimality conditions are satisfied.

● Keep Solver Solution

○ Restore Orignal Values

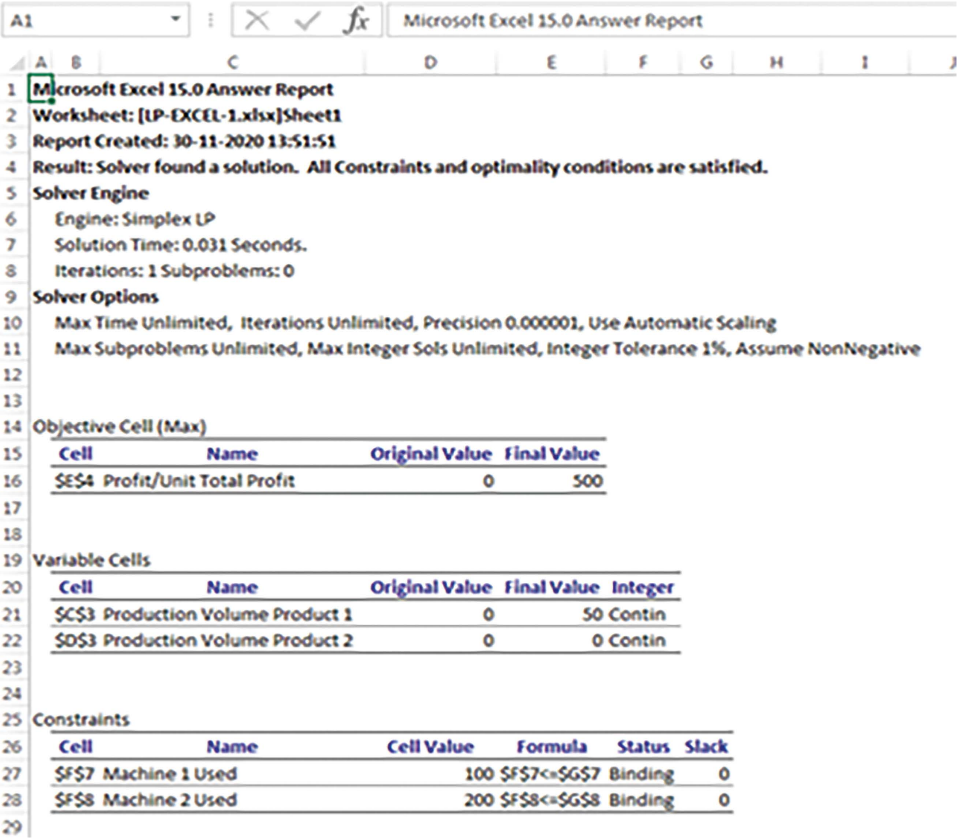
□ Return to Solver Parameters Dialog

□ Outline Reports

Solver found a solution. All Constraints and optimality conditions are satisfied.

When the GRG engine is used, Solver has found at least a local optimal solution. When Simplex LP is used, this means Solver has found a global optimal solution.

Figure 17.10 Screenshot after clicking Solve button in the dropdown menu of Figure 17.9



A1 Microsoft Excel 15.0 Answer Report

Microsoft Excel 15.0 Answer Report
Worksheet: [LP-EXCEL-1.xlsx]Sheet1
Report Created: 30-11-2020 13:51:51
Result: Solver found a solution. All Constraints and optimality conditions are satisfied.
Solver Engine
Engine: Simplex LP
Solution Time: 0.031 Seconds.
Iterations: 1 Subproblems: 0
Solver Options
Max Time Unlimited, Iterations Unlimited, Precision 0.000001, Use Automatic Scaling
Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 1%, Assume NonNegative

Objective Cell (Max)

Cell	Name	Original Value	Final Value
$E$4	Profit/Unit Total Profit	0	500

Variable Cells

Cell	Name	Original Value	Final Value	Integer
$C$3	Production Volume Product 1	0	50	Contin
$D$3	Production Volume Product 2	0	0	Contin

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$F$7	Machine 1 Used	100	$F$7<=$G$7	Binding	0
$F$8	Machine 2 Used	200	$F$8<=$G$8	Binding	0

Figure 17.11 Screenshot of results of LP model after clicking the Answer option under Report and clicking the OK button in Figure 17.10

From Figure 17.11, the objective function value and the values of the decision variables, that is, the production volume of product 1 and that of product 2 of the linear programming model, are as follows.

