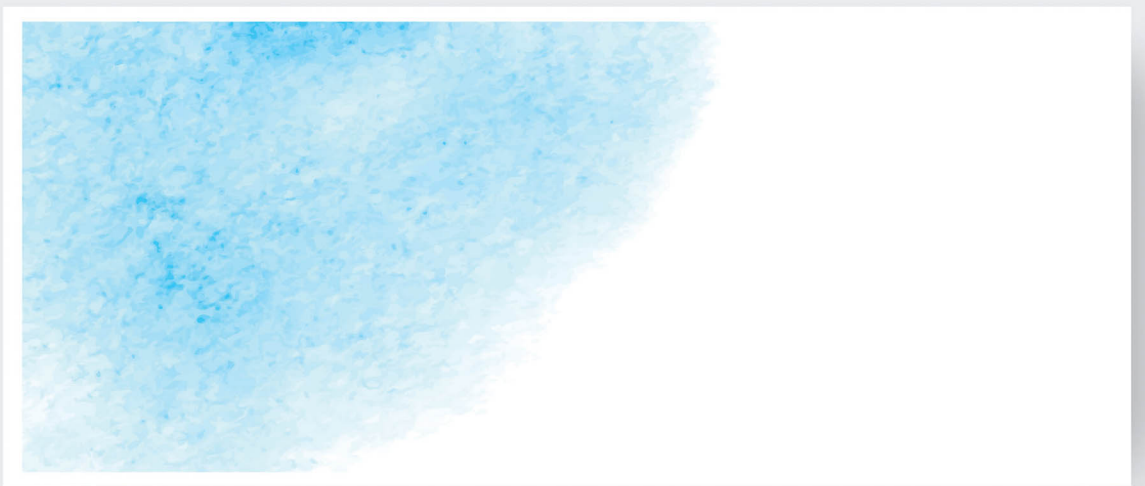
Quantitative Techniques in Business, Management and Finance

A Case-Study Approach



Umeshkumar Dubey
D P Kothari
G K Awari

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Quantitative Techniques in Business, Management and Finance

A Case-Study Approach

Quantitative Techniques in Business, Management and Finance

A Case-Study Approach

Umeshkumar Dubey

Tulsiramji Gaikwad-Patil College of Engineering & Technology
Nagpur, Maharashtra State, India

D P Kothari

S B Jain Institute of Technology, Management & Research
Nagpur, Maharashtra State, India

G K Awari

Tulsiramji Gaikwad-Patil College of Engineering & Technology
Nagpur, Maharashtra State, India

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20160808

International Standard Book Number-13: 978-1-4987-6946-4 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: Dubey, Umeshkumar, author. | Kothari, D. P. (Dwarkadas Pralhaddas), 1944- author. | Awari, G. K., author.
Title: Quantitative techniques in business, management and finance : a case-study approach / Umeshkumar Dubey, D P Kothari, and G K Awari.
Description: Boca Raton : Chapman & Hall/CRC, 2016. | Includes bibliographical references and index.
Identifiers: LCCN 2016022252 | ISBN 9781498769464 (alk. paper)
Subjects: LCSH: Business enterprises--Finance. | Decision making--Methodology. | Quantitative research.
Classification: LCC HG4026 .D83 2016 | DDC 650.072/1--dc23
LC record available at https://lccn.loc.gov/2016022252

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

The application of quantitative techniques has become increasingly important because management principles are applied in organisations like hospitals, educational systems and non-enterprise management administration.

The motivation for writing this book has come from the lack of a contextually relevant and comprehensive book on quantitative techniques which are sufficient to provide a rather substantial course of study for some semesters. The book is intended for students of management, accountancy, chartered and cost accountancy and economics. Most chapters have been developed from the lecture notes used in teaching and tested in classes over several years.

We have kept the presentation simple and stimulating; nevertheless, the treatment of topics is detailed and up-to-date so that experts (researchers, etc.) in the field can use the book as a text or reference.

The book includes chapters with examples, figures, cases and MCQs; a glossary; and a bibliography. Every chapter has a set of problems which include conceptual, descriptive and design problems.

Furthermore, we have embedded relevant case studies to illustrate the fundamental concepts.

To enhance understanding of the subject matter by students belonging to different disciplines, our approach is conceptual rather than mathematical. Each chapter contains a sufficiently large number of review questions, and some chapters contain self-practice problems with answers to help readers in self-evaluation. Explanations are richly illustrated with numerous interesting and varied business-oriented examples. A large number of business-oriented problems, both solved and practice, have been added, thus creating a bank of problems.

To provide an opportunity for students to gain more skills to apply these concepts, a variety of problems have been included in each chapter. To facilitate ready recall of the concepts and formulae discussed in each chapter, a brief summary is provided at the end of each chapter.

Acknowledgements

In preparing the text of this book, we have benefitted immensely by referring to many books and publications. We express our gratitude to such authors, publications and publishers; many of them have been listed in the bibliography. If anybody is left out inadvertently, we seek their pardon.

The authors are highly thankful to Dr Siddharthavinayaka P Kane, honourable vice chancellor of the Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, India, for his constant encouragement and help while developing this text.

We express our deep sense of gratitude towards Dr D G Wakde, Dr L B Bhuyar, Dr P G Khot and Dr P L Neulkar, who have supported us and have been a source of encouragement.

Finally, we thank CRC Press, Taylor & Francis Group, especially Ms Astha Sharma and Mr Alexander Edwards for facilitating the work.

We are indeed grateful to Mrs Pinky Dubey for her timely help in drawing all the diagrams and formatting the text, without which this book would not have seen the light of the day.

We hope that the book will serve the purpose of its readers and that we will continue to get their support and suggestions. Suggestions and comments to improve the book in content and style are always welcome and will be appreciated, acknowledged and incorporated in future editions of the book.

Dr Umeshkumar Dubey
Dr D P Kothari
Dr G K Awari
Nagpur, India

Authors

Umeshkumar Dubey is presently head of the Vidarbha Bahuuddeshiya Shikshan Sanstha, MBA Department at the Tulsiramji Gaikwad-Patil College of Engineering and Technology, Nagpur, India. Prior to this, he served for 12 years as an academic at the Rashtrasant Tukdoji Maharaj University, Nagpur, and ICFAI University, Dehradun. Dr. Dubey is a Life Member of the Indian Society for Technical Education by approval of the executive council for promoting quality and standards in technical education. He has been nominated as a member of the Research and Development Board and the Internal Quality Assurance Cell in the organisation. He has also held the position of controller of examinations in the organisation. Dr. Dubey is a PhD supervisor in statistics in the Faculty of Science and in business management and business administration in the Faculty of Commerce at Rashtrasant Tukdoji Maharaj University, Nagpur. He has convened international conferences on 'Emerging Trends in Business Management' and 'Sustaining and Enhancing Competitiveness in Today's Business Scenario', jointly organised by Rashtrasant Tukadoji Maharaj Nagpur University's Department of Business Management and the MBA Department of Tulsiramji Gaikwad at the Patil College of Engineering and Technology, Nagpur. Dr. Dubey earned a PhD from Rashtrasant Tukdoji Maharaj University, Nagpur, in statistics, specialising in operations research; an MBA from IGNOU, specialising in operations management; a second MBA from RTM Nagpur University, Nagpur, specialising in marketing and human resource management; an Advanced Diploma in Computer Programming and System Management; and a Computer Programming Course in C and C++. He is currently pursuing a second PhD in human resource management from RTM Nagpur University, Nagpur. He has taught statistics, quantitative techniques, business research and operation management for 17 years at undergraduate and postgraduate levels. He was the recipient of an "Excellent Personality Award" from the minister of state for higher and technical education, government of Maharashtra Mantralaya, Mumbai. He has been appointed as an editor of the online statistics journal *InterStat* (ISSN 1941-689X).

D P Kothari is presently director of research at the S B Jain Institute of Technology, Management & Research, Nagpur. He earned his BE (Electrical) in 1967, ME (Power Systems) in 1969 and PhD in 1975 from the Birla Institute of Technology and Sciences, Pilani, Rajasthan. From 1969 to 1977, he was involved in teaching and development of several courses at the Birla Institute of Technology and Sciences, Pilani. Prior to the role of director of research at the Gaikwad Patil group of institutions, Dr Kothari served as vice chancellor at the Vellore Institute of Technology, director in-charge, deputy director (administration) and head of the Centre of Energy Studies at the Indian Institute of Technology, Delhi, and principalat the Visvesvaraya Regional College of Engineering, Nagpur. He was visiting professor at the Royal Melbourne Institute of Technology, Melbourne, Australia, during 1982–1983 and 1989, for two years. He was also NSF Fellow at Purdue University, Indiana, in 1992. Dr Kothari, who is a recipient of the Most Active

Researcher Award, has published and presented 780 research papers in various national as well as international journals and at conferences, guided 50 PhD scholars and 68 MTech students, and authored 45 books in various allied areas. He has delivered several keynote addresses and invited lectures at both national and international conferences. He has also delivered 42 video lectures on YouTube with over 45,000 hits. Dr. Kothari is a fellow of the National Academy of Engineering, Indian National Academy of Science, Institution of Engineers and Institute of Electrical and Electronics Engineers, and an honorary fellow of the International Society for Technology in Education.

His many awards include the National Khosla Award for Lifetime Achievements in Engineering (2005) from the Indian Institue of Technology, Roorkee. The University Grants Commission (government of India) bestowed on him their National Swami Pranavandana Saraswati Award (2005) in the field of education for his outstanding scholarly contributions. He is also the recipient of the 2009 Lifetime Achievement Award conferred by the World Management Congress, New Delhi, for his contribution to the areas of educational planning and administration. Recently he received an Excellent Academic Award at the Indian Institute of Technology, Guwahati, by NPSC-2014.

G K Awari earned a Bachelor of Engineering from Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, Maharashtra, India, in 1991 and an ME from Thapar University, Patiala, Punjab, India, in 1995, in mechanical engineering. He completed his PhD at Sant Gadgebaba Amravati University, Amravati, Maharashtra, India, in 2007. He has more than 22 years of teaching experience at undergraduate/postgraduate and doctorate levels. He has taught various subjects such as computer graphics, research methodology, automation engineering, operation research, fluid power and machine design. His area of interest is graphical modelling of computational fluid dynamics. He has more than 51 international journal publications, 25 international conference publications and 15 national conference publications to his name. He is a reviewer for various renowned international journals. Currently, 6 research scholars are pursuing doctoral research work under his supervision, and 12 candidates have been honoured with a PhD in mechanical engineering under his supervision. He is also a recipient of the Best Paper Award and Golden Educationist Award. He is a member of the Board of Studies in Mechanical Engineering at Goa University, Goa, S.G. Amravati University, Amravati, and Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur. He is currently the principal of Tulsiramji Gaikwad-Patil College of Engineering, Nagpur. He has authored four books in various disciplines.

1

Quantitative Decision Making – An Overview

1.1 Introduction

A person managing a production unit, farm, factory or domestic kitchen has to coordinate people, machines and money against several constraints, like those of time, cost and space, in order to achieve the organisation's objectives in an efficient and effective manner. The manager has to analyse the situation on a continuous basis; determine the objectives; identify the best option from the set of available alternatives; and implement, coordinate, evaluate and control the situation continuously to achieve these objectives.

Organisations of today have become increasingly complex, and hence managerial decision making has become even more complex. As a result, management is becoming more of a science than an art.

As the complexity of organisations and the business environment has made the process of decision making difficult, managers cannot take decisions on the basis of subjective factors like their experience, observation or evaluation anymore. Decisions need to be based on thorough analysis of data that reveals relationships, indicates trends and shows rates of change in the relevant variables.

Quantitative methods provide ways to collect, present, analyse and interpret the available data meaningfully. They are the powerful tools through which managers can accomplish their predetermined objectives, like profit maximisation, cost minimisation or efficient and effective use of production capacities. The study of quantitative methods has a wide range of applications, especially in business.

The information needed by the decision maker or owner to make effective decisions was much less extensive. Thus, he made decisions based on his past experience and intuition only. The reasons are

1. The marketing of the product was not a problem because customers were personally known to the business owner.
2. Test marketing of the product was not essential because the owner used to know the choice and need of the customers just by interaction.
3. The owner used to work with his workers at the shop floor. He knew all of them personally.
4. Progress on the work was being made daily at the work centre. Thus, production records were not required.

Quantitative methods are used in decision making for the following reasons:

1. Complexity of today's managerial activities, which involve constant analysis in setting objectives, seeking alternatives and implementing, coordinating, controlling and evaluating the result
2. Availability of different tools for quantitative analysis of complex problems

1.2 Meaning of Quantitative Techniques

Quantitative methods are those statistical and programming techniques which help decision makers solve many problems, especially those concerning business and industry. These methods involve the use of numbers, symbols, mathematical expressions and other elements of quantities and serve as supplements to the judgement and intuition of the decision makers.

In other words, quantitative methods are those techniques that provide the decision makers with systematic and powerful means of analysis, based on quantitative data, for achieving predetermined goals.

1.2.1 Concept of Statistics

In the view of laypeople, statistics is 'data'. Some statisticians called it the science of counting averages and estimations (Bowley, 1926), the science of numerical data (Ya-Lun-Chou 1972) and so forth. As a method, statistics may be defined as a process of the collection of data and analysis for drawing conclusions.

1.2.2 Concept of Operations Research

Operations research is a quantitative technique that is applied in the process of decision making, in the scientific analysis of problems, which involves systematic operations. The term *operations research* is coined by combining two different terms, *operation* and *research*. *Operation* means 'action applied', and *research* is finding 'unknown facts and information scientifically'. Thus, operations research can be defined as the application of certain tools for dealing with decision-making problems.

1.3 Evolution of Quantitative Techniques

The utility of quantitative methods was realised long ago, and the science of mathematics is probably as old as human society. However, with the evolution of industrial engineering, scientific methodologies that were prominent earlier in the natural sciences were found to be applicable to management functions – planning, organisation and controlling of operations.

In the late nineteenth century, Fredrick W Taylor proposed an application of the scientific method to an operations management problem, namely, the productivity of men shovelling ore.

Prior to this development, there was a strong belief that the largest shovel a worker could fill and carry was the shovel size which maximises the output. Taylor questioned the validity of the assumption and performed a series of experiments to disprove it. He determined that the only variable that was significant was the combined weight of the shovel and its load. Too much weight on the shovel would result in the worker getting tired soon and moving slowly, while too little load would result in too many trips. The experiments concluded that productivity could be raised substantially by loading the shovel properly.

Another significant contribution to the evolution of the quantitative methods was made by Henry L Gantt, who devised a chart, known as the Gantt chart, to schedule production activities. Prior to his work, production was a haphazard exercise. The jobs processed through one stage of production used to wait for days for acceptance in the next processing centre. The Gantt chart scheduled each job from one machine to another, and minimised the delays in between.

1.4 Classification of Quantitative Methods

The wide variety of quantitative methods that are available in modern times can be broadly classified into two groups:

1. Statistical techniques
2. Programming techniques

Statistical techniques are used in conducting the statistical inquiry concerning a certain phenomenon. They include various methods, ranging from the collection of data to the task of interpreting the collected data. The methods of collecting, classifying and tabulating statistical data; the calculation of various statistical measures like mean, standard deviation and coefficient of correlation; the methodologies of analysing and interpreting data; and finally, the task of deriving inferences and judging their reliability are examples of statistical techniques.

Programming techniques (also known as operations research techniques) are used by many decision makers in modern times. They were first designed to tackle defence and military problems and are now being used to solve business problems. They include a variety of techniques, like

1. Linear programming
2. Games theory
3. Simulation
4. Network analysis
5. Queuing theory

These techniques involve the building of mathematical models that relate the relevant variables in a situation to the outcomes, and yield solutions to problems in terms of the values of the variables involved. The following sections describe the steps involved in the application of the programming techniques.

1.5 Data Collection

1.5.1 Statistical Data

Statistical data refers to the numerical description of quantitative aspects of things in the form of counts or measurements.

1.5.2 Statistical Methods

The large volume of numerical information or data gives rise to the need for systematic methods which can be used to collect, organise or classify, present, analyse and interpret the information effectively for the purpose of making wise decisions.

1.5.3 Data Collection

The following data is required for the given purposes:

1. Population of the city
2. Number of individuals who are getting income
3. Daily income of each earning individual

1.5.4 Organisation of Data

The data so obtained should now be organised in different income groups. This will reduce the bulk of the data.

Once the mathematical model that represents the problem situation is constructed, the next step is to collect the appropriate data required by the model. Data can come from various sources, like well-maintained records, tests or experiments and even from the hunches based on past experiences. Collection of data is an important step in the decision-making process as it affects the output of the model significantly. Data collection takes significant time. For instance, a moderate-sized linear programming model with 30 decision variables and 20 constraints involves as many as 10,000 data elements that are to be identified.

1.6 Characteristics of Data

It is probably more common to refer to data in the quantitative form as statistical data. But not all numerical data is statistical. In order that numerical descriptions may be called statistics, they must possess the following characteristics:

1. They must be aggregate of facts, for example, single unconnected figures cannot be used to study the characteristics of the phenomenon.
2. They should be affected to a marked extent by a multiplicity of causes; for example, in social services the observations recorded are affected by a number of factors (controllable and uncontrollable).

3. They must be enumerated or estimated according to reasonable standard of accuracy; for example, in the measurement of height, one may measure correctly up to 0.01 cm. The quality of the product is estimated by certain tests on small samples drawn from a big lot of products.
4. They must have been collected in a systematic manner for a predetermined purpose. Facts collected in a haphazard manner and without a complete awareness of the object will be confusing and cannot be made the basis of valid conclusions. For example, collected data on price serves no purpose unless one knows whether he wants to collect data on wholesale or retail prices and what are the relevant commodities in view.
5. They must be placed in relation to each other. That is, data collected should be comparable; otherwise, these cannot be placed in relation to each other. For example, statistics on the yield of crop and use of fertiliser are related, but these yields cannot have any relation with the statistics on the health of the people.
6. They must be numerically expressed. That is, any facts to be called statistics must be numerically or quantitatively expressed. Quantitative characteristics such as beauty and intelligence cannot be included in statistics unless they are quantified.

1.7 Types of Statistical Data

An effective managerial decision concerning a problem on hand depends on the availability and reliability of statistical data. Statistical data can be broadly grouped into two categories:

1. Secondary or published data
2. Primary or unpublished data

The secondary data is that which has already been collected by another organisation and is available in the published form. You must first check whether any such data is available on the subject matter of interest and make use of it, since it will save considerable time and money. But the data must be scrutinised properly since it was originally collected perhaps for another purpose. The data must also be checked for reliability, relevance and accuracy.

A great deal of data is regularly collected and disseminated by international bodies, such as the World Bank, Asian Development Bank, International Labour Organization and Secretariat of the United Nations; government and its many agencies, such as the Reserve Bank of India and Census Commission; ministries, such as the Ministry of Economic Affairs and Commerce Ministry; private research organisations; and trade associations.

1.7.1 Arriving at the Solution

Once the data has been collected, managers can use the data as input and attempt to solve the model. The solution provides an answer to the problem under consideration, but only under the condition represented by the model. For example, suppose a cooler manufacturer

develops an inventory model to calculate the quantity of raw materials to be kept in stock (by considering the ordering costs, carrying costs and stock-out costs, which are described in operations management), keeping in view the demand for the product in the summer season. If the manufacturer sticks to that solution in winter also, it may end up with excessive stocks and cash shortages. Hence, managers need to review the solutions periodically to check whether they are appropriate. Managers should also try to alter inputs and the model to analyse the corresponding changes in the output. This process of altering the inputs and studying the changes in the output is called sensitivity analysis.

1.7.2 Presentation

The organised data may now be presented by means of various types of graphs or other visual aids. Data presented in an orderly manner facilitates statistical analysis.

The final step in the process is to present the solution to the top management. This step includes a full explanation of the findings and an effort to correlate them to the solution. The presentation should specify the conditions under which the solution can be used. It should also point out the weakness of the underlying assumptions so that the top management can know the risks involved in employing the model to generate results.

1.7.3 Analysis

On the basis of systematic presentation in a tabular form or graphical form, determine the average income of an individual and the extent of disparities that exist. This information will help to get an understanding of the phenomenon (i.e. income of individuals).

1.7.4 Interpretation

All the above steps may now lead to drawing conclusions which will aid in decision making – a policy decision for improvement of the existing situation.

1.8 Classification of Quantitative Techniques

Figure 1.1 shows a tree diagram of Quantitative Techniques.

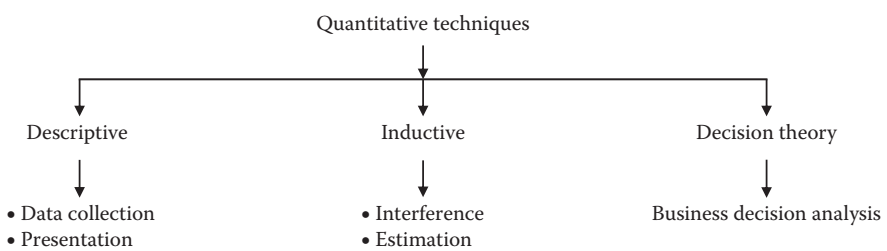


FIGURE 1.1
Classification of quantitative techniques.

1.8.1 Descriptive Statistics

There are statistical methods which are used for re-arranging, grouping and summarising sets of data to obtain better information or facts, and thereby a better description of the situation can be made.

Example 1.1: Changes in the Price Index

Yield by rice and so forth are frequently illustrated using different types of charts and graphs. These devices summarise large quantities of numerical data for easy understanding. Various types of averages can also reduce a large mass of data to a single descriptive number.

1.8.2 Inductive Statistics

This is concerned with the development of some criteria which can be used to derive information about the nature of the members of entire groups, also called population or universe, from the nature of the small portion, also called sample of the group.

Samples are drawn instead of a complete enumeration for the following reasons:

1. The number of units in the population may not be known.
2. The population units may be too many in number or widely dispersed. Thus, complete enumeration is extremely time-consuming, and at the end of a full enumeration, so much time is lost that the data becomes obsolete by that time.
3. It may be too expensive to include each population item.

1.8.3 Statistical Decision Theory

Statistical decision theory deals with analysing complex business problems with alternative courses of action or strategies and possible consequences. Basically, it is to provide more concrete information concerning these consequences, so that the best course of action can be identified from the alternative courses of action.

Statistical decision theory relies heavily not only on the nature of the problem at hand, but also on the decision environment. Basically, there are four different states of decision environment, as given in Table 1.1.

TABLE 1.1

States of Decision Environment

State of Decision	Consequence
Certainty	Deterministic
Risk	Probabilistic
Uncertainty	Unknown
Conflict	Influenced by an opponent

1.9 Methodology of Quantitative Techniques

Managers in an organisation generally observe various decision-making environments, define problems in those environments, develop models that seek to solve those problems, select the inputs of data necessary for solutions, find the optimum solutions to the problems and implement the solutions.

1.9.1 Steps

1. Observing the organisational environment is the first step in the process of solving a problem. The procedure involved in the process of observing an organisational environment may be as simple as visiting an enterprise and listening to a manager describe a problem. The process may also be as complex as spending months gathering data on production and distribution for the output of a manufacturing unit.
2. The person observing the environment should necessarily be a good listener, who is familiar with the process of gathering data. He or she should also be able to distinguish the problem from its environment details.
3. Defining the problem is the next step in the process. A manager identifies the problem by analysing the environment and defining the problem in specific terms. The problem definition determines the factors relevant to the solutions and should also isolate the factors that are under the control of the management.
4. Once the problem is clearly defined, the last step is to construct a mathematical model that represents the organisational situations. The model should enable the manager to forecast the crucial factors that affect the solution of the problem.

A model is said to be good if it is a concise and precise representation of the problem, is easily modifiable and cannot be easily misconstrued. Developing a model that represents the real-world problem is not an easy task. Managers need to review, test and refine their models on a regular basis.

1.10 Various Statistical Methods

1.10.1 Measure of Central Tendency

For proper understanding of quantitative data, it should be classified and converted into a frequency distribution, the number of times or frequency with which a particular data occurs in the given mass of data. This type of condensation of data reduces its bulk and gives a clear picture of its structure.

1.10.1.1 Mean

The mean is the common arithmetic average. It is computed by dividing the sum of the values of the observations by the number of items observed.

1.10.1.2 Median

The median is that item which lies exactly halfway between the lowest and highest value when the data is arranged in an ascending or descending order.

1.10.1.3 Mode

The mode is the central value (or item) that occurs most frequently. When the data is organised as a frequency distribution, the mode is that category which has the maximum number of observations.

1.10.2 Measures of Dispersion

Central tendency measures the most typical value around which most values in the distribution tend to converge. However, there are always extreme values in each distribution. These extreme values indicate the spread or dispersion of the distribution. The measures of this spread are called 'measures of dispersion' or 'variation' or 'spread'. Measures of dispersion would tell you the number of values which are substantially different from the mean, median or mode. The commonly used measures of dispersion are range, mean deviation and standard deviation.

The data may spread around the central tendency in a symmetrical or asymmetrical pattern. The measures of the direction and degree of symmetry are called measures of the skewness. Another characteristic of the frequency distribution is the shape of the peak, when it is plotted on a graph paper. The measures of the 'peakedness' are called Measures of Kurtosis.

1.10.3 Correlation

Correlation coefficient measures the degree to which the change in one variable (the dependent variable) is associated with a change in the other variable (independent variable). For example, as a marketing manager, you would like to know if there is any relation between the amount of money you spend on advertising and the sales you achieve. Here, a sale is the dependent variable and advertising budget is the independent variable. Correlation coefficient, in this case, would tell you the extent of relationship between these two variables, whether the relationship is directly proportional (i.e. an increase or decrease in advertising is associated with an increase or decrease in sales), inverse (i.e. an increase in advertising is associated with a decrease in sales, and vice versa) or there is no relationship between the two variables.

1.10.4 Regression Analysis

For determining a causal relationship between two variables, you may use regression analysis. Using this technique, you can predict the dependent variables on the basis of the independent variables. In 1970, the National Council of Applied and Economic Research (NCAER) predicted the annual stock of scooters using a regression model in which real personal disposable income and the relative weighted price index of scooters were used as independent variables.

1.10.5 Time-Series Analysis

A time series consists of a set of data (arranged in some desired manner) either recorded at successive points in time or covering successive periods of time. The changes in such types

of data from time to time are considered as the result of the combined impact of a force that is constantly at work. This force has four components:

1. Editing time-series data
2. Secular trend
3. Periodic changes, cyclical changes and seasonal variations
4. Irregular or random variations

With time-series analysis, you can isolate and measure the separate effects of these forces on the variables. Examples of these changes can be seen if you start measuring the increase in the cost of living; the increase of population over a period of time; the growth of agricultural food production in a country over the last 15 years; the seasonal requirement of items; and the impact of floods, strikes, wars and so on.

1.10.6 Index Numbers

An index number may be described as a specialised average designed to measure the change in a group of related variables over a period of time. Index numbers are stated in the form of percentages. For example, if we say that the index of prices is 105, it means that prices have gone up by 5% compared to the point of reference, called the base year. If the prices of the year 2015 are compared with those of 2005, the year 2015 would be called 'given or current year' and the year 2005 would be termed the 'base year'. Index numbers are also used in comparing production, sales price, volume employment, and so forth, changes over a period of time, relative to a base.

1.10.7 Sampling and Statistical Inference

In many cases, due to shortage of time, cost or non-availability of data, only a limited part or section of the universe (or population) is examined to

1. Get information about the universe as clearly and precisely as possible
2. Determine the reliability of the estimates

This small part or section selected from the universe is called the sample, and the process of selecting such a section (or part) is called sampling.

A scheme of drawing samples from the population can be classified into two broad categories: random and non-random sampling.

1.10.7.1 Random Sampling

Here, the drawing of elements from the population is random, and selection of an element is made in such a way that every element has an equal chance (probability) of being selected.

1.10.7.2 Non-Random Sampling

Here, the drawing of elements from the population is based on the choice or purpose of the selector.

Sampling analysis through the use of various 'tests', namely Z-normal distribution, Student's t-distribution, F-distribution and χ^2-distribution, makes it possible to derive inferences about population parameters with specified levels of significance and given degrees of freedom.

1.11 Advantages of Quantitative Methods

1. Quantitative methods like regression analysis help managers predict the unknown value of a variable from the known variables. This method can be used to predict sales for a period on the basis of past data. The cost of production can be estimated for a particular production level.
2. Quantitative methods like hypothesis testing are used to test the significance of a population parameter on the basis of a sample. Practically, it is not possible to estimate a parameter by surveying the entire population, and hence managers rely mainly on the sample data and then check whether the sample data represents the population.
3. Quantitative methods like decision theory help managers make decisions under uncertainty, taking various courses of action into consideration. Decision theory can be used by managers to decide whether to outsource a function. The method also gives the outcome of a decision in monetary terms.
4. Quantitative methods like linear programming help managers allocate scarce resources in an optimum manner while solving the problems involved in scheduling, product mix determination, and so on.

1.11.1 Definiteness

The study of statistics helps us to present general statements in a precise and definite form. Statements of facts conveyed numerically are more precise and convincing than those stated quantitatively. For example, the statement 'the literacy rate as per the 1981 census was 36% compared to 29% for the 1971 census' is more convincing than stating simply that 'literacy in our country has increased'.

1.11.2 Condensation

The new data is often unwieldy and complex. The purpose of statistical methods is to simplify a large mass of data and present meaningful information from it. For example, it is difficult to form a precise idea about the income position of the people of India from the data of individual income in the country. The data will be easy to understand and more precise if it can be expressed in the form of per capita income.

1.11.3 Comparison

The object of statistics is to enable comparisons between the past and present results with a view to ascertaining the reasons for changes which have taken place and the effect of such changes in the future. Thus, if one wants to appreciate the significance of figures, then he or she must compare them with others of the same kind. For example, the statement 'per

capita income has increased considerably' will not be meaningful unless some comparison of figures of the past is made. This will help in drawing conclusions as to whether the standard of living of the people of India is improving.

1.11.4 Policy Formulation

Statistics provides the basic material for framing policies not only in business, but also in other fields. For example, data on birth and mortality rate not only helps in assessing future growth in the population, but also provides necessary data for framing a scheme of family planning.

1.11.5 Hypothesis Testing

Statistical methods are useful in formulating and testing a hypothesis or assumption or statement and developing new theories. For example, the hypothesis 'whether a student has benefitted from a particular medium of instruction' can be tested by using an appropriate statistical method.

1.11.6 Prediction

For framing suitable policies or plans, and then for implementation, it is necessary to have the knowledge of future trends. Statistical methods are highly useful for forecasting future events. For example, for a businessman to decide how many units of an item should be produced in the current year, it is necessary for him to analyse the sales data of past years.

1.12 Application of Quantitative Techniques in Business Management

1.12.1 Management

1.12.1.1 Marketing Management

1. Analysis of marketing research information
2. Statistical records for building and maintaining an extensive market
3. Sales forecasting

1.12.1.2 Production Management

1. Production planning, control and analysis
2. Evaluation of machine performance
3. Quality control requirements
4. Inventory control measures

1.12.1.3 Finance Management

1. Financial forecast and budget preparation
2. Financial investment decisions

3. Selection of securities
4. Auditing function
5. Credit policies, credit risk and delinquent accounts

1.12.1.4 Personnel Management

1. Labour turnover rate
2. Employment trends
3. Performance appraisal
4. Wage rates and incentive plans

1.12.2 Economics

1. Measurement of gross national product and input/output analysis
2. Determination of business cycle, long-term growth and seasonal fluctuations
3. Comparison of market prices, cost and profits of individual firms
4. Analysis of population, land economics and economic geography
5. Operational studies of public utilities
6. Formulation of appropriate economic policies and evaluation of their effect

1.12.3 Research and Development

1. Development of new product lines
2. Optimal use of resources
3. Evaluation of existing products

1.12.4 Natural Science

1. Diagnosing the disease based on data like temperature, pulse rate and blood pressure
2. Judging the efficacy of a particular drug for curing a certain disease
3. Study of plant life

1.13 Limitations of Quantitative Techniques

1. Quantitative methods involve the use of mathematical models, equations and other mathematical expressions, which are derived on the basis of several assumptions. Such underlying assumptions that are present in the current problem may or may not be relevant to some other problems. If this caution is not taken into consideration, wrong applications of quantitative methods may yield disastrous solutions.

2. Quantitative methods are usually expensive, as they require the services of specialised people. Even big organisations use quantitative methods to a limited extent because many applications are not worth their costs. Instead of using quantitative methods, managers generally prefer to exercise their intuition and judgement to make decisions.
3. The accuracy of the solution attained by quantitative methods is also hindered by pitfalls like inadequacy of data, inconsistency in definitions, selection of the wrong sample, improper choice of the method, inappropriate comparisons and improper presentations.
4. Quantitative methods cannot be used to study qualitative phenomena as they do not take the intangible and non-measurable human factors into consideration. For instance, the methods make no allowances for intangible factors such as skill, attitude and vigour of managers in taking decisions. However, the methods can be successfully applied indirectly by first converting intangible expressions into quantitative terms. For example, the intelligence of a manager can be quantified by providing different scores to difference qualifications.

1.14 Summary

As the process of decision making is becoming increasingly difficult, managers can no longer afford to make decisions on the basis of subjective factors like experience, observation or evaluation alone. They have to use objective and quantitative methods to collect, present, analyse and meaningfully interpret the available data to arrive at proper solutions.

Quantitative methods involve the use of numbers, symbols, mathematical expressions and other quantitative elements and serve as supplements to the subjective intuition of the decision makers. The utility of these methods was realised long ago.

Quantitative methods are broadly classified into two groups: Statistical techniques and Programming techniques. Statistical techniques include all the statistical methods that range from the collection of data to the task of interpreting the collected data. Programming techniques involve the construction of mathematical models that relate the relevant variables in a situation to the outcome.

This chapter discussed the historical evolution of quantitative methods and the methodology of quantitative methods. The chapter also discussed the advantages and limitations of quantitative methods.

There is an ever-increasing demand for managers with numerate ability as well as literary skills, so that they can present numerate data and information which requires analysis and interpretation, but more importantly, they can quickly scan and understand analysis provided both from within the firm and by outside organisations. In the competitive and dynamic business world, those enterprises which are most likely to succeed, and indeed survive, are those which are capable of maximising the use of the tools of management, including quantitative techniques.

This chapter has attempted to describe the meaning and use of various quantitative techniques in the field of business and management. The importance and complexity of the decision-making process has resulted in the wide application of quantitative techniques in the diversified field of business and management. With the evolution of more

powerful computing techniques, users of these techniques are encouraged to explore new and more sophisticated methods of data analysis. A quantitative approach in decision making, however, does not totally eliminate the scope of qualitative or judgement ability of the decision maker.

REVIEW QUESTIONS

1. Think of any major decision you made recently. Recall the steps taken by you to arrive at the final decision. Prepare a list of those steps.
2. Comment on the following statements:
   a. 'Statistics are numerical statement of facts, but all facts numerically stated are not statistics'.
   b. 'Statistics is the science of averages'.
3. What is the type of the following models?
   a. Frequency curves in statistics
   b. Motion films
   c. Flow chart in production control
   d. Family of equations describing the structure of an atom
4. List at least two applications of statistics in each functional area of management.
5. What factors in modern society contribute to the increasing importance of the quantitative approach to management?
6. Describe the major phases of statistics. Formulate a business problem and analyse it by applying these phases.
7. Explain the distinction between
   a. Static and dynamic models
   b. Analytical and simulation models
   c. Descriptive and prescriptive models
8. Describe the main features of the quantitative approach to management.

2

Arranging Data

2.1 Meaning of Data

Data is a collection of related observations, facts or figures. Managers collect facts such as number of units produced per day by each department or by each worker, and these facts and figures of production can be called data.

2.2 Types of Data

2.2.1 Published Data

Published data can be obtained in the form of bulletins, reports and so forth from various government agencies. For example, data relating to monetary and banking activities (industrial production index, price index, etc.) can be obtained from Reserve Bank of India (RBI) bulletins and finance reports published by RBI every month. Many research organisations and private sources provide useful data to managers.