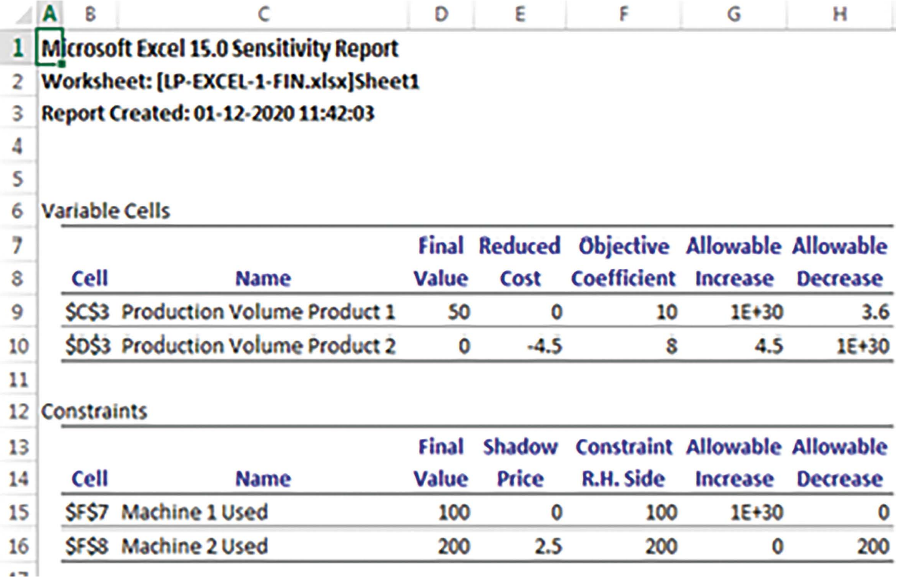
Optimised objective function value $\left(Z_{Opt}\right) = ₹\,500$
Production volume of product 1 $\left(X_1\right) = 50\,\textit{units}$
Production volume of product 2 $\left(X_2\right) = 0\,\textit{unit}$

This combination of the production volumes of product 1 and product 2 fully utilises the maximum machine hours under machine 1 and machine 2.

A screenshot of the sensitivity analysis is shown in Figure 17.12, and a screenshot of the limits report is shown in Figure 17.13.

Example 17.7

A marketing executive is in the process of media planning. The executive considered two media, television and newspaper. The cost of broadcasting an advertisement one time on television is ₹ 4500, and the cost of an advertisement in the newspaper per insertion is ₹ 1495. The reach indices of the advertisement on television in rural and urban areas are 3 and 10, respectively. The reach indices of an advertisement in the newspaper in rural and urban areas are 1 and 2, respectively. The executive wants to have minimum total reach indices of 60 and 160 in rural and urban areas, respectively. The management has given a combined budget constraint of ₹ 1,00,000 per week for advertisements. Develop a linear programming model to find the frequency of advertisements per week for television and newspaper such that the minimum expected



Microsoft Excel 15.0 Sensitivity Report
Worksheet: [LP-EXCEL-1-FIN.xlsx]Sheet1
Report Created: 01-12-2020 11:42:03

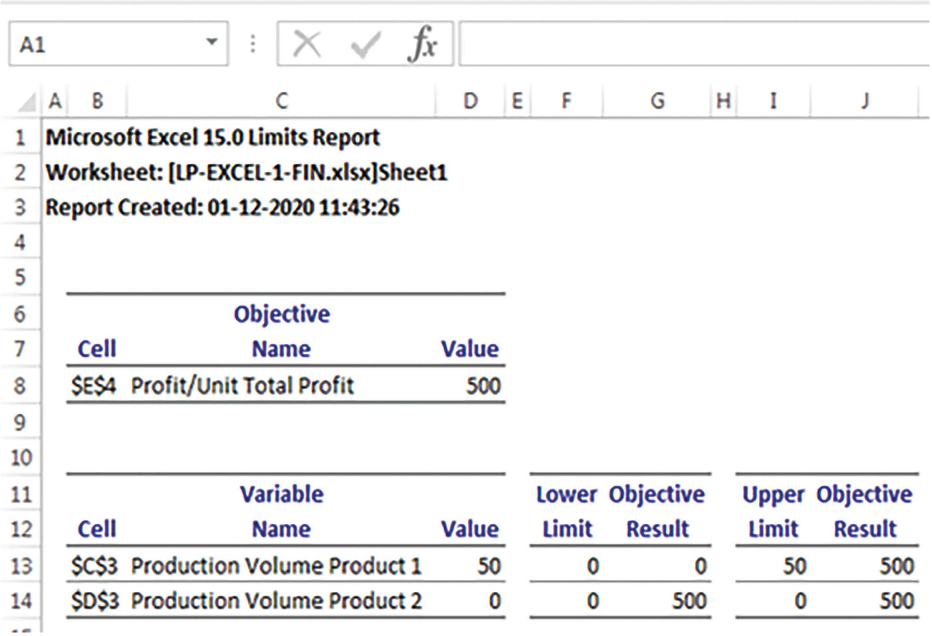
Variable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
\$C\$3	Production Volume Product 1	50	0	10	1E+30	3.6
\$D\$3	Production Volume Product 2	0	-4.5	8	4.5	1E+30