2.2.1.1 Published Sources

There are a number of national and international organisations which collect statistical data and publish their findings in statistical reports periodically. Some of the national organisations which collect, compile and publish statistical data are the Central Statistical Organization (CSO), National Sample Survey Organisation (NSSO), Office of the Registrar General and Census Commissioner of India, Labour Bureau, Federation of Indian Chambers of Commerce and Industry, Indian Council of Agricultural Research (ICAR), *Economic Times* and *Financial Express*. Some of the international agencies which provide valuable statistical data on a variety of socio-economic and political events are the United Nations (UN), World Health Organization (WHO), International Labour Organization (ILO), International Monetary Fund (IMF) and World Bank.

2.2.2 Unpublished Data

Managers have to go in for first-hand data collection by the way of sample survey or a census. The information may be obtained by using various methods like observation, personal interview or questionnaires.

2.2.2.1 Unpublished Sources

All statistical data need not be published. A major source of statistical data produced by government, semi-government, private and public organisations is based on the data drawn from internal records. This data based on internal records provides authentic statistical data and is much cheaper than primary data. Some examples of the internal records include employees' payroll, the amount of raw materials, cash receipts and the cash book. It may be pointed out that it is very difficult to access unpublished information.

2.2.3 Primary Data

Primary data is the first-hand data collected by the researcher personally. Data used in statistical study is termed either 'primary' or 'secondary' depending on whether it was specifically for the study in question or for some other purpose. When the data used in a statistical study is collected under the control and supervision of the investigation, such data is referred to as primary data.

2.2.3.1 Editing Primary Data

Once the questionnaires have been filled and the data collected, it is necessary to edit this data. Editing of data should be done to ensure completeness, consistency, accuracy and homogeneity.

2.2.3.1.1 Completeness

Each questionnaire should be complete in all respects; that is the respondent should have answered each and every question. If some important questions have been left unanswered, attempts should be made to contact the respondent and get the response. If, despite all efforts, answers to vital questions are not given, such questionnaires should be dropped from final analysis.

2.2.3.1.2 Consistency

Questionnaires should also be checked to see that there are no contradictory answers. Contradictory responses may arise due to wrong answers filled in by the respondents or because of carelessness on the part of the investigator in recording the data. A respondent's unwillingness to answer a specific question can result in non-response or a contradictory response, where the respondent completes the rest of the questions other than those he or she is uncomfortable with, refusal to complete the rest of the questionnaire or deliberate falsification. For example, questions such as `Would you resort to stealing things in a hotels or supermarkets if you knew there were no hidden cameras?' are virtually sure to attract stereotyped responses or refusals from participants.

2.2.3.1.3 Accuracy

The questionnaire should also be checked for the accuracy of information provided by the respondent. It may be pointed out that this is the most difficult job of the investigator and at the same time the most important one. If inaccuracies are permitted, this would lead to misleading results. Inaccuracies may be checked by random cross-checking.

2.2.3.1.4 Homogeneity

It is equally important to check whether the questions have been understood in the same sense by all the respondents. For instance, if there is a question on income, it should be very clearly stated whether it refers to weekly, monthly or yearly income. If it is left ambiguous, then respondents may give different responses and there will be no basis for comparison because we may take some figures which are valid for monthly income and some for annual income.

2.2.4 Secondary Data

When the data is not collected by the investigator, but is derived from other sources, then such data is referred to as 'secondary data'.

Secondary data is collected from other available sources, collected by organisations in the form of financial statements, sales reports, cash flow data, production schedules, budgets and so forth.

2.2.4.1 Precautions in the Use of Secondary Data

A careful scrutiny must be made before using published data. The user should be extra cautious in using secondary data, and he should not accept it at its face value. The reason may be that such data is full of errors because of bias, inadequate sample size, errors of definitions, computational errors and so forth. Therefore, before using such data, the following aspects should be considered.

2.2.4.1.1 Suitability

The investigator must ensure that the data available is suitable for the purpose of the inquiry on hand. The suitability of data may be judged by comparing the nature and scope of investigation.

2.2.4.1.2 Reliability

It is of utmost importance to determine how reliable is the data from secondary sources and how confidently we can use it. In assessing the reliability, it is important to know whether the collecting agency is unbiased, whether it has a representative sample, whether the data has been properly analysed and so on.

2.2.4.1.3 Adequacy

Data from secondary sources may be available, but its scope may be limited, and therefore this may not serve the purpose of investigation. The data may cover only a part of the requirement of the investigator or may pertain to a different time period.

Only if the investigator is fully satisfied on all the above-mentioned points should he proceed with this data as the starting point for further analysis.

2.3 Primary versus Secondary Data

See Table 2.1 for differences between primary and secondary data.

TABLE 2.1

Differences between Primary and Secondary Data

Primary Data	Secondary Data
Data which is primary in the hands of one	Data which is secondary in the hands of another
Data collected by himself or through his agent	Data was already used by someone before

2.4 Classification of Data

This is the process of arranging data into sequences and groups according to its common characteristics. Here, data is arranged according to the points of similarities and dissimilarities. It is like the process of sorting the mail in a post office, where the mail for different destinations is placed in different compartments after it has been carefully sorted out from the huge heap. For example, data collected in a consumer survey can be classified along characteristics like age, gender, education, income and so forth.

2.4.1 Methods of Classification

1. Geographical: Area-wise or region-wise
2. Chronological: According to occurrence of an event in time
3. Qualitative: Depending on characteristics and attributes
4. Quantitative: According to magnitudes

2.4.1.1 Geographical Classification

In this type of classification, data is classified according to area or region. For example, when we consider production of wheat country-wise, this would be called geographical classification. The listings of individual entries are generally done in alphabetical order or according to size to emphasise the importance of a particular area or region.

2.4.1.2 Chronological Classification

When the data is classified according to the time of its occurrence, this is known as chronological classification. For example, sales figures of a company for 4 years are given in Table 2.2.

TABLE 2.2

Classification of Data according to Time

Year	Sales (Rs. in Thousands)
2002–2003	271
2003–2004	227
2004–2005	471
2005–2006	161

2.4.1.3 Qualitative Classification

When the data is classified according to some attributes (distinct categories) which are not capable of measurement, this is known as qualitative classification. In a simple (or dichotomous) classification, an attribute is divided into two classes, one possessing the attribute and the other not possessing it. For example, we may classify population on the basis of employment, that is the employed and unemployed. Similarly, we can have manifold classifications when an attribute is divided to form several classes. For example, the attribute 'education' can have different classes, such as primary, middle, higher secondary and university.

2.4.1.4 Quantitative Classification

When the data is classified according to some characteristics that can be measured, this is called quantitative classification. For example, the employees of a company may be classified according to their monthly salaries. Since quantitative data is characterised by different numerical values, the data represents the values of a variable. Quantitative data may be further classified into one or two types: discrete or continuous. The term *discrete* data refers to quantitative data that is limited to certain numerical values of a variable, for example the number of employees in an organisation or the number of machines in a factory.

2.5 Data Collection

2.5.1 Population

This is the entire collection of entities that a manager is trying to study.

2.5.2 Sample

This is a fraction of the population that represents the entire population in its characteristics proportionately. For example, when a magazine conducts an opinion poll among 1000 individuals from all over India, with a view to know the general opinion of Indians towards politicians, then all Indians is the target population and 1000 individuals represent the population and is referred to as the sample.

In another example, to estimate the potential market for a new innovation, managers in a research department may study 1000 consumers in a particular territory. The managers make sure that this sample contains the consumers belonging to all cross sections (income, religion, education and locality) of the society.

2.5.3 Testing the Validity of Data

This is testing the data for adequacy and reliability, as it influences the quality of the final decision.
Managers can pose the following questions:

1. Data origin?
2. Is the source reliable?

3. Does the data support or contradict the previous decisions?
4. Are the conclusions derived from the data?
5. What is the size of the sample? Does it represent the entire population under consideration for decision making?

2.5.4 Complete Enumeration or Census Survey or Census

This is a method in which the entire population is taken up and information is collected relating to all the units of the population. For example, government undertakes complete enumeration of all citizens once every 27 years, and information is collected with respect to each and every person living in India.

2.5.5 Sample Method

This is a method in which enumeration of part of the population or universe is taken up and information is gathered regarding this selected part. This is commonly adopted in an organisation to check the quality of the finished products of a manufacturing unit.

2.5.6 Methods of Collecting Primary Data

2.5.6.1 Observation Method

In the observation method, the investigator asks no questions, but he simply observes the phenomenon under consideration, and records the necessary data. Sometimes individuals make the observation; on other occasions, mechanical and electronic devices do the job.

In the observation method, it may be difficult to produce accurate data. Physical difficulties on the part of the observer may result in errors. Because of these limitations in the observation method, the questionnaire method is more widely used for collecting data. In the questionnaire method, the investigator draws up a questionnaire containing all the relevant questions which he wants to ask of his respondents and accordingly records the responses. The questionnaire method may be conducted through personal interview or by mail or telephone.

2.5.6.2 Personal Interviews

In this method, the interviewer sits face-to-face with the respondent and records his responses. The information is likely to be more accurate and reliable because the interviewer can clear up doubts and cross-check the respondents. This method is time-consuming and can be very costly if the number of respondents is large and widely distributed.

2.5.6.3 Questionnaire Method

In this method, a list of questions (questionnaire) is prepared and mailed to respondents. The respondents are expected to fill in the questionnaire and send it back to the investigator. Sometimes, mail questionnaires are placed in respondents' hands through other means, such as attaching them to consumer products or putting them in newspapers or magazines. This method can be easily adopted where the field of investigation is very vast and the respondents are spread over a wide geographical area. But it can only be

adopted where the respondents are literate and can understand written questions and answer them.

2.5.6.3.1 Telephone

In this method, the investigator asks the relevant questions of the respondents over the telephone. This method is less expensive, but it has limited application since only those respondents can be interviewed who have telephone; moreover, very few questions can be asked over the telephone.

The questionnaire is a very efficient and fast method of collecting data. But it has very serious limitations, as it may be extremely difficult to collect data on certain sensitive aspects, such as income, age or personal life details, which the respondents may not be willing to share with the investigator. This is so with other methods also; different people may interpret the questions differently, and consequently, there may be errors and inaccuracies in data collection.

2.5.6.3.2 Designing a Questionnaire

The success of collecting data through a questionnaire depends mainly on how skilfully and imaginatively the questionnaire has been designed. A badly designed questionnaire will never be able to gather the relevant data. In designing the questionnaire, some of the important points to be kept in mind are given here.

1. **Cover letter.** Every questionnaire should contain a cover letter. The cover letter should highlight the purpose of the study and assure the respondent that all responses will be kept confidential. It is desirable that some inducement or motivation is provided to the respondent for better responses. The objectives of the study and questionnaire design should be such that the respondent derives a sense of satisfaction through his or her involvement.
2. **Number of questions should be kept to a minimum.** The fewer the questions, the greater the chances of getting a better response and of having all the questions answered. Otherwise, the respondent may feel disinterested and provide inaccurate answers, particularly towards the end of the questionnaire. Informing the questions, the investigator has to take into consideration several factors, such as the purpose of study, time and resources available. As a rough indication, the number of questions should be from 21 to 47. If the number of questions is more than 21, it is desirable that the questionnaire be divided into various parts to ensure clarity.
3. **Questions should be simple, short and unambiguous.** The questions should be simple, short, easy to understand and such that their answers are unambiguous. For example, if the question is 'Are you literate?' the respondent may have doubts about the meaning of literacy. To some, literacy may mean a university degree, whereas to others, even the capacity to read and write may mean literacy. Hence, it is desirable to specify whether you have passed (a) high school, (b) graduation or (c) post-graduation.

   Questions can be of a yes/no type or multiple choice, depending on the requirement of the investigator. Open-ended questions should generally be avoided.
4. **Questions of a sensitive or personal nature should be avoided.** The questions should not be such as would require the respondent to disclose any private, personal or confidential information. For example, questions relating to sales, profits,

marital happiness and so forth should be avoided as far as possible. If such questions are necessary in the survey, an assurance should be given to the respondent that the information provided will kept strictly confidential and will not be used at any cost to their disadvantage.

5. **Answers to questions should not require calculations.** The questions should be framed in such a way that their answers do not require any calculations.
6. **Logical arrangement.** The questions should be logically arranged so that there is a continuity of responses and the respondent does not feel the need to refer back to the previous questions. It is desirable that the questionnaire should begin with some introductory questions, followed by vital questions crucial to the survey, and ending with some light questions so that the overall impression of the respondent is a good one.
7. **Cross-check and footnotes.** The questionnaire should contain some questions which act as a cross-check to the reliability of the information provided. For example, when a question relating to income is asked, it is desirable to include the question 'Are you an income tax payee?'

2.6 Data Presentation Devices

2.6.1 Tables

In tables, the data is classified based on observation time and magnitude, or some other characteristics of the variable.

2.6.2 Tabulation

This is the logical listing of related quantitative data in vertical columns and horizontal rows with

1. Sufficient explanatory and qualifying words
2. Phrases and statements in the form of titles
3. Headings and explanatory notes – to make clear the full meaning, context and origin of the data

2.6.3 Uses of Tabulation

1. To summarise or condense data
2. To help managers analyse the relationships and trend in the collected data
3. To represent all the available data in the least possible space without losing its clarity

2.6.4 Objectives of Tabulation

1. To condense complex data
2. To show a trend

3. To display huge volumes of data in less space
4. To be self-explanatory
5. To highlight key characteristics of data
6. To facilitate comparison of data elements
7. To help decision making using statistical methods
8. To serve as a reference for future decisions

2.6.5 Parts of an Ideal Table

1. Table number – identity
2. Title – idea about the nature of data
3. Captions – headings of vertical column(s) that explain the mode of classification, such as
   a. Time
   b. Quantity
   c. Region
4. Stubs – headings explaining the basis for classifying the rows
5. Body – where row and column headings (captions and stubs) explain the data
6. Footnote – other information to explain the data
7. Source – source of information

2.7 Graphs

These are pictorial representation of the data. They are a non-quantitative form of presentation. The quantities are also indicated along with them. The magnitude of the data is depicted visually through the proportional size of the diagram or graph.

Here, the collected data is represented by various types of geometrical devices, such as points, lines, bars, multi-dimensional figures and pictorials. Graphs, maps and charts are drawn to 'scale' to give accurate data to managers.

2.7.1 Types of Graphs or Charts

Line charts, bar charts and pie charts are all examples of graphs.

2.7.1.1 Rules for Constructing the Line Graph

1. Place time on the x-axis (horizontal) and the variable on the y-axis (vertical). The unit of time in which the variable under consideration is measured should be clearly stated in the title.
2. Begin the y-axis with zero and select a suitable scale so that the entire data is accommodated in the space available. On the arithmetic scale, equal magnitude must be represented by equal distances.

This requirement is true for both the x-axis and y-axis, but for each separately.

3. Corresponding to the time factor, plot the value of the variable and join the various points by straight lines (and not with curves). The main points on the graph should not be indicated by circles or crosses; rather, dots should be used so that they disappear into line.
4. Join the various points with straight lines, not curves.
5. If on one graph more than one variable is shown, they should be distinguished by the use of thick or thin dotted lines or different colours.

   Every graph should be given a suitable title.
6. Lettering on the graph, that is the indication of years, units and so forth, should be done horizontally and not vertically.

2.7.1.2 Bar Chart

Business and statistical data can be presented through bar charts. Bar charts are one-dimensional because the magnitude of the data is represented by the length of the bar. Bar charts consist of a group of equidistant rectangular bars, each representing one class interval of a given data in the table. The width of the bar has no relevance; only the length of the bar is important (Figure 2.1).

2.7.1.2.1 Types of Bar Chart

1. Vertical bar chart
2. Horizontal bar chart
3. Multiple bar diagram or compound bar diagram

2.7.1.2.2 Guidelines for Constructing a Bar Chart

1. The length of the bars should be proportionate to the data they represent.
2. All bars should rise from the same baseline.
3. Uniformity in the width of the bars need not be maintained, as width is irrelevant in a bar chart.
4. The scale should be selected depending on the highest value in the table.
5. The scale should be arranged from left to right.

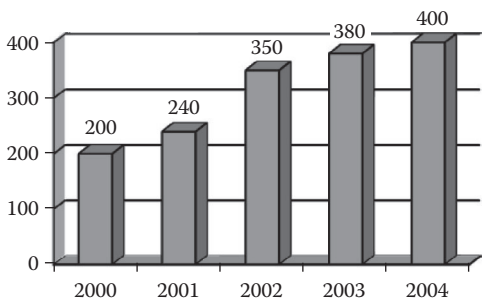


FIGURE 2.1
Bar chart.

6. The gaps between the bars should be uniform.
7. The bars should be arranged from left to right.
8. Mention the (magnitude) figure on the bar to enable the manager to see the magnitude at a glance.

Exercise

A company shows the export and import values for four consecutive years between 2005 and 2008. Draw a compound bar chart showing the export/import values by putting the year on the x-axis and export/import values on the y-axis (Figure 2.2).

2.7.1.3 Pie Diagram

A pie diagram is a circle which is divided into various segments, and each segment represents the percentage contribution of the various components to the total. Managers use it to compare many components simultaneously.

2.7.1.3.1 Steps for Drawing a Pie Diagram

Express the value of each category as a percentage of the total (360°) in a circle representing the whole (i.e. 100%).

Exercise

Present the following data on a pie diagram (Table 2.3 and Figure 2.3).
Formula:

$$\text{Degree of each part} = \text{Part} \times 360/\text{Total}$$

$$= \text{Part} \times 3.6$$

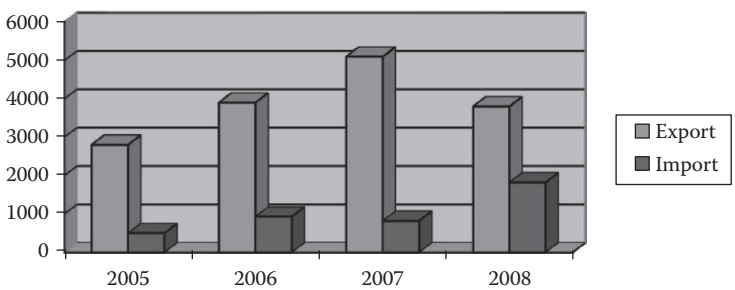


FIGURE 2.2
Compound bar chart.

TABLE 2.3

Data of Share in Five Classes

Class	A	B	C	D	E	Total
Share (%)	70	30	10	5	5	120

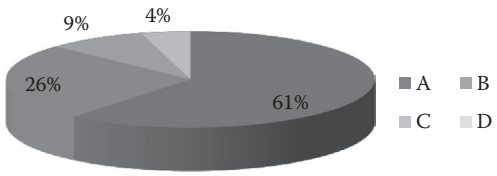


FIGURE 2.3
Pie diagram.

2.8 Frequency Distribution

Frequency distribution is the table in which raw data is tabulated, by dividing it into classes of convenient size and computing the number of data elements falling within each pair of class boundaries.

It is a tabular form that organises data into classes, that is groups of values having the same characteristics of data.

Example 2.1: Need for Frequency Distribution

The raw data relating to the ages of 30 employees of a statistics department are as shown in Table 2.4.

The ages given in Table 2.4 are in an arbitrary manner. It is very difficult for the human resources manager to grasp any trend from the raw data. Logical arrangement of the data is necessary to compress the data and help the manager see to which age group an employee belongs and what the frequency of each age group is.

2.8.1 Discrete Frequency Distribution

The process of preparing a frequency distribution is very simple. In the case of discrete data, place all possible values of the variable in ascending order in one column, and then prepare another column of 'tally' marks to count the number of times a particular value of the variable is repeated. To facilitate counting, blocks of five tally marks are prepared and some space is left in between the blocks. The frequency column refers to the number of tally marks a particular class will contain. To illustrate the construction of a discrete frequency distribution, consider a sample study in which 46 families were surveyed to find the number of children per family. The data obtained are

2 3 4 2 4 5 4 3 5 6 4 3 4 5 6 4 3 4 2 4 5 7 4
2 4 5 6 4 4 4 5 6 4 5 7 4 5 7 5 6 6 7 1 6 6

TABLE 2.4
Ages of 30 Employees of a Statistics Department

26	37	36	27	43	32	31	27	37	47
47	27	33	27	32	47	46	47	32	32
37	27	32	33	47	27	37	44	34	47

TABLE 2.5

Frequency Distribution Table

No. of Children	No. of Families (Tally Marks)	Frequency
7	𝍸	4
6	𝍸 𝍸	8
5	𝍸 𝍸	9
4	𝍸 𝍸 𝍸	15
3	𝍸	4
2	𝍸	4
1	𝍸	2
Total		46

To condense this data into a discrete frequency distribution, we use tally marks, as shown in Table 2.5.

2.8.2 Continuous Frequency Distribution

In constructing the frequency distribution for continuous data, it is necessary to clarify some of the important terms that are frequently used.

Class: A group of objects with a common property.

Class frequency: The number of observations falling within a particular class is called its class frequency or simply frequency. Total frequency (sum of all the frequencies) indicates the total number of observations considered in a given frequency distribution.

Class interval: The class interval represents the width (span or size) of a class. The width may be determined by subtracting the lower limit of one class from the lower limit of the following class (alternatively, successive upper limits may be used). For example, if the two classes are 10–20 and 20–30, the width of the class interval would be the difference between the two successive lower limits, that is 20 – 10 = 10, or the difference between the upper limit and lower limit of the same class, that is 20 – 10 = 10.

Class limits: The range of values of a given class. Class limits denote the lowest and highest values that can be included in the class. The two boundaries (i.e. lowest and highest) of a class are known as the lower limit and the upper limit of the class. For example, in the class 50–60, 50 is the lower limit and 60 is the upper limit, or we can say that there can be no value in that class which is less than 50 and more than 60.

Class mark: The middle of a class interval. Frequency distributions can be constructed with classes of qualitative attributes.

Class midpoint: The sum of two successive lower limits divided by 2. Therefore, it is the value lying halfway between the lower and upper class limits. In the example taken above, the midpoint would be (10 + 20)/2 = 15 corresponding to the class 10–20.

2.8.2.1 Types of Class Interval

1. **Exclusive method.** The class intervals are so arranged that the upper limit of one class is the lower limit of the next class.
2. **Inclusive method.** In this method, the upper limit of one class is included in that class itself. The following examples illustrate this point.

3. **Open-ended.** In an open-ended distribution, the lower limit of the very first class and upper limit of the last class are not given. Where there is a big gap between the minimum and maximum values, the open-ended distribution can be used, such as in income distributions. The income disparities of residents of a region may vary between Rs. 100 and Rs. 10,000 per month. In such a case, we can form classes like less than Rs. 1000, 1000–2000, 2000–3000, 3000–4000 and 4000 and above.

$$\text{Correction factor} = \frac{\text{Lower limit of second class} - \text{Upper limit of the first class}}{2}$$

2.8.2.2 Selection of Class Intervals

The number of classes should not be too small or too large. However, there is no hard-and-fast rule about it. If the number of observations is smaller, the number of classes formed should be towards the lower side of this limit, and when the number of observations increases, the number of classes formed should be towards the upper side of the limit. If possible, the widths of the intervals should be numerically simple. It is desirable to have classes of equal width. However, in the case of distributions having a wide gap between the minimum and maximum values, classes with an unequal class interval can be formed, like with income distribution.

The class interval should be determined after taking into consideration the minimum and maximum values and the number of classes to be formed.

2.8.3 Cumulative Frequencies

Cumulative frequencies cumulate the frequencies, starting at either the lowest or the highest value. The cumulative frequencies of a given class interval thus represent the total of all the previous class frequencies, including the class against which it is written.

2.8.4 Relative Frequencies

Very often, the frequencies in a frequency distribution are converted to relative frequencies to show the percentage for each class. If the frequency of each class is divided by the total number of observations (total frequency), then this proportion is referred to as relative frequency.

2.9 Histogram

A histogram is a series of rectangles, the width of each being proportional to the range of values within a class and the height being proportional to the number of items falling in the class. While constructing the histogram, the variable is always placed on the x-axis and the frequencies depending on it on the y-axis. The distance for each rectangle on the x-axis shall remain the same in case the class intervals are uniform throughout. The y-axis represents the frequencies of each class, which constitute the height of its rectangle.

When class intervals are equal, place frequency on the y-axis and the variable on the x-axis and construct adjacent rectangles. In such a case, the height of the rectangles will be proportional to the frequencies. When the class intervals are unequal, a correction for unequal class interval must be made.

When the widths of classes in a frequency distribution are equal, the widths of the bars are uniform. The length of the bar is proportionate to the number of data elements in the class it represents.

Exercise

Draw the histogram for the data in Table 2.6 (Figure 2.4).

2.9.1 Relative Frequency Histogram

When a histogram is constructed using relative frequency, it is called a relative frequency histogram. It represents the relative size of each class with the total.

TABLE 2.6

Data of Sales in Rupees Ten Thousand in Factories

Sales (Rs. in Thousands)	No. of Factories	Sales (Rs. in Thousands)	No. of Factories
10–20	10	50–60	70
20–30	20	60–70	50
30–40	30	70–80	20
40–50	100	80–90	5

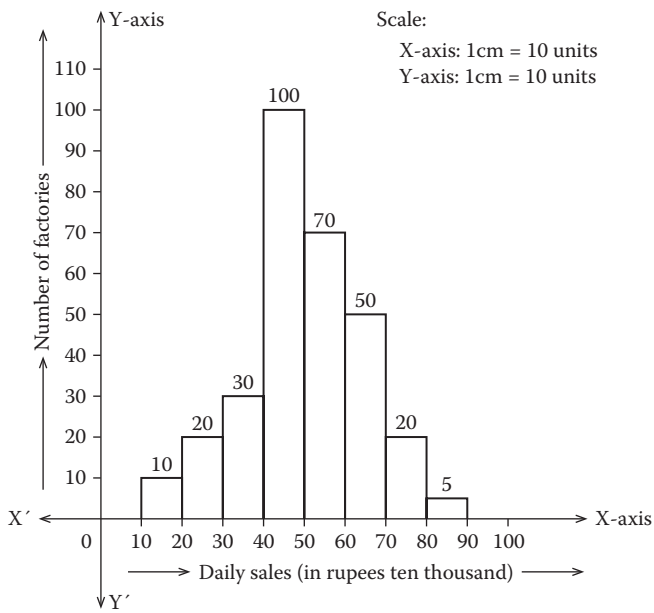


FIGURE 2.4
Histogram.

2.9.2 Absolute Histogram

This represents the number of data items.

Example 2.2: Educational Qualification of 200 Employees

See Table 2.7.

2.9.3 Difference between a Relative Frequency Histogram and an Absolute Histogram

See Table 2.8 for the differences between a relative frequency histogram and an absolute histogram.

2.10 Frequency Polygon

A frequency polygon is a graph of the frequency distribution. It has more than four sides. It is particularly effective in comparing two or more frequency distributions.

1. We may draw a histogram of the given data and then join by straight lines the midpoints of the upper horizontal side of each rectangle with the adjacent ones. The figure so formed is called a frequency polygon.

TABLE 2.7

Data of Educational Qualification of 200 Employees

Qualification	Frequency	Relative Frequency
SSC	15	0.075
HSC	13	0.065
BCom	20	0.100
BA	53	0.265
BE	50	0.250
BSc	26	0.130
BBA	23	0.115
Total	200	1.000

TABLE 2.8

Relative Frequency Histogram vs. an Absolute Histogram

Relative Frequency Histogram	Absolute Histogram
1. When a histogram is constructed using relative frequency, it is called a relative frequency histogram.	1. When a histogram is constructed using absolute frequency, it is called an absolute histogram.
2. The scale of the y-axis is the number of observations in each class as a fraction of the total number of observations.	2. The scale of the y-axis is the absolute number of observations in each class.
3. The relative frequency histogram shows the relative size of each class with the total.	3. The absolute histogram represents the number of data items.

2. Another method of constructing a frequency polygon is to take the midpoints of the various class intervals and then plot the frequency corresponding to each point and join all these points by straight lines.

   A frequency polygon represents graphically both simple and relative frequency distributions. For constructing a frequency polygon, the frequencies are marked on the y-axis. The values of variables are placed on the x-axis. Dots are put on the graph against the class marks to represent the frequencies. These dots are connected by drawing straight lines, forming a frequency polygon.

Exercise

Draw the frequency polygon for the data in Table 2.6 (Figure 2.5).

2.11 Frequency Curve

When the straight lines are smoothed by adding classes and data points, this is called a frequency curve. The number of classes and observations is increased, and the dots are connected by curves, to get a frequency curve. A smoothed frequency curve can be drawn through the various points of the polygon. The curve is drawn freehand in such a manner that the area included under the curve is approximately the same as that of the polygon.

Exercise

Draw the frequency curve for the data of Table 2.6 (Figure 2.6).

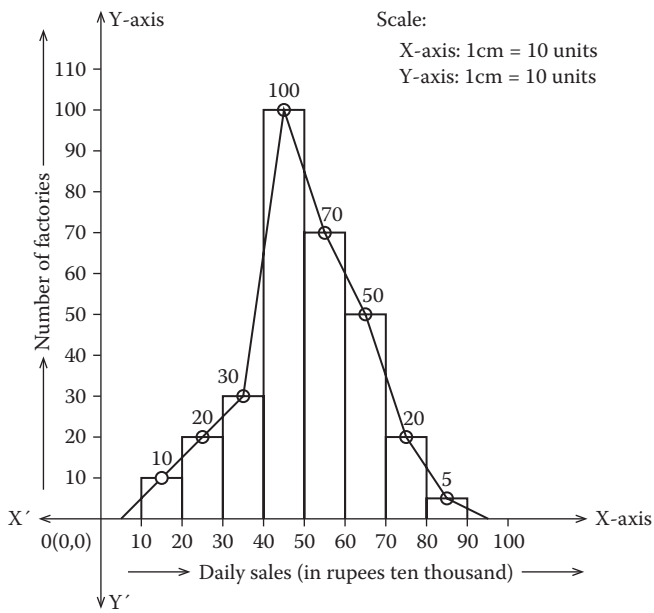


FIGURE 2.5
Frequency polygon.

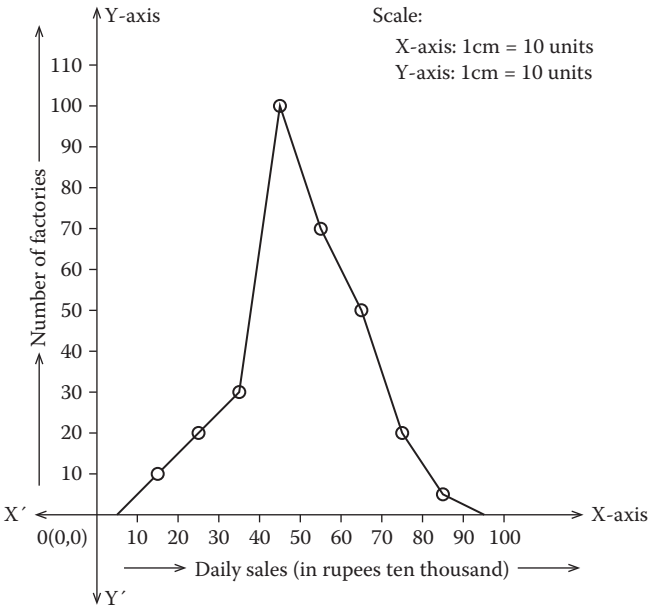


FIGURE 2.6
Frequency curve.

2.12 Cumulative Frequency Distribution

This is a tabular display of data showing how many observations lie above or below certain values, rather than simply showing the number of items within intervals.

2.12.1 Ogive or Cumulative Frequency Curve

An ogive is the graphical presentation of a cumulative frequency distribution, and therefore when the graph of such a distribution is drawn, it is called a cumulative frequency curve or ogive.

2.12.1.1 Less than Ogive

In the 'less than' method, we start with the upper limits of the classes and add the frequencies. When these frequencies are plotted, we get a rising curve. Here, the upper limits of the various classes are placed on the x-axis and the frequencies obtained by the process of cumulating the preceding frequencies on the y-axis. By joining these points, we get less than ogive.

Exercise

Draw a less than ogive for the data in Table 2.9.
The less than ogive curve is shown in Figure 2.7.

TABLE 2.9

Data of Sales in Rupees Thousand in Factories

Sales (Rs. in Thousands)	No. of Factories	Sales (Rs. in Thousands)	No. of Factories
10–20	10	<20	10
20–30	20	<30	30
30–40	30	<40	60
40–50	40	<50	100

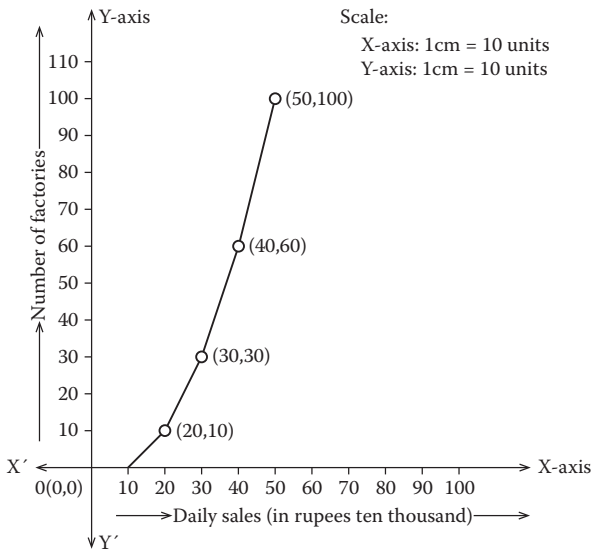


FIGURE 2.7
Less than ogive.

2.12.1.2 More than Ogive

In the `more than' method, we start with the lower limits of the classes and from the total frequencies we subtract the frequency of each class. When these frequencies are plotted, we get a declining curve. More than ogive can be drawn by taking the lower limits on the x-axis and the cumulative frequency on the y-axis. By joining these points, we get more than ogive (Table 2.10).

The more than ogive curve is shown in Figure 2.8.

TABLE 2.10

Data of Sales in Rupees Thousand in Factories

Sales (Rs. in Thousands)	No. of Factories	Sales (Rs. in Thousands)	No. of Factories
10–20	10	>10	100
20–30	20	>20	80
30–40	30	>30	50
40–50	40	>40	10

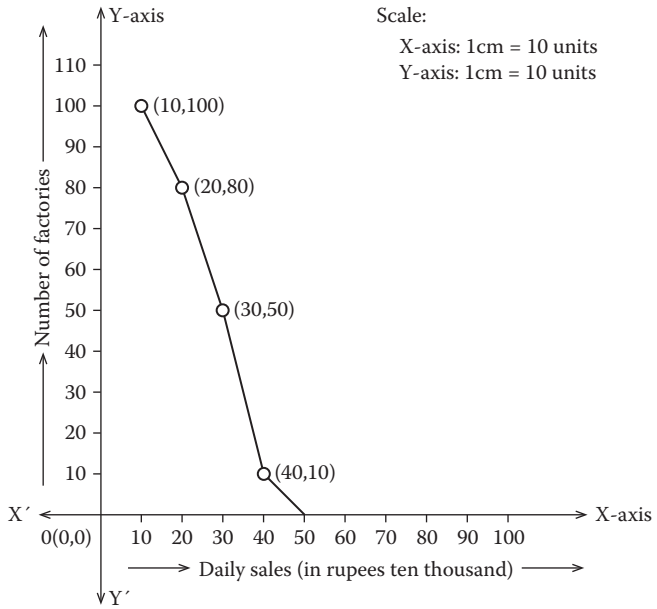


FIGURE 2.8
More than ogive.

The shape of a less than ogive curve would be a rising one, whereas the shape of a more than ogive curve should be a falling one.

2.13 Skewness and Kurtosis

These are the two characteristics of data sets. They provide useful trends and patterns in the data, represented as frequency distribution curves.

2.13.1 Skewness

Skewness is the extent to which a distribution of data points is concentrated at one end or the other, or the lack of symmetry in the curve.

2.13.1.1 Symmetrical Curves

A curve is said to be symmetrical when a vertical line drawn from the centre of the curve to the x-axis divides the area under the curve into equal parts.

2.13.1.2 Skewed Curve

A curve is said to be skewed when the values in the frequency distribution are concentrated more towards the left or right side of the curve, that is the values are not equally distributed from the centre of the curve.

2.13.1.3 Positively Skewed Curve

A curve is said to be positively skewed when the tail of the curve is more stretched towards the right side.

2.13.1.4 Negatively Skewed Curve

A curve is said to be negatively skewed when the tail of the curve is more stretched towards the left side.

2.13.2 Kurtosis

Kurtosis is the degree of peakness of a distribution of points; that is it measures the peakedness of a distribution. Two curves with the same central location and dispersion may have different degrees of kurtosis, that is curves with different Kurtosis but the same central location.