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
\$F\$7	Machine 1 Used	100	0	100	1E+30	0
\$F\$8	Machine 2 Used	200	2.5	200	0	200

Figure 17.12 Screenshot of sensitivity analysis of Example 17.6



Microsoft Excel 15.0 Limits Report
Worksheet: [LP-EXCEL-1-FIN.xlsx]Sheet1
Report Created: 01-12-2020 11:43:26

	Objective	
Cell	Name	Value
$E$4	Profit/Unit Total Profit	500

	Variable		Lower	Objective	Upper	Objective
Cell	Name	Value	Limit	Result	Limit	Result
$C$3	Production Volume Product 1	50	0	0	50	500
$D$3	Production Volume Product 2	0	0	500	0	500

Figure 17.13 Screenshot of limits report of Example 17.6

total reach indices in rural and urban areas are satisfied subject to the given budget constraint.

Solution

Cost of broadcasting advertisement one time on television = ₹ 4500
Cost of advertisement in newspaper per insertion = ₹ 1495
Reach index of advertisement through television in rural area = 3
Reach index of advertisement through television in urban area = 10
Reach index of advertisement through newspaper in rural area = 1
Reach index of advertisement through newspaper in urban area = 2
Minimum required total reach index in rural area = 60
Minimum required total reach index in urban area = 160
Combined budget constraint per week for advertisements = ₹ 1,00,000
X_1 be the frequency of advertisement in television
X_2 be the frequency of advertisement in newspaper

A linear programming model to find the frequency of advertisements per week for television and newspaper such that the minimum expected total reach indices in rural and urban areas are satisfied subject to the given budget constraint is presented as follows.

$$Minimise\, Z = 4500X_1 + 1495X_2$$

Subject to

$$3X_1 + X_2 \geq 60$$

$$10X_1 + 2X_2 \geq 160$$

$$4500X_1 + 1495X_2 \leq 100000$$

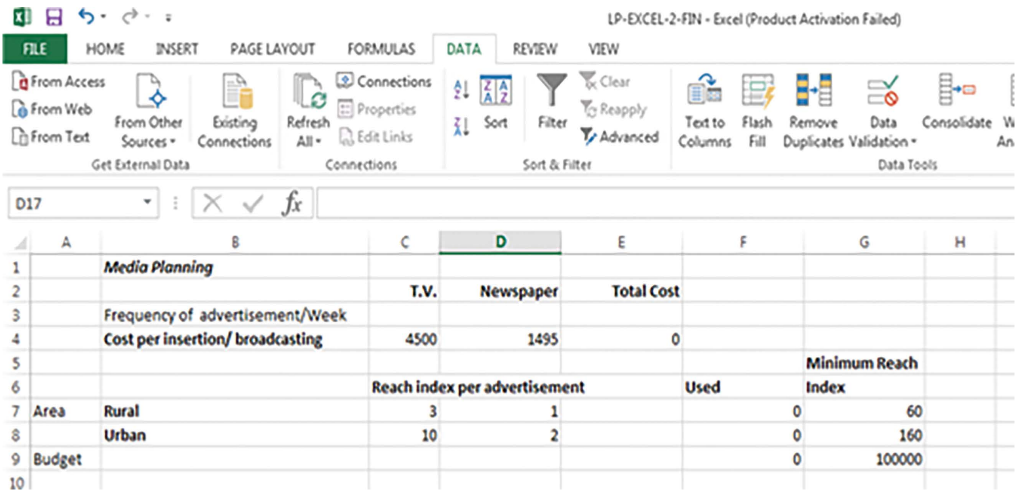
where $X_1 \geq 0$ *and* $X_1 \geq 0$

In this model, the objective function minimises the total cost per week of advertisements on television and in the newspaper. The first constraint ensures that the total reach index in the rural area is at least equal to 60. The second constraint ensures that the total reach index in the urban area is at least equal to 160. The third constraint ensures that the total advertising expenditure does not exceed ₹ 1,00,000.

The data for this problem are entered in an Excel sheet as shown in Figure 17.14.

The definitions of the cells in Figure 17.14 are presented in Table 17.11. One should note the = symbol used in each item in the last column of Table 17.11, which signifies the formula used in the respective cell. Even data like cost of advertisement on television one time in cell C4 is given in formula form, =4500. The same is applicable for all other data in addition to the usage of formulas to compute the total cost of the advertisements in cell E4, the formula for the total reach index in rural area in cell F7, the formula for the total reach index in the urban area in cell F8, the formula for the total cost of advertisements in cell F9, the minimum reach index in the rural area in cell G7, the minimum reach index in the urban area in cell G8, and the combined maximum budget for the advertisements in cell G9.