2.14 Summary

Statistical data is a set of facts expressed in quantitative form. The use of facts expressed as measurable quantities can help a decision maker to arrive at better decisions. Data can be obtained through a primary source or secondary source. When the data is collected by the investigator himself, it is called primary data. When the data has been collected by others, it is known as secondary data. A frequency distribution is the principal tabular summary of either discrete or continuous data. The frequency distribution may show actual, relative or cumulative frequencies. Actual and relative frequencies may be charted as a histogram, bar chart or frequency polygon. Two graphs of cumulative frequencies are less than ogive and more than ogive.

Presentation of data is provided through tables and charts.

REVIEW QUESTIONS

1. Distinguish between primary and secondary data. Discuss the various methods of collecting primary data. Indicate the situations in which each of these methods should be used.
2. Discuss the validity of the statement 'A secondary source is not as reliable as a primary source'.
3. Discuss the appropriateness of the methods of collecting data by
   a. Mailed questionnaire
   b. Personal interviews
4. Explain the advantages of direct personal investigation compared with the other methods generally used in collecting data.
5. Compare the different methods used in the collection of statistical data. Explain the importance of determining a statistical unit in the collection of data.

6. Discuss the various sources of secondary data. Point out the precautions to be taken while using such data.
7. Explain what precautions must be taken while drafting a useful questionnaire.
8. As the personnel manager in a particular industry, you are asked to determine the effect of increased wages on output. Draft a suitable questionnaire for this purpose.
9. If you were to conduct a survey regarding smoking habits among students of a degree college, what method of data collection would you adopt? Give reasons for your choice.
10. Explain the main points that you would keep in mind while editing primary data.
11. Distinguish between primary source and secondary source of statistical data. What precautions would you take before using data from a secondary source?
12. Discuss the validity of the statement 'A secondary source is not as reliable as a primary source'.
13. Define *secondary data*. State its chief sources and point out the dangers involved in its use and what precautions are necessary before using it.
14. Describe the primary and secondary methods of collecting data. In what special circumstances are the two methods suitable?
15. What do you understand by classification and tabulation? Discuss their importance.
16. Explain the terms *classification* and *tabulation*. Point out their importance in a statistical investigation. What precautions would you take in tabulating statistical data?
17. Explain the general principles of classification of data for forming an empirical frequency distribution of one variable.
18. What are the requisites of a good table? State the rules that serve as a guide in tabulating statistical material?
19. What are the chief functions of tabulation? What precautions would you take in tabulating statistical data?
20. Outline the considerations you will bear in mind in the construction of a frequency distribution.
21. Discuss the importance and drawbacks of diagrammatic representation of data. Discuss the usefulness of diagrammatic representation of facts.
22. What are the merits and demerits of diagrammatic representation of statistical data? Write short notes on any three important methods used for diagrammatic representation.
23. a. What, in your opinion, are the tests of a good diagram?
    b. Discuss the merits and demerits of diagrammatic representation of statistical data.
24. Explain briefly the purposes served by the diagrammatic representation of data.
    a. Charts or graphs are more effective in attracting attention than are any of the other methods of presenting data. Do you agree? Give reasons and illustrations.
    b. Diagrams help us visualise the whole meaning of a numerical complex at a single glance. Comment. What points should be taken into consideration while presenting a table diagrammatically?

3

Measures of Central Tendency

3.1 Introduction

One of the most important objectives of statistical analysis is to get one single value that describes the characteristic of the entire mass of unwieldy data. Such a value is called the central value, or an 'average'. The word *average* is very commonly used in day-to-day conversation. For example, we often talk of average students in a class, average height or weight, average income and so forth.

The objective here is to find one representative value which can be used to locate and summarise the entire set of varying values. This one value can be used to make many decisions concerning the entire set. We can define measures of central tendency (or location) to find some central value around which the data tend to cluster.

3.2 Significance of Measures of Central Tendency

The significance is to get one single value that describes the characteristics of the entire group. The central value is measured by condensing the mass of data into one single value, enabling us to get a bird's-eye view of the entire data. Thus, one value can represent thousands, lakhs and even millions of values.

To facilitate comparison, the central value is measured by reducing the mass of data into one single figure, enable comparisons to be made. Comparison can be made either at time points or over a period of time. For example, we can compare the percentage results of the students of different colleges on a certain examination.

Measures of central tendency, that is considering the mass of data in one single value, enable us to get an idea of the entire data. For example, it is impossible to remember the individual incomes of millions of earning people of any country. But if the average income is obtained, we get one single value that represents the entire population.

Measures of central tendency also enable us to compare two or more sets of data to facilitate comparison. For example, the average sales figures of June may be compared with the sales figures of previous months.

3.3 Properties of Good Measures of Central Tendency

A good measure of central tendency should possess, as far as possible, the following properties.

1. Easy to understand
2. Rigidly defined
3. Mathematically expressed (have a mathematical formula)
4. Readily comprehensible
5. Calculated based on all observations
6. Least affected by extreme fluctuations in sampling data
7. Suitable for further mathematical treatment
8. Simple to compute
9. Uniquely defined

In addition to the above requisites, a good average should also retain a maximum number of characteristics of the data; it should be a nearest value to all the data elements. Averages should be calculated for homogenous data, that is ages, sales and so forth.

Below are some of the important measures of central tendency which are commonly used in business and industry.

1. Mathematical averages
   a. Arithmetic mean
   b. Geometric mean
   c. Harmonic mean
2. Positional averages
   a. Median
   b. Mode

Of the above, arithmetic mean, median and mode are the widely used averages. Figure 3.1 shows the types of averages.

3.4 Arithmetic Mean

The arithmetic mean (AM) or mean is the simplest and most frequently used average. Arithmetic mean is represented by notation $\bar{x}$ (read x-bar).

3.4.1 Calculating the Mean from Ungrouped Data

The mean $\bar{x}$ of a collection of observations $x_1, x_2, \ldots, x_n$ is given by

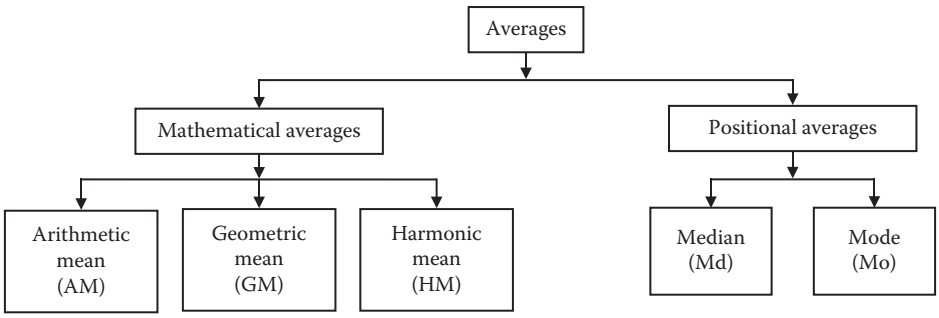


FIGURE 3.1
Types of averages.

$$\bar{x} = \left(\frac{1}{n}\right)(x_1 + x_2 + \ldots + x_n)$$

$$= \sum x/n$$

$$= \left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i$$

where:

$\bar{x}$ = sample mean
i = set of natural numbers
$\sum_{i=1}^{n} x_i$ = sum of values of all observations
n = number of elements
$\sum$ indicates that the values of x are summed together

When the mean is calculated for the entire population, it is known as the population arithmetic mean (μ). n is the number of elements (observations) in the population. Then, $\mu = \sum x/n$.

1. Calculation of arithmetic mean – individual series

Exercise

The haemoglobin levels of the 10 women given here are 12.5, 13, 10, 11.5, 11, 14, 9, 7.5, 10 and 12. Find the mean haemoglobin of the samples.

Solution

$$AM(\bar{x}) = \frac{12.5 + 13.10 + \ldots + 12}{10} = 11.05$$

Example 3.1: Absentee List of Drivers of the Transport Department over a Span of 60 Days

Table 3.1 shows the absentee list of drivers of the transport department.

Solution

When a manager wants to know the average number of days a driver is on leave in 60 days, he can calculate the mean of the ungrouped data as follows:

$$\bar{x} = \frac{\sum x}{n} = \frac{8+6+6+7+4+5+6+2+4+6}{10}$$

$$= 54/10$$

$$= 5.4 \text{ days per driver out of 60 days}$$

Here, the mean is calculated by adding every observation separately in no set order. This is ungrouped data. One can calculate the mean using the above method for limited values. But the task becomes difficult while calculating average for vast data, say for 5000 employees. In such cases, a frequency distribution of the data will be helpful to a manager, and the mean should be calculated using a different method.

Exercise

Find the mean for the following data:
2500, 2700, 2400, 2300, 2550, 2650, 2750, 2600, 2400

Solution

If we compute the arithmetic mean, then

$$\bar{X} = \frac{2500+2700+2400+2300+2550+2650+2750+2450+2600+2400}{10}$$

$$= \frac{25300}{10} = \text{Rs. } 2530$$

Therefore, the average monthly salary is Rs. 2530.

2. Calculation of arithmetic mean – step deviation method

$$\bar{x} = A + \frac{\sum d}{N}$$

TABLE 3.1

Absentee List of Drivers of the Transport Department

Driver	1	2	3	4	5	6	7	8	9	10
No. of days on leave	8	6	6	7	4	5	6	2	4	6

where:
$\bar{x}$ = arithmetic mean
A = assumed mean
$\sum d$ = sum of deviations
N = number of items

Exercise

Find the mean heights of the students.

Heights (inches):	64	65	66	67	68	69	70	71	72	73
No. of students:	1	6	10	22	21	17	74	5	3	1

Solution

Let the assumed mean (A) = 68.
Table 3.2 shows the calculation of the arithmetic mean (deviation method).

$$\bar{x} = A + \frac{\sum fdx}{\sum f}$$

$$\bar{x} = 68 + \frac{13}{100}$$

$$\bar{x} = 68 + 0.13$$

$$\bar{x} = 68.13$$

3. Calculation of arithmetic mean – continuous series (grouped data)

$$\bar{x} = \frac{\sum fx}{\sum f}$$

TABLE 3.2

Calculation of Arithmetic Mean (Deviation Method)

Heights (inches) (x)	No. of Students (f)	Deviation (dx = x – A = x – 68)	fdx
64	1	–4	–4
65	6	–3	–18
66	10	–2	–20
67	22	–1	–22
68 (A)	21	0	0
69	17	1	17
70	14	2	28
71	5	3	15
72	3	4	12
73	1	5	5
	N = $\sum f$ = 100		$\sum fdx$ = 13

Exercise

Calculate the mean age of the group of people shown in Table 3.3.

Solution

Table 3.4 shows the calculation of the mean.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{3762.5}{135} = 27.87$$

Therefore, the mean age of the people in this group will be 27.87 years.

Example 3.2

The data in Table 3.5 relate to the monthly sales of 200 firms.

Solution

For calculation of the arithmetic mean, see Table 3.6.

TABLE 3.3

Number of People

Class Interval Age	No. of People
15–20	15
20–25	20
25–30	40
30–35	60

TABLE 3.4

Calculation of Mean

Class Interval of Age	No. of People (f)	Midpoint = $\frac{\text{Lower Limit} + \text{Upper Limit}}{2}$	fx
15–20	15	17.5	262.5
20–25	20	22.5	450
25–30	40	27.5	1100
30–35	60	32.5	1950
	N = ∑f = 135		∑fx = 3762.5

TABLE 3.5

Monthly Sales of 200 Firms

Monthly Sales (Rs. in Thousands)	No. of Firms	Monthly Sales (Rs. in Thousands)	No. of Firms
300–350	5	550–600	25
350–400	14	600–650	22
400–450	23	650–700	7
450–500	50	700–750	2
500–550	52		

TABLE 3.6

Calculation of Arithmetic Mean

Monthly Sales (Rs. in Thousands)	Midpoint (x)	No. of Firms (f)	fx
300–350	325	5	1,625
350–400	375	14	5,250
400–450	425	23	9,775
450–500	475	50	23,750
500–550	525	52	27,300
550–600	575	25	14,375
600–650	625	22	13,750
650–700	675	7	4,725
700–750	725	2	1,450
		N = 200	∑fx = 102,000

$$\bar{X} = \frac{\sum fx}{N} = \frac{102000}{200} = 510$$

Hence, the average monthly sales are Rs. 510.

4. Step deviation method

$$\bar{x} = A + \frac{\sum fd'}{\sum f} \times i$$

where:

$\bar{x}$ = arithmetic mean
A = assumed mean
$\sum fd'$ = total of product of the step deviation and frequencies
$\sum f$ = total number of frequencies
i = common factor in x

Exercise

Calculate the mean for the following data:

Wages (Rs.):	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90
Frequency:	1	4	10	22	30	35	10	7	1

Solution

Let the assumed mean (A) = 55.
See Table 3.7 for calculation of the mean.

TABLE 3.7

Calculation of Mean (Step Deviation Method)

Class Interval	Frequency (f)	Midvalue (x)	dx = x – A	d = dx ÷ 10	fd′
0–10	1	5	–50	–5	–5
10–20	4	15	–40	–4	–16
20–30	10	25	–30	–3	–30
30–40	22	35	–20	–2	–44
40–50	30	45	–10	–1	–30
50–60	35	55 (A)	0	0	0
60–70	10	65	10	1	10
70–80	7	75	20	2	14
80–90	1	85	30	3	3
N = $\sum$f = 120					$\sum$fd′ = –98

$$AM(\bar{x}) = A + \frac{\sum fd'}{\sum f} \times i$$

$$= 55 + \frac{(-98)}{120} \times 10$$

$$= 55 - 8.17$$

$$= 46.83$$

Merits of arithmetic mean

1. Easy to calculate and simple to understand.
2. Concept is familiar and clear to most people.
3. Based on all the observations.
4. Capable of further algebraic treatment.
5. Not necessary to arrange the observation in ascending–descending order.
6. It is rigidly defined.
7. Every data set has one and only one mean.
8. Provides a good basis for comparison. For example, if a manager wants to compare the performance of salesmen of four different regions of the state, the arithmetic average provides the correct basis for assessing the relative efficiency of the regions.

Demerits of arithmetic mean

1. Highly affected by the extreme items. The mean of 2, 6 and 301 is 103, and none of the values are adequately represented by the mean 103.
2. It may be affected by highly fluctuating values that are not far from other values of the group. Observe that if the units produced in a day by five workers are as shown in Table 3.8, the mean units produced per day is

TABLE 3.8

Number of Units Produced by Workers in a Day

Workers	1	2	3	4	5
Units	22	23	21	24	4

$$\mu = \sum x/n = (22 + 23 + 21 + 24 + 4)/5 = 18.8 \text{ units}$$

When the mean units are calculated omitting the fifth worker (i.e. 5), the mean is 22.5 units. Thus, one extreme value (4) has affected the mean. Hence, it is more appropriate to calculate the mean excluding the extreme value in order to make it more representative.

3. In some cases, the arithmetic mean may give misleading impressions. For example, the average number of patients admitted in a hospital is 9.8 per day. Here, the mean is useful information, but it does not represent the actual item.
4. It is very difficult to find the actual mean (using $\mu = \sum x/n$).
5. It can hardly be located by inspection.
6. It cannot be calculated even if one value is missing.
7. We cannot calculate the mean for a data set with open-ended classes at either end of the scale. A class that allows either the upper or lower end of a quantitative classification scheme to be limitless is called as open-ended class.

3.4.2 Mathematical Properties of Arithmetic Mean

Because the arithmetic is defined operationally, it has several useful mathematical properties. Some of these are as follows:

1. The sum of deviations of the observations from the arithmetic mean is always zero. Symbolically, it is

$$\sum \left(X - \bar{X}\right) = 0$$

It is because of this property that the mean is characterised as a point of balance; that is the sum of positive deviations from mean is equal to the sum of the negative deviations from mean.

2. The sum of squared deviations of the observations from the mean is minimum; that is the total of the squares of the deviations from any other value than the mean value will be greater than the total sum of squares of deviations from the mean. Symbolically,

$$\sum \left(X - \bar{X}\right)^2 \text{ is a minimum}$$

3. The arithmetic means of several sets of data may be combined into a single arithmetic mean for the combined sets of data. For two sets of data, the combined arithmetic mean may be defined as

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

where:
$\bar{X}_{12}$ = combined mean of two sets of data
X_1 = arithmetic mean of the first set of data
X_2 = arithmetic mean of the second set of data
N_1 = number of observations in the first set of data
N_2 = number of observations in the second set of data

If we have to combine three or more than three sets of data, then the same formula can be generalised as

$$\bar{X}_{123\ldots} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3 + \ldots}{N_1 + N_2 + N_3 + \ldots}$$

The arithmetic mean has the great advantages of being easily computed and readily understood. This is due to the fact that it possesses almost all the properties of good measures of central tendency. No other measure of central tendency possesses so many properties. However, the arithmetic mean has some disadvantages. The major disadvantage is that its value may be distorted by the presence of extreme values in a given set of data. A minor disadvantage is when it is used for open-ended distribution since it is difficult to assign a midpoint value to an open-ended class.

3.4.3 Weighted Arithmetic Mean

The weighted mean is calculated taking into account the relative importance of each of the values to the total value. Consider, for example, the manufacturing company in Table 3.9 that employs three grades of labour (unskilled, semi-skilled and skilled) to produce two products. When the company wants to know the average cost of labour per hour for each product, the simple arithmetic average of the labour wage will be

$$\bar{x} = \left(\sum x\right)\Big/ n$$
$$= (10 + 15 + 20)/3 = \text{Rs. } 15/\text{hour}$$

TABLE 3.9
Labour Capital Involved in Manufacturing Two Products

		Labour Hours per Unit	
Class of Labour	**Wage per Hour (Rs.) (x)**	**Product 1**	**Product 2**
Unskilled	10	2	6
Semi-skilled	15	3	2
Skilled	20	5	1

When the above average wage per hour is taken to calculate the labour cost of one unit of Product 1, the value would be 15 (2 + 3 + 5) = Rs. 150.

And, labour cost for one unit of Product 2 is 15 (6 + 2 + 1) = Rs. 135.

But when calculated using simple arithmetic averages, these values are incorrect, as they do not take into consideration the fact that different amounts of each class of labour are used. The correct value (cost per one unit) can be determined in the following manner:

For Product 1, the total labour cost per unit = (10 × 2) + (15 × 3) + (20 × 5)
= Rs. 165

And, cost per hour = Rs. 165/(2 + 3 + 5)
Rs. 16.5/hour

For Product 2, the total labour cost per unit = (10 × 6) + (15 × 2) + (20 × 1)
= Rs. 110

And, cost per hour = Rs. 110/(6 + 2 + 1)
= Rs. 12.22/hour

The arithmetic mean, as discussed earlier, gives equal importance (or weight) to each observation. In some cases, all the observations do not have the same importance. When this is so, we compute the weighted arithmetic mean. The weighted arithmetic mean can be defined as

$$\bar{X}_W = \frac{\sum WX}{\sum W}$$

where:
$\bar{X}_W$ = weighted arithmetic mean
W = weights assigned to the variable X

You are familiar with the use of weighted averages to combine several grades that are not equally important.

Exercise

Assume that the grades consist of one final examination and two midterm assignments. If each of the three grades is given a different weight, then the procedure is to multiply each grade (X) by its appropriate weight (W). If the final examination is 50% of the grade and each midterm assignment is 25%, then the weighted arithmetic mean is given as follows:

$$\bar{X}_W = \frac{\sum WX}{\sum W} = \frac{W_1X_1 + W_2X_2 + W_3X_3}{W_1 + W_2 + W_3}$$

$$= \frac{50X_1 + 25X_2 + 25X_3}{50 + 25 + 25}$$

TABLE 3.10

Calculation of Weighted Arithmetic Mean

	Grade (X)	Weight (W)	WX
Final examination	80	50	4000
First assignment	95	25	2375
Second assignment	85	25	2125
		$\sum W = 100$	$\sum WX = 8500$

Suppose you got 80 on the final examination, 95 on the first midterm assignment and 85 on the second midterm assignment; then

$$\bar{X}_W = \frac{50(80) + 25(95) + 25(85)}{100}$$

$$= \frac{4000 + 2375 + 2125}{100} = \frac{8500}{100} = 85$$

Table 3.10 shows the calculation of the weighted arithmetic mean.

$$\bar{X}_W = \frac{\sum WX}{\sum W} = \frac{8500}{100} = 85$$

The concept of weighted arithmetic mean is important because the computation is the same as that used for averaging ratios and determining the mean of grouped data.

The weighted mean is especially useful in problems relating to the construction of index numbers.

3.5 Median

The median, as the name suggests, is the middle value of a series arranged in any order of magnitude.

A second measure of central tendency is the median. Median is that value which divides the distribution into two equal parts. Fifty percent of the observations in the distribution are above the value of median, and the other 50% of the observations are below this value. The median is the value of the middle observation when the series is arranged in order of size or magnitude. If the number of observations is odd, then the median is equal to one of the original observations. If the number of observations is even, then the median is the arithmetic mean of the two middle observations; for example if the incomes of seven persons are 1100, 1200, 1350, 1500, 1550, 1600 and 1800, then the median income would be Rs. 1500. Suppose one more person joins and his income is Rs. 1850; then the median income of eight persons would be

$$\frac{1500 + 1550}{2} = 1525$$

(Since the number of observations is even, the median is the arithmetic mean of the fourth and fifth persons.)

As it is distinct from the arithmetic mean, which is calculated from the value of every item in the series, it is called a positional average. The term *position* refers to the place of value in the series. The median is just the 50th percentile value below which 50% of the values in a sample fall. The object of median is therefore not merely to fix a value that shall be representative of a set, but also to establish a dividing line separating the higher from the lower values.

1. It is especially useful in the case of open-ended classes since only the position and not the values of items must be known.
2. It is not influenced by the magnitude of extreme deviations from it. For example, the median of 10, 20, 30, 40 and 50 would be 30, whereas the mean is 50. Hence, very often when extreme values are present in a set of observations, the median is a more satisfactory measure of the central tendency than the mean.
3. It is not appropriate average in dealing with quantitative data, that is where ranks are given or there are other types of items that are not counted or measured but are scored.

3.5.1 Calculating the Median from Ungrouped Data

If the data set contains an odd number of items, the middle item of the array is the median. If there is an even number of items, the median is the average of the two middle items. If the total of the frequencies is odd, say n, the value of [(n + 1)/2]th item gives the median, and when the total of the frequencies is even, say 2n, then nth and (n + 1)th are two central items and the arithmetic mean of those two items gives the median.

If a group of N observations is arranged in ascending or descending order of magnitude, then the middle value is called the median of these observations.

The middle value divides the number of observations in the data into two equal parts.

It is also called the positional average.

1. For individual observation

$$\text{Median} = \text{value of}\left(\frac{N+1}{2}\right)\text{item}$$

Exercise

The following data gives the weight of seven people in pounds. What is the median of the data?

158, 167, 143, 169, 172, 146, 151

Solution

Table 3.11 shows computation of the median.

TABLE 3.11

Computation of Median

Serial No.	Size of Items, Ascending Order (x_i)		Size of Items, Descending Order (x_i)
1	143		172
2	146		169
3	151		167
4	158	← Median →	158
5	167		151
6	169		146
7	172		143

$$\text{Median} = \text{value of}\left(\frac{N+1}{2}\right)\text{th item}$$

$$\text{Median} = \text{value of}\left(\frac{7+1}{2}\right)\text{th item}$$

$$= \frac{8}{2} = 4\text{th items}$$

Therefore, the fourth item corresponds to 158.
Thus, the median = 158.

2. Calculation of median – discrete series (ungrouped data)

Exercise

Calculate the median for the following:

Income/day (x_i):	100	150	200	250	300	350
No. of households (f_i):	05	19	03	11	06	09

Solution

Table 3.12 shows calculation of the median.

TABLE 3.12

Calculation of Median

x_i	f_i	c.f.	
100	05	05	
150	19	05 + 19 = 24	
200 ← **median**	03	24 + 03 = 27	← **C.f. of median**
250	11	27 + 11 = 38	
300	06	38 + 06 = 44	
350	09	44 + 09 = 53	
	$N = \sum f_i = 53$		

Now,

$$\text{Median} = \text{size of}\left(\frac{N+1}{2}\right)\text{th item}$$

$$= \text{size of}\left(\frac{53+1}{2}\right)\text{th item}$$

$$= 27\text{th item}$$

27th item corresponds to c.f. 27 for the value 200 of x_i.
So, median = 200.

Exercise

Find the median of the following:

X_i:	–1	2	3	4	5	6
f_i:	–7	12	17	19	21	24

Solution

Calculation of the median is shown in Table 3.13.
Now,

$$\text{Median} = \text{size of}\left(\frac{N+1}{2}\right)\text{th item}$$

$$= \text{size of}\left(\frac{100+1}{2}\right)\text{th item}$$

$$= \text{size of } 50.5\text{th item}$$

50.5 lies in c.f. 55 for the value 4.
So, median = 4.

TABLE 3.13

Calculation of Median

x_i	f_i	c.f.	
1	7	7	
2	12	19	
3	17	36	
4 ← **median**	19	55	← **c.f. of median**
5	21	76	
6	24	100	
	$N = \sum f_i = 100$		

3. Calculation of median – continuous series (grouped data)

$$\text{Median } (M) = l + \left(\frac{\frac{N}{2} - \text{c.f.}}{f} \right) \times i$$

where:
M = median
N = total frequency
c.f. = cumulative frequency of the class proceeding the median class
i = width of the median class
f = frequency of the median class
l = lower limit of the median class

Exercise

Calculate the median of the following:

Income/day (x_i):	100–150	150–200	200–250	250–300	300–350	350–400
No. of households (f):	05	19	03	11	06	09

Solution

Calculation of the median is shown in Table 3.14.
Now,

$$\text{Median} = \text{size of} \left(\frac{N+1}{2} \right) \text{th item}$$

$$= \text{size of} \left(\frac{53+1}{2} \right) \text{th item}$$

$$= 27\text{th item}$$

27 lies in c.f. 27 for the value 200–250.
So, median class = 200–250.

TABLE 3.14

Calculation of Median

x_i	f_i	c.f.	
100–150	05	05	
150–200	19	05 + 19 = 24	← **c.f. of the class preceding median class**
200–250 ← **median**	03 ← **f**	24 + 03 = 27	← **c.f. of median**
250–300	11	27 + 11 = 38	
300–350	06	38 + 06 = 44	
350–400	09	44 + 09 = 53	
	$N = \sum f_i = 53$		

Hence,

$$M = l + \left(\frac{\frac{N}{2} - c.f.}{f} \right) \times i$$

$$M = 200 + \left(\frac{26.5 - 24}{3} \right) \times 50$$

$$= 241.66$$

Therefore, the median is 241.66.

Merits of median

1. The median is not strongly affected by the extreme or abnormal values. Hence, median is a better average than mean.
2. It is useful in the case of open-ended and unequal classes.
3. It can be located by inspection.
4. It can be determined graphically.
5. It is defined rigidly.
6. The median is easy to understand, and it can be computed from any kind of data (even for grouped data with open-ended classes, but excluding the case when the median falls in the open-ended class).
7. The median can also be calculated for qualitative data.

Demerits of median

1. It is a time-consuming process, as it is required to arrange the data before calculating the median.
2. It is not based on all observations.
3. It is not capable of further algebraic treatment.
4. It is affected more by sampling fluctuations.
5. It is difficult to compute the median for a data set with a large number of observations. Therefore, to find an estimate of a large population, mean is easier to use than median.

3.5.2 Mathematical Properties of Median

The important mathematical property of the median is that the sum of the absolute deviations about the median is a minimum. In symbols, $\sum |X - \text{Median}| = \text{a minimum}$.

Although the median is not as popular as the arithmetic mean, it does have the advantage of being both easy to determine and easy to explain.

As illustrated earlier, the median is affected by the number of observations rather than the values of the observations; hence, it will be less distorted as a representative value than the arithmetic mean.

An additional advantage of median is that it may be computed for an open-ended distribution.

The major disadvantage of median is that it is a less familiar measure than the arithmetic mean. However, since median is a positional average, its value is not determined by each and every observation. Also, median is not capable of algebraic treatment.

3.6 Quantiles

Quantiles are the related positional measures of central tendency. These are useful and frequently employed measures of non-central locations. The most familiar quantiles are the quartiles, deciles and percentiles.

3.6.1 Quartiles

Quartiles are those values which divide the total data into four equal parts. Since three points divide the distribution into four equal parts, we will have three quartiles. Let us call them Q_1, Q_2 and Q_3. The first quartile, Q_1, is the value such that 25% of the observations are smaller and 75% of the observations are larger. The second quartile, Q_2, is the median; that is 50% of the observations are smaller and 50% of the observations are larger. The third quartile, Q_3, is the value such that 75% of the observations are smaller and 25% of the observations are larger.

For grouped data, the following formulae are used for quartiles:

$$Q_j = L + \frac{j\frac{N}{4} - pcf}{f} \times i \quad \text{for } j = 1, 2, 3$$

where:

L = lower limit of the quartile class
pcf = preceding cumulative frequency to the quartile class
f = frequency of the quartile class
i = size of the quartile class

3.6.2 Deciles

Deciles are those values which divide the total data into 10 equal parts. Since nine points divide the distribution into 10 equal parts, we will have nine deciles denoted by $D_1, D_2, \ldots, D_9$.

For grouped data, the following formulae are used for deciles:

$$D_k = L + \frac{k\frac{N}{10} - pcf}{f} \times i \quad \text{for } k = 1, 2, \ldots, 9$$

where the symbols have the usual meaning and interpretation.

3.6.3 Percentiles

Percentiles are those values which divide the total data into 100 equal parts. Since 99 points divide the distribution into 100 equal parts, we shall have 99 percentiles denoted by

$$P_1, P_2, \ldots, P_{99}$$

For grouped data, the following formulae are used for deciles:

$$P_I = L + \frac{I\frac{N}{100} - pcf}{f} \times i \quad \text{for } I = 1, 2, \ldots, 99$$

Exercise

The grouped data in Table 3.15 relate to the profits of 100 companies during the year 1987–1988.

Calculate Q_1, Q_2 (median), D_6 and P_{90} from the given data and interpret these values.

Solution

The calculation of Q_1, Q_2, D_6 and P_{90} is shown in Table 3.16.

Q_1 = size of (N/4)th observation = (100/4) 25th observation

which lies in the class 40–50.

$$Q_1 = L + \frac{\frac{N}{4} - pcf}{f} \times i = 40 + \frac{25 - 12}{18} \times 10 = 40 + 7.22 = 47.22$$

TABLE 3.15

Profits of 100 Companies

Profits (Rs. in Lakhs)	No. of Companies	Profits (Rs. in Lakhs)	No. of Companies
20–30	4	60–70	15
30–40	8	70–80	10
40–50	18	80–90	8
50–60	30	90–100	7

TABLE 3.16

Calculation of Q_1, Q_2, D_6 and P_{90}

Profit (Rs. in Lakhs)	No. of Companies (f)	c.f.
20–30	4	4
30–40	8	12
40–50	18	30
50–60	30	60
60–70	15	75
70–80	10	85
80–90	8	93
90–100	7	100

This value of Q_1 suggests that 25% of the companies earn an annual profit of Rs. 47.22 lakh or less.

Median or Q_2 = size of (N/2)nd observation = 100/2th observation = 50th observation

which lies in the class 50–60.

$$Q_2 = L + \frac{\frac{N}{2} - pcf}{f} \times i = 50 + \frac{50 - 30}{30} \times 10 = 50 + 6.67 = 56.67$$

This value of Q_2 (or median) suggests that 50% of the companies earn an annual profit of Rs. 56.67 lakh or less, and the remaining 50% of the companies earn an annual profit of Rs. 56.67 lakh or more.

D_6 = size of (6N/10)th observation = (6*100/10) the observation = 60th observation

which lies in the class 50–60.

$$D_6 = L + \frac{\frac{6N}{10} - pcf}{f} \times i = 50 + \frac{60 - 30}{30} \times 10 = 50 + 10 = 60$$

Thus, 60% of the companies earn an annual profit of Rs. 60 lakh or less and 40% of companies earn Rs. 60 lakh or more.

P_{90} = size of (90N/100)th observation = (90*100/100)th observation = 90th observation

which lies in the class 80–90.

$$P_{90} = L + \frac{\frac{90N}{100} - pcf}{f} \times i = 80 + \frac{90 - 85}{10} \times 10 = 80 + 5 = 85$$

This value of 90th percentiles suggests that 90% of the companies earn an annual profit of Rs. 85 lakh or less and 20% of companies earn Rs. 85 lakh or more.

3.7 Mode

It is the value which has the highest frequency in the data. Mode is defined as the value of the variable which occurs most frequently in the data set. The mode of the distribution is around which the items tend to be most heavily concentrated. It is denoted by Z or M_o.

The mode is the typical or commonly observed value in a set of data. It is defined as the value which occurs most often or with the greatest frequency. The dictionary meaning of the term *mode* is 'most usual'. For example, in the series of numbers 3, 4, 5, 5, 6, 7, 8, 8, 8 and 9, the mode is 8 because it occurs the maximum number of times.

1. Calculation of mode – individual series

Exercise

Calculate mode for data on the amount of sugar purchases by people from a shop:

3, 1, 7, 4, 1, 2, 5, 3, 4, 6, 5, 5, 4, 4, 3, 5, 2, 4

Solution

Calculation of mode is shown in Table 3.17.
Therefore, the item 4 occurs the maximum number of times, that is 5.
Thus, mode = 4.

NOTE: When there are two or more values having the same maximum frequency, the mode is said to be ill-defined. Such a series is also known as bi-model or multi-model.

Exercise

Calculate mode from the following data of marks obtained by 10 students:

10, 27, 24, 12, 27, 27, 20, 18, 15, 30

Solution

Calculation of mode is shown in Table 3.18.
Therefore, item 27 occurs the maximum number of times, that is 3.
Thus, the modal marks are 27.

2. Calculation of mode – continuous series (grouped data)

Calculating the mode from ungrouped data

Table 3.19 shows the weights of 18 workers of an organisation. The mode is 68, as it repeats five times (more than other values).

TABLE 3.17

Calculation of Mode

Size of Item	Number of Times It Occurs (Tally Marks)	Frequency (f_i)
1	II	2
2	II	2
3	III	3
Mode → 4	𝍸	5
5	IIII	4
6	I	1
7	I	1
		$N = \sum f_i = 18$

TABLE 3.18

Calculation of Mode

Size of Item	Tally Marks	Frequency (f_i)
10	|	1
12	|	1
15	|	1
18	|	1
20	|	1
24	|	1
27	|||	3
30	|	1
		$N = \sum f_i = 10$

TABLE 3.19

Weights of 18 Workers

58	60	62	56	59	56	68	68	70
68	58	59	60	69	68	68	63	61

$$M_o \text{ or } Z = L + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times i$$

where:

L = lower limit of the modal class
f_0 = frequency of modal class
f_1 = frequency of the class preceding the modal class
f_2 = frequency of the class succeeding the modal class
i = width of the class interval

Exercise

Identify the mode in the following frequency table:

Income/day (x_i):	100–150	150–200	200–250	250–300	300–350	350–400
No. of households (f_i):	05	19	03	11	06	09

Solution

Here, by inspection

$$f_0 = 19, f_1 = 5, f_2 = 3$$

The class interval (150–200) has the maximum frequency, that is 19.
Therefore, the modal class is 150–200.
Now,

$$M_o = L + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times i$$

$$= 150 + \frac{19 - 5}{2(19) - 5 - 3} \times 50 = 173.33$$

Hence, the mode is 173.33.

Merits of mode

1. Mode can be used as a central location for qualitative as well as quantitative data.
2. It is not affected by extreme values. It can also be used even when the classes are open-ended.
3. It is usually an actual value, as it occurs most frequently in the series.
4. Its value can be determined graphically.

Demerits of mode

1. At times, a data set contains no value that occurs more than once. Further, all values in a data set might occur an equal number of times; that is all items have the same frequency.
2. It cannot always be determined, as some data sets contain two, three or many modes, making it difficult to interpret them.
3. It is not based on all observations.
4. It is not capable of further algebraic treatment.

3.8 Relationship among Mean, Median and Mode

A distribution in which mean, median and mode coincide is known as a symmetrical (bell-shaped) distribution. If a distribution is skewed (i.e. not symmetrical), then mean, median and mode are not equal. In a moderately skewed distribution, a very interesting relationship exists among mean, median and mode. In such type of distributions, it can be proved that the distance between the mean and median is approximately one-third of the distance between the mean and mode. This is shown below for two types of such distributions.

In the case of a symmetrical distribution, the mean, median and mode coincide. However, according to Karl Pearson, if the distribution is moderately asymmetrical, the mean, median and mode are related in the following manner:

$$\text{Mean} - \text{Median} = (\text{Mean} - \text{Mode})/3$$

Thus,

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

Similarly, we can express the approximate relationship for median in terms of mean and mode. Also, this can be expressed for mean in terms of median and mode. Thus, if

we know any of the two values of the averages, the third average value can be determined from this approximate relationship.

Exercise

For a moderately skewed distribution in which the mean and median are 35.4 and 34.3, respectively, calculate the value of the mode.

Solution

To compute the value of the mode, we use the approximate relationship

$$\begin{aligned}\text{Mode} &\approx 3\,\text{Median} - 2\,\text{Mean}\\ &= 3\,(34.3) - 2\,(35.4)\\ &= 102.9 - 70.8 = 32.1\end{aligned}$$

Therefore, the value of the mode is 32.1.

Exercise

Median = 139.69
Mean = 139.51
Calculate the mode.

Solution

Mode = 3Median – 2 mean

Given: Median = 139.69
Mean = 139.51

Substituting the values, we get

$$\begin{aligned}\text{Mode} &= 3\,(139.69) - 2\,(139.51)\\ &= 419.07 - 279.02\\ &= 40.05\end{aligned}$$

3.9 Comparison of Mean and Median

See Table 3.20 for a comparison of the arithmetic mean and median.

3.10 Geometric Mean

Managers often come across quantities that change over a period of time, and may need to know the average rate of change over this period. Arithmetic mean is inaccurate in tracing such a change. Hence, a new measure of central tendency is needed to calculate the change rate – the 'geometric mean'.

TABLE 3.20

Comparison of Arithmetic Mean and Median

Serial No.	Arithmetic Mean	Median
1.	It is calculated value, and not based on position in the series.	It is especially useful in the case of open-ended classes, since only the position, and not values of items, must be known. The median is also recommended if the distribution has unequal classes, since it is much easier to compute than the mean.
2.	It is affected by the value of every item in the series.	It is not influenced by the magnitude of extreme deviations from it.
3.	It cannot be graphically ascertained.	It can be determined graphically.
4.	Being determined by a rigid formula, it lends itself to subsequent algebraic treatment better than the median.	It is not capable of further algebraic treatment.

TABLE 3.21

Growth Rate of Textile Units

Year	1	2	3	4	5
Growth rate (%)	7	8	10	12	18

$$\text{GM} = \sqrt[n]{\text{product of all values}}$$

$$= \sqrt[n]{x_1, x_2, \ldots\ldots, x_n}$$

$$= \left(x_1, x_2, x_3 \ldots x_n\right)^{1/n}$$

where n is the number of values.

Geometric mean is applicable in many cases. Its use in calculating the growth rates of a textile unit in the southern region for the last 5 years is given in Table 3.21.

$$\text{The geometric mean} = \sqrt[n]{x_1, x_2, \ldots\ldots, x_n}$$

where $x_1, x_2, x_3, \ldots, x_n$ are the terms of the growth factor and are equal to 1 + (rate/100).

$$\text{GM} = \sqrt[5]{1.07 \times 1.08 \times 1.10 \times 1.12 \times 1.18} = 1.1093$$

1.1093 is the average growth factor.
The growth rate is calculated as 1.1093 – 1 = 0.1093.
Then, 0.1093 × 100 = 10.93.
So, the growth rate is 10.93% per year.

The geometric mean, like the arithmetic mean, is a calculated average. The geometric mean (GM) of a series of numbers, $X_1, X_2, \ldots, X_n$, is defined as

$$GM = \sqrt{X_1\, X_2\, X_3 \ldots X_n}$$

or the nth root of the product of N observations.

When the number of observations is three or more, the task of computation becomes quite tedious. Therefore, a transformation into logarithms is useful to simplify calculations. If we take logarithms of both sides, then the formula for GM becomes

$$\text{Log } M = \frac{1}{N}\left(\log X_1 + \log X_2 + \ldots + \log X_n\right)$$

And therefore,

$$GM = \text{Antilog}\left(\frac{\sum \log X}{N}\right)$$

For the grouped data, the geometric mean is calculated with the following formula:

$$GM = \text{Antilog}\left(\frac{\sum f \log X}{N}\right)$$

where the notation has the usual meaning.

Geometric mean is especially useful in the construction of index numbers. It is an average, and is most suitable when large weights have to be given to small values of observations, and small weights to large values of observations. This average is also useful in measuring the growth of a population.

The following example illustrates the use and computations involved in the geometric mean.

Example 3.2: Computation of Geometric Mean

A machine was purchased for Rs. 50,000 in 2004. Depreciation on the diminishing balance was charged at 40% in the first year, 25% in the second year and 15% per annum during the next 3 years. What is the average depreciation charged during the whole period?

Solution

Since we are interested in finding the average rate of depreciation, the geometric mean will be the most appropriate average.

Calculation of the geometric mean is shown in Table 3.22.

$$GM = \text{Antilog}\left(\frac{\sum \log X}{N}\right)$$

$$= \text{Antilog}\left(\frac{9.44144}{5}\right)$$

$$= \text{Antilog } 1.8883 = 77.32$$

The diminishing value being Rs. 77.32, the depreciation will be 100 – 77.32 = 22.6%.

TABLE 3.22

Calculation of Geometric Mean

Year	Diminishing Value (for a Value Rs. 100) (X)	log X
2004	100–40 = 60	1.77815
2005	100–25 = 75	1.87506
2006	100–15 = 85	1.92941
2007	100–15 = 85	1.92941
2008	100–15 = 85	1.92941
		$\sum \log X = 9.44144$

Merits of geometric mean

1. The geometric mean is very useful in averaging ratios and percentages.
2. It helps in determining the rates of increase and decrease.
3. It is capable of further algebraic treatment, so that a combined geometric mean can easily be computed.

Demerits of geometric mean

1. Compared with the arithmetic mean, the geometric mean is more difficult to compute and interpret.
2. The geometric mean cannot be computed if any observation has either a zero value or negative.

3.11 Harmonic Mean

Harmonic mean is based on the reciprocals of numbers averaged. It is defined as the reciprocal of the arithmetic mean of their reciprocals of the given individual observations. Thus, by definition,

$$HM = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_n}}$$

where X_1, X_2, X_3 and so forth refer to various items of the variable, and N refers to the total number of items.

Exercise

Mohan Plastics Ltd. got a raw material delivery order from Blowplast, Inc. However, the condition was that the delivery had to be made within 4 hours, failing which the order would be considered cancelled. Robert, the salesman at Mohan, was assigned the

responsibility to make the delivery. Robert had to be careful not to exceed the 80 kph speed limit; otherwise, he would be flouting the traffic rules. The marketing manager asked him not to go below 60 kph, as there was a risk of the order being cancelled. Robert divided his journey time into 4 hours. He travelled the first quarter of the distance at the speed of 50 kph, the second quarter of the distance at 65 kph and the last quarter of the distance at 55 kph.

He was successful in delivering the product on time. If his average speed was 60.5 kph, what was his speed when he covered the third quarter of the distance?

Solution

Let the speed of Robert's vehicle in the first hour, second hour, third hour and fourth hour be X_1, X_2, X_3 and X_4, respectively.

Let the average speed of Robert's whole journey from Mohan to Blowplast be $H_0 = 60.5$.

From the given information in the problem, we have

$$X_1 = 50$$

$$X_2 = 65$$

$$X_4 = 55$$

$$H_0 = 60.5$$

$$N = 4$$

After inserting the values in the formula for calculating the harmonic mean, we get

$$HM = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_n}}$$

$$60.5 = \frac{4}{\frac{1}{50} + \frac{1}{65} + \frac{1}{X_3} + \frac{1}{55}}$$

$$\frac{1}{4}\left[\frac{1}{50} + \frac{1}{65} + \frac{1}{X_3} + \frac{1}{55}\right] = \frac{1}{4}\left[0.0535664 + \frac{1}{X_3}\right] = \left[\frac{1}{60.5}\right]$$

$$\frac{1}{4}\left[0.0535664\right] + \frac{1}{4}\left[\frac{1}{X_3}\right] = 0.0165289$$

$$0.0133916 + \frac{1}{4X_3} = 0.0165289$$

$$\frac{1}{4X_3} = 0.0031373$$

$$\frac{1}{X_3} = 0.0125492$$

$$X_3 = 79.68 \approx 80$$

Thus, in the third hour Robert travelled at the speed of 80 kph.

The harmonic mean is a measure of central tendency for data expressed as rates, such as kilometres per hour, tonnes per day and kilometres per litre. The harmonic mean is defined as the reciprocal of the individual observations and can be represented by the following formula:

$$HM = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_N}} = \frac{N}{\sum\left(\frac{1}{X}\right)}$$

For example, the harmonic mean of 2, 3 and 4 is

$$HM = \frac{3}{\frac{1}{2} + \frac{1}{3} + \frac{1}{4}} = \frac{3}{\frac{13}{12}} = \frac{36}{13} = 2.77$$

For grouped data, the formula becomes

$$HM = \frac{N}{\sum\left(\frac{f}{X}\right)}$$

The harmonic mean is useful for computing the average rate of the increase of profits, or the average speed at which a journey has been performed, or the average price at which an article has been sold. Otherwise, its field of application is really restricted.

Exercise

In a factory, a unit of work is completed by A in 4 minutes, by B in 5 minutes, by C in 6 minutes, by D in 10 minutes and by E in 12 minutes. Find the average number of units of work completed per minute.

Solution

Calculations for computing the harmonic mean are shown in Table 3.23.

Hence, the average number of units computed per minute is 6.25.

The harmonic mean, like arithmetic mean and geometric mean, is computed from each and every observation. It is especially useful for averaging rates.

However, harmonic mean cannot be computed when one or more observations have a zero value or when there are either positive or negative observations. In dealing with business problems, harmonic mean is rarely used.

TABLE 3.23

Calculations for Computing Harmonic Mean

X	1/X
4	0.250
5	0.200
6	0.167
10	0.100
12	0.083
	$\sum 1/X = 0.8$

3.12 Summary

Central tendency is a statistical measure that takes one number as the representative of a group. It is an economical estimate of the general characteristics of the group. There are three main measures: mean, median and mode. Mean is the average of all the observations. The median is the middle of a distribution: half the observations are above the median and half are below the median. The median is less sensitive to extreme scores than the mean. This makes it a better measure than the mean for a data set with extreme values. The mode is the most frequently occurring observation in a distribution.

Measures of central tendency give one of the very important characteristics of data. Any one of the various measures of central tendency may be chosen as the most representative or typical measure. The arithmetic mean is widely used and understood as a measure of central tendency. The concepts of weighted arithmetic mean, geometric mean and harmonic mean are useful for specific types of applications. The median is generally a more representative measure for open-ended distributions and highly skewed distributions. The most demanded or customary value is needed.

REVIEW QUESTIONS

1. a. Distinguish between arithmetic mean and median?
   b. Define *measure of central te*ndency.
   c. What are the applications of central tendency?
   d. State the importance of median.
2. What are the various measures of central tendency studied in this unit? Explain the difference between them.
3. Discuss the mathematical properties of arithmetic mean and median.
4. Review the advantages and disadvantages of each measure of central tendency.
5. Explain how you will decide which average to use in a particular problem.
6. What are quantiles? Explain and illustrate the concepts of quartiles, deciles and percentiles.

SELF-PRACTICE PROBLEMS

1. Median = 139.69, mean = 139.51. Calculate the mode.
2. Find the mean of the following:

Marks:	0–10	10–20	20–30	30–40	40–50	50–60	60–70
Students:	6	5	8	15	7	6	3

3. The frequency distribution of weight in grams of mangoes of a given variety is given below. Calculate arithmetic mean and median and interpret your answer.

Weight in grams	410– 419	420– 429	430– 439	440– 449	450–459	460– 469	470– 479
No. of mangoes	14	20	42	54	45	18	7

4. What is quartile deviation? Calculate quartile deviation and its coefficient on the basis of the following data:

Marks	10	11	12	13	14	15
No. of students	4	6	7	8	6	4

5. Calculate median and arithmetic mean from the following data:

   0.5 men get less than Rs. 5

   12 men get less than Rs. 10

   22 men get less than Rs. 10

   30 men get less than Rs. 20

   36 men get less than Rs. 25

   40 men get less than Rs. 30

6. A 3-month study of the phone calls received by Small Company yielded the following information:

Number of Calls per Day	**No. of Days**	**Number of Calls per Day**	**No. of Days**
100–200	3	600–700	10
200–300	7	700–800	9
300–400	11	800–900	8
400–500	13	900–1000	4
500–600	27		

   Compute the arithmetic mean, median and mode.

7. From the following distribution of travel time of 213 days of work of a firm's employee, find the modal travel time.

Travel Time (minutes)	**No. of Days**	**Travel Time (minutes)**	**No. of Days**
<80	213	<40	85
<70	210	<30	50
<60	195	<20	18
<50	156	<10	2

8. The mean monthly salary paid to all employees in a company is Rs. 1600. The mean monthly salaries paid to technical employees are Rs. 1800 and Rs. 1200, respectively. Determine the percentage of technical and non-technical employees of the company.
9. The geometric mean of 10 observations on a certain variable was calculated to be 16.2. It was later discovered that one of the observations was wrongly recorded as 10.9 when in fact it was 21.9. Apply appropriate corrections and calculate the correct geometric mean.
10. The following table shows the income distribution of a company.

Income (Rs.)	No. of Employees	Income (Rs.)	No. of Employees
1200–1400	8	2200–2400	35
1400–1600	12	2400–2600	18
1600–1800	20	2600–2800	7
1800–2000	30	2800–3000	6
2000–2200	40	3000–3200	4

Determine (a) the mean income, (b) the median income, (c) the mean, (d) the income limits for the middle 50% of the employees, (e) the seventh deciles (D_7) and (f) the eightieth percentile (P_{80}).

4

Measures of Variation and Skewness

4.1 Introduction

Average alone cannot adequately describe a set of observations, unless all the observations are the same. It is necessary to describe the variability of dispersion of the observations. Also, in two or more distributions the central value may be the same, but still there can be wide disparities in the formation of the distribution. Measures of variation help us in studying this important characteristic of a distribution, that is the extent to which the items vary, from one another and from some central value.

A measure of variation or dispersion describes the spread or scattering of the individual values around the central value. To illustrate the concept of variation, let us consider the data given in Table 4.1.

Since the average sales are similar, it may be observed that in Firm A, daily sales are the same irrespective of the day, whereas there is less amount of variation in the daily sales for Firm B and greater amount of variation in the daily sales for Firm C. Therefore, different sets of data may have the same measure of central tendency but differ greatly in terms of variation.

4.1.1 Significance of Measuring Variation

1. Measuring variability determines the reliability of an average by pointing out how far an average is representative of the entire data.
2. Another purpose of measuring variability is to determine the nature and cause of variation in order to control the variation itself.
3. Measures of variation enable comparisons of two or more distributions with regard to their variability.
4. Measuring variability is of great importance to advance statistical analysis. For example, sampling or statistical inference is essentially a problem in measuring variability.

4.1.2 Absolute versus Relative Measures of Variation

Table 4.2 shows absolute versus relative measures of variation.

TABLE 4.1

Daily Sales of Firms A, B and C

Firm A Daily Sales ($)	Firm B Daily Sales ($)	Firm C Daily Sales ($)
5,000	5,050	4,900
5,000	5,025	3,100
5,000	4,950	2,200
5,000	4,835	1,800
5,000	5,140	13,000
$\bar{X}_A = 5{,}000$	$\bar{X}_B = 5{,}000$	$\bar{X}_C = 5{,}000$

TABLE 4.2

Absolute vs. Relative Measures of Variation

Absolute Measures of Variation	Relative Measures of Variation
Expressed in terms of the original data	Useful in case the two sets of data are expressed in different units of measurement
Does not involve comparison between two sets of data	Involves comparison between two sets of data having the same unit of measurement but with different means

4.2 Range

4.2.1 For Ungrouped Data

The range is defined as the difference between the numerical largest value and the smallest numerically value in a set of data.

Symbolically,

$$R = L - S$$

where

R = range
L = largest value
S = smallest value

4.2.2 For Grouped Data

The range may be approximated as the difference between the upper limit of the largest class and the lower limit of the smallest class.

The relative measure corresponding to range, called the coefficient of range, is obtained by applying the following formula:

$$\text{Coefficient of range} = \frac{L - S}{L + S}$$

Merits

1. Range is the simplest to understand and easiest to compute.
2. It takes minimum time to calculate the value.
3. It is used in industry, for the quality control of products. For example, when the weight of a spare exceeds a particular range, the entire production line is checked, to correct the deviation in the weight of the spare produced.
4. It is used in the construction of charts and for quality control.
5. It is used in studying the fluctuations in financial and share markets.

Demerits

1. Range is not based on each and every item of the distribution. Its computation is based on the highest and lowest values and ignores the nature of dispersion among other values of observations in the data set.
2. It is influenced by extreme values and hence fluctuates from sample to sample of a population, even though the values that fall in between the highest and lowest values are similar.
3. Range cannot tell us anything about the capture of the distribution within the two extreme observations.
4. Range is not useful for comparison if the observations are in different units.
5. It cannot be computed for frequency distributions with open-ended classes.
6. It fails to explain the character of the distribution within two extreme observations (i.e. L and S).

4.2.3 Coefficient of Range

$$\text{Coefficient of range} = \frac{\text{Largest value} - \text{Smallest value}}{\text{Largest value} + \text{Smallest value}}$$

$$= \frac{L - S}{L + S}$$

Exercise

Compute the range and coefficient of range for the number of employees in an organisation in various age groups (Table 4.3).

TABLE 4.3

Employees in an Organisation in Various Age Groups

Class (years)	46–50	51–55	56–60	61–65	66–70	Total
No. of employees	6	17	40	27	10	100

Solution

$$\text{Range} = (\text{Upper limit of the highest class}) - (\text{Lower limit of the lowest class})$$
$$= 70 - 46 = 24$$

And,

$$\text{Coefficient of range} = \frac{(L - S)}{(L + S)}$$
$$= \frac{24}{136} = 0.1765$$

4.2.4 Interquartile Range

4.2.4.1 Need for Interquartile Range

Range is based on two extreme observations. Hence, it fails to explain the scatter within the range. So, when these extreme observations are discarded, the limited range would be more accurate and representative of the entire data.

4.2.4.2 Definition of Interquartile Range

Range calculated based on the middle 50% of observations is called interquartile range. It is calculated from observations obtained after discarding one quartile of the observations at the lower end and another quartile of the observations at the upper end of the distribution.

So, it is the difference between the third quartile and first quartile.

The quartiles (Q_1, Q_2, Q_3) are the highest values, in each of the first three of the four parts, of the distribution.

$$\text{Interquartile range} = Q_3 - Q_1$$

In symmetrical distribution, the quartiles Q_3 and Q_1 are equidistant from the median, that is

$$\text{Median} - Q_1 = Q_3 - \text{Median}$$

4.2.5 Semi-Interquartile Range or Quartile Deviation

Quartile deviation (QD) is one-half of the interquartile range. The quartile deviation is computed by taking the average of the difference between the third quartile and the first quartile. In symbols, this can be written as

$$QD = \frac{Q_3 - Q_1}{2}$$

where:
Q_1 = First quartile
Q_2 = Third quartile

4.2.5.1 For Ungrouped Data

$$\text{Lower quartile } Q_1 = \left(\frac{(N+1)}{4}\right)\text{th observation}$$

$$\text{Upper quartile } Q_3 = \left(\frac{3(N+1)}{4}\right)\text{th observation}$$

where N = Total number of observations.

4.2.5.2 For Grouped Data

$$Q_1 = L_1 + \frac{\left(\frac{1}{4}N - C\right)}{f} \times h$$

$$Q_3 = L_3 + \frac{\left(\frac{3}{4}N - C\right)}{f} \times h$$

where:
L_1 → The lower boundary of the first quartile class Q_1
L_3 → The lower boundary of the third quartile class Q_3
N → Total cumulative frequency
f → Frequency of the quartile class
h → Class interval (width)
C → Cumulative frequency of the class, just above the quartile class

Merits

1. It is superior to range.
2. It has a special utility in measuring variations in the case of open-ended distributions, one in which the data may be ranked but measured quantitatively.
3. It is useful in erratic or badly skewed distributions, and quartile deviation is not affected by the presence of an extreme value.

Demerits

1. It ignores 50% of the items, that is the first 25% and last 25%, as the value of the quartile deviation does not depend on every item of the series. It cannot be regarded as a good method of measuring dispersion.
2. It is not capable of mathematical manipulation.

3. Its value is very much affected by sampling fluctuation.
4. It is, in fact, not a measure of dispersion, as it really does not show the scatter around an average but rather a distance on a scale, that is, quartile deviation is not itself measured from an average, but it is a positional average.

4.2.5.3 Coefficient of Quartile Deviation

The coefficient of quartile deviation is the relative measure of quartile deviation. It is used to compare the degree of variation in different distributions.

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Exercise

Compute the quartile deviation from the data in Table 4.4.

Solution

See the computation in Table 4.5.

$$Q_1 = \text{size of}\left(\frac{N+1}{4}\right)\text{th item} = \frac{195+1}{4} = 49\text{th item}$$

Size of 49th item = 60
Hence, $Q_1 = 60$.

$$Q_3 = \text{size of } 3\left(\frac{N+1}{4}\right)\text{th item} = \frac{3\times 195}{4} = 147\text{th item}$$

Size of 147th item = 60
Hence, $Q_3 = 63$.

$$QD = \frac{Q_3 - Q_1}{2} = \frac{63-60}{2} = 1.5$$

TABLE 4.4

Heights of 195 Candidates

Height (inches)	58	59	60	61	62	63	64	65	66
No. of candidates	15	20	32	35	33	22	20	10	8

TABLE 4.5

Computation of Quartile Deviation

Height (inches)	58	59	60	61	62	63	64	65	66	
f_i	15	20	32	35	33	22	20	10	8	N = 195
Cumulative frequency (c.f.)	15	35	67	102	135	157	177	187	195	

Exercise

Compute the coefficient of quartile deviation from the data in Table 4.6.

Solution

See the computation in Table 4.7.

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$Q_1 = \text{size of} \left(\frac{N}{4}\right)\text{th item} = \frac{109}{4} = 27.25\text{th item}$$

Q_1 lies in class 12–16

$$Q_1 = L + \frac{\frac{N}{4} - c.f}{f} \times i$$

where:

L = 12
N/4 = 27.25
c.f. = 16
f = 18
i = 4

TABLE 4.6

Profit in Rs. in Crores of 109 Industries

Profit (Rs. in Crores)	No. of Industries	Profit (Rs. in Crores)	No. of Industries
4–8	6	24–28	12
8–12	10	28–32	10
12–16	18	32–36	6
16–20	30	36–40	2
20–24	15		

TABLE 4.7

Computation of Quartile Deviation

Profits (Rs. in Crores)	No. of Industries	c.f.
4–8	6	6
8–12	10	16
12–16	18	34
16–20	30	64
20–24	15	79
24–28	12	91
28–32	10	101
32–36	6	107
36–40	2	109

$$Q_1 = 12 + \frac{(27.25 - 16)}{18} \times 4 = 12 + \frac{11.25}{18} \times 4 = 12 + 25 = 14.5$$

$$Q_3 = \text{size of} \left(\frac{3N}{4} \right) \text{th item} = \frac{3 \times 109}{4} = 81.75 \text{th item}$$

Q_3 lies in class 24–28

$$Q_3 = L + \frac{\frac{3N}{4} - c.f}{f} \times i$$

where:
L = 24
3N/4 = 81.75
c.f. = 79
f = 12
i = 4

$$Q_3 = 24 + \frac{(81.75 - 79)}{12} \times 4 = 24 + \frac{11}{12} = 24 + 0.917 = 24.917$$

where:
Q_1 = 14.5
Q_3 = 24.917

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{24.917 - 14.5}{24.917 + 14.5} = \frac{10.417}{39.417} = 0.264$$

Exercise

Compute the quartile deviation for the data given in Table 4.8.

Solution

See the calculation in Table 4.9.

TABLE 4.8

Weight of 70 Employees

Weight (kg)	55–60	60–65	65–70	70–75	75–80
No. of employee	10	18	14	16	12

TABLE 4.9

Calculation of Quartile Deviation

Weight (kg)	55–60	60–65	65–70	70–75	75–80
No. of employee (f_i)	10	18	14	16	12
c.f.	10	28	42	58	70

First, we find quartile class for Q_1

$$Q_1 = \frac{1}{4} \times N = \frac{1}{4} \times 70$$

$$= 17.5\text{th observation}$$

This observation will fall in class (60–65)
$\therefore\ L_1 = 60,\ C = 10,\ f = 18,\ h = 5$
Now, we find quartile deviation Q_1

$$Q_1 = 60 + \frac{17.5 - 10}{18} \times 5$$

$$= 62.083$$

$$Q_3 = \frac{3}{4} \times N = \frac{3}{4} \times 70$$

$$= 52.5\text{th observation}$$

This observation will fall in class (70–75)

$$Q_3 = 70 + \frac{52.5 - 42}{16} \times 5$$

$$= 73.28$$

$$\therefore \text{Quartile Deviation} = \frac{Q_3 - Q_1}{2} = \frac{73.28 - 62.0823}{2} = 5.5985$$

$$\therefore \text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{73.28 - 62.0823}{73.28 + 62.0823} = 0.0827$$

Merits of Quartile Deviation

1. It can be used as a measure of variation, for open-ended distributions.
2. It is a better measure of variation, for highly skewed distribution or distribution with extreme values, as quartile deviation is not affected by the presence of extreme values.

Limitations of Quartile Deviation

1. As the quartile deviation is calculated using only 50% of the total observations, it cannot be regarded as a good measure of variation.
2. Quartile deviation is not a real measure of variation, as it does not measure the scatter of observations from the average. It is only a positional average.

4.3 Mean Deviation or Average Deviation

The average deviation some timed called mean deviation (MD). It is the average difference between the item in a distribution and the median or mean of that series.

To study the formation of a distribution we should take the deviation from an average.

4.3.1 Discrete Series

The formula is

$$MD = \frac{\Sigma fi\,|D|}{N}$$

Exercise

Calculate mean deviation and coefficient of mean deviation from the data in Table 4.10.

Solution

The calculation is shown in Table 4.11.
The formula for mean deviation is

$$MD = \frac{\Sigma fi\,|D|}{N}$$

where:
$|D|$ = Deviation from median

$$MD = \frac{36}{48}$$

$$\therefore MD = 0.75$$

TABLE 4.10

Marks Scored by 5 Students out of 20

Roll no.	10	11	12	13	14
Marks scored	3	12	18	12	3

TABLE 4.11

Calculation of Mean Deviation and Coefficient of Mean Deviation

x_i	**Frequency (f_i)**	$\|D\| = m - x_i = 12 - x_i$	$f_i\|D\|$	**c.f.**
10	3	2	6	3
11	12	1	12	15
12	18	0	0	33
13	12	1	12	45
14	3	2	6	48
	$N = \sum f_i = 48$		$\sum f_i\|D\| = 36$	

Calculation of the median is as follows:

$$\text{Median} = \text{size of}\left(\frac{N+1}{2}\right)\text{th item}$$

$$= \text{size of}\left(\frac{48+1}{2}\right)\text{th item}$$

$$= \text{size of}\left(\frac{49}{2}\right)\text{th item}$$

$$= \text{size of } 24.5\text{th item}$$

The size of the 24.5th item = 12; therefore, median = 12

The calculation of the coefficient of mean deviation is

$$\text{Coefficient of mean deviation} = \frac{\text{Mean deviation}}{\text{Median}} = \frac{0.75}{12} = 0.0625$$

Exercise

Calculate the mean deviation and coefficient of mean deviation for the data in Table 4.12.

Solution

Calculation of the mean deviation is shown in Table 4.13.
Calculation of the mean deviation:

$$\text{MD} = \frac{\sum \text{fi}\,|\,D\,|}{N} = \frac{460}{50} = 9.2$$

TABLE 4.12

Marks Scored by 50 Students in Statistics

Marks scored	5	15	25	35	45
No. of students	5	8	15	16	6

TABLE 4.13

Calculation of Mean Deviation and Coefficient of Mean Deviation

Marks (x_i)	**No. of Students (f_i)**	**$\|D\| = m - x_i = 25 - x_i$**	**$f_i\|D\|$**	**c.f.**
5	5	20	100	5
15	8	10	80	13
25	15	0	0	28
35	16	10	160	44
45	6	20	120	50
	$N = \sum f_i = 50$		$\sum f_i\|D\| = 460$	

Calculation of the median:

$$\text{Median} = \text{size of}\left(\frac{N+1}{2}\right)\text{th item}$$

$$= \text{size of}\left(\frac{50+1}{2}\right)\text{th item}$$

$$= \text{size of}\left(\frac{51}{2}\right)\text{th item}$$

$$= \text{size of } 25.5\text{th item}$$

Therefore, the size of the 25.5th item = 25 and median = 25

Calculation of the coefficient of mean deviation:

$$\text{Coefficient of mean deviation} = \frac{\text{Mean Deviation}}{\text{Median}} = \frac{9.2}{25} = 368$$

4.3.2 Continuous Series

The formula is

$$MD = \frac{\sum fi\,|D|}{N} \times c$$

Exercise

Compute the mean deviation from the data in Table 4.14.

Solution

Calculation of the mean deviation is shown in Table 4.15.

Calculation of mean deviation:

$$MD = \frac{\sum fi\,|D|}{N} \times c$$

where:

c (class width) = 10
N = 60

TABLE 4.14

Marks Scored by 60 Students

Marks	0–10	10–20	20–30	30–40	40–50	50–60	60–70
No. of students	4	6	10	20	10	6	4

TABLE 4.15

Calculation of Mean Deviation and Coefficient of Mean Deviation

Marks (x_i)	No. of Students (f_i)	c.f.	m	$d = \frac{m-35}{10}$	|D|	$f_i\|D\|$
0–10	4	4	$\frac{0+10}{2}=5$	–3	3	12
10–20	6	10	15	–2	2	12
20–30	10	20	25	–1	1	10
30–40	20	40	35	0	0	0
40–50	10	50	45	1	1	10
50–60	6	56	55	2	2	12
60–70	4	60	65	3	3	12
	$N = \sum f_i = 60$					$\sum f_i\|D\| = 68$

Therefore,

$$MD = \frac{68}{60} \times 10 = 11.33$$

Calculation of median:

$$\text{Median} = \text{size of } \left(\frac{N}{2}\right)\text{th item}$$

$$= \text{size of } \left(\frac{60}{2}\right)\text{th item}$$

$$= \text{size of } 30\text{th item}$$

Therefore, the size of the 30th item = 35 and median = 35

Calculation of coefficient of mean deviation:

$$\text{Coefficient of mean deviation} = \frac{\text{Mean Deviation}}{\text{Median}}$$

$$= \frac{11.33}{35}$$

$$= 0.3237$$

Merits of Mean Deviation

1. It is simple to understand and easy to compute.
2. It is based on each and every item of data.
3. It is less affected by an extreme item.
4. Since deviations are taken from a central value, comparisons of the formation of different distributions can easily be made.

Limitations of Mean Deviation

1. Algebraic signs are ignored while taking the deviations of the item.
2. This method may not give us very accurate results.

Exercise

Calculate the mean deviation from the (1) arithmetic mean, (2) mode and (3) median with respect to the marks obtained by nine students (given in Table 4.16) and show that the mean deviation from the median is minimum.

If the marks are doubled (converted out of 50), will the variation in marks increase? Give reasons.

Solution

Calculation of mean:

$$\bar{x} = \frac{\text{Sum of the obs.}}{9} = \frac{7+4+10+9+15+12+7+9+7}{9} = \frac{80}{9} = 8.89$$

Calculation of median:

$$\text{Median} = \text{size of}\left(\frac{N+2}{2}\right)\text{th item} = \frac{9+2}{2} = 5\text{th item}$$

For calculating median, arrange the items in ascending order:

Marks: 4, 7, 7, 7, 9, 10, 10, 12, 15

The size of the fifth item = 9. Hence, median = 9.
Calculation of mode:

Mode = 7 (since 7 is repeated, the maximum number of times, i.e. 3)

See Table 4.17 for computation of mean deviation.

$$\text{Mean deviation}\left(\text{from mean}\right) = \frac{\sum |D|}{N} = \frac{21.1}{9} = 2.34$$

$$\text{Mean deviation}\left(\text{from median}\right) = \frac{\sum |D|}{N} = \frac{21}{9} = 2.33$$

$$\text{Mean deviation}\left(\text{from mode}\right) = \frac{\sum |D|}{N} = \frac{23}{9} = 2.56$$

TABLE 4.16

Marks Scored by 9 Students out of 25

Marks (out of 25)	7	4	10	9	15	12	7	9	7

TABLE 4.17

Calculation of Mean Deviation from Arithmetic Mean, Median and Mode

Marks (x_i)	Deviations from Mean $\|D\|$	Deviations from Median $\|D\|$	Deviations from Mode $\|D\|$
7	1.9	2	0
4	4.9	5	3
10	1.1	1	3
9	0.1	0	2
15	6.1	6	8
12	3.1	3	5
7	1.9	2	0
9	0.1	0	2
7	1.9	2	0
$\sum x_i = 80$	$\sum \|D\| = 21.1$	$\sum \|D\| = 21$	$\sum \|D\| = 23$

From these calculations, it is clear that the mean deviation is least from the median.

When the marks are doubled, that is converted out of 50, the variation in marks would also increase, and in fact would be doubled. The reason is that when all the values of a series are multiplied by a certain constant, say 2, the value of the mean, median, mode, mean deviations, standard deviation and so forth would also be increased and would be obtained by multiplying with that constant.

Exercise

Calculate the mean deviation (taking deviations from the mean) from the data in Table 4.18.

Solution

Calculation of the mean deviation is shown in Table 4.19.

$$\bar{x} = \frac{\sum fixi}{N} = \frac{96}{16} = 6 \qquad MD = \frac{\sum fi|D|}{N} = \frac{24}{16} = 1.5$$

TABLE 4.18

Data for Mean Deviation

x_i	2	4	6	8	10
f_i	1	4	6	4	1

TABLE 4.19

Calculation of Mean Deviation Taking Deviations from the Mean

x_i	f_i	$f_i x_i$	$\|D\| = x_i - \bar{x}$	$f_i\|D\|$
2	1	2	4	4
4	4	16	2	8
6	6	36	0	0
8	4	32	2	8
10	1	10	4	4
	$N = 16$	$\sum f_i x_i = \mathbf{96}$		$\sum f_i\|D\| = 24$

Example 4.10

Calculate the mean deviation from the mean of the data shown in Table 4.20.

Solution

Calculation of the mean deviation is shown in Table 4.21.

The formula is

$$MD = \frac{\sum fi\,|D|}{N} \times C$$

$$\bar{x} = A + \frac{\sum fid'}{N} \times C$$

We have from Table 4.21,

$A = 35, \sum f_i\, d' = -8, N = 50, C = 10$

Substituting the value, we get

$$\bar{x} = 35 + \frac{(-8)}{50} \times 10 = 33.4$$

Now,

$\sum f_i|D| = 64, N = 50, C = 10$

$$\therefore MD = \frac{64}{50} \times 10$$

$$= 12.80$$

TABLE 4.20

Data of Marks Scored by 50 Students

Marks	0–10	10–20	20–30	30–40	40–50	50–60	60–70
No. of students	6	5	8	15	7	6	3

TABLE 4.21

Calculation of Mean Deviation Taking Deviations from Mean

Marks	**fi**	**M**	$\mathbf{d' = \frac{m-A}{10}, A = 35}$	**fi d′**	$\mathbf{|d'| = \frac{m-5}{10}}$	**fi|d′|**
0–10	6	5	–3	–18	3	18
10–20	5	15	–2	–10	2	10
20–30	8	25	–1	–8	1	8
30–40	15	35	0	0	0	0
40–50	7	45	1	7	1	7
50–60	6	55	2	12	2	12
60–70	3	65	3	9	3	9
	$N = \sum f_i = 50$			$\sum fi\, d' = -8$		$\sum f_i\|d'\| = 64$

4.4 Standard Deviation

The concept of standard deviation (SD) was introduced by Karl Pearson in 1823. Also known as root mean square deviation, standard deviation provides an average distance for each element from the mean.