Clicking the Data button in Figure 17.14, entering all the required data in the boxes, and selecting suitable options of linear programming on the response screen gives a display as shown in Figure 17.15. One should note the fact that the Min option is selected in Figure 17.15 against To, because the objective function of this model is a minimisation type. Clicking the Solve button in Figure 17.15, clicking the Answer option in the response screen, and then clicking the OK button gives the results of the model as shown in Figure 17.16.

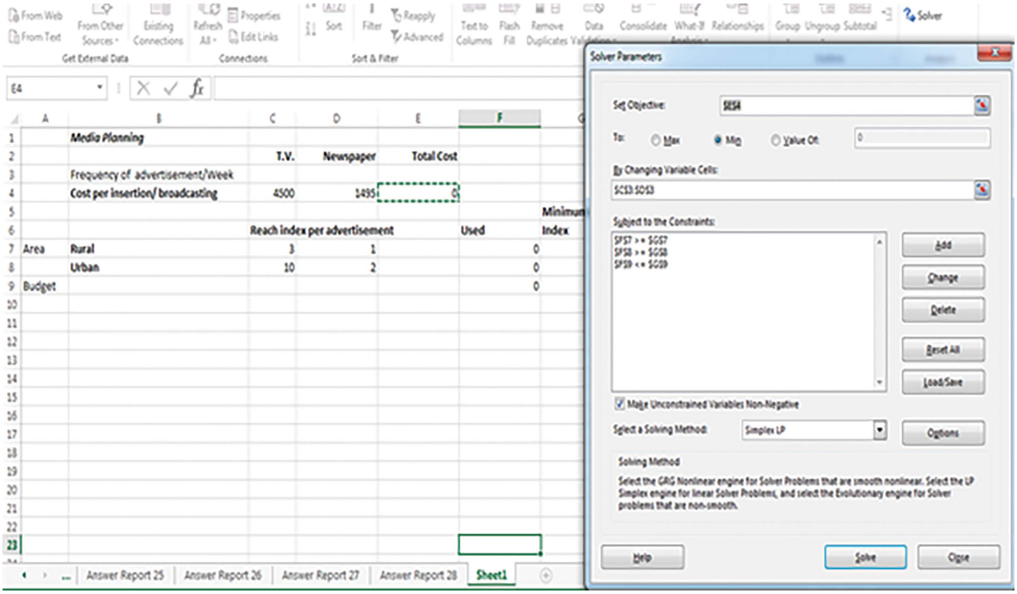


	A	B	C	D	E	F	G	H
1		***Media Planning***						
2			**T.V.**	**Newspaper**	**Total Cost**			
3		Frequency of advertisement/Week						
4		**Cost per insertion/ broadcasting**	4500	1495	0			
5							**Minimum Reach**	
6			**Reach index per advertisement**			**Used**	**Index**	
7	Area	**Rural**	3	1		0	60	
8		**Urban**	10	2		0	160	
9	Budget					0	100000	
10								

Figure 17.14 Screenshot of data for Example 17.7 in linear programming format in Excel sheet

Table 17.11 Cell Definitions of Figure 17.14

S. No.	*Cell Address*	*Definition*	*Remark*
1-2	C3:D3	Frequency of advertisement in Television and Newspaper, respectively.	Initially, these cells will have any value
3-4	C4:D4	Cost per broadcasting/insertion in Television and Newspaper, respectively.	C4: =4500 D4: =1495
5	E4	Formula of objective function (Total cost)	=C4*C3+D4*D3
6-7	C7:C8	Reach indices of the television and newspaper in rural area, respectively.	C7: =3 C8: =1
8-9	D7:D8	Reach indices of the television and newspaper in urban area, respectively.	D7: =10 D8: =2
10-11	F7:F8	Total reach indices of rural and urban areas, respectively, achieved.	F7: =C7*C3+D7*D3 F8: =C8*C3+D8*D3
12	F9	Total cost spent on the advertisements.	F9: =E4
13-14	G7:G8	Minimum reach indices to be realized in rural and urban areas, respectively.	G7: =60 G8: =160
15	G9	Combined maximum budget available for advertisements.	G9: =100000



	A	B	C	D	E	F	G
1		*Media Planning*					
2			T.V.	Newspaper	Total Cost		
3		Frequency of advertisement/Week					
4		Cost per insertion/ broadcasting	4500	1495	0		
5							Minimum
6			Reach index per advertisement			Used	Index
7	Area	Rural	3	1		0	
8		Urban	10	2		0	
9	Budget					0	