It is the positive square root of the mean of the squared deviations of values from the arithmetic mean. It is denoted by σ (sigma). It is also known as root mean square deviation. It is the square root of the means of the squared deviation from the arithmetic mean.

It measures the absolute dispersion or variability of dispersion; the greater the amount of dispersion or variability, the greater the standard deviation.

If $\bar{x}$ is the means of $x_1, x_2,\ldots,x_n$, then

$$\sigma = \sqrt{\frac{1}{N}\left\{\left(x_1-\bar{x}\right)^2+\left(x_2-\bar{x}\right)^2+\ldots+\left(x_n-\bar{x}\right)^2\right\}}$$

$$= \sqrt{\sum_{i=1}^{N}\frac{\left(x_i-\bar{x}\right)^2}{N}}$$

4.4.1 Individual Series

Exercise

Find the standard deviation of (Rs.) 7, 9, 16, 24, 26.

Solution

Calculation of the standard deviation is shown in Table 4.22.

Calculation from Actual Mean: $\bar{x} = \frac{\Sigma x_i}{N} = \frac{82}{5} = \text{Rs. } 16.40 \quad \sigma = \sqrt{\frac{\Sigma\left(x_i-\bar{x}_i\right)^2}{N}}$

$$\sigma = \sqrt{\frac{293.20}{5}} = \text{Rs. } 7.66$$

Calculation for Assumed Mean: $\sigma = \sqrt{\frac{\Sigma dx_i^{\,2}}{N}-\left(\frac{\Sigma dx_i}{N}\right)^2}$

$$\sigma = \sqrt{\frac{294}{5}-\left(\frac{2}{5}\right)^2} = \text{Rs. } 7.66$$

Exercise

Table 4.23 shows marks obtained by students in Quantitative Methods. Find the standard deviation.

TABLE 4.22

Calculation of Standard Deviation from Actual Mean and Assumed Mean

A. Calculate from actual mean

Variants, Rs. x_i	Deviation from Actual Mean (16.40), $x = x_i - \bar{x}$	x_i^2
7	–9.4	88.36
9	–7.4	54.76
16	–0.4	0.16
24	7.6	57.76
26	9.6	92.16
$\Sigma x_i = 82$		$\Sigma x_i^2 = 293.20$

B. Calculate from assumed mean

Variants, Rs. x_i	Deviation from Assumed Mean (16), $dx_i = x_i = A$	dx_i^2
7	–9	81
9	–7	49
16	0	0
24	8	64
26	10	100
$\Sigma x_i = 82$	$\Sigma dx_i = 2$	$\Sigma dx_i^2 = 294$

TABLE 4.23

Marks Scored by 10 Students

Roll no.	1	2	3	4	5	6	7	8	9	10
Marks	43	48	65	57	31	60	37	48	78	59

Solution

Calculation of the standard deviation is shown in Table 4.24.

$$\bar{x} = \frac{\Sigma xi}{N} = \frac{526}{10} = 50.6 \cong 51$$

$$\text{S.D.}(\sigma) = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2} = \sqrt{\frac{1826}{10} - \left(\frac{26}{10}\right)^2} = 13.260$$

4.4.2 Discrete Series

Exercise

Calculate the standard deviation from the data in Table 4.25.

TABLE 4.24

Calculation of Standard Deviation

Roll No.	Marks (x_i)	$d = x_i - \bar{x} = x_i - 50$	d^2
1	43	–7	49
2	48	–2	4
3	65	15	225
4	57	7	49
5	31	–19	361
6	60	10	100
7	37	–13	169
8	48	–2	4
9	78	28	784
10	59	9	81
N = 10	$\Sigma x_i = 526$	$\Sigma d = 26$	$\Sigma d^2 = 1826$

TABLE 4.25

Data for Calculation of Standard Deviation

Class size	3.5	4.5	5.5	6.5	7.5	8.5	9.5
Frequency	3	7	22	60	85	32	8

Solution

Calculation of the standard deviation is shown in Table 4.26.

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{362}{217} - \left(\frac{128}{217}\right)^2} = 1.146$$

4.4.3 Step Deviation Method

Exercise

Find the standard deviation from the data in Table 4.27.

TABLE 4.26

Calculation of Standard Deviation (Discrete Series)

Size of Item (x_i)	f_i	$d = x_i - A = x_i - 6.5$	d^2	$f_i d$	fid^2
3.5	3	–3	9	–9	27
4.5	7	–2	4	–14	28
5.5	22	–1	1	–22	22
6.5	60	0	0	0	0
7.5	85	1	1	85	85
8.5	32	2	4	64	128
9.5	8	3	9	24	72
	$N = \Sigma f_i = 217$			$\Sigma f_i d = 128$	$\Sigma fid^2 = 362$

TABLE 4.27

Data for Calculation of Standard Deviation

x	4.5	14.5	24.5	34.5	44.5	54.5	64.5
f	1	5	12	22	17	9	4

TABLE 4.28

Calculation of Standard Deviation (Step Deviation Method)

x_i	f_i	$d' = \frac{xi - 34.5}{10}$	d'^2	fid'^2	fid'^2
4.5	1	–3	9	–3	9
14.5	5	–2	4	–10	20
24.5	12	–1	1	–12	12
34.5	22	0	0	0	0
44.5	17	1	1	17	17
54.5	9	2	4	18	36
64.5	4	3	9	12	36
	$\sum f_i = 70$			$\sum fid' = 22$	$\sum fid'^2 = 130$

Solution

Calculation of the standard deviation is shown in Table 4.28.

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times C \quad = \sqrt{\frac{130}{70} - \left(\frac{22}{70}\right)^2} \times 10 = 13.3$$

4.4.4 Continuous Series

Exercise

Calculate the standard deviation given 300 telephone calls, according to their duration in seconds (Table 4.29).

Solution

Calculation of the standard deviation is shown in Table 4.30.

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times C = \sqrt{\frac{665}{300} - \left(\frac{137}{300}\right)^2} \times 30$$

$$= \sqrt{2.217 - (0.457)^2} \times 30 = 1.42 \times 30 = 42.6$$

TABLE 4.29

Telephone Calls according to Their Duration in Seconds

Duration	0–30	30–60	60–90	90–120	120–150	150–180	180–210
No. of calls	9	17	43	82	81	44	24

TABLE 4.30

Calculation of Standard Deviation (Continuous Series)

Duration	f	M	$d' = \frac{m-a}{30}$	d'^2	fd'	fd'^2
0–30	9	15	–3	9	–27	81
30–60	17	45	–2	4	–34	68
60–90	43	75	–1	1	–43	43
90–120	82	105	0	0	0	0
120–150	81	135	1	1	81	81
150–180	44	165	2	4	88	176
180–210	24	195	3	9	72	216
	$N = \sum f = 300$				$\sum fd' = 137$	$\sum fd'^2 = 665$

4.5 Variance

Variance is calculated using the sum of the squared distances between the mean and each observation, divided by the total number of elements in the distribution (population). While calculating variance, the differences (deviations) are squared to make them positive. The square of the standard deviation is called variance. Therefore, variance = σ^2. The standard deviation and variance become larger as the variability or spread within the data becomes greater. More important, it is readily comparable with other standard deviations, and the greater the standard deviation, the greater the variability.

Remarks

1. If the data represents a sample of size N from a population, then it can be proved that the sum of the squared deviations is divided by (N – 1) instead of N. However, for large sample sizes, there is very little difference in the use of (N – 1) or N in computing the standard deviations.
2. The standard deviation is commonly used to measure variability, while all other measures have rather special users.
3. In addition, it is the only measure possessing the necessary mathematical properties to make it useful for advanced statistical work.

4.5.1 For Grouped and Ungrouped Data

For Grouped Data	For Ungrouped Data
$\sigma^2 = \frac{\Sigma f_i (X_i - \bar{x})^2}{N}$	$\sigma^2 = \frac{\Sigma (X_i - \bar{x})^2}{N}$
	$= \frac{\Sigma X_i^2}{N} - \bar{x}^2$

where:

σ^2 → Variance
X_i → The value of observation
$\bar{x}$ → Mean
N → Total cumulative frequency
f_i → Frequency of a class

The following formulae for standard deviation are mathematically equivalent to the above formula and are often more convenient to use in calculations.

$$\sigma = \sqrt{\frac{\Sigma f_i x_i^2}{N} - \left(\frac{\Sigma f_i x_i}{N}\right)^2} \qquad = \sqrt{\frac{\Sigma f_i x_i^2}{N} - \bar{x}^2}$$

$$= \sqrt{\frac{\Sigma f_i d^2}{N} - \left(\frac{\Sigma f_i d}{N}\right)^2} \times i$$

where:

$$d = \frac{x - A}{i}$$

4.6 Coefficient of Variation

A frequently used relative measure of the variation is the coefficient of variation (CV). This measure is simply the ratio of the standard deviation to the mean expressed as a percentage.

$$\text{Coefficient of variation (\%)} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{\sigma}{\bar{x}} \times 100$$

When the coefficient of variation is less in the data, it is said to be less variable or more consistent.

Use of coefficient of variation

1. It measures the spread of a set of data, as a proportion of its mean.
2. It is used in problem situations, where we want to compare the variability, homogeneity, stability, uniformity and consistency of two or more data sets.

3. Remember: The data set for which the coefficient of variation is greater is said to be more variable, that is less consistent or less homogenous.

Exercise

Compute the variance, standard deviation and coefficient of variation given the profitability of 50 companies (Table 4.31).

Solution

Calculation of the variance and coefficient of variation is shown in Table 4.32.

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{870}{50} = 17.4 \qquad \text{Variance}\left(\sigma^2\right) = \frac{\Sigma f_i \left(x_i - \bar{x}\right)^2}{\Sigma f_i} = \frac{1962}{50} = 39.24$$

Now, S.D. of profits of 50 companies is

$$\text{S.D.}(\sigma) = \sqrt{\text{Variance}} = \sqrt{\sigma^2} = \sqrt{39.24} = 6.26$$

$$\text{Coefficient of Variation (\%)} = \frac{\text{S.D.}}{\text{Mean}} \times 100 = \frac{\text{S.D.}}{\bar{x}} \times 100 = \frac{6.26}{17.4} \times 100 = 0.3598 \times 100 = 35.98\%$$

Exercise

The scores of two batsmen, A and B, in 10 innings during certain seasons are shown in Table 4.33. Find which batsman is more consistent in scoring.

TABLE 4.31

Profitability of 50 Companies

Profit (%)	10	15	20	25	30
Number of companies (f_i)	15	10	15	6	4

TABLE 4.32

Calculation of Variance and Coefficient of Variation

x_i	f_i	$x_i f_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
10	15	150	–7.4	54.76	821.40
15	10	150	–2.4	5.76	57.60
20	15	300	2.6	6.76	101.40
25	6	150	7.6	57.76	346.56
30	4	120	12.6	158.76	635.04
Total	N = 50	$\sum x_i f_i = 870$			$\sum f_i(x_i - \bar{x})^2 = 1962$

TABLE 4.33

Scores of Two Batsmen A and B in 10 Innings

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

Solution

Computation of the coefficient of variation is shown in Table 4.34.

For x-series	For y-series
$\bar{x} = \frac{\Sigma x}{N} = \frac{460}{10} = 46$	$\bar{y} = \frac{\Sigma y}{N} = \frac{500}{10} = 50$
Standard deviation for Batsman A	Standard deviation for Batsman B
$\sigma_A = \sqrt{\frac{\Sigma dx^2}{N}} = \sqrt{\frac{6500}{10}} = 25.5$	$\sigma_B = \sqrt{\frac{\Sigma dy^2}{N}} = \sqrt{\frac{5968}{10}} = 24.43$
$CV = \frac{\sigma_A}{\bar{x}} \times 100$	$CV = \frac{\sigma_B}{\bar{y}} \times 100$

Since the coefficient of variation is more for Batsman A, Batsman B is more consistent in scoring.

TABLE 4.34

Computation of Coefficient of Variation

Score of A (x_i)	**Deviation from Mean ($dx_i = x_i - A = x_i - 46$)**	**Square Deviation (dx_j^2)**
32	–14	196
28	–18	324
47	1	1
63	17	289
71	25	625
39	–7	49
10	–36	1296
60	14	196
96	50	2500
14	–32	1024
$\sum x_i = 460$	$\sum dx_i = 0$	$\sum dx_j^2 = 6500$
Score of B (y_j)	**Deviation from Mean ($dy_j = y_j - A = y_j - 50$)**	**Square Deviation dy_j^2**
19	–31	961
31	–19	361
48	–2	4
53	3	9
67	17	289
90	40	1600
10	–40	1600
62	12	144
40	–10	100
80	30	900
$\sum y_j = 500$	$\sum dy_j = 0$	$\sum dy_j^2 = 5968$

Exercise

Table 4.35 shows the marks obtained by 40 students of a class. Calculate the coefficient of variation.

Solution

Calculation of the coefficient of variation is shown in Table 4.36.

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

$$\bar{x} = A + \frac{\Sigma fidi'}{N} \times i$$

where:

$A = 62$
$\sum f_i di' = -70$
$N = 40$
$i = 5$

TABLE 4.35

Marks Obtained by 40 Students of Class

Marks	No. of Students	Marks	No. of Students
80–84	1	50–54	6
75–79	1	45–49	6
70–74	1	40–44	6
65–69	4	35–39	6
60–64	4	30–34	0
55–59	7	25–29	1

TABLE 4.36

Calculation of Coefficient of Variation

Marks	f_i	Mid-Point (m)	$d_i' = (m - A)/5, A = 62$	$f_i d_i'$	$fidi'^2$
80–84	1	82	4	4	16
75–79	1	77	3	3	9
70–74	1	72	2	2	4
65–69	4	67	1	4	4
60–64	4	62	0	0	0
55–59	7	57	–1	–7	7
50–54	6	52	–2	–12	24
45–49	6	47	–3	–18	54
40–44	6	42	–4	–24	96
35–39	3	37	–5	–15	75
30–34	0	32	–6	0	0
25–29	1	27	–7	–7	49
	$N = \sum f_i = 40$			$\sum f_i d_i' = -70$	$\sum fidi'^2 = 338$

$$\bar{x} = 62 - \frac{70}{40} \times 5 = 62 - 8.75 = 53.25$$

$$\sigma = \sqrt{\frac{\Sigma fidi'^2}{N} - \left(\frac{\Sigma fidi'}{N}\right)^2} \times i \qquad = \sqrt{\frac{338}{40} - \left(\frac{-70}{40}\right)^2} \times 5$$

$$\sigma = \sqrt{8.45 - 3.06} \times 5 = \sqrt{5.39} \times 5 = 2.322 \times 5 = 11.61$$

$$\bar{x} = 53.25, \sigma = 11.61$$

$$CV = \frac{11.61}{53.25} \times 100 = 21.80\%$$

4.7 Bienayme–Chebyshev Rule

This rule was developed by Russian mathematicians Bienayme and P L Chebyshev. It says that whatever may be the shape of a distribution, at least 75% of the values in the population will fall within ±2 standard deviations of the mean and at least 89% will fall within ±3 standard deviations of the mean.

4.7.1 Statement of the Bienayme–Chebyshev Rule

The rule states that the percentage of data observations lying within ±k standard deviations of the mean is at least $(1-(1/k^2))\times 100$.

4.7.2 Application

This formula applies to differences greater than one standard deviation about the mean, and k must be greater than 1.

In the case of a symmetrical bell-shaped curve (Figure 4.1), we can say that

1. Approximately 68% of the observations in the population fall within ±1 standard deviations from the mean.
2. Approximately 95% of the observations in the population fall within ± 2 standard deviations from the mean.
3. Approximately 99% of the observations in the population fall within ±3 standard deviations from the mean.

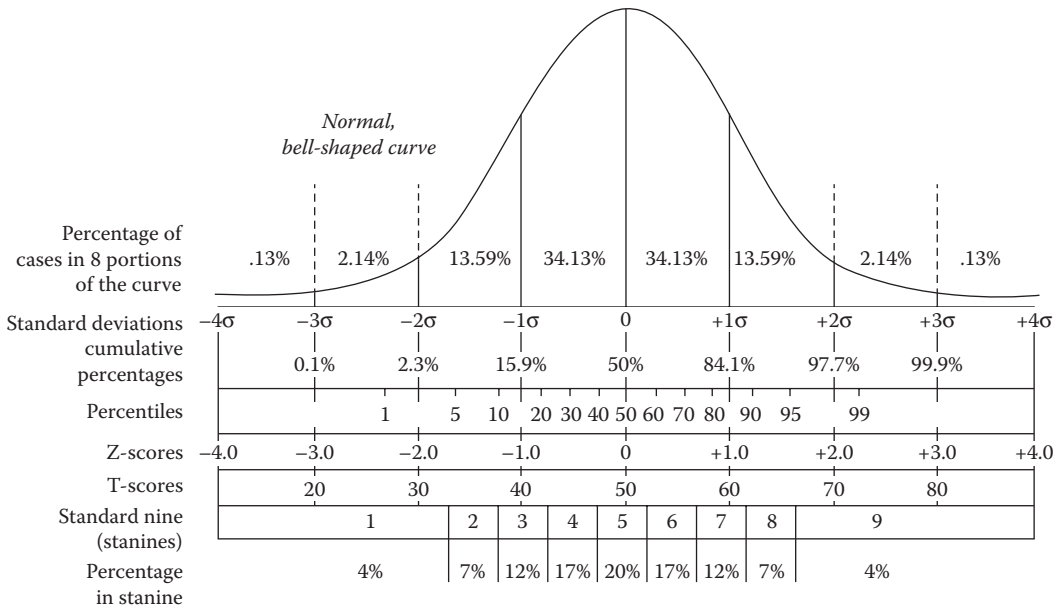


FIGURE 4.1
Diagrammatic representation of the Bienayme–Chebychev rule for a bell-shaped curve.

4.8 Skewness

The measures of central tendency and variation do not reveal all the characteristics of a given set of data. For example, two distributions may have the same mean and standard deviation but may differ widely in the shape of their distribution. Either the distribution of data is symmetrical or it is not. If the distribution of data is not symmetrical, it is called asymmetrical or skewed. Thus, skewness refers to the lack of symmetry in distribution.

A simple method of detecting the direction of skewness is to consider the tails of the distribution.

Rules

1. Data is symmetrical when there are no extreme values in a particular direction so that low and high values balance each other (Figure 4.2). In this case, mean = median = mode.
2. If the longer tail is towards the lower value or left-hand side, the skewness is negative (Figure 4.3). Negative skewness arises when the mean is decreased by some extremely low values, thus making mean < median < mode.

If the longer tail of the distribution is towards the higher values or right-hand side, the skewness is positive (Figure 4.4). Positive skewness occurs when the mean is increased by some unusually high values, thereby making mean > median > mode.

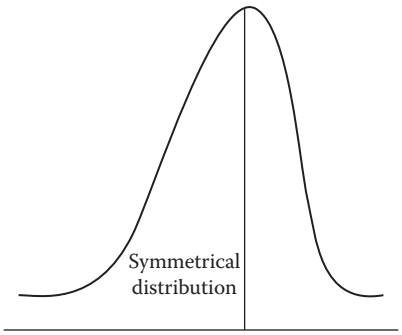


FIGURE 4.2
Symmetrical distribution.

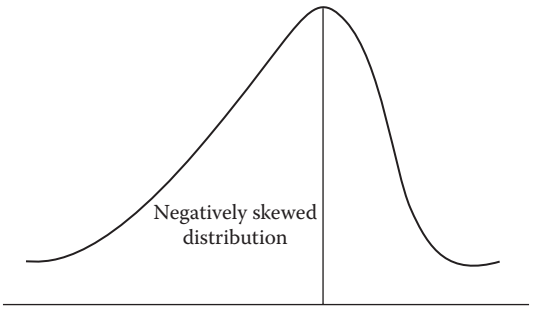


FIGURE 4.3
Negatively skewed distribution.

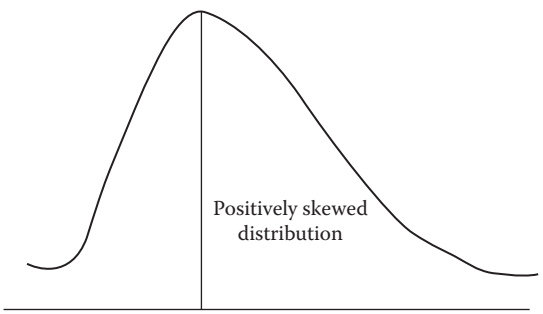


FIGURE 4.4
Positively skewed distribution.

4.8.1 Relative Skewness

In order to make comparisons between the skewness in two or more distributions, the coefficient of skewness (SK) (given by Karl Pearson) can be defined as

$$SK = \frac{\text{Mean} - \text{Mode}}{SD}$$

If the mode cannot be determined, then using the approximate relationship, Mode = 3 Median – 2 Mean, the above formula reduces to

$$\text{SK} = \frac{3(\text{Mean} - \text{Median})}{\text{SD}}$$

If the value of this coefficient is zero, the distribution is symmetrical; if the value of the coefficient is positive, it is a positively skewed distribution, or if the value of the coefficient is negative, it is a negatively skewed distribution. In practice, the value of this coefficient usually lies between –1 and +1.

When we are given open-ended distributions where extreme values are present in the data or positional measures such as median and quartiles, the following formula for the coefficient of skewness (given by Bowley 1926) is more appropriate.

$$\text{SK} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

Again, if the value of this coefficient is zero, it is a symmetrical distribution. For a positive value, it is a positively skewed distribution, and for a negative value, it is a negatively skewed distribution.

4.9 Summary

Measures of dispersion indicate how much the data in a given set of numerical data is spread out. The simplest of these measures is the range. The range is the distance between the largest and the smallest value in the data set. Of all measures of dispersion, the range is the most sensitive to extreme values. Similar in spirit to the range, but not affected by outliers, is the interquartile range, that is the difference between the third and first quartiles. The interquartile range is the range of the middle half of the data. The mean deviation and the standard deviation indicate how much the average value in the data set differs from the mean, whereas the mean deviation is just that the mean distance of the measurements from the mean and the standard deviation is usually calculated by first finding the variance and then extracting the square root of the variance. Although more difficult to compute than the other measures of dispersion, the standard deviation and the variance are the most useful and accurate methods.

Also, we have shown how the concepts of measures of variation and skewness are important. The measures of variation considered were the range, average deviation, quartile deviation and standard deviation. The concept of coefficient of variation was used to compare relative variations of different data. The skewness was used in relation to lack of symmetry.

REVIEW QUESTIONS

1. What do you mean by measures of dispersion?
2. Differentiate between absolute and relative measures of dispersion.
3. What do you mean by coefficient of variation?

4. Give the importance of measures of variability in quantitative decision making.
5. Write the advantages and disadvantages of all measures of variation.
6. What is the meaning of relative variation? When it is required?
7. Explain the term *dispersion*. What purpose does a measure of dispersion serve? Distinguish between absolute and relative measure of dispersion.
8. a. What are the requisites of a good measure of dispersion? In the light of these, comment on some of the well-known measures of dispersion.
   b. What is dispersion? Explain what you understand by absolute and relative dispersion. Describe some of the measures of relative dispersion known to you.
9. Describe the various measures of dispersion known to you and compare their properties.
10. Define *mean deviation*. How does it differ from standard deviation?
11. Define the mean deviation, standard deviation and interquartile range of a frequency distribution. Why is the standard deviation usually chosen as a measure of dispersion? Give an example for which you would prefer an alternative measure of dispersion.
12. What do you understand by skewness? How will you measure skewness?
13. a. What is skewness? How does it differ from dispersion? Describe the various measures of skewness.
    b. Explain the term *skewness* as applied to a frequency distribution and describe the various measures of skewness known to you.
14. Explain the importance of measures of skewness and dispersion, and comment on the various measures of skewness. Which measure is generally preferred and why?

SELF-PRACTICE PROBLEMS

1. Find the arithmetic mean and standard deviation of the following data. Also find the coefficient of variation.

Age (less than)	10	20	30	40	50	60	70	80	90	100
No. of persons	16	32	52	75	102	112	117	127	132	143

2. Compute the following data:
   106, 30, 20, 110, 27, 41, 112, 22, 106, 78, 74, 109, 67, 96, 26, 32
   a. Range
   b. Coefficient of range
   c. Quartile deviation
3. Calculate the range and its coefficient from the following data:
   54, 47, 19, 17, 76, 85, 29, 32

4. Calculate the mean deviation from the following data:

x_i	10	20	30	40	50	60	70
f_i	5	8	16	9	8	3	1

5. Calculate the range and its coefficient from the following data:
   Price of Gold per 10 g from Monday to Saturday

Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
160	158	170	142	176	187

   [Answer: Range = Rs. 45, coefficient of range = 0.137]
6. Compute the quartile deviation and mean deviation from the following data:

Height (inches)	58	59	60	61	62	63	64	65	66
No. of students	15	20	32	35	33	22	20	10	8

   [Answer: $Q_1 = 60$, $Q_2 = 61$, $Q_3 = 63$, MD = 1.71]
7. From the following table compute the quartile deviation:

Size	4–8	8–12	12–16	16–20	20–24	24–28	28–32	32–36	36–40
Frequency	6	10	18	30	15	12	10	6	2

   [Answer: QD = 5.2]
8. Calculate the mean deviation and standard deviation from the following data:

Profits (Rs.)	**No. of Firms**	**Profits (Rs.)**	**No. of Firms**
5000–6000	10	0 to 1,000	4
4000–5000	15	–1000 to 0	6
3000–4000	30	–2000 to –1000	8
2000–3000	10	–3000 to –2000	10
1000–2000	5		

   [Answer: MD = 2096, σ = 2534]
9. From the prices of shares X and Y, given below, state which is more stable in value, by calculating the coefficient of variation:

X	55	54	52	53	56	58	52	50	51	49
Y	108	107	105	105	106	107	104	103	104	101

   [Answer: Shares X, CV = 4.99; shares Y, CV = 2.90; shares Y are more stable]

10. Calculate the mean deviation and the standard deviation of the following data:

Age (Years)	**0–10**	**10–20**	**20–30**	**30–40**	**40–50**	**50–60**	**60–70**	**70–80**
No. of persons	15	15	23	22	25	10	5	10

11. From the following data, calculate the mean and standard deviation:

Age Group	**No. of Employees**	**Age Group**	**No. of Employees**
Below 20	20	40–45	109
20–25	26	45–50	84
25–30	44	50–55	66
30–35	60	55 and above	10
35–40	101		

What inference will you draw from the above?
[Answer: $\bar{x} = 39.5$, $\sigma = 9.55$]

5

Probability Theory

5.1 Introduction

The concept of probability originated in the seventeenth century and has become one of the most fascinating and debatable subjects in recent years. Probability has gained a lot of importance, and the mathematical theory of probability has become the basis for statistical applications in the areas of management, space technology, and the like.

In fact, most people use probability in their day-to-day lives without being aware of it. Statements like 'It may rain today', 'Probably I will continue with the same job', 'India might win the cricket series against Pakistan' and so forth are examples of the usage of probability in day-to-day life.

Various business decisions in real life are made under situations when a decision maker is very uncertain as to what will happen after the decisions are made.

The concept of probability was used by gamblers during the early days in games of chance, such as throwing a die, drawing a card from the deck or tossing a coin. In these games of chance, there is an uncertainty regarding the face of the die that will appear in a throw or the card that will appear in a draw or the face of a coin that will appear when it is tossed. Although there is an uncertainty concerning the outcome of any particular throw or any particular drawing, there is a predictable long-term outcome. For example, if a die is thrown many times, experimental studies have shown that the probability of a number to appear is one-sixth (as the die has six faces).

Business activities are much associated with future changes. Future is the most uncertain element. Therefore, policy makes try to make it less uncertain by using probability techniques.

Uncertainty is part and parcel of human life. Weather, stock market prices and product quality are nothing but some of the areas where commenting on the future with certainty becomes impossible. Decision making in such areas is facilitated through format and precise expressions for the uncertainties involved. Study of rainfall, spelt out in a form amenable for analysis, may render the decision on water management easy. Intuitively, we see that if there is a high chance of a large quantity of rainfall in the coming year, we may decide to use more rainfall water for power generation and irrigation this year. We may also take some steps regarding flood control. However, in order to know how much water to release for different purposes, we need to quantify the chances of different quantities of rainfall in the coming year.

Similarly, formal and precise expressions of stock market prices and product quality uncertainties may help to analyse and facilitate decisions on portfolio and sales planning,

respectively. Probability theory provides us with the ways and means to attain formal and precise expressions for uncertainties involved in different situations.

5.2 Basic Concepts

5.2.1 Experiment

Experiment refers to a process which results in different possible outcomes or observations. It describes an act which can be repeated under some given conditions. It is an operation that produces outcomes which can be observed.

Example 5.1: Inspecting a Light Bulb

When inspecting a light bulb, the outcome is either defective or non-defective.

5.2.2 Random Experiment

A random experiment is any well-defined process of observing a given chance phenomenon through a series of trials that are finite or infinite, each of which leads to a single outcome.

The observation in a random experiment involves a chance phenomenon that is not under controlled conditions.

An experiment in which all the possible results are known in advance, but none of them can be predicted with certainty, is called a random experiment.

The results of those experiments depend on chance, such as tossing a coin or throwing dice. If all the possible outcomes are known in advance and none of the outcomes can be predicted with certainty, then such an experiment is called a random experiment.

5.2.3 Outcome

The result of a random experiment is called an outcome.

5.2.4 Sample Space

The set of all possible outcomes of a random experiment is called the sample space. It is denoted by S.

5.2.5 Event

An event is one more possible outcomes of an experiment or trial or an observation. An event is used to denote a phenomenon that occurs with every realization of a set of conditions. For example, tossing a coin is a trial and getting heads or tails is an event.

5.2.6 Certain Event

If tossing of a coin, then it is certain that either heads or tails must happen. It is an event which contains all the sample points.

5.2.7 Impossible Event

If the happening of the event is not certain, then there is no chance of its `happening'. That is, probability of the complementary even is unity. Hence, if the probability of an even is zero, then that even is impossible. It is an event which does not contain any sample point. It is a null set (Φ).

5.2.8 Compound Event

If two or more events occur in connection with each other, they are a compound event. For example, the sum of the two numbers on the face of two dice is less than 8.

5.2.9 Complement of an Event

A complementary event is the number of unfavourable cases in an experiment.

Let there be two events, P and Q. Then P is called the complementary even of Q and vice versa. If P and Q are mutually exclusive (when both cannot happened simultaneously in a single trial) and exhaustive (when their totality includes all the possible outcomes of a random experiment). For example, when a dice is thrown, occurrence of an odd number (1, 3, 5) and an even number (2, 4, 6) are complementary events.

5.2.10 Mutually Exclusive Events

Events are said to be mutually exclusive if they cannot occur at the same time, when an experiment is performed. That is, the acceptance of one precludes the acceptance of another. Also, mutually exclusive events are those which do not overlap, when represented in a Venn diagram. (A Venn diagram is a set that can be represented by an arbitrary plane figure, and its relationship with other set(s) can be explained by the combination of two or more such figures.)

Two events are said to be mutually exclusive or incompatible when both cannot happen simultaneously in a single trial or the occurrence of any one of them precludes the occurrence of the other. For example:

1. In throwing a die, all six faces, numbered 1–6, are mutually exclusive since if any one of these faces comes up, the possibility of another in the same trial is ruled out.
2. A person may be either alive or dead at a point in time—he cannot be both alive and dead at the same time.

   X and Y are the two shortlisted candidates. The vacancy is for one post only. After their interviews, it is announced that X was selected. So, Y immediately left in dismay. Why? Because there was only one vacancy. Therefore, selection of X meant the rejection of Y. The selection of X and Y are mutual events; that is both cannot occur together.

5.2.11 Independent Events

Two or more events are said to be independent when the outcome of one does not affect and is not affected by the other. Two events A and B are independent events if the occurrence of one event, say A, is in no way related to the occurrence of another event, say B. For example, let A stand for a firm's spending a large amount of money on advertisement

and B for its showing an increase in sales. Advertising does not guarantee higher sales, but the probability that the firm will show an increase in sales will be higher if A has taken place.

5.2.12 Dependent Events

These are events in which the occurrence or non-occurrence of one event, in any one trial, affects the probability of other events in other trials.

Two events A and B are dependent events if the occurrence of one event, say A, is related to the occurrence of another event, say B.

5.2.13 Exhaustive Events

The total number of possible outcomes is known as an exhaustive event. Events are said to be exhaustive when their totality includes all the possible outcomes of a random experiment. For example, in throwing two dice, the exhaustive number of cases is 36; in tossing a coin, there are two exhaustive cases. While tossing a die, the possible outcomes are 1, 2, 3, 4, 5 and 6, and hence the exhaustive number of cases is 6.

5.2.14 Favourable Event

The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the events. In other words, these events are favourable for experiment. For example, in drawing a card from a pack of cards, the number of cases favourable for drawing an ace is 4, for drawing a spade is 13 and for drawing a red card is 26.

5.2.15 Equally Likely Event

The outcomes of a trial are said to be equally likely if, taking into consideration all the relevant evidence, there is no reason to expect one in preference to the other. In other words, those events have an equal chance of happening. Equally likely events convey the notion that each outcome of an experiment has the same chance of appearing as any other. Events are said to be equally likely events when one does not occur more often than the others.

5.2.16 Sample Spaces

This is the set of all possible outcomes of a random experiment, denoted by S. Sample spaces can be either finite or infinite, depending on whether the number of sample points is finite or infinite. The set S of all possible outcomes (or elementary events) of a given experiment is called the sample space of the experiment.

5.3 Probability

The development of the theory of probability dates back to the seventeenth century. The probability formulae and techniques were developed by Jacob Bernoulli (1713), De Moivre

(1718) and Thomas Bayes (1764). Most of these were concerned with the application of the theory of permutations and combinations to the calculation of probabilities associated with various dice and card games.

Probability means 'a chance'. It can be defined as an expression of chance of occurrence of an event. Probability is used in making viable predictions, suitable decisions, convenient planning and operational policies. Thus, probability is one of the significant contributors to the science of quantitative techniques.

The degree of uncertainty can be measured numerically with the help of probability.

Probability theory is used to analyse data for decision making.

1. The insurance industry uses probability to calculate premium rates.
2. A stock analyst/investor uses probability to estimate the returns of the stocks.
3. A project manager uses probability in decision making.

Probability is the chance of occurrence or a number assigned to the occurrence of an event, in a sample space. Probability means 'a chance'. The probability of an event equals the number times it happens divided by the number of opportunities.

It can be defined as an expression of the chance of occurrence of an event. Business activities are much associated with future changes. The future is the most uncertain element. Therefore, it is still a dream to forecast the future with 100% certainty in any decision problem. The probability theory provides a media of coping with uncertainty. Probability is used to make viable predictions, suitable decisions, convenient planning and operational policies. Thus, probability is one of the significant contributors to the science of quantitative techniques.

The degree of uncertainty can be measured numerically with the help of probability. Thomas Bayes (1702–1761) introduced the concept of inverse probability.

5.3.1 Classical Probability

Classical probability is based on the assumption that each event is equally likely to occur. This is an a priori assumption (the term *a priori* refers to something that is known by reason alone), and the probability based on this assumption is known as a priori probability. This approach employs abstract mathematical logic and hence is also called 'abstract' or 'mathematical' probability. This is the reason for considerable use of familiar objects like cards, coins and dice, where the answer can be stated in advance before picking a card, tossing a coin or throwing a die, respectively.

5.3.1.1 Definition of Classical Probability

If a random experiment results in N exhaustive, mutually exclusive and equally likely outcomes, out of which f are favourable to the happening of an event E, then the probability of occurrence of E, usually denoted by P (E), is given by

$$P = P(E) = \frac{f}{N}$$

This is a rather complex way of defining something that may seem intuitively obvious, but it can be used in tossing a coin and throwing a die examples.

$$\text{Probability of an event } (E) \text{ happening} = \frac{\text{Number of favourable outcomes}}{\text{Total number of outcomes}}$$

The scale of probability extends from 0 to 1. When p = 0, it denotes impossibility of the event taking place, that is, the even cannot take place. For example, the probability of throwing 7 with a single die is 0. When p = 1, it denotes certainty, that is, the event is bound to take place.