Solver Parameters

Set Objective: $E$4

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells: $C$3:$D$3

Subject to the Constraints:

$F$7 >= $G$7
$F$8 >= $G$8
$F$9 <= $G$9

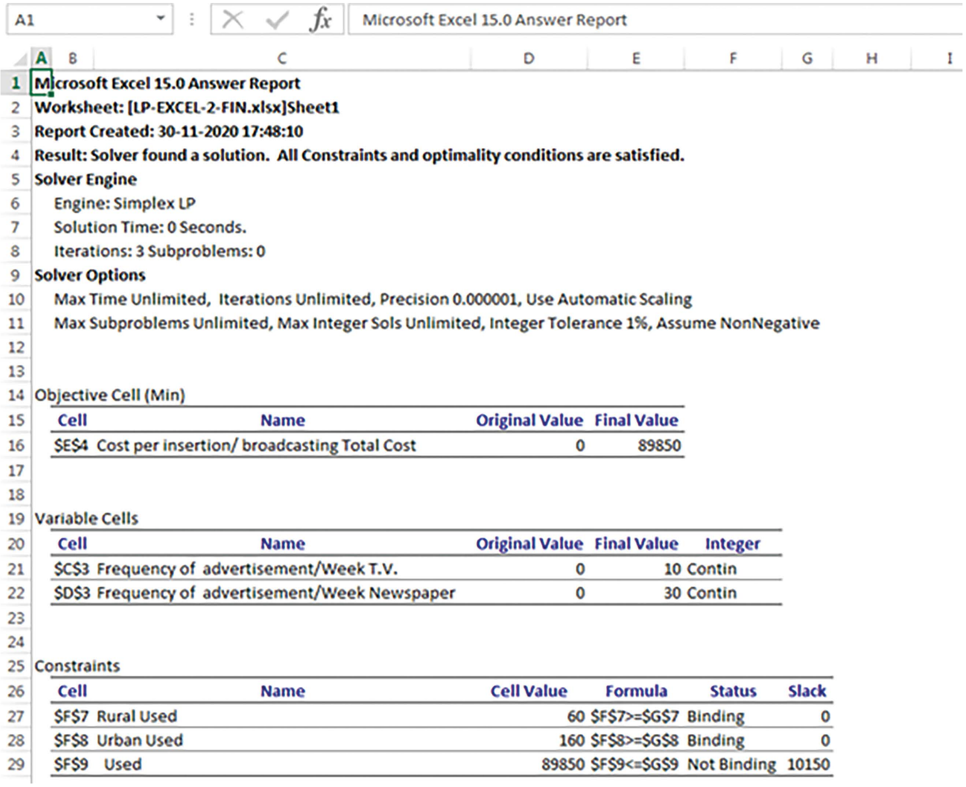
☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Simplex LP

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 17.15 Screenshot after clicking Data button, filling data and options in the response screen



Microsoft Excel 15.0 Answer Report
Worksheet: [LP-EXCEL-2-FIN.xlsx]Sheet1
Report Created: 30-11-2020 17:48:10
Result: Solver found a solution. All Constraints and optimality conditions are satisfied.
Solver Engine
Engine: Simplex LP
Solution Time: 0 Seconds.
Iterations: 3 Subproblems: 0
Solver Options
Max Time Unlimited, Iterations Unlimited, Precision 0.000001, Use Automatic Scaling
Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 1%, Assume NonNegative

Objective Cell (Min)

Cell	Name	Original Value	Final Value
$E$4	Cost per insertion/ broadcasting Total Cost	0	89850

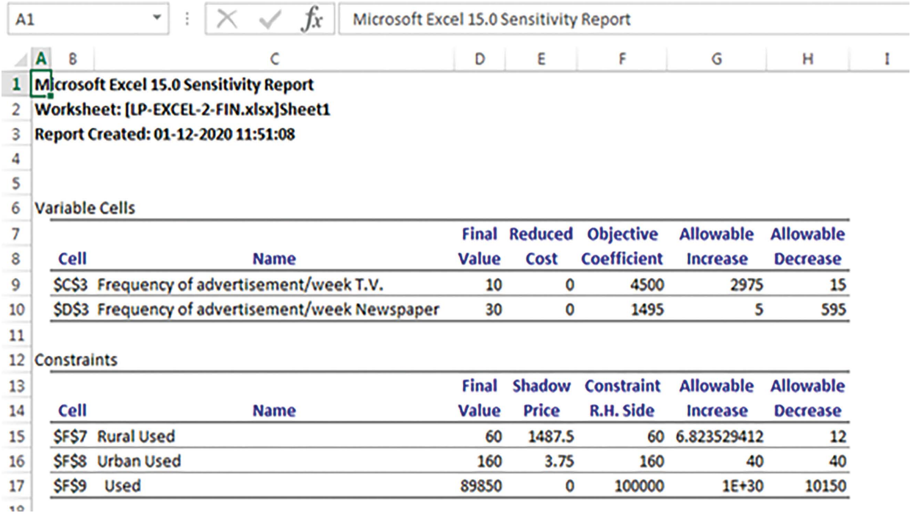
Variable Cells

Cell	Name	Original Value	Final Value	Integer
$C$3	Frequency of advertisement/Week T.V.	0	10	Contin
$D$3	Frequency of advertisement/Week Newspaper	0	30	Contin

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$F$7	Rural Used	60	$F$7>=$G$7	Binding	0
$F$8	Urban Used	160	$F$8>=$G$8	Binding	0
$F$9	Used	89850	$F$9<=$G$9	Not Binding	10150

Figure 17.16 Screenshot after clicking Solve button in Figure 17.15, clicking the Answer option in the response screen, then clicking the OK button



Microsoft Excel 15.0 Sensitivity Report
Worksheet: [LP-EXCEL-2-FIN.xlsx]Sheet1
Report Created: 01-12-2020 11:51:08

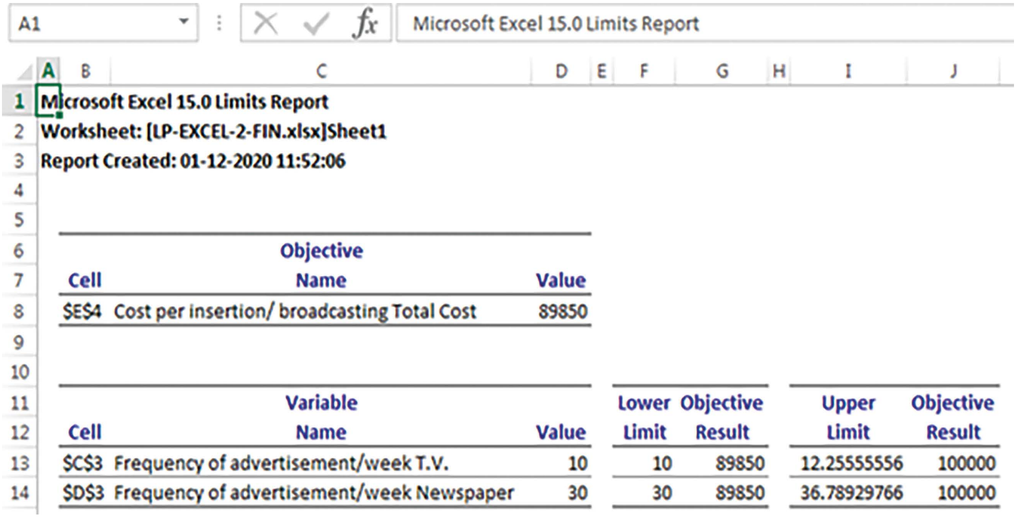
Variable Cells

		Final	Reduced	Objective	Allowable	Allowable
Cell	Name	Value	Cost	Coefficient	Increase	Decrease
$C$3	Frequency of advertisement/week T.V.	10	0	4500	2975	15
$D$3	Frequency of advertisement/week Newspaper	30	0	1495	5	595

Constraints

		Final	Shadow	Constraint	Allowable	Allowable
Cell	Name	Value	Price	R.H. Side	Increase	Decrease
$F$7	Rural Used	60	1487.5	60	6.823529412	12
$F$8	Urban Used	160	3.75	160	40	40
$F$9	Used	89850	0	100000	1E+30	10150

Figure 17.17 Screenshot of sensitivity analysis of Example 17.7



Microsoft Excel 15.0 Limits Report
Worksheet: [LP-EXCEL-2-FIN.xlsx]Sheet1
Report Created: 01-12-2020 11:52:06

	Objective	
Cell	Name	Value
$E$4	Cost per insertion/ broadcasting Total Cost	89850

	Variable		Lower	Objective	Upper	Objective
Cell	Name	Value	Limit	Result	Limit	Result
$C$3	Frequency of advertisement/week T.V.	10	10	89850	12.25555556	100000
$D$3	Frequency of advertisement/week Newspaper	30	30	89850	36.78929766	100000

Figure 17.18 Screenshot of limits report of Example 17.7

17.7.1 Summary of Results

From Figure 17.16, the objective function value and the values of the decision variables, that is, the weekly frequency of an advertisement on television and in the newspaper of the linear programming model, are as follows.

Optimised objective function value $\left(Z_{Opt}\right) = ₹\ 89{,}850$ (Total cost of advertisements)
Weekly frequency of advertisement on television $\left(X_1\right) = 10$
Weekly frequency of advertisement in newspaper $\left(X_2\right) = 30$

This combination of the frequencies of advertisements exactly fulfils the minimum required reach indices in rural and urban areas.