The basic assumption underlying the classical theory is that the outcomes of a random experiment are equally likely.

Probability is the ratio of the number of favourable cases to the total number of equally likely cases. If the probability of occurrence of A is denoted by p(a), then by this definition we have

$$p(a) = \frac{\text{No. of favourable cases}}{\text{Total number of equally likely cases}}$$

Classical probability is the ratio of the number of equally likely possible outcomes favourable for an event to the total number of possible outcomes.

Classical probability is often called a priori probability because if we keep using orderly examples of unbiased dice, fair coins and so forth, we can state the answer in advance (a priori) without rolling a die, tossing a coin and so on.

5.3.2 Relative Frequency

5.3.2.1 Relative Frequency of Occurrence Approach

The relative frequency of occurrence approach defines the probabilities as either

1. The proportions of times that an event occurs in the long run when the conditions are stable, or
2. The observed relative frequency of an event in a very large number of trials.

In this approach, the probability of an event happening is calculated knowing how often the event has happened in the past. In other words, this method uses the relative frequencies of past occurrences as probabilities. For instance, suppose that an organisation knows from past data that about 25 of its 300 employees entering every year leave the organisation due to good opportunities elsewhere. Then, the organisation can predict the probability of employee turnover for this reason as

$$25/300 = 1/12 = 0.083$$

Another characteristic of probabilities established by the relative frequency of occurrence approach can be illustrated by tossing a fair coin 1000 times. In this case, it is found that the proportion of getting either heads or tails is more initially, but as the number of tosses increases, both heads and tails become equally likely and the probability of the event showing heads or tails is 0.5. Thus, accuracy is gained as the experiment is repeated and the number of observations is more. But, the limitation of this approach is the consumption of time and cost for such large repetitions and additional observations.

Moreover, predicting probability using this approach becomes a blunder if the prediction is not based on sufficient data.

The relative frequency is defined as the proportion of times an event occurs, if the experiment is repeated several times under similar conditions. For example, consider the distribution of salaries in a company for March 2014 shown in Table 5.1. For a subsequent month, the salaries are likely to have the same distribution, unless employees leave or have their salaries raised or new people join. Hence, we have the probabilities shown in Table 5.2 obtained from the above relative frequencies.

These probabilities show the chance that an employee chosen at random will be in a particular salary class. For example, the probability of an employee's salary being Rs. 17,000–20,000 is 14%.

5.3.3 Limitation of the Classical Approach

5.3.3.1 Limitations of Classical Approach to Probability

The definition cannot be applied if it is not possible to make a simple enumeration of cases which can be considered equally likely. For example, if a person jumps from the top of

TABLE 5.1

Distribution of Salaries in a Company

Salaries (Rs. in Thousands)	Frequency	Relative Frequency (%)
5,000–8,000	10	$\frac{10}{50}\times 100 = 20\%$
8,000–11,000	11	$\frac{11}{50}\times 100 = 22\%$
11,000–14,000	10	$\frac{10}{50}\times 100 = 20\%$
14,000–17,000	10	$\frac{10}{50}\times 100 = 20\%$
17,000–20,000	7	$\frac{7}{50}\times 100 = 14\%$
20,000–23,000	2	$\frac{2}{50}\times 100 = 4\%$
	$N = \sum f = 50$	Total 100%

TABLE 5.2

Calculation of Probabilities Obtained from Relative Frequencies

Salaries (Rs. in Thousands)	Probability
5,000–8,000	20%
8,000–11,000	22%
11,000–14,000	20%
14,000–17,000	20%
17,000–20,000	14%
20,000–23,000	4%
Total	100%

Mount Everest, the probability of his or her survival will not be 50%, since survival and death, that is the two mutually exclusive and exhaustive outcomes, are not equally likely.

The classical approach to probability assumes a world that does not exist or is highly hypothetical in its assumptions. It assumes situations that are very unlikely but could conceivably happen. The limitations of this approach are as follows:

1. The classical definition is applicable only when the trials are equally likely or equally probable. For instance, the probability that a candidate attending an interview will succeed is not 50% since the two possible outcomes, that is success and failure, are not equally likely.
2. The classical definition is applicable only when the exhaustive number of cases in a trial is finite.
3. The classical definition is applicable only when the events are mutually exclusive.

Thus, the classical approach to probability is useful in card games, dice games, tossing coins and the like, but has serious problems when it is applied to less orderly decision problems that are encountered in the area of management. Probabilities of occurrence, such as an employee resigning from a job before his or her retirement age or the delay in delivery of a product to a nearby customer, cannot be predicted using this approach.

5.3.4 Subjective Probability

Subjective probabilities are those assigned to events by the manager or the researcher based on the past experiences or occurrences or on the evidence available. It may be an educated guess or intuition. At higher levels of managerial decisions, when the decision making becomes very important, specific and is demanded to be unique, managers use subjective probability. The approach was introduced by Frank Ramsay in 1926.

Here, the probability of an event is based on personal judgement.

5.3.5 Marginal or Unconditional Probability

This is the ratio of the number of possible outcomes favourable to Event A to the number of possible outcomes.

5.3.6 Empirical Probability

Probability derived from past experience is called empirical probability. It is used in the preparation of insurance mortality tables, which are based on past experience. It is also used in the analysis of most practical business problems.

5.4 Probability Rules

Figure 5.1 shows the rules of probability.

1. If A is an event with probability denoted by P (A), then the probability of the entire sample space S is 1, that is P (S) = 1. Area of sample space rectangle = 1.

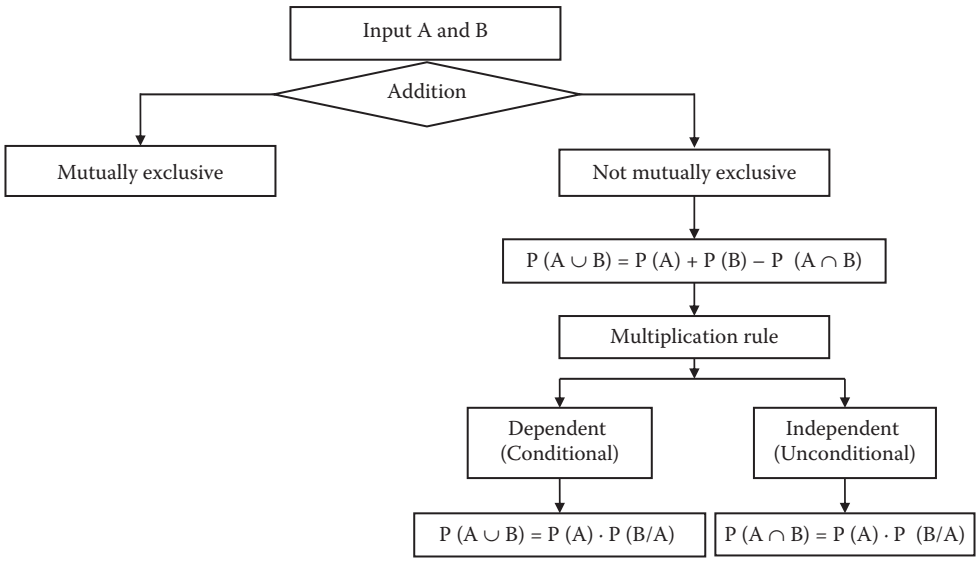


FIGURE 5.1
Rules of probability.

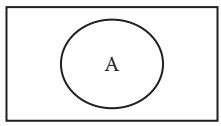


FIGURE 5.2
Event.

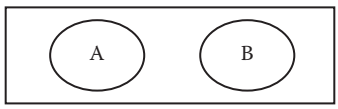


FIGURE 5.3
A and B are mutually exclusive events.

Event A is represented within the rectangle (Figure 5.2). Its minimum possible area of A = 0 and maximum possible area of A = 1. The probability of Event A must be greater than or equal to 0 or equal to 1 or 100%, that is $0 \leq P(A) \leq 1$. Probability cannot be negative. As the probability of the sample space is 1, the probability of an event contained in the sample space should be less than or equal to 1.

2. If A and B are mutual exclusive events, then the probability of (A or B) is equal to the sum of the probability of A and B.

$$P(A \text{ or } B) = P(A) + P(B)$$

Therefore, P (A and B) = 0 as A and B are mutually exclusive (Figures 5.3 and 5.4).

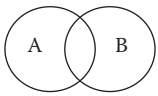


FIGURE 5.4
A and B are non-mutually exclusive events.

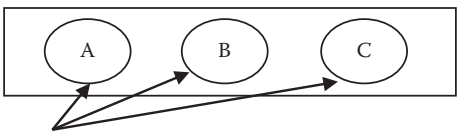


FIGURE 5.5
P (A or B or C) = P (A) + P (B) + P (C).

5.4.1 Additional Rule (Mutually Exhaustive Events)

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$

A representation by Venn diagram is shown in Figure 5.5.

Example 5.2

Suppose in a throw of a fair die,
A = getting 1, B = getting 2, C = getting 3
For a die, equally likely outcomes = 6.

$$\therefore P(A \text{ or } B \text{ or } C) = \frac{3}{6} \tag{5.1}$$

$$P(A) + P(B) + P(C) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} \tag{5.2}$$

From Equations 5.1 and 5.2 we get the required result.

5.4.1.1 Addition Theorem

Statement: If two events A and B are mutually exclusive, the probability of the occurrence of either A or B is the sum of the individual probability of A and B.
Symbolically,

$$P(A \text{ or } B) = P(A) + P(B)$$

5.4.1.2 Sample Space

The set S of all possible outcomes (or elementary events) of a given experiment is called the sample space of the experiment. Note that

$$P(A \text{ or } B) = P(A) + P(B)$$

or

$$P(A \cup B) = P(A) + P(B)$$

where $A \cup B$ is the union of Events A and B.

Proof: If an event A can happen in a_1 ways and B in a_2 ways, then the number of ways in which either event can happen is $a_1 + a_2$. If the total number of possibilities is n, then by definition the probability of either the first or the second event happening is

$$\frac{a_1 + a_2}{n} = \frac{a_1}{n} + \frac{a_2}{n}$$

but

$$\frac{a_1}{n} = P(A) \text{ and } \frac{a_2}{n} = P(B)$$

Hence, P (A or B) = P (A) + P (B).

The theorem can be extended to three or more mutually exclusive events. Thus, P (A or B or C) = P (A) + P (B) + P (C).

5.4.2 Additional Rule (Not Mutually Exhaustive Events)

Statement: If two events are not mutually exclusive, the probability of one of them occurring is the sum of the marginal probabilities of the event minus the joint probability of the occurrence of the events:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

This is represented by a Venn diagram in Figure 5.6.

5.4.2.1 Multiplication Theorem

Statement: If two events A and B are independent, the probability that they both will occur is equal to the product of their individual probability.

Symbolically, if A and B are independent, then

$$P(A \text{ and } B) = P(A) \times P(B)$$

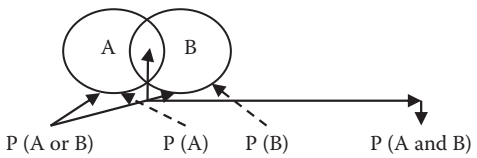


FIGURE 5.6
P (A or B) = P (A) + P (B) – P (A and B).

The theorem can be extended to three or more independent events; thus,

$$P(A, B \text{ and } C) = P(A) \times P(B) \times P(C)$$

Proof: If Event A can happen in n_1 ways, of which a_1 are successful, and Event B can happen in n_2 ways, of which a_2 are successful, we can combine each successful event in the first case with each successful event in the second case.

Thus, the total number of successful happenings in both cases is $a_1 \times a_2$.

Similarly, the total number of possible cases is $n_1 \times n_2$. Then by definition, the probability of the occurrence of both events is

$$\frac{a_1 \times a_2}{n_1 \times n_2} = \frac{a_1}{n_1} \times \frac{a_2}{n_2}$$

but

$$\frac{a_1}{n_1} = P(A) \text{ and } \frac{a_2}{n_2} = P(B)$$

$$\therefore P(A \text{ and } B) = P(A) \times P(B)$$

In a similar way, the theorem can be extended to three or more events.

5.4.3 Multiplication Rule (Independent Events)

The joint probability of two independent events is equal to the product of their marginal probabilities:

$$P(A \text{ and } B) = P(A) \times P(B).$$

For example, if the probability that A will be alive in 10 years is 0.6 and the probability that B will be alive in 10 years is 0.8, then the probability that they will both be alive in 20 years is $0.6 \times 0.8 = 0.48$.

5.4.4 Multiplication Rule (Dependent Events)

The joint probability of two events A and B which are dependent is equal to the probability of A multiplied by the probability B given that A has occurred:

$$P(A \text{ and } B) = P(A) \cdot P(A/B) \text{ or } P(B \text{ and } A) = P(B) \cdot P(A/B)$$

This formula is derived from the formula of conditional probability of dependent events.

$$P\left(\frac{B}{A}\right) = \frac{P(A \text{ and } B)}{P(A)}$$

5.4.5 Axioms to Probability

The probability of an event ranges from zero to one. If the event cannot take place, its probability will be zero, and if it is certain, that is bound to occur, its probability will be 1.

The probability of the entire sample space is 1, that is p (S) = 1.

If A and B are mutually exclusive (or disjoint) events, then the probability of occurrence of either A or B, denoted by P (A ∪ B), is given by

$$P(A \cup B) = P(A) + P(B)$$

5.4.6 Addition Theorem

Statement: If two events A and B are mutually exclusive, the probability of the occurrence of either A or B is the sum of the individual probability of A and B.

Symbolically,

$$P(A \text{ or } B) = P(A) + P(B)$$

Note:

$$P(A \text{ or } B) = P(A) + P(B)$$

or

$$P(A \cup B) = P(A) + P(B)$$

where A ∪ B is the union of Events A and B.

Proof: If an event A can happen in a_1 ways and B in a_2 ways, then the number of ways in which either event can happen is $a_1 + a_2$. If the total number of possibilities is n, then by definition the probability of either the first or the second event happening is

$$\frac{a_1 + a_2}{n} = \frac{a_1}{n} + \frac{a_2}{n}$$

but

$$\frac{a_1}{n} = P(A) \text{ and } \frac{a_2}{n} = P(B)$$

Hence, P (A or B) = P (A) + P (B).

The theorem can be extended to three or more mutually exclusive events.

Thus, P (A or B or C) = P (A) + P (B) + P (C).

The addition rule is used for mutually exclusive events, to find the probability of occurrence of one or the other of two joint events.

5.4.7 Multiplication Theorem

Statement: If two events A and B are independent, the probability that they both will occur is equal to the product of their individual probabilities.

Symbolically, if A and B are independent, then

$$P(A \text{ and } B) = P(A) \times P(B)$$

The theorem can be extended to three or more independent events; thus,

$$P(A, B \text{ and } C) = P(A) \times P(B) \times P(C)$$

Proof: If an event A can happen in n_1 ways of which a_1 are successful and the event B can happen in n_2 ways of which a_2 are successful, we can combine each successful event in the first case with each successful event in the second case.

Thus, the total number of successful happenings in both cases is $a_1 \times a_2$.

Similarly, the total number of possible cases is $n_1 \times n_2$. Then by definition, the probability of the occurrence of both events is

$$\frac{a_1 \times a_2}{n_1 \times n_2} = \frac{a_1}{n_1} \times \frac{a_2}{n_2}$$

but

$$\frac{a_1}{n_1} = P(A)$$

and

$$\frac{a_2}{n_2} = P(B)$$

$$\therefore P(A \text{ and } B) = P(A) \times P(B)$$

In a similar way, the theorem can be extended to three or more events.

The multiplication rule helps to find the joint probability of particular outcomes or events.

5.5 Conditional Probability

When we are computing the probability of a particular event A given information about the occurrence of another event B, this probability is referred to as conditional probability. It is the probability of occurrence of Event A subject to the occurrence of a previous event, say B.

If A and B are independent, then the occurrence of A is no way related to the occurrence or non-occurrence of Event B, and

$$P(A/B) = P(A)$$

If A and B are dependent, then

$$P\left(\frac{A}{B}\right) = \frac{P(A \text{ and } B)}{P(B)}$$

5.5.1 Dependent Events

Two events A and B are said to be dependent when B can occur only when A is known to have occurred (or vice versa). The probability attached to such an event is called the 'conditional probability' and is denoted by P (A/B), or the probability of A given that B has occurred.

Note: The multiplication theorem is not applicable in the case of dependent events.

Statement: If two events A and B are dependent, then the conditional probability of B given A is

$$P\left(\frac{B}{A}\right) = \frac{P(AB)}{P(A)}$$

Proof: Suppose a_1 is the number of cases for the simultaneous happening of A and B out of $a_1 + a_2$ cases in which A can happen with or without the happening of B.

$$P\left(\frac{B}{A}\right) = \frac{a_1}{a_1 + a_2}$$

$$= \frac{\frac{a_1}{n}}{\frac{(a_1 + a_2)}{n}}$$

$$= \frac{P(AB)}{P(A)}$$

Similarly, it can be shown that

$$P\left(\frac{A}{B}\right) = \frac{P(AB)}{P(B)}$$

The general rule of multiplication in its modified form, in terms of conditional probability, becomes

$$P\ (A \text{ and } B) = P\ (B) \times P\ (A/B)$$

or

$$P\ (A \text{ and } B) = P\ (A) \times P\ (B/A)$$

For three events A, B and C, we have

$$P(ABC) = P(A) \times P(B/A) \times P(C/AB)$$

That is, the probability of the occurrence of A, B and C is equal to the probability of A times the probability of B, given that A has occurred, times the probability of C, given that both A and B have occurred.

5.5.2 A Priori or Prior Probability

Probability before revision by Bayes' rule is called a priori or prior probability because it is determined before the sample information is taken into account.

5.5.3 Posterior or Revised Probability

It is obtained by revising the prior probabilities in the light of the additional information gained. Posterior probabilities are always conditional probabilities, the conditional event being the sample information. Thus, a priori probability, which is unconditional probability, becomes a posterior probability, which is a conditional probability by using Bayes' rule.

5.5.4 Bayes' Theorem

In modern-day business, there is an interesting number of instances where the probability of any particular event happening has to change in order to make profits. For example, a home appliances retailer calculates that it would be wise to stock his showroom with microwave ovens to the extent of 20% of his available shelf space. But actually, he finds out that the sales for microwave ovens are showing poor progress due to an increase in the electricity tariff.

It is therefore important at this stage for the retailer to recalculate the probability of a microwave oven selling under the new circumstances. This would help him in making a more profitable product mix decision for his showroom.

Here, we find that some probabilities were changed after the people involved (the retailer) got additional information (information about increased electricity tariff). The new probability thus obtained is known as posterior probability. Since probabilities can be revised as new information is gathered, the study of probability is of great significance in managerial decision making.

The concept of posterior probabilities was founded by the eighteenth-century British Presbyterian reverend Thomas Bayes. His formula for determining posterior probabilities under dependence is as follows:

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

This is also known as Bayes' theorem. It helps us to find the conditional probability of one event (A) occurring, given that another (B) has already occurred.

The terms *prior* and *posterior* refer to the time when information is collected. Before information is obtained, we have prior probabilities. Bayes' theorem provides a means of calculating posterior probabilities from prior probabilities.

Statement: Bayes' theorem deals with specific mutually exclusive events that have prior probabilities. It allows us to calculate the probability of an event, say A_1, given that Event B has already occurred with a known probability, P (B). The probability P (A_1/B) is called posterior (or revised) probability. It is denoted by P (A_1/B).

Therefore, for any event A_1, Bayes' theorem has the form

$$P\left(\frac{A_1}{B}\right)=\frac{P(A_1 \text{ and } B)}{P(B)}$$

Note: *Prior* and *posterior* refer to the time when information is collected.

5.5.5 Application of Bayes' Theorem

1. Bayes' theorem provides a means of calculating posterior probabilities from prior probabilities.
2. Bayes' theorem allows us to revise prior probabilities in the light of new information. The results of the formula are posterior probabilities.

Exercise

A person who sells newspapers wants to find out the chances that on any day, he will be able to sell more than 100 copies of *Indian Express*. From his diary, where he has recorded the daily sales of the last year, he finds that out of 365 days, on 73 days he had sold 85 copies, on 146 days he had sold 95 copies, on 60 days he had sold 105 copies and on 86 days he had sold 110 copies of the *Indian Express*. Can you help him find the required probability?

Solution

Taking the relative frequency approach we find the data shown in Table 5.3.
Thus, the number of days when his sales were more than 100 is (60 + 86) days = 146 days.
Hence, the required probability is $\frac{146}{365}=0.4$.

Exercise

Consider your locality, where out of 5000 people residing, 1200 are above 30 years of age and 3000 are female. Out of the 1200 that are above 30, 200 are female. Suppose, after a person is chosen you are told that the person is a female. What is the probability that she is above 30?

TABLE 5.3

Relative Frequency

Sales	No. of Days (Frequency)	Relative Frequency
85	73	73/365
95	146	146/365
105	60	60/365
110	86	86/365
	365	

Solution

We define the following events:
A = a person chosen is above 30 years
B = a person chosen is female
We are interested in Event A, given that B has occurred. If we denote this event by A/B, we want to find P (A/B).
Out of the 1200 persons who are above 30, 200 are females.
Therefore, out of 5000 people in the locality, 200 possess the characteristics of being female as well as above 30 years. Using a notation similar to the last example, we define Event AB:
AB: Event that a person is both a female and above 30 years of age
We derive from the data given that

$$P(A)=\frac{1200}{5000}, P(B)=\frac{3000}{5000} \text{ and } P(AB)=\frac{200}{5000}$$

To find the probability that, given a female has been chosen, she will be above 30, we see that out of 3000 females in the total population, only 200 females are above 30. Thus, the required probability is

$$\frac{200}{3000}, \text{ i.e. } P\left(\frac{A}{B}\right)=\frac{200}{3000}$$

We may note that

$$P\left(\frac{A}{B}\right)=\frac{200}{3000}=\frac{200}{5000}\times\frac{5000}{3000}=\frac{200}{5000}/\frac{3000}{5000}=\frac{P(AB)}{P(B)}$$

Exercise

In a class, there are 20 boys and 10 girls. Three students are selected at random. What is the probability that one boy and two girls are selected?

Solution

Number of ways of selecting 1 boy from 20 boys = C_1^{20}
Number of ways of selecting 2 girls from 10 girls = C_2^{10}
Number of ways of selecting 3 students from 30 students = C_3^{30}

The probability that one boy and two girls are selected is

$$\frac{C_1^{20}\times C_2^{10}}{C_3^{30}}=\left[C_1^{20}\times C_2^{10}\right]\div C_3^{30}$$

$$=\left[\frac{20!}{(20-1)!1!}\times\frac{10!}{(10-2)!2!}\right]\div\left[\frac{30!}{(30-3)!3!}\right]$$

$$= \left[\frac{20 \times 19!}{19! \times 1} \times \frac{10 \times 9 \times 8!}{8! \times 2 \times 1}\right] \div \left[\frac{30 \times 29 \times 28 \times 27!}{27! \times 3 \times 2}\right]$$

$$= \left[20 \times \frac{10 \times 9}{2}\right] \div \left[\frac{30 \times 29 \times 28}{3 \times 2}\right]$$

$$= \left[20 \times \frac{10 \times 9}{2} \times \frac{3 \times 2}{30 \times 29 \times 28}\right]$$

$$= \frac{10 \times 9}{29 \times 28} = \frac{45}{203} = 0.22$$

or

$$\left(C_1^{20} \times C_2^{10}\right) \div C_3^{30}$$

$$= \left[20 \times \frac{10 \times 9}{2}\right] \div \left[\frac{30 \times 29 \times 28}{3 \times 2}\right]$$

$$= (10 \times 10 \times 9) \div (10 \times 29 \times 14)$$

$$= 0.22$$

Exercise

A speaks truth in 60% of cases and B in 70% of cases. In what percentage of cases are they likely to contradict each other in stating the same fact?

Solution

They will contradict each other only if one of them speaks the truth and the other speaks a lie.

The probability that A speaks the truth and B a lie is

$$\frac{60}{100} \times \frac{30}{100} = \frac{9}{50}$$

The probability that B speaks the truth and B a lie is

$$\frac{70}{100} \times \frac{40}{100} = \frac{14}{50}$$

Therefore, the total probability is

$$\frac{9}{50} + \frac{14}{50} = \frac{23}{50}$$

Thus, the percent of cases in which they contradict each other is

$$\frac{23}{50}\times 100 = 46\%$$

Exercise

A bag contains four white, five red and six green balls. Three balls are drawn at random. What is the probability that a white, a red and a green ball are drawn?

Solution

The total number of balls in the bag is 4 + 5 + 6 = 15.
Three balls can be drawn out of 15 in C_3^{15} ways.
The probability that a white, a red and a green ball are drawn is

$$\frac{C_1^4 \times C_1^5 \times C_1^6}{C_3^{15}}$$

$$= \frac{120}{C_3^{15}} \quad \left\{\because C_1^n = n\right\}$$

$$= 120 \div C_3^{15}$$

$$= 120 \div \frac{15\times 14\times 13}{1\times 2\times 3}$$

$$= \frac{120}{455} = \frac{24}{91}$$

Exercise

If from a pack of 52 cards 2 cards are drawn, find the probability that one may be a king and the other is a queen when

1. The cards are replaced after the draw
2. The cards are not replaced after the draw

Solution

Two cards can be drawn from 52 in C_2^{52} ways. A king can be drawn in C_1^4 ways. A queen can be drawn in C_1^4 ways.

1. The cards are replaced after the draw. That is, each of the former can be associated with each of the latter.
   A king and a queen can be drawn in $C_1^4 \times C_1^4$ ways.

$$p = \frac{\text{No. of favourable cases}}{\text{Total no. of equally likely cases}}$$

$$= \frac{C_1^4 \times C_1^4}{C_2^{52}}$$

$$= C_1^4 \times C_1^4 \div C_2^{52}$$

$$= 4 \times 4 \div \left(\frac{52 \times 51}{1 \times 2} \right) \quad \left\{ \because C_1^n = n \right\}$$

$$= 16 \div 26 \times 51$$

$$= \frac{16}{26 \times 51} = \frac{8}{13 \times 51}$$

$$= \frac{8}{663} = 0.012$$

5.6 Set Theory

5.6.1 Power of Set

The number of elements in a set is called the power of set and is denoted by $|A|$ for Set A. This may be a finite number of elements in a given set A, such that A = {1, 2, 3, 4, 5} has a definite number of elements and its power is 5.

When we describe a set B = {x/x is an integer}, Set B has an indefinite number of elements and is called an infinite set.

When there are no elements in a set, it is called a 'null' or 'empty' set and is denoted $\phi = \{\}$. For clarification, even {0} is not a null set, as it has one element, 'zero', and its power is one.

Similarly, when the elements are not well defined, they do not make a set, such as 'intelligent boys and 'pretty girls'.

5.6.2 Elementary Concepts of Set

5.6.2.1 Universal Set

A set describing all the objects that are possible points of interest in a problem is called a universal set. It is the largest set of elements.

Thus, A = {x/x indicates world population} is a universal set with all human beings of the world, as its members or elements, whereas B = {x/x is the female population of the world} is not a universal set, as it does not contain the total population or the male population. It is represented by S.

5.6.2.2 Subset of a Set

Set A is called the subset of B if all the elements of A are also the elements of B. This is denoted by the symbol $\subset$. If A = {2, 4, 6, 8}, B = {4, 6} and C = {3}, then all the elements of B are the elements of A. Hence, B is a subset of A and symbolically, $B \subset A$ or $B \subseteq A$.

Where all of the elements of C are not the elements of A or B, we write

$$C \not\subset A, C \not\subset B$$

Thus, we can say that

1. Every set is a subset of itself.
2. An empty set is subset of every set.
3. The total number of all possible subsets of a given set containing n elements is 2^n.
4. Power set: The collection of all possible subsets of a given set A is called the power set of A and is denoted P (A).

Thus, if A = [1, 2], then

$$P(A) = \{ \phi, [1],[2],[1,2],[2,1]\}$$

and

$$P(\phi) = \phi.$$

5.6.2.3 Equality of Two Sets

When all the elements of one set are also elements of the second set and they are equal in value, both sets are called equal; that is sets A and B are equal if $A \subset B$ and $B \subset A$.

Let

$$A = \{11, 12, 13, 14, 15\}$$

$$B = \{15, 14, 13, 12, 11\}$$

and $$C = \{15, 13, 12, 14, 11\}$$

Then $A \subset B$, $B \subset A$, $B \subset C$, and $A = B = C$.

5.6.2.4 Complement of a Set

The complement of Set A is the set of all of the elements of a universal set which do not belong to A. The complement is denoted by A′ or $\bar{A}$.

Thus, let the universal set S = {1, 3, 5, 7, 9, 11, 13} and A = {3, 7, 11}.

Then complement set $A = A' = \bar{A} = \{1, 5, 9, 13\}$.

We can thus say (in a broad sense) that

1. The complement of a universal set is a null set.
2. The complement of a null set is a universal set.

5.6.2.5 Difference of Two Sets

A set of elements belonging to A but not to B is called a difference set.

Thus, $A - B = \{x/x \in A; x \notin B\}$.

5.6.2.6 Cardinal Number of a Finite Set

The number of distinct elements of a finite set is called its cardinal number and is denoted n (A).

5.6.3 Operations of Sets

5.6.3.1 Union of Two Sets

The union of two sets A and B is the set of elements which belong to A or B or both. It is denoted by $A \cup B$.

Hence, if A = {1, 3, 5, 7, 9} and B = {7, 9, 11}, then $A \cup B$ = {1, 3, 5, 7, 9, 11}.

Thus, $A \cup B = \{x/x \in A \text{ or } x \notin B\}$.

5.6.3.2 Intersection of Two Sets

The intersection of two sets A and B is a set of elements that belong to both A and B. It is denoted by $A \cap B$. Thus, if A = {1, 3, 5, 7, 9} and B = {3, 5, 9, 11}, then $A \cap B$ = {3, 5, 9}.

It can be written symbolically as $A \cap B = \{x/x \in A \text{ and } x \in B\}$.

However, if $A \cap B = \phi$, then A and B are called disjoint sets.

5.6.3.3 Difference of Two Sets

The difference of two sets A and B is the set of elements which belong to A but do not belong to B. It is denoted as A – B.

Hence, $A - B = \{x/x \in A \text{ and } x \notin B\}$. For example, if A = {1, 3, 5, 7, 9} and B = {3, 5, 7, 11}, then A – B = {1, 9}.

We can also establish that

$$A - B = \{x/x \in A; x \notin B\}$$

$$= \{x/x \in A; x \in \overline{B}\}$$

$$= A \cap \overline{B}$$

5.7 Venn Diagram

A set can be represented by an arbitrary plane figure and its relationship with other sets can be explained by the combination of two or more such figures (Figures 5.7 and 5.8). All elements of a set can be considered to be the points contained within the boundary of the figure. The universal set U is generally denoted by a rectangle. Such figures are known as Venn diagrams after the English logician John Venn (1834–1883).

5.7.1 Universal Set

A rectangle denotes the universal set S and is drawn as shown in Figure 5.9.

FIGURE 5.7
Outcomes of coin toss.

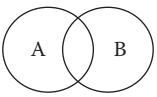


FIGURE 5.8
Outcomes of selection.

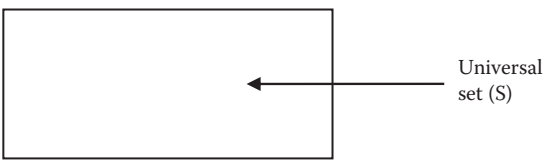


FIGURE 5.9
Universal set.

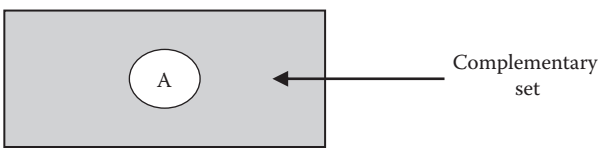


FIGURE 5.10
Complementary set.

5.7.2 Complementary Set

Set A, whose elements do not belong to a universal set, is a complementary set and is denoted as given in Figure 5.10 (shaded area).

5.7.3 Union of Two Sets

If A and B are represented by two circles and if we have to find the union of the two sets, that is common elements to both sets, we denote it as the shaded area in Figure 5.11.

5.7.4 Intersection of Two Sets

Since the intersection of two sets is the set belonging to A or B or both, the Venn diagram representation can be drawn as shown in Figure 5.12, with the shaded area indicating the intersection of A and B, that is $A \cap B$.

5.7.5 Difference of Two Sets

The set of elements belonging to Set A but not to Set B is the set of difference. It can be represented as given in Figure 5.13.

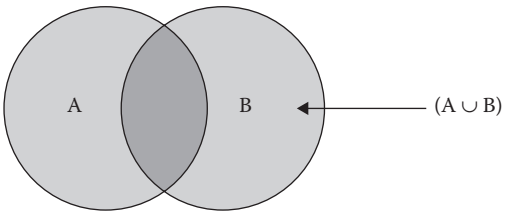


FIGURE 5.11
Union of two sets.

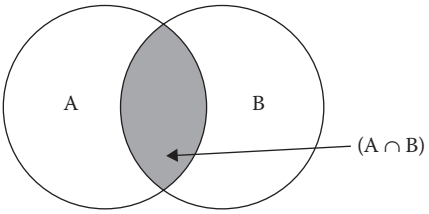


FIGURE 5.12
Intersection of two sets.

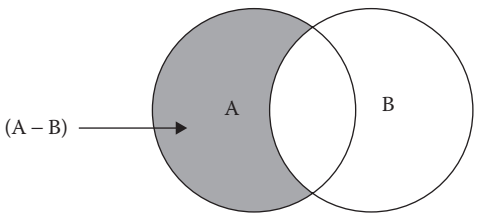


FIGURE 5.13
Difference of two sets.

5.7.6 Enhanced Application of the Venn Diagram

When we represent the relationship of more than two sets, we can use a Venn diagram to advantage as follows: Set A ∪ (B ∩ C) means the union of Set A with the intersection of Sets B and C and is represented as the shaded area in Figure 5.14. Other applications are indicated in Figure 5.15.

5.8 Fundamental Laws of Operation

Identity law

$A \cup \phi = A$

$A \cup S = S$

$A \cap \phi = \phi$

$A \cap S = A$

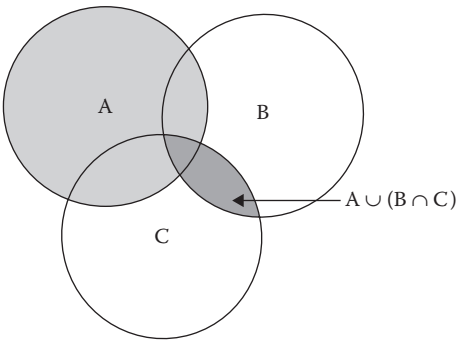


FIGURE 5.14
Set A ∪ (B ∩ C).

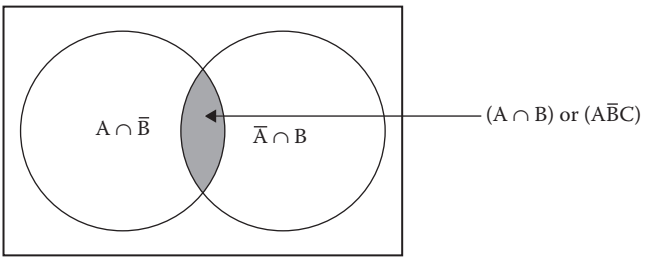


FIGURE 5.15
(A ∩ B) or A $\bar{B}$ C.

Complement law

$A \cup A' = S$
$A \cap A' = \phi$
$(A')' = A$

Idepotent law

$A \cup A = A$
$A \cap A = A$

Cumulative law

$A \cup B = B \cup A$
$A \cap B = B \cap A$

Exercise

Eight boys and two girls are to be seated on chairs arranged in a row for a photograph. If the arrangement is made at random, find the probability that the two girls (1) occupy the chairs at the end and (2) will be together.

Solution

A total of 10 students are to be arranged in a row. The can be arranged in $c_{10}^{10} = 10!$ different ways

$$\therefore n(S) = 10!$$

1. Let A be the two girls occupying seats at the end.
   In the two end seats, the girls can be arranged in P_2^2 ways. In the remaining eight chairs, eight boys can be arranged in $P_8^8 = 8!$ ways.

$$\therefore n(A) = 2!\,8!$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{2!8!}{10!} = \frac{2\times1\times8!}{10\times9\times8!} = \frac{1}{45}$$

2. Let B be the two girls being together.
   Consider one group of two girls. This group and the remaining eight boys can be arranged in $P_9^9 = 9!$ ways. The two girls themselves can be arranged in $P_2^2 = 2!$ ways.

$$\therefore n(B) = 2!\,8!$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{2!9!}{10!} = \frac{2\times1\times9!}{10\times9!} = \frac{1}{5}$$

Exercise

Consider a bag containing four white and five black balls. A man draws three balls at random. What is the probability that all three are black?

Solution

$$P(\text{All three are black}) = \frac{\text{Favourable events}}{\text{Total events}}$$

$$= \frac{C_3^5}{C_3^9} = C_3^5 \div C_3^9$$

$$\because C_r^n = \frac{n!}{n-r!r!} = \frac{5!}{(5-3)!3!} \div \frac{9!}{(9-3)!3!}$$

$$= \frac{5!}{2!3!} \div \frac{9!}{6!3!} = \frac{5!}{2!3!} \times \frac{6!3!}{9!}$$

$$= \frac{5\times4\times3\times2!}{2!} \times \frac{6!}{9\times8\times7\times6!} = \frac{5}{42}$$

Exercise

Consider a bag containing five white and seven black balls. If two balls are drawn, what is the probability that one is white and the other is black?

Solution

$$P(1 \text{ is white and other is black}) = \frac{\text{Favourable events}}{\text{Total events}}$$

$$= \frac{C_1^5 \times C_1^7}{C_2^{12}} = \frac{35}{66}$$

Exercise

What is the chance of drawing a diamond face card in a draw from a pack of 52 well-shuffled cards?

Solution

A diamond card can be drawn out of $13 = C_1^{13} = 13$ ways.
Therefore, the probability of drawing a diamond card is 13/52 = ¼.

Exercise

If A = {3, 4, 6, 8} and B = {6, 8, 10, 11}, show the representation of A and B by Venn diagram.

Solution

The required Venn diagram is given in Figure 5.16.

Exercise

For each of the following, what relation must hold between sets A and B? Draw the Venn diagram.

1. $A \cap B = B$

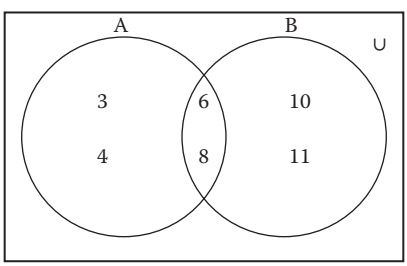


FIGURE 5.16
Venn diagram.

Solution

For this, the relation between A and B is $B \subseteq A$ (Figure 5.17).

2. $A \cup B = A$

Solution

For this, the relation between A and B is $B \subseteq A$ (Figure 5.18).

Exercise

Let A, B and C be sets such that $A \subseteq B$, $A \subseteq C$, $B \cap C \subseteq A$ and $A \subseteq (B \cap C)$. Draw a Venn diagram.

Solution

Here, $(B \cap C) \subseteq A$ and $A \subseteq (B \cap C)$. This means $A = (B \cap C)$. Hence, $A \subseteq B$ and $A \subseteq C$ (Figure 5.19).

Exercise

Prove that $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ using Venn diagrams.

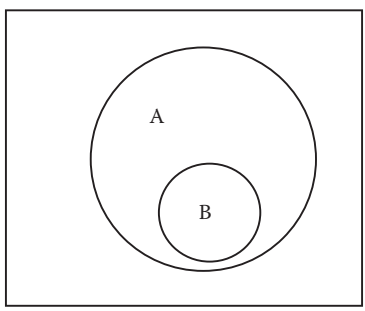


FIGURE 5.17
$A \cap B = B$.

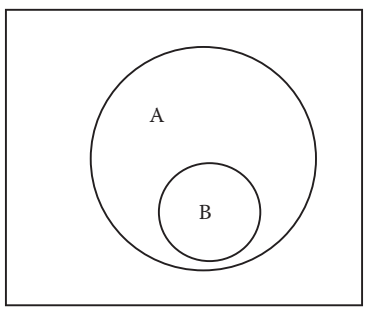


FIGURE 5.18
$A \cup B = A$.

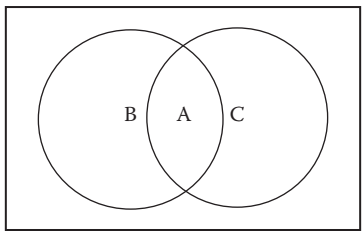


FIGURE 5.19
$A \subseteq B$, $A \subseteq C$, $B \cap C \subseteq A$ and $A \subseteq (B \cap C)$.

Solution

Let

$$A = \{2, 3, 7, 8\} \tag{5.3}$$

$$B = \{5, 6, 7, 8\} \tag{5.4}$$

$$C = \{3, 4, 5, 8\} \tag{5.5}$$

$$\therefore (B \cap C) = \{5, 8\} \tag{5.6}$$

(using Equations 5.4 and 5.5)

$$\therefore A \cup (B \cap C) = \{2, 3, 5, 7, 8\}$$

(using Equations 5.3 and 5.6)
Now,

$$(A \cup B) = \{2, 3, 5, 6, 7, 8\} \tag{5.7}$$

and

$$(A \cup C) = \{2, 3, 4, 5, 7, 8\} \tag{5.8}$$

$$\therefore (A \cup B) \cap (A \cup C) = \{2, 3, 5, 7, 8\}$$

(using Equations 5.7 and 5.8)

In Figure 5.20, $B \cap C$ is represented by areas 5 and 8, and A is represented by areas 2, 3, 7 and 8. So, $A \cup (B \cap C)$ is represented by areas 2, 3, 5, 7 and 8.

Again, areas 2, 3, 5, 6, 7 and 8 represent $A \cup B$, and areas 2, 3, 4, 5, 7 and 8 represent $A \cup C$.

So, areas 2, 3, 5, 7 and 8 represent $(A \cup B) \cap (A \cup C)$.

This proves the assertion.

Exercise

A bag contains six red and eight black balls. If two balls are drawn at random one by one, find the probability of getting

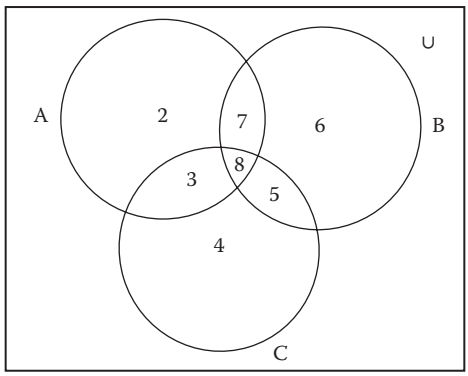


FIGURE 5.20
$(A \cup B) \cap (A \cup C)$.

1. Both balls of different colours
2. Both balls of the same colour

Solution

Two balls are drawn one by one. Hence, we have two draws. These are independent. Assuming that after the first draw, the ball is not replaced in the bag, for the first draw, n (S) = 14, and for the second draw, n (S) = 13.

Let
R_1 = First draw we get red ball
B_1 = First draw we get black ball
R_2 = Second draw we get red ball
B_2 = Second draw we get black ball

1. The probability that both are different colours is

$$P(R_1 \cap B_2) + P(B_1 \cap R_2) = P(R_1) \times P(B_2) + P(B_1) \times P(R_2)$$

$$= \frac{6}{14} \times \frac{8}{13} + \frac{8}{14} \times \frac{6}{13} = \frac{96}{182} = \frac{48}{91}$$

2. The probability that both are the same colour is

$$P(\text{Both red}) + P(\text{Both black})$$

$$= P(R_1 \cap R_2) + P(B_1 \cap B_2)$$

$$= P(R_1) \times P(R_2) + P(B_1) \times P(B_2)$$

$$= \frac{6}{14} \times \frac{5}{13} + \frac{8}{14} \times \frac{7}{13} = \frac{86}{182} = \frac{43}{91}$$

Exercise

One ticket is drawn at random from a bag containing 30 tickets numbered 1–30. Represent the sample space and the event of drawing a ticket containing a number which is a prime; also, find the number of elements in them.

Solution

Let S be the sample space and E be the event of occurrence of a prime number.
Then,

$$S = \{1, 2, 3, 4, 5, 6, \ldots, 30\}$$

and

$$E = \{2, 3, 5, 7, 11, 13, 17, 19, 23, 29\}$$

Also, n (S) = 30 and n (E) = 10.

Exercise

If a leap year is selected at random, what is the chance it will contain 53 Sundays?

Solution

A leap year has 367 days, that is 52 complete weeks and two more days.
These two days will be two consecutive days of the week.
A leap year will have 53 Sundays if out of the two consecutive days of a week selected at random one is a Sunday.
Let S be the sample space and E be the event that out of the two consecutive days of a week, one is Sunday; then,
S = {(Sunday, Monday), (Monday, Tuesday), (Tuesday, Wednesday), (Wednesday, Thursday), (Thursday, Friday), (Friday, Saturday), (Saturday, Sunday)}

$$\therefore n(S) = 7$$

and
E = {(Sunday, Monday, Tuesday), (Friday, Saturday, Sunday), (Saturday, Sunday, Monday)}

$$\therefore n(E) = 3$$

Therefore, the required probability is

$$P(E) = \frac{n(E)}{n(S)} = \frac{3}{7}$$

Exercise

A bag contains seven black and six white balls. Three balls are drawn at random. Find the probability that

1. All three balls are white
2. All three balls are black
3. One ball is black and two balls are white

Solution

Let S be the sample space, E_1 be the event of getting three white balls, E_2 be the event of getting three black balls and E_3 be the event of getting one black ball and two white balls.

Therefore, n (S) = number of two ways of selecting 3 balls out of 13 (i.e. 7 black + 6 white):

$$C_3^{13} = \frac{13 \times 12 \times 11}{1 \times 2 \times 3} = 286$$

1. n (E_1) = number of ways of selecting three white balls out of five:

$$C_3^5 = \frac{5 \times 4 \times 3}{1 \times 2 \times 3} = 10$$

Therefore, P (getting three white balls) is

$$\frac{n(E_1)}{n(S)}$$

$$= \frac{10}{286}$$

$$= \frac{5}{143}$$

2. n (E_2) = number of ways of selecting three black balls out of eight:

$$C_3^8 = \frac{8 \times 7 \times 6}{1 \times 2 \times 3} = 56$$

Therefore, P (getting three red balls) is

$$\frac{n(E_2)}{n(S)}$$

$$= \frac{56}{286}$$

$$= \frac{28}{143}$$

3. $n(E_3)$ = number of ways of selecting one black ball out of seven and two white balls out of six:

$$C_1^7 \times C_2^6$$

$$= 7 \times (6 \times 5)/(2 \times 1)$$

$$= 105$$

Therefore, P (getting one black and two white balls) is

$$\frac{n(E_3)}{n(S)}$$

$$= \frac{105}{286}$$

Exercise

A card is drawn from a well-shuffled pack of playing cards. What is the probability that it is either a spade or an ace?

Solution

The equiprobable sample space of the experiment of drawing a card from a well-shuffled pack of playing cards consists of 52 sample points.

Let A be the event that the card drawn is a spade.

Then, A consists of 13 sample points and

$$P(A) = \frac{13}{52}$$

Let B be the event that the card drawn is an ace.

Then, B consists of four aces and

$$P(B) = \frac{4}{52}$$

Now, $(A \cap B)$ consists of only one sample point, the ace of spades, and

$$P(A \cap B) = \frac{1}{52}$$

Hence, P (the card drawn is either a spade or an ace) is

$$P(A \cup B)$$

$$= P(A) + P(B) - P(A \cap B) \text{ (by addition rule)}$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52}$$

$$= \frac{16}{52}$$

$$= \frac{4}{13}$$

Exercise

A class consists of 20 boys and 30 girls, of which half the boys and half the girls have blue eyes. Find the probability that a student chosen at random is a boy or has blue eyes.

Solution

The equiprobable sample space of the experiment of choosing a student from the class consists of 20 + 30 = 50 sample points.

Let A be the event that the student selected at random is a boy. Then, A consists of 20 sample points and

$$P(A) = \frac{20}{50}$$

Let B be the event that the student selected at random has blue eyes. Then, B consists of 10 + 15 = 25 sample points and

$$P(B) = \frac{25}{50}$$

Event $A \cap B$ consists of 10 boys who have blue eyes. Hence,

$$P(A \cap B) = \frac{10}{50}$$

Therefore, P (the student selected at random is a boy or has blue eyes) is

$$P(A \cup B)$$

$$= P(A) + P(B) - P(A \cap B)$$

$$= \frac{20}{50} + \frac{25}{50} - \frac{10}{50}$$

$$= \frac{7}{10}$$

Exercise

The odds in favour of A hitting a target are 5:6, and those of B are 4:5. Find the probability of the target being hit at all, when they both try.

Solution

The target will be hit when

A alone hits, that is Event A occurs.
B alone hits, that is Event B occurs.
Both A and B hit, that is Event AB occurs.

Hence, it is a problem of a compound event which consists of both the exclusive and inclusive events.

Thus, according to the rules of both addition and subtraction, the required probability is given by

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

where

$$P(A) = \frac{5}{11}, P(B) = \frac{4}{9}, \text{ and } P(AB) = \frac{5}{11} \times \frac{4}{9} \ldots . \{P(AB) = P(A) \times P(B)\}$$

$$\therefore P(A \cup B) = \frac{5}{11} + \frac{4}{9} - \left(\frac{5}{11} \times \frac{4}{9}\right)$$

$$= \frac{89}{99} - \frac{20}{99} = \frac{69}{99}$$

Exercise

Suppose that a bag contains 40 balls, of which 12 are red, 5 are white, 9 are yellow and 14 are black. Find the probability of getting a white and a black ball.

Solution

The probability of getting a white ball is 5/40.
The probability of getting a black ball is 14/40.
The probability of getting a white and a black ball is calculated thus:

$$\frac{5}{40} + \frac{14}{40} = \frac{5+14}{40} = \frac{19}{40}$$

Exercise

Two cards are randomly drawn from a pack of 52 cards and thrown away. What is the probability of drawing an ace in a simple draw from the remaining 50 cards?

Solution

There are three cases when two cards have been drawn from a pack of 52 cards:

Case 1: There is no ace in the cards thrown away; the probability of getting an ace out of $50 = \frac{4}{50}$.
Case 2: There is one ace in the cards thrown away; the probability of getting an ace out of $50 = \frac{3}{50}$.
Case 3: There are two aces in the cards thrown away; the probability of getting an ace out of $50 = \frac{2}{50}$.

Hence, the required probability is

$$\frac{4}{50} + \frac{3}{50} + \frac{2}{50}$$

$$= \frac{9}{50}$$

Exercise

A bag contains six white, five black, four yellow and five red balls. What is the probability of getting a black or red ball at random in a single draw of one?

Solution

Total balls = 6 + 5 + 4 + 5 = 20
Probability of getting a black ball = 5/20
Probability of getting a red ball = 5/20
Therefore, the probability of getting a black or a red ball is

$$= \frac{5}{20} + \frac{5}{20}$$

$$= \frac{10}{20}$$

$$= 0.5$$

Exercise

What is the probability of getting a total of either five or seven in a single throw with two dice?

Solution

Total ways N = 36
Let A = the sum of 5, which can be obtained as (1, 4), (2, 3), (3, 2), (4, 1).
So, favourable ways to A = 4.

$$\therefore P(A) = \frac{4}{36}$$

Let B = the sum of 7, which can be obtained as (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1). So, favourable ways to B = 6.

$$\therefore P(A) = \frac{6}{36}$$

$$\therefore P(5 \text{ or } 7) = P(A \cup B) = P(A) + P(B)$$

$$= \frac{4}{36} + \frac{6}{36} = \frac{10}{36} = \frac{5}{18}$$

Exercise

A card is drawn at random from a well-shuffled pack of 52 cards. Find the probability of getting

1. A jack or a queen or a king
2. A three of hearts or three of diamonds

Solution

1. In a pack of 52 cards, we have four jacks, four queens and four kings.

Now, clearly a jack and a queen and a king are mutually exclusive events.
Also,

$$P(\text{a jack}) = \frac{C_1^4}{C_1^{52}} = \frac{4}{52} = \frac{1}{13}$$

$$P(\text{a queen}) = \frac{C_1^4}{C_1^{52}} = \frac{4}{52} = \frac{1}{13}$$

$$P(\text{a king}) = \frac{C_1^4}{C_1^{52}} = \frac{4}{52} = \frac{1}{13}$$

Therefore, by the addition theorem of probability,
P (a jack or a queen or a king) = P (a jack) + P (a queen) + P (a king)

$$= \frac{1}{13} + \frac{1}{13} + \frac{1}{13} = \frac{3}{13}$$

2. P (three of hearts or three of diamonds) = P (three of hearts) + P (three of diamonds)

$$= \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{3}{52}$$

Exercise

Bag A contains three white and four red balls and Bag B contains five white and six red balls. One ball is drawn at random from one of the bags and it is found to be red. Find the probability that it was drawn from Bag B.

Solution

Let $E_1 \equiv$ the event of a ball being drawn from Bag A, $E_2 \equiv$ the event of a ball being drawn from Bag B, and $E \equiv$ the event of a ball being red.

Since both bags are equally likely to be selected,

$$P(E_1) = P(E_2) = \frac{1}{2}, P\left(\frac{E}{E_1}\right) = \frac{4}{7} \text{ and } P\left(\frac{E}{E_2}\right) = \frac{6}{11}$$

Therefore, the required probability is

$$P\left(\frac{E_2}{E}\right) = \frac{P(E_2) \times P\left(\frac{E}{E_2}\right)}{P(E_1) \times P\left(\frac{E}{E_1}\right) + P(E_2) \times P\left(\frac{E}{E_2}\right)}$$

$$= \frac{\frac{1}{2} \times \frac{6}{11}}{\frac{1}{2} \times \frac{4}{7} + \frac{1}{2} \times \frac{6}{11}} = \frac{\frac{3}{11}}{\frac{2}{7} + \frac{3}{11}} = \frac{21}{43}$$

Exercise

What is the chance of drawing a diamond face card from a pack of 52 well-shuffled cards?

Solution

A diamond card can be drawn out of $13 = C_1^{13} = 13$ ways.
Therefore, the probability of drawing a diamond card is 13/52 = ¼.

5.9 Summary

The concepts of probability were first developed to assist gamblers in games of chance, such as throwing dice or tossing a coin. However, the concepts were later applied in several areas of management and space technology.

Probability concepts can be broadly classified into three major types based on the approaches to the study of probability theory: classical approach, relative frequency approach and subjective approach. This chapter discussed the rules of probability. It also discussed Bayes' theorem of probability.

The need to develop a formal and precise expression for uncertainty in decision making led to different approaches to probability measurement. These approaches, namely, the classical, relative frequency and subjectivists' approaches, arose mainly to cater to different types of situations where we face uncertainties. Certain results in probability theory which are helpful in this context have been presented. In the final section of this unit, we showed a method to revise the probability estimate as added information on the outcome of the experiment becomes available.

REVIEW QUESTIONS

1. Give the meaning of probability.
2. Bring out the properties of probability.
3. Define probability and bring out the importance of probability.
4. Define mutually exclusive events.
5. Distinguish between independent and dependent events.
6. What do we mean by compound probability?
7. a. Define probability and explain the importance of this concept to statistics.
   b. State and explain Bayes' theorem.
8. Define the probability of the occurrence of an event. State and prove the addition theorem of probability.
9. Explain what do you understand by the term *probability*. State and prove the addition and multiplication theorems of probability.
10. Explain the concepts of independent and mutually exclusive events in probability. State the theorem of total and compound probability.

SELF-PRACTICE PROBLEMS

1. A card is drawn at random from a pack of cards. What is the probability that it is either a heart or the queen of spades?
   [Answer: 14/52]
2. A bag contains three red, four black and two white balls. What is the probability of drawing a red and a white ball, each ball being put back after it is drawn?
   [Answer: 4/27]
3. What is the chance that a leap year, selected at random, will contain 53 Sundays?
   [Answer: 2/7]
4. A bag contains five white and four black balls. They are drawn out one by one. Find the chance that the balls drawn will be alternatively white and black.
   [Answer: 1/126]
5. Two cards are drawn at random from the ordinary pack. If either of them is a king or if both are kings, both cards are replaced; otherwise, they are not replaced. Another card is then drawn at random. What is the probability that it is a king?
   [Answer: 0.08]

6

Statistical Decision Theory

6.1 Introduction

In every sphere of our life, we need to make various kinds of decisions. The ubiquity of decision problems, together with the need to make good decisions, has led many people, from different times and fields, to analyse the decision-making process. The analysis varies with the nature of the decision problem, so that any classification base for decision problems provides us with a means to segregate the decision analysis literature.

A necessary condition for the existence of a decision problem is the presence of alternative ways of action. Each action leads to a consequence through a possible set of outcomes, the information on which might be known or unknown. One of several ways of classifying decision problems has been based on this knowledge about the information on outcomes.

This chapter discusses the framework of decision making and how decisions can be made under uncertainty. It also explains how one can use marginal analysis and decision trees for decision making.

Decision making is not an easy task for the management of an industry because of the complexity of a business situation in the case of several existing business management and industrial problems. In many business situations, a lot of reasoning and thinking are involved in the business decision making. In this context, there is a need for business decision theory and its analysis under different business situations.

The types of classification are

1. The information on outcomes is rare deterministic and is known with certainty (decision making under certainty).
2. The information on outcomes is probabilistic, with the probabilities known or unknown (decision making under uncertainty).

The theory that has resulted from analysing decision problems in uncertain situations is commonly referred to as decision theory. Business decision making, and its analysis, is of vital importance for industry management.

6.2 Decision Theory

Every manager has a set of objectives that are to be achieved through the process of decision making. These objectives may be related to an increase in profit, a reduction of

cost, an increase in turnover, the outsourcing of activities, an increase in market share and so forth. Decision theory can be applied to any type of decision situation. The decision may involve short-range or long-range consequences.

6.2.1 Certain Key Issues in Decision Theory

1. Decisions can be viewed as either independent decisions or decisions figuring in the whole sequence of decisions that are made over a period of time.
2. Depending on the planning horizon under consideration, we have either a single-stage decision problem or a sequential decision problem. In real life, all decisions, past, present and future, can be considered sequential. The problem becomes combinatorial, and hence difficult to solve.
3. Valid assumptions in most cases help to reduce the number of stages, and make the problem tractable.

6.2.2 Applications of Business Decision Making

1. It is concerned with how to assist the management of an industry in making business decisions.
2. It provides a meaningful conceptual framework for improved business decision making.
3. Select the most profitable investment portfolio. Decide whether to invest in a new plant, research programmes, marketing facilities, even risky orders, etc.
4. It involves evolving modified business operation measures from time to time under varying business situations for achieving the maximised gains.
5. Its goal is to minimise the total cost of production and other business interests and earn goodwill in the market by meeting its social responsibilities.
6. Its goal is to maximise social welfare and the accruing total gains of business through the industry's business operations, but not at the cost of society.

6.2.3 Framework for Decision Making

Some of the characteristics common for most decision theory problems are the states of nature.

It is very difficult to predict the exact outcome resulting from a decision made to solve a problem. Hence, every alternative course of action carries some uncertainty (as the success of this course of action is not within the domain of a manager making the decision, and he can only give a subjective probability of its occurrence).

Managers try to overcome this uncertainty by attaching some probabilities to the various events that might take place.

6.2.4 Decision Making under Uncertainty

This is a problem situation where the probabilities associated with the occurrences of different states of nature are not known to the manager; that is, he does not have any prior data that could give the probability of occurrence of a particular state of nature.

In business situations, there are many such problems that a manager has to solve. The strategy selected by a manager to find a solution depends to a great extent on him and the policy of the organisation. Thus, a manager has to assign some probability values to these states in order to find the expected value under uncertainty.

6.2.5 Concept of Business Decision Making and Business Decision

Business decision making is an important function of the management of an industry. Business decisions cannot be determined by the management of an industry, and business decision theories cannot be propounded and described without having a clear concept of business decision making.

It is a fundamental process of determining the most effective business operational strategy or course of action. It is required to meet the preset objective of the business decision problem of an industry. With the help of the business-making process, the management of an industry can determine a practical, feasible and optimal business operational strategy. Such a decision of industrial management is known as its business decision.

6.3 Determinants

The management of an industry has to determine its rational business decisions from time to time under given business decision constraints. It has to analyse the causes and consequences of its business operations at regular intervals.

Determinants of business decisions are discussed in the next subsections.

6.3.1 Business Environment

The business decisions made in such a business environment are found to be deterministic in nature. They do not involve competitive or probabilistic phenomena. Under certainty, the business decisions are determined by using the simplex method of linear programming or input/output analysis or activity analysis. If the business outcome gain or loss cannot be predetermined with certainty due to existing competitive, risky or probabilistic phenomena, then the business decisions made in such a business environment are found to be probable in their nature. The existing business situation is known as a stochastic or risky business situation.

The business decisions cannot be made with certainty under a business situation because of its uncertainty. In this case, the subjective judgement, that is the experience and skill of a business decision maker, is involved in determining the most suited business decision to the given business decision problem.

Under such a business situation of uncertainty, business decisions of the most suitable nature are determined by techniques such as matrix games and minimax/maximin/minimax regret criteria.

6.3.2 Business Objective

The available solution to the given business decision problem under given postulates of business constraints determines the required business decision, which is not possible in the absence of a preset business objective.

6.3.3 Alternative Course of Action/Strategies

The manager analysing the problem has to select one of the two or more alternatives available to him. He must select an alternative that best achieves the objective of the organisation. For example, a manager facing a stock-out problem may have alternatives like

1. Studying the nature of demand
2. Maintaining additional inventory of finished goods
3. Adopting just-in-time delivery systems

The business decision problem of an industry exists only when a set of two or more limited or unlimited courses of action/strategies are available.

6.3.4 Decision Pay-Off or Pay-Off Matrix

Depending on the strategy of the decision maker, each selected alternative or combination of alternatives will result in positive or negative costs known as pay-off.

It is a matrix representation of employed courses of action/strategies by the industry and their states of nature. The business outcome is represented for each combination of course of action/strategy and state of nature.

The pay-off matrix is important for formulating a business decision problem. The decision pay-offs may be a fixed pay-off or a random variable. In the former case, it is of a deterministic nature, and in the latter, it is of a probabilistic nature.

6.3.5 Decision Criteria

The decision criteria/rules are important in determining business decisions of an industry under given hypotheses formulated on the basis of various decision theories. These decision criteria act as a guide in the choice of the best course of action. They may be economic and quantitative or qualitative in nature.

Business decisions are taken to fulfil criteria such as maximising profit, minimising cost and minimising cycle time.

The degree of uncertainty is considered for the following business situations:

1. Business decision making under certainty (deterministic business situation)
2. Business decision making under risk (stochastic business situation)
3. Business decision making under uncertainty (undeterministic business situation)

6.3.6 Miscellaneous Factors

1. Various market competition
2. Business location factor
3. Financial position of an industry
4. Consumer's behavioural trends and patterns
5. Nature and volume of business
6. Cost of capital investments
7. Business period

8. Means of business courses of action/strategies
9. Interference of government (non-business group) and so forth

These factors are taken into account to analyse the reliability element of determined business decisions in a given period

1. To endorse the required time-bound modifications
2. To have best business performance
3. To achieve its business ends without sacrificing its long-run business interest and goodwill
4. To accomplish its social goal of maximising social welfare

6.4 Business Decision Theory under Certainty

Here, the business decision can be determined by the deterministic decision model when all the determinants (elements) of a business decision problem are known, that is

1. Business environment
2. Business outcome or decision pay-off of each course of action/strategy
3. Business objective
4. Set of alternative courses of action/strategies
5. Decision pay-off matrix
6. Decision criteria
7. Miscellaneous factors of business decision

Exercise

Determine the best strategy from the given decision pay-off (gain) matrix shown in Table 6.1.

Solution

Apply the deterministic decision model approach (Table 6.2).
where:

TABLE 6.1

Decision Pay-Off (Thousands of Rupees)

	Objective			
Strategy	O_1	O_2	O_3	O_4
S_1	50	80	90	70
S_2	110	60	40	210
S_3	30	40	80	40

TABLE 6.2

Calculation of Total Pay-Off (Thousands of Rupees)

	Objective					
Strategy	**O_1**	**O_2**	**O_3**	**O_4**	**Total Pay-Off (Gain) P**	**Remarks**
S_1	50	80	90	70	$\mathbf{P}_1 = 290$	
S_2	110	60	40	210	$P_2 = 420$	**Optimum strategy** $= S_2$
S_3	30	40	80	40	$P_3 = 190$	

$$P_1 = 50 + 80 + 90 + 70 = 290$$

$$P_2 = 110 + 60 + 40 + 210 = 420$$

$$P_3 = 30 + 40 + 80 + 40 = 190$$

Therefore, the required best (optimum) strategy is S_2, where the total pay-off P is maximum, that is $P_2 = 420$ (thousands of rupees).

Exercise

Determine the best strategy from the given decision pay-off (loss) matrix shown in Table 6.3.

Solution

Apply the deterministic decision model approach (Table 6.4).
where:

TABLE 6.3

Decision Pay-Off (Thousands of Rupees)

	Objective		
Strategy	**O_1**	**O_2**	**O_3**
S_1	20	25	35
S_2	35	40	45
S_3	30	40	40
S_4	50	20	30

TABLE 6.4

Calculation of Total Pay-Off (Thousands of Rupees)

	Objective				
Strategy	**O_1**	**O_2**	**O_3**	**Total Pay-Off (Loss) P**	**Remarks**
S_1	20	25	35	$P_1 = 80$	Optimum strategy $= S_1$
S_2	35	40	45	$P_2 = 120$	
S_3	30	40	40	$P_3 = 110$	
S_4	50	20	30	$P_4 = 100$	

$$P_1 = 20 + 25 + 35 = 80$$

$$P_2 = 35 + 40 + 45 = 120$$

$$P_3 = 30 + 40 + 40 = 110$$

$$P_4 = 50 + 20 + 30 = 100$$

Therefore, the required best (optimum) strategy is S_1, where total pay-off P is maximum, that is $P_1 = 80$ (thousands of rupees).

6.5 Business Decision Theory under Risk (Stochastic Business Situation)

Here, the business decision can be determined by the probabilistic decision model (or stochastic model) under the existing competitive or risky business situation in which

1. The decision maker often knows the probability of occurrence of each state of nature (event) and corresponding act.
2. And also the pay-off of each state of Nature (event) and corresponding Act.

If the decision pay-off matrix of each state of nature and corresponding act without or with the probability of each state of nature is given, then the business decision, that is the choice of best (optimum) act (strategy) is determined by using either of the two decision criteria given below:

1. Expected monetary value criterion (EMV criterion)
2. Expected opportunity loss criterion (EOL criterion)

6.5.1 EMV Criterion

In both cases of given decision pay-off matrix of many states of nature (events) and corresponding acts (or strategies) without or with the probability of each state of nature (event), the EMV of each act is calculated by the formulae given below.

6.5.1.1 Without Given Probability of Each State of Nature (P_j, Not Given)

$$EMV_i = \sum_{j=1}^{n} a_{ij}, \quad i = 1, 2, \ldots, m.$$

where:

EMV_i = EMV of ith act (or strategy)
a_{ij} = pay-off of ith act (or strategy) corresponding to jth state of nature (event)
$\sum$ = summation
i = number of act (or strategy) or i = 1, 2, …, m
j = number of states of nature (events) or j = 1, 2, …, n

6.5.1.2 With Given Probability of Each State of Nature (P_j, Given)

$$EMV_i = \sum_{j=1}^{n} a_{ij} P_j, \quad i = 1, 2, \ldots, m.$$

where:

EMV_i = expected monetary value of ith act (or strategy)
a_{ij} = pay-off of ith act (or strategy) corresponding to jth state of nature (event)
P_j = probability of occurrence of nature (event), $P_j \geq 0$, $\Sigma P_j \leq 1$
$\sum$ = summation
i = number of act (or strategy) or i = 1, 2, …, m
j = number of states of nature (events) or j = 1, 2, …, n

Exercise

The decision pay-offs in thousands of rupees for each act and event are given in Table 6.5.

Determine the best (optimum) act by EMV criterion in the two cases (1) without the given probability of each event and (2) with the given probability of each event.

Solution

By EMV criterion

1. Without the given probability of each event case (Table 6.6) where:

$$EMV_X = (-30) + 300 + 500 = 770$$
$$EMV_Y = (-60) + (-200) + 700 = 440$$
$$EMV_Z = 300 + (-60) + 400 = 640$$

TABLE 6.5

Decision Pay-Off (Thousands of Rupees)

	Events		
Acts	**A**	**B**	**C**
X	–30	300	500
Y	–60	–200	700
Z	300	–60	400
Probability of events (P)	0.4	0.5	0.4

TABLE 6.6

Computation of Total EMV (Thousands of Rupees)

	Events				
Acts	**A**	**B**	**C**	**Total EMV**	**Remark**
X	–30	300	500	$EMV_X = 770$	Optimum act = X
Y	–60	–200	700	$EMV_Y = 440$	
Z	300	–60	400	$EMV_Z = 640$	

TABLE 6.7

Calculation of Total EMV (Thousands of Rupees)

	Events				
Acts	**A**	**B**	**C**	**Total EMV**	**Remark**
X	(–30) × 0.4	300 × 0.5	500 × 0.4	338	Optimum act = X
Y	(–60) × 0.4	(–200) × 0.5	700 × 0.4	156	
Z	300 × 0.4	(–60) × 0.5	400 × 0.4	220	
Probability of events (P)	0.4	0.5	0.4		

Here, the maximum total EMV is obtained by Act X. Therefore, the best (optimum) act is X, which yields Total $EMV_{max} = 770$ (thousands of rupees).

2. With the given probability of each event case (Table 6.7)
   Using the formula of EMV criterion,

$$EMV_i = \sum_{j=1}^{n} a_{ij} P_j$$

Therefore,

$$EMV_X = (-30) \times 0.4 + 300 \times 0.5 + 500 \times 0.4 = 338$$

$$EMV_Y = (-60) \times 0.4 + (-60) \times 0.5 + 700 \times 0.4 = 156$$

$$EMV_Z = 300 \times 0.4 + (-60) \times 0.5 + 400 \times 0.4 = 220$$

Here, the maximum total EMV is obtained by Act X. Therefore, the best (optimum) act is X, which yields Total $EMV_{max} = 338$ (thousands of rupees).

6.5.2 EOL Criterion

In both cases of given decision pay-off matrix of many states of nature (events) and corresponding acts (or strategies) without or with the probability of each state of nature (event), the EOL of each act is computed by the formulae given below.

6.5.2.1 Without Given Probability of Each State of Nature (P_j, Not Given)

$$EOL_i = \sum_{j=1}^{n} l_{ij}, \quad i = 1, 2, \ldots, m.$$

where:

EOL_i = EOL of the ith act (or strategy)

l_{ij} = opportunity loss of the ith act (or strategy) corresponding to the jth state of nature (event) or

l_{ij} = maximum pay-off of the jth state of nature (event) – pay-off of the ith act (or strategy) corresponding to the jth state of nature (event) or

$$l_{ij} = M_j - a_{ij}, \; a_{lj} \le M_j \le a_{mj}$$

$\sum$ = summation
i = number of act (or strategy) or i = 1, 2, …, m
j = number of states of nature (events) or j = 1, 2, …, n

6.5.2.2 With Given Probability of Each State of Nature (P_j, Given)

$$EOL_i = \sum_{j=1}^{n} l_{ij} P_j, \quad i = 1, 2, \ldots, m.$$

where:
EOL_i = EOL of the ith act (or strategy)
l_{ij} = opportunity loss of the ith act (or strategy) corresponding to the jth state of nature (event) or
l_{ij} = maximum pay-off of the jth state of nature (event) – pay-off of the ith act (or strategy) or

$$l_{ij} = M_j - a_{ij},\ a_{lj} \le M_j \le a_{mj}$$

P_j = probability of occurrence of nature (event), $P_j \ge 0$, $\Sigma P_j \le 1$
$\sum$ = summation
i = number of act (or strategy) or i = 1, 2, …, m
j = number of states of nature (events) or j = 1, 2, …, n

Exercise

The decision pay-offs in thousands of rupees for each act and event are given in Table 6.8.