A screenshot of the sensitivity analysis is shown in Figure 17.17, and another screenshot for the limit report is shown in Figure 17.18.

Summary

- Linear programming is a mathematical programming technique which optimises a measure of performance of a system of interest under a given set of constraints imposed by the management.
- The linearity assumption deals with the linear relationship between the quantity of a resource used and the level of an activity.
- The divisibility assumption of a linear programming model relates to the nature of values permitted for the decision variables.
- The non-negativity assumption of a linear programming model ensures that negative values of the decision variables are not permitted in the model.
- The additivity assumption of a linear programming model signifies the aggregation of the output levels of the individual activities to arrive the total output for a given combination of activity levels.

- The decision variables of a linear programming model are the level of activities of the decision environment.
- If the number of decision variables in the linear programming model is only two, one can use graphical method to solve that model.
- The Solver function in Excel is used to solve linear programming models, which come under the decision-making environment of operations research.

Keywords

- Linear programming is a mathematical programming technique which optimises a measure of performance of a system of interest under a given set of constraints imposed by the management.
- The simplex method was developed by George Dantzig in 1947 to solve linear programming problems and is an iterative procedure that begins with an initial feasible solution and continues until an optimality condition is reached.
- A slack variable is a variable in a less than or equal to constraint, which equates the left-hand side of that constraint to the right-hand side of that constraint.
- A surplus variable is a variable in a greater than or equal to constraint, which equates the left-hand side of that constraint to the right-hand side of that constraint.
- An artificial variable is introduced in a ≥ or = constraint just to serve as the basic variable and should not be present in the final solution.
- A solver function is used to solve linear programming models, which come under the decision-making environment of operations research.

Review Questions

1. List and explain the assumptions of linear programming.
2. What are the components of linear programming? Explain them using an example.
3. Solve the following linear programming model using the graphical method.

   Maximise $Z = 40X_1 + 25X_2$
   Subject to
   $2X_1 + 3X_2 \leq 45$

   $3X_1 + 2X_2 \leq 35$

   $5X_1 + 3X_2 \leq 65$

   X_1 and $X_2 \geq 0$

4. Solve the following linear programming problem using the graphical method.

   Maximise $Z = 50X_1 + 20X_2$
   Subject to
   $4X_1 + 2X_2 \leq 55$

   $2X_1 + 6X_2 \leq 60$

   $4X_1 + 6X_2 \leq 45$

   X_1 and $X_2 \geq 0$

5. Solve the following LP problem using the simplex method.

   $Maximise\,Z = 2X_1 + 5X_2 - 2X_2$

   Subject to

   $X_1 - 3X_2 + X_3 \leq 45$

   $3X_1 + 2X_2 + 4X_3 \leq 35$

   X_1, X_2 and $X_3 \geq 0$

6. Solve the following LP problem using the simplex method in Excel.

   $Maximise\,Z = 5X_1 + 5X_2 + 2X_2$

   Subject to

   $2X_1 + 4X_2 + X_3 \leq 35$

   $X_1 + 2X_2 + 4X_3 \leq 20$

   $X_1, X_2\ and\ X_3 \geq 0$

7. Solve the following LP problem using the simplex method in Excel.

   $Maximise\ \mathrm{Z} = 12X_1 + 15X_2 - 1X_2$

   Subject to

   $2X_1 - X_2 + 3X_3 \leq 45$

   $3X_1 + X_2 + 2X_3 \leq 35$

   $X_1, X_2\ and\ X_3 \geq 0$

8. Solve the following LP problem using the simplex method in Excel.

   $Minimise\ \mathrm{Z} = 15X_1 + 12X_2 + 10X_2$

   Subject to

   $4X_1 + 2X_2 + 3X_3 \geq 30$

   $2X_1 + X_2 + 4X_3 \geq 25$

   X_1, X_2 and $X_3 \geq 0$

9. Two plants make up a multi-plant organisation, and the finished commodities are transported from those plants to three different markets. Plants P1 and P2 have weekly supply volumes of 1200 and 1500 units, respectively. Markets M1, M2, and M3 have weekly demands of 900 units, 800 units, and 500 units, respectively. The costs of producing one unit of the product at plants P1 and P2 are 500 and 400 rupees, respectively. The following table provides an overview of the per-unit shipping costs from the plants to the markets.
   Summary of shipping cost (in rupees per unit)

		Market		
		M_1	M_2	M_3
Plant	P_1	50	40	45
	P_2	50	25	35

a. Develop a linear programming model for this problem in order to estimate the shipping amounts from the plants to the markets in a way that minimises the overall cost while taking account of supply and demand limitations.

b. Use Excel to solve the model.

References

1. Panneerselvam, R., *Operations Research* (2nd edition), PHI Learning Private Limited, New Delhi, 2006.
2. https://New%20folder%20back-4-12-2020/Solve%20problems%20with%20linear%20programming%20and%20Excel%20-%20FM.html [November 29, 2020].

Annexures

Annexure 1: F Table for $\alpha = 0.10$

Denominator	*Numerator Degrees of Freedom (n_1)*																		
Degrees of Freedom(n2)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	*α*
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17		3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52

(*Continued*)

Annexure 1 (Continued)

Denominator	*Numerator Degrees of Freedom (n_1)*																		
Degrees of Freedom(n2)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	*α*
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.03	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
α	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Annexure 2: F Table for $\alpha = 0.05$

Denominator	*Numerator Degrees of Freedom (n_1)*																	
Degrees of Freedom(n_2)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	*α*
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
α	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Annexure 3: F Table for $\alpha = 0.01$

Denominator	*Numerator Degrees of Freedom (n_1)*																		
Degrees of Freedom(n_2)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	∞
1	4052.0	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	5.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.36	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Annexure 4: Area Under Standard Normal Distribution From Its Mean

Z	*00*	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*.09*
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Annexure 5: Values of Student's *t* Statistic (tα)

Degrees of Freedom	*α*								
	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.320	318.310
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893
6	0.265	0.727	1.440	1.943	2.447	3.143	3.707	4.317	5.208
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.019	4.785
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144

(*Continued*)

Annexure 5 (Continued)

Degrees of Freedom	*α*								
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090

Annexure 6: Value of Chi-Square Statistic for a Given Significance Level and Degrees of Freedom

Degrees of Freedom	*α*							
	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.0000393	0.000157	0.000982	0.00393	3.841	5.024	6.635	7.879
2	0.0100	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.0717	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.056	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819

Degrees of Freedom	*α*							
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.484	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.772	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.706	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.535	37.485	40.482	43.118	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.292	116.321
90	59.196	61.754	65.646	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

Annexure 7: Critical Values of *d* for Kolmogorov-Smirnov One-Sample Test

Sample Size (n)	*Level of Significance (α)*		
	0.10	*0.05*	*0.01*
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.642	0.708	0.828
4	0.564	0.624	0.733
5	0.510	0.565	0.669
6	0.470	0.521	0.618
7	0.438	0.486	0.577
8	0.411	0.457	0.543
9	0.388	0.432	0.514
10	0.368	0.410	0.490
11	0.352	0.391	0.468
12	0.338	0.375	0.450
13	0.325	0.361	0.433

(*Continued*)

Annexure 7 (Continued)

14	0.314	0.349	0.418
15	0.304	0.338	0.404
16	0.295	0.328	0.392
17	0.286	0.318	0.381
18	0.278	0.309	0.371
19	0.272	0.301	0.363
20	0.264	0.294	0.356
25	0.240	0.270	0.320
30	0.220	0.240	0.290
35	0.210	0.230	0.270
Above 35	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

Annexure 8: Critical Values of *r* of Run Test

Any observed value of *r* which is less than equal to the smaller value or greater than or equal to the large value for a given n_1 and n_2 is significant at $\alpha = 0.05$.

n_2																			
n_1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
2											2	2	2	2	2	2	2	2	2
											-	-	-	-	-	-	-	-	-
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
				9	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
			9	10	10	11	11	-	-	-	-	-	-	-	-	-	-	-	-
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
		-	9	10	11	12	12	13	13	13	13	-	-	-	-	-	-	-	-
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
		-	-	11	12	13	13	14	14	14	14	15	15	15	-	-	-	-	-
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
		-	-	11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
		-	-	-	13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
		-	-	-	13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
		-	-	-	13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
	-	-	-	-	13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
	-	-	-	-	-	15	16	17	18	19	19	20	20	21	21	22	22	23	23

Annexure 8: Contd.

n_2																			
n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
	-	-	-	-	-	15	16	17	18	19	20	20	21	22	22	23	23	23	24
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	
	-	-	-	-	-	15	16	18	18	19	20	21	22	22	23	23	24	24	25
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
	-	-	-	-	-	-	17	18	19	20	21	21	22	23	23	24	25	25	25
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
	-	-	-	-	-	-	17	18	19	20	21	22	23	23	24	25	25	26	26
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
	-	-	-	-	-	-	17	18	19	20	21	22	23	24	25	25	26	26	27
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
	-	-	-	-	-	-	17	18	20	21	22	23	23	24	25	26	26	27	27
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14
	-	-	-	-	-	-	17	18	20	21	22	23	24	25	25	26	27	27	28

Index