Determine the best (optimum) act by EOL criterion in the two cases (1) without the given probability of each event and (2) with the given probability of each event.

Solution

Apply the EOL criterion:

1. Without given probability of each event
   The formula is

$$EOL_i = \sum_{j=1}^{n} l_{ij}, \quad i = 1, 2, \ldots, m.$$

TABLE 6.8

Decision Pay-Off (Thousands of Rupees)

	Events		
Acts	**A**	**B**	**C**
1	60	220	300
2	30	60	340
3	–20	300	500
4	–30	400	450
Probability of event (P)	0.3	0.6	0.4

TABLE 6.9

Calculation of Total EOL (Thousands of Rupees)

	Events				
Acts	**A**	**B**	**C**	**Total EOL**	**Remark**
1	60 – 60 = 0	400 – 220 = 180	500 – 300 = 200	0 + 180 + 200 = 380	
2	60 – 30 = 30	400 – 60 = 340	500 – 340 = 160	30 + 340 + 160 = 530	
3	60 – (–20) = 80	400 – 300 = 100	500 – 500 = 0	80 + 100 + 0 = 180	
4	60 – (–30) = 90	400 – 400 = 0	500 – 450 = 50	90 + 0 + 50 = 140	Optimum act = 4

TABLE 6.10

Computation of Total EOL (Thousands of Rupees)

	Events				
Acts	**A**	**B**	**C**	**Total EOL**	**Remark**
1	(60 – 60) × 0.3 = 0	(400 – 220) × 0.6 = 108	(500 – 300) × 0.4 = 80	0 + 108 + 80 = 188	
2	(60 – 30) × 0.3 = 9	(400 – 60) × 0.6= 204	(500 – 340) × 0.4 = 64	9 + 204 + 64 = 277	
3	[60 – (–20)] × 0.3 = 24	(400 – 300) × 0.6= 60	(500 – 500) × 0.4 = 0	24 + 60 + 0 = 94	
4	[60 – (–30)] × 0.3 = 27	(400 – 400) × 0.6 = 0	(500 – 450) × 0.4 = 20	27 + 0 + 20 = 47	Optimum act = 4
Probability event (P)	0.3	0.6	0.4		

$$l_{ij} = M_j - a_{ij}$$

The calculation of the total EOL (Table 6.9) shows that the best (optimum) act is 4, which results in total $EOL_{min} = 140$ (thousands of rupees).

2. With given probability of each event case
   The formula is

$$EOL_i = \sum_{j=1}^{n} l_{ij} P_j, \quad i = 1, 2, \ldots, m.$$

$$l_{ij} = M_j - a_{ij}$$

The computation of the total EOL (Table 6.10) shows that the best (optimum) act is 4, which results in total $EOL_{min} = 47$ (thousands of rupees).

6.6 Business Decision Theory under Uncertainty

In uncertain business situations in which the decision maker knows the decision pay-off of each state of nature (event) and corresponding act but does not know the probabilities of the states of nature (events), the business decision can be determined by the following criteria:

1. Maximin
2. Minimax

3. Maximax
4. Laplace
5. Hurwitz Alpha
6. Regret

The choice to use any one of the above decision criteria by the decision maker depends on his or her personal judgement, experience and the company's policy.

6.6.1 Maximin Criterion

This is a conservative approach to arrive at a business decision. Here,

1. The minimum pay-off for each act of the given events is determined in the given decision pay-off matrix.
2. The maximum of the determined minimum pay-offs is determined in it. The corresponding act to the determined maximin pay-off is the determined best (optimum) act out of given acts in the decision pay-off matrix.
3. In this case, the pay-off should be cost or any other economic variable of which minimum value is to be maximised by the best act.

Exercise

Determine the best act by the Maximin Criterion from the decision pay-off (cost) shown in Table 6.11.

Solution

Use the maximin criterion (Table 6.12).

This computation table of maximin (Table 6.12) shows that act a_2 results in maximin = –2. Therefore, the best act is a_2.

TABLE 6.11

Pay-Off

	States of Nature			
Acts	X_1	X_1	X_3	X_4
a_1	–3	–4	6	4
a_2	–2	1	9	5
a_3	–4	–5	7	8

TABLE 6.12

Computation of Maximin

	States of Nature						
Acts	X_1	X_2	X_3	X_4	**Min**	**Maximin**	**Remark**
a_1	–3	–4	6	4	–4		
a_2	–2	1	9	5	–2	–2	Best act = a_2
a_3	–4	–5	7	8	–5		

6.6.2 Minimax Regret Criteria (Savage Principle) or Criterion of Pessimism or Wald's Criterion

In this criterion, the decision maker ensures that he or she should earn no less or pay no more than some specified amount. Thus, he or she selects the alternative that represents the maximum of the minima or the minimum of the maxima in the case of loss pay-offs in each case of profit.

6.6.2.1 Working Method

Locate the minimum or maximum in the case of profit pay-off in the case of loss (or cost) values corresponding to each alternative.

Select an alternative with the best-anticipated pay-off maximum value for profit and minimum value for loss or cost.

As the decision maker is conservative about the future and anticipates the worst possible outcome minimum for profit and maximum for cost or loss, this is called a pessimistic decision criterion or Wald's criterion.

This decision criterion was developed by L J Savage. He pointed out that the decision maker might experience regret after the decision has been made and the states of nature, that is events, have occurred.

Thus, the decision maker should attempt to minimise regret before actually selecting a particular alternative/strategy.

Steps

1. Determine the amount of regret corresponding to each alternative for every state of nature. The regret for the jth event corresponding to the ith alternative is given by

$$j^{th}\ \text{regret} = \left(\text{maximum pay-off} - i^{th}\ \text{pay-off}\right)\ \text{for the}\ i^{th}\ \text{event}$$

2. Determine the maximum regret amount for each alternative.
3. Choose the alternative which corresponds to the minimum of the above maximum regrets.

It is an optimistic approach to arrive at a business decision. Here,

1. The pay-off for each act of the given events is determined in the given decision pay-off matrix.
2. The minimum of the determined maximum pay-offs is determined in it.
3. The corresponding act to the determined minimax pay-off is the determined best (optimum) act out of given acts in the decision pay-off matrix. In this case, the pay-off should be a profit or any other economic variable of which the maximum value is to be minimised by the best act.

Exercise

Determine the best act by minimax criterion from the decision pay-off (cost) shown in Table 6.13.

TABLE 6.13

Pay-Off

	States of Nature			
Acts	X_1	X_2	X_3	X_4
a_1	6	–3	8	11
a_2	–2	4	7	–3
a_3	5	1	–6	4

TABLE 6.14

Calculation of Maximin

	States of Nature						
Acts	X_1	X_2	X_3	X_4	**Max**	**Minimax**	**Remark**
a_1	6	–3	8	11	11		
a_2	–2	4	7	–3	7		
a_3	5	1	–6	4	5	5	Best act = a_3

Solution

Use the maximin criterion (Table 6.14).

This calculation table of maximin (Table 6.14) shows that act a_3 results in maximin = 5. Therefore, the best act is a_3.

6.6.3 Maximax Criterion

Here, the decision maker ensures that he or she should not miss the opportunity to achieve the greatest possible pay-off or lowest possible cost.

He or she selects the alternative decision choice or course of action that represents the maximum of the maxima or the minimum of the minima pay-off consequences or outcomes.

6.6.3.1 Working Method

1. Find the maximum or minimum pay-off values corresponding to each alternative or course of action.
2. Select an alternative with the best-anticipated pay-off maximum value for profit and minimum value for cost.

Since in this criterion the decision maker selects an alternative with the largest or lowest possible pay-off value, it is called an optimistic decision criterion.

It is the most optimistic approach to arrive at a business decision. Here,

1. The maximum pay-off for each act of the given event is determined in the given pay-off matrix.
2. The maximum of the determined maximum pay-offs is determined in it.
3. The corresponding act to the determined maximax pay-off is the determined best optimum act out of given acts in the decision pay-off matrix.

TABLE 6.15

Pay-Off

	States of Nature			
Acts	X_1	X_2	X_3	X_4
a_1	6	4	8	10
a_2	9	13	16	–2
a_3	1	9	21	5
a_4	–6	4	–3	2

TABLE 6.16

Computation of Maximax

	States of Nature						
Acts	X_1	X_2	X_3	X_4	**Max**	**Maximax**	**Remarks**
a_1	6	4	8	10	10		
a_2	9	13	16	–2	16		
a_3	1	9	21	5	21	21	Best act = a_3
a_4	–6	4	–3	2	4		

In this case, the pay-off should be a profit or any other economic variable of which maximum value is to be maximised by the best act.

Exercise

Determine the best act by maximax criterion from the pay-off (profit) shown in Table 6.15.

Solution

Use the maximax criterion (Table 6.16).

The computation table of maximax (Table 6.16) shows that act a_3 results in maximax = 21. Therefore, the best act is a_3.

6.6.4 Equally Likely Decision (Laplace Criterion)

As the probabilities of states of nature are not known, it is assumed that all states of nature will occur with equal probability; that is, each state of nature is assigned an equal probability.

As states of nature are mutually exclusive and collectively exhaustive, so the possibility of each of these must be 1/(number of states of nature).

6.6.4.1 Working Method

1. Assign an equal probability value to each state of nature by the formula 1/(number of states of nature).
2. Calculate the expected or average pay-off for each alternative course of action, by adding all the pay-offs and dividing by the number of possible states of nature or by applying the formula.

   Probability of state of nature j and P_{ij} pay-off value for the combination of the alternative and state of nature j

3. Select the best expected pay-off maximum value for profit and minimum value for cost.

   This criterion is not of much practical utility, as the decision maker is generally not completely ignorant of the various states of nature and the chances of their occurrence.

   It is useful when there is no information about the probability of occurrence of various events and the decision pay-off matrix is given.

Here,

1. The equal probability for each event is assumed on the basis of equal likelihood and the total expected pay-off (total EMV) of each act is calculated in the given decision pay-off matrix.
2. The maximum total EMV, that is total EMV_{max}, is determined in it.
3. The corresponding act to the determined total EMV_{max} is the determined best (optimum) act out of given acts in the decision pay-off matrix.

Here, the pay-off should be a profit or any other economic variable of which the total EMV is to be maximised by the best act.

Exercise

Determine the best act by the Laplace criterion from the pay-off (profit) in Table 6.17.

Solution

Use the Laplace criterion. On the basis of equal likelihood, all five events are assumed to have equal probabilities of occurrence. Therefore, each event is assigned a probability of 0.2. The computation table of the total EMV is given in Table 6.18.

This calculation table of total EMV (Table 6.18) shows that a_3 is the best act because it gives the maximised total EMV, that is Total $EMV_{max} = 5.8$.

6.6.5 Criterion of Realism (Hurwicz Alpha Criterion)

Here, a rational decision maker should be neither completely optimistic nor pessimistic, and therefore must display a mixture of both.

Leonid Hurwicz introduced the idea of a coefficient of optimism, denoted by α, to measure the decision maker's degree of optimism.

TABLE 6.17

Pay-Off

	Events				
Acts	X_1	X_2	X_3	X_4	X_5
a_1	5	1	–6	4	7
a_2	–3	7	10	2	7
a_3	8	4	3	5	4
a_4	9	6	–4	6	3

TABLE 6.18

Calculation of Total EMV

	Events						
Acts	X_1	X_2	X_3	X_4	X_5	**Total EMV**	**Remark**
a_1	$5 \times 0.2 = 1.0$	$1 \times 0.2 = 0.2$	$(-6) \times 0.2 = -1.2$	$4 \times 0.2 = 0.8$	$7 \times 0.2 = 1.4$	2.2	
a_2	$(-3) \times 0.2 = -0.6$	$7 \times 0.2 = 1.4$	$10 \times 0.2 = 2.0$	$2 \times 0.2 = 0.4$	$7 \times 0.2 = 1.4$	4.6	
a_3	$8 \times 0.2 = 1.6$	$4 \times 0.2 = 0.8$	$3 \times 0.2 = 0.6$	$5 \times 0.2 = 1.0$	$4 \times 0.2 = 0.8$	5.8	Best act = a_3
a_4	$9 \times 0.2 = 1.8$	$6 \times 0.2 = 1.2$	$-4 \times 0.2 = -0.8$	$6 \times 0.2 = 1.2$	$3 \times 0.2 = 0.6$	4.0	
Probability of events (P)	0.2	0.2	0.2	0.2	0.2		

This coefficient lies between 0 and 1, where 0 represents a completely pessimistic attitude about the future and 1 is a completely optimistic attitude about the future. Thus, if α is the coefficient of optimism, then (1 – α) represents the coefficient of pessimism.

According to Leonid Hurwicz, the maximin and maximax criteria are two extremes on the scale of optimism. The first criterion represents dismal pessimism, and the other represents wild optimism.

In between these two, there may be a criterion which determines the decision with the help of the degree of optimism (α), where $0 \leq \alpha \leq 1$.

When $\alpha = 0$, then there exists a maximin criterion (total pessimism). On the other hand, when $\alpha = 1$, then there exists a maximax criterion (total optimism).

According to this criterion, the decision index of each act is calculated from the given decision pay-off matrix by the formula

$$D_1 = \alpha \times M_1 + (1 - \alpha) \times m_1$$

where:

D_1 = decision index of ith act
α = coefficient of optimism
M_1 = maximum pay-off from any of the outcomes resulting from the ith act
m_1 = minimum pay-off from any of the outcomes resulting from the ith act

The calculated highest values decision index determines the best act in a given decision pay-off, where the value of α is given (assumed).

Exercise

Determine the best act by the Hurwicz criterion from the decision pay-off shown in Table 6.19, given $\alpha = 0.6$.

TABLE 6.19

Pay-Off

	Events		
Acts	X_1	X_2	X_3
a_1	35	25	20
a_2	50	30	15
a_3	70	35	13

TABLE 6.20

Calculation of Decision Index

	Events				
Acts	X_1	X_2	X_3	**Decision Index (D)**	**Remark**
a_1	35	25	20	$0.6 \times 35 + 0.4 \times 20 = 29.0$	
a_2	50	30	15	$0.6 \times 50 + 0.4 \times 15 = 36.0$	
a_3	70	35	13	$0.6 \times 70 + 0.4 \times 13 = 47.2$	Best act = a_3

Solution

Use the Hurwicz criterion (Table 6.20).

$$D_1 = \alpha \times M_1 + (1 - \alpha) \times m_1$$

$$a = 0.6,\ 1 - a = 1 - 0.6 = 0.4$$

The calculation table of decision index (Table 6.20) shows that the best act is a_3, which represents the highest-valued decision index, that is 47.2.

6.6.6 Regret Criterion

This criterion is between Maximin and Maximax Criteria. According to this, the 'expected opportunity loss' of each act, of corresponding event, is computed in the decision pay-off matrix and thus the expected opportunity loss (EOL) or regret table is obtained by the following formula.

$$l_{ij} = M_j - a_{ij}$$

where:

l_{ij} = opportunity loss of the ith act corresponding to the jth event
M_j = maximum pay-off of the jth event
a_{ij} = pay-off of the ith act corresponding to the jth event

The maximised value of EOL of each act is determined in the regret criterion table. The corresponding act is the best act which gives the minimum value of the determined maximised values of EOL of given acts in the computed regret criterion table.

Exercise

Determine the best act by the regret criterion from the decision pay-off shown in Table 6.21.

TABLE 6.21

Pay-Off

	Events			
Acts	S_1	S_2	S_3	S_4
a_1	1	–2	8	–4
a_2	–6	9	–3	–3
a_3	–11	5	16	13

TABLE 6.22

Calculation of Regret Criterion

	EOLs of Events					
Acts	S_1	S_2	S_3	S_4	EOL_{max}	**Remark**
a_1	1 – 1 = 0	9 – (–2) = 11	16 – 8 = 8	13 – (–4) = 17	17	
a_2	1 – (–6) = 7	9 – 9 = 0	16 – (–3) = 19	13 – (–2) = 15	19	
a_3	1 – (–11) = 12	9 – 5 = 4	16 – 16 = 0	13 – 13 = 0	12	Best act = a_3

Solution

Use the regret criterion (Table 6.22).

$$l_{ij} = M_j - a_{ij}$$

where:

l_{ij} = opportunity loss of the ith act corresponding to the jth event
M_j = maximum pay-off of the jth event
a_{ij} = pay-off of the ith act corresponding to the jth event

The calculation table of EOL or the regret criterion shows that the best act is a_3, which results in a minimum value of 12 for the EOL_{max} values of given acts a_1, a_2 and a_3.

6.7 Decision Tree Analysis

This is a diagrammatic presentation of a decision process. It is an excellent tool for

1. Making financial or number-based decisions where a lot of complex information needs to be taken into account
2. Providing an effective model in which alternative decisions and the implications of making those decisions can be laid down and evaluated
3. Helping the managers to get an accurate, balanced picture of the risks and rewards that can result from a particular decision
4. Evaluating the risk in decisions concerning investments, new product launch, outsourcing and so forth

Guidelines to Draw a Decision Tree

1. Start with a decision that needs to be made. This decision is represented by a small square; usually decision trees are drawn from left to right.
2. Each possible solution is represented by a line drawn from the decision square diverging to the right, and the solution is written along the line. These lines end with a result.
3. When the result of making that decision is uncertain, draw a small circle to represent that result/event.

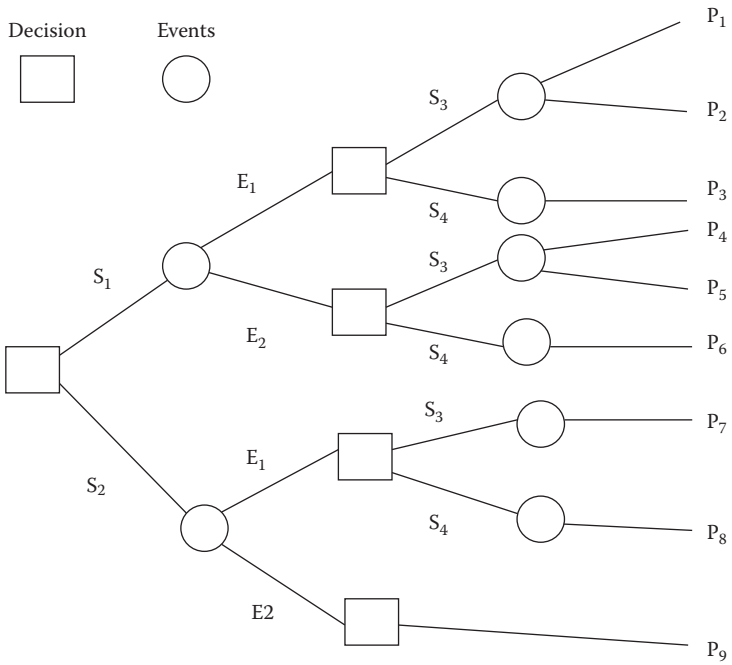


FIGURE 6.1
Decision tree model.

4. If the result is a decision, draw another square. Here, squares represent decisions and circles represent uncertainty (event) or random factors.
5. When the solution is complete at the end of the line, leave it blank. Figure 6.1 shows what a decision tree looks like.
6. Review this to find out whether any other solutions or outcomes can be considered for further evaluation. Then, prepare a final decision tree diagram.

Example 6.1: Construction of Decision Tree

A textile company is making plans to either launch a new product or consolidate its existing products. The company can launch new products in the market in two ways: (1) through detailed product development and (2) through rapid product development. If the company wants to consolidate, then it would do so by either strengthening its existing products through advertising and promotion or reaping the benefits of the brand name of the company without making any additional investments. For this decision, the company has employed a market research firm to find the market reaction of its products. The market research firm, after conducting a survey of the company's products, found that the market may have three reactions, good, average and poor, and accordingly calculated the profits for each reaction.

If the company goes for the detailed product development for launching new products, then it can make a profit of Rs. 1,000,000 when the market reaction is good, Rs. 50,000 when it is average and Rs. 2,000 it is when poor, and the probabilities of such reactions are 0.4, 0.4 and 0.2, respectively.

If the company goes for rapid product development for launching new products, then the company can make a profit of Rs. 800,000 when the market reaction is good,

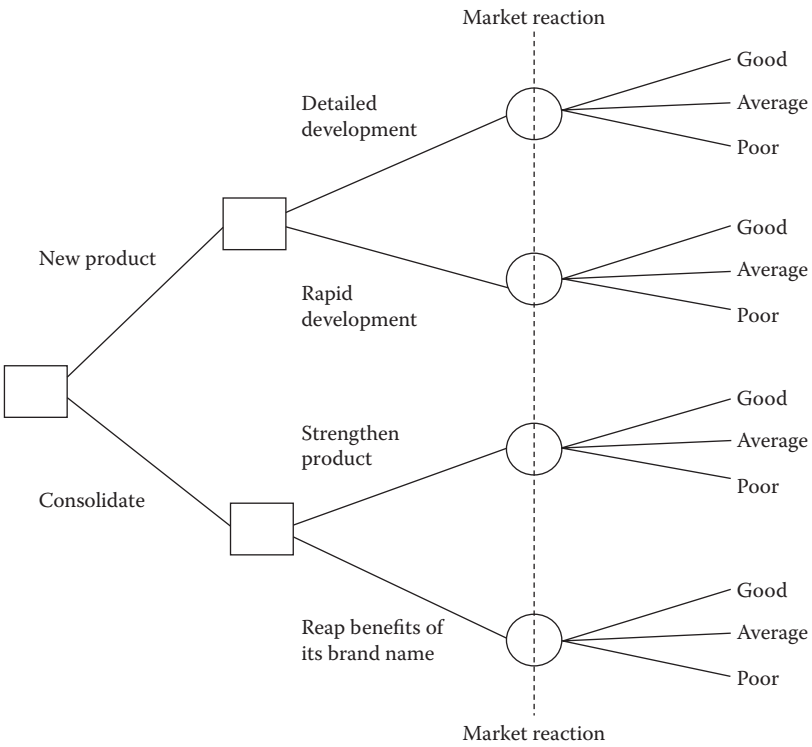


FIGURE 6.2
Decision tree model for textile company.

Rs. 25,000 when it is average and Rs. 2,000 when it is poor, and the probabilities of such reactions are 0.2, 0.1 and 0.7, respectively.

If the company strengthens its existing products for consolidation, then it can make a profit of Rs. 300,000 when the market reaction is good, Rs. 20,000 when it is average and Rs. 6,000 when it is poor, and the probabilities of such reactions are 0.2, 0.4 and 0.4, respectively.

If the company goes for consolidation of its existing products without making any additional investments, that is reaping the existing products, then the company can make a profit of Rs. 20,000 when the market reaction is good, Rs. 9,000 when it is average and Rs. 6,000 when it is poor, and the probabilities of such reactions are 0.3, 0.2 and 0.5, respectively. The detailed development cost is Rs. 150,000, and the rapid development cost is Rs. 80,000, while the cost for strengthening the existing products is Rs. 30,000. Now, the company has to decide whether to launch new products or consolidate existing products. Figure 6.2 shows the decision tree for this problem.

Exercise

A textile company is contemplating the introduction of a revolutionary new product with new packaging to replace the existing product at a much higher price (S_1) or a moderate change in the composition of the existing product with new packaging at a small increase in price (S_2) or a small change in the composition of the existing, except the word *new*, with a negligible increase in price (S_3). The three possible states of nature or events are (1) high increase in sales (N_1), (2) no change in sales (N_2) and (3) decrease in sales (N_3). The marketing department of the company worked out the pay-offs in terms

TABLE 6.23

Pay-Off Matrix

	States of Nature		
Strategies	N_1	N_2	N_3
S_1	800,000	400,000	250,000
S_2	600,000	550,000	0
S_3	400,000	400,000	400,000

of yearly net profits for each of the strategies of three events (expected sales). This is represented in Table 6.23.

Which strategy should the concerned executive choose on the basis of the following criteria?

1. Maximin
2. Maximax
3. Minimax regret
4. Laplace

Solution

The pay-off matrix is rewritten as follows:

1. Maximin criterion (Table 6.24)
   The maximum of the column minima is 400,000. Hence, the company should adapt strategy S_3.
2. Maximax criterion (Table 6.25)
   The maximum of the column maxima is 800,000. Hence, the company should adopt strategy S_1.

TABLE 6.24

Pay-Off Matrix (Maximin Criterion)

	Strategies		
Status of Nature	S_1	S_2	S_3
N_1	800,000	600,000	400,000
N_2	400,000	550,000	400,000
N_3	250,000	0	400,000
Column minimum	250,000	0	400,000

TABLE 6.25

Pay-Off Matrix (Maximax Criterion)

	Strategies		
Status of Nature	S_1	S_2	S_3
N_1	800,000	600,000	400,000
N_2	400,000	550,000	400,000
N_3	250,000	0	400,000
Column maximum	800,000	600,000	400,000

TABLE 6.26

Pay-Off Matrix (Minimax Regret Criterion)

	Strategies		
Status of Nature	**S_1**	**S_2**	**S_3**
N_1	800,000 – 800,000 = 0	800,000 – 600,000 = 200,000	800,000 – 400,000 = 400,000
N_2	550,000 – 400,000 = 150,000	550,000 – 550,000 = 0	550,000 – 400,000 = 150,000
N_3	400,000 – 250,000 = 250,000	400,000 – 0 = 400,000	400,000 – 400,000 = 0
Maximum opportunity loss	250,000	400,000	400,000

TABLE 6.27

Pay-Off Matrix (Laplace Criterion)

Strategy	**Expected Return (Rs.)**	
S_1	1/3(800,000 + 400,000 + 250,000)	= 483,333.33
S_2	1/3(600,000 + 550,000 + 0)	= 383,333.33
S_3	1/3(400,000 + 400,000 + 400,000)	= 400,000

3. Minimax regret criterion (Table 6.26)
   Hence, the company should adopt the minimum opportunity loss strategy, S_1.
4. Laplace criterion (Table 6.27)
   Since we do not know the probabilities of states of nature, assume that they are equal. For this example, we would assume that each state of nature has a one-third probability of occurrence.
   Since the largest expected return is from strategy S_1, the executive must select strategy S_1.

6.8 Summary

In today's competitive environment, a manager has to fulfil various objectives, and in this process, he or she has to make decisions that are in line with the objectives. Some of the characteristics common for most decision theory problems are alternative courses of action, uncertainty and pay-off. In a business context, decisions are made under uncertainty or with perfect information. This chapter discussed various strategies adopted by managers while making decisions under uncertainty.

Decision theory provides us with the framework and methods for analysing decision problems under uncertainty. A decision problem under uncertainty is characterised by different alternative course of action and uncertain outcomes corresponding to each action. We have used the criterion of maximising the EMV. Thus, EMV basically assumes that the decision marker is risk neutral. We have seen how instead of maximising the EMV, we can maximise the expected preference, and thereby consider the decision maker's attitude towards risk. Also, we have examined certain other criteria that are helpful in making decisions, when the probabilities of occurrence of the outcomes are not known.

REVIEW QUESTIONS

1. Briefly explain business decision.
2. Briefly explain the stages of business decision.
3. Describe in brief different decision-making criteria and discuss the Bayesian criterion in detail.

SELF-PRACTICE PROBLEMS

1. Apply the minimax regret criterion to solve the following decision problem:

	E_1	E_2	E_3	E_4
A_1	30	–2	70	20
A_2	300	70	5	10
A_3	100	160	–10	5

2. Apply the EMV criterion to solve the following decision problem:

	A_1	A_2	A_3	Probability
E_1	20	–15	–120	0.2
E_2	300	450	500	0.8
E_3	600	750	550	0.1

3. The research department of Hindustan Lever has recommended to its marketing department the launch of a shampoo and has provided three different types. The marketing manager has to decide on one of the types of shampoo to be launched under the following estimated pay-offs for various levels of sales:

	Estimated Levels of Sales		
Type of Shampoo	**16,000**	**12,000**	**6,000**
Amla shampoo	40	20	20
Dove shampoo	50	25	6
Sunsilk shampoo	45	30	5

What will be the marketing manager's decision when each of the following criteria are applied?

a. Maximin
b. Minimax
c. Maximax
d. Laplace
e. Regret

7

Linear Programming and Problem Formulation

7.1 Introduction

Basically, statistics are classified as descriptive statistics and programmed statistics.

Descriptive statistics include mean, median, mode, range, interquartile range, quartile deviation, mean deviation, variances and standard deviation. Programmed statistics include operation research, χ^2-test and decision theory. Operation Research includes the linear programming problem (LPP), cybernetics, organisation theorem, behavioural sciences, general system theory, logic, mathematics and statistics.

Linear programming was developed by George B Dantzing during World War II to plan expenditure and returns in such a way that the costs to the U.S. Army decreased and the losses to the enemy increased. Today, linear programming has become the central technique of operational research and is used to determine an optimum schedule of interdependent activities in view of available resources.

In many activities, we often face the problem of optimisation (maximisation or minimisation) of profit, sales, cost labour and so forth, subject to a set of constraints. The constraints may concern demand for commodities, availability of raw materials, storage space, variation in costs and so on. In such situations where conditional optimisation is required, we may adopt the linear programming model (LPM).

It is a technique useful in solving decision-making problems which involve maximising a linear objective function subject to a set of linear constraints.

7.2 Linear Programming Problem

7.2.1 Linearity

The equation $y = a + bx$ is a linear, and the equation $y = a + bx + cx^2$ is parabolic and not linear. In equation $y = a + bx$, all variables have a power of 1, meaning they possess linearity.

1. LPP is a recently developed branch of mathematics.
2. George Dantzing formulated general LPP.
3. The acronym LP is a combination of two different terms: *linear* and *programming*.
4. Linearity signifies the nature of the relationship between variables; it is defined as a straight line and exhibits the pattern of a straight line when plotted on a graph.

5. Programming is a process or procedure to be carried out in a systematic and predetermined way to achieve an objective. It is an activity that passes through a series of logical sequences under a set of condition to reach a particular goal.

7.2.2 Definition of LPP

Linear programming, one of the techniques of operations research, has been applied to a wide range of business and economic problems. It is a technique for determining an optimum schedule of interdependent activities in view of available resources. LPP is also useful in solving problems in finance, budgeting and investment.

Mathematician Kohlar has defined linear programming as a method of planning and operation, involved in the construction of a model, of a real situation, containing the following elements:

1. Variables representing the available choices
2. Mathematical expressions
3. Relation of the variables to the controlling conditions
4. Reflection of the criteria to be used in measuring the benefits derivable from each of several possible plans
5. Establishment of the objective

7.2.3 Features of LPP

1. The relationship between variables is linear.
2. The objective function should be linear in its variable.
3. The equation or inequation which represents the limitation of resources is also linear.
4. Mathematic techniques are designed, developed and operated on mathematical principle.
5. The programme is optimised.
6. Concern is optimised by minimising some quantity and maximising some quantity.
   a. Minimising quantities = cost, expenses, losses and so forth
   b. Maximising quantities = profit, sale
7. It deals with interdependent activity and can influence the behaviour of other elements.
8. Restrictions or constraints exist in the problem to some extent.
9. It deals only with quantitative elements.

7.2.4 Importance of LPP

Linear programming technique is extensively used in management, administration, trade and industry. Its uses are so extensive that wherever there is a need for planning under uncertainty, choosing the best course of action from various alternatives, linear programming is applied. With the development of scientific management, managers started

exploring new techniques to maximise returns and minimise costs. This policy of modern management made linear programming an important technique in business. Its importance as a technique can be studied in the following areas.

7.2.4.1 Modern Management

The administration and organisation of trade or non-trade concerns includes a congregation of different departments, sections, branches or units. Some of the departments play a key role in the achievement of organisational goals, and others are supportive. Its common features are that the resources, such as capital, raw materials, skilled employees and machinery, are scarce and needs are more. The technique of linear programming allocates the limited resources in a logical way so that key departments can function more perfectly.

Scientific management approaches the problems with standardisation of task and minimisation of cost. Optimisation of the capacity of humans and machinery, tools, equipment and so forth is another objective of scientific management. Such issues can be determined with the help of linear programming.

Top executives continuously encounter problems when several levels exist in a managerial hierarchy with different departments, where they work with different objectives and styles. Multiple and conflicting objectives of key departments confuse the policy makers and planners planning the activities of each department to achieve institutional goals. The linear programming technique gives a rational and logical knowledge of the role of different departments, so that the policy makers can identify the key departments.

Coordination and control functions can also be made effective by choosing the most suitable course of action with the logic of linear programming.

Assigning tasks to employees working with different time schedules around the clock, such as TV, radio, railways and police departments, can be made more convenient, economical and suitable with the technique of linear programming. This has very specific applications where staff are fewer and tasks to be accomplished are more, and in preparing time schedules for different groups in different ways, without dissatisfying any group of employees.

7.2.4.2 Industry

Industry frequently experiences a common problem of less raw material, less number of available working hours of machines and so forth, compared with the requirements of production. If the capacity and demand in the market for products x and y is the availability of 100 units of raw materials, and the working capacity of machines is less than 100 units, this situation creates a problem of allocation of limited resources and time. Under such a situation, the linear programming technique suggests the most profitable channels to divert limited resources, so that profit can be maximised.

1. Product mix is another area where linear programming can be applied to determine the most profitable product mix. It suggests the products which can give maximum returns if produced to the optimum level.
2. Productivity can be improved by determining an ideal and profitable combination of inputs to be purchased and stored for optimum utilisation. It helps to reduce the costs of idle storage and maximises productivity per unit.

7.2.4.3 Other Uses

1. Linear programming can be applied in the determination of the investment pattern of different assets. Portfolio managers use this technique in the analysis of risk and return of different securities and assets.
2. The technique can be used for military and police operations during emergency conditions or political crisis.
3. Linear programming helps to minimise the cost of transportation of products to different destinations within and outside of the industry.
4. Advertising agencies and sale promotion departments apply the technique of linear programming to select a suitable medium to advertise, such as the press, TV or posters, under the given limitation and advantages of each medium.
5. It can be used for the determination of an economical distribution system that will minimise transportation costs between two points of travel or the transport of humans and material.

7.2.5 Applications of Linear Programming

1. Blending problems
2. Production planning
3. Oil refinery management
4. Distribution
5. Financial and economic planning
6. Manpower planning
7. Farm planning
8. Selection of a product mix to maximise the profit
9. Determination of the capital budget which maximises the net present value (NPV) of the firm
10. Choice of mixing short-term financing which minimises the cost subject to certain funding constraints

It is an optimisation technique that is useful not only in industry and business but also in non-profit organisations.

7.2.6 Requirements of an LPP

1. The problem must have a well-defined single objective to achieve. For example, a firm manufactures two types of furniture, chairs and tables. The firm, in order to apply linear programming concepts for its production problems, should have a major objective, say profit maximisation.
2. There must be alternative courses of action, one of which will achieve the objective. For example, the firm may allocate its production resources, like raw materials and manufacturing capabilities, to chairs and tables in the ratio of 1:1, 1:2, 2:1 or some other ratio.

3. The decision variables must be continuous in nature. That is, the number of chairs and tables to be manufactured is flexible enough to take any non-negative values within a range.
4. Resources must be limited in supply, and achievement of the objective function is restricted by these constraints. That is, the furniture plant has limited manpower and machine-hours available. Hence, the more it allocates for chairs, the fewer tables it can make.

   The objective and constraints must be linear functions. That is, the manager should be able to express the firm's objectives and limitations (constraints) in the form of linear mathematical equations or inequalities. For example, suppose a manager has been given resources for use (i.e. available man-hours, quantity of raw material, machine time, etc.).

   a. The quantity of available resources is known to the manager.
   b. The objective is to determine the resources required or utilisation to optimise the goal of the firm, which can be profit maximisation, sales maximisation, cost optimisation and so forth.
   c. This situation requires a search for the variables that affect the objective and are subject to certain constraints. With the help of decision variables, their constraints, equalities or inequalities can be formed.
   d. They provide the optimal solution to achieve the objective.

7.2.7 Formulation of LPP

Formulation is important because it gives the meaning of the business decision problem. It is used to mean the process of converting the oral description into mathematical expressions, which represent the relevant relationships among decision factors, objectives and restrictions on the use of resources.

The general LPP with n decision variables and m constraints can be stated in the following form:

Let z be a linear function defined by

$$z = c_1x_1 + c_2x_2 + \ldots + c_nx_n \tag{7.1}$$

$x_1, x_2, x_3, \ldots, x_n$ are decision variables.

where c_j ($j = 1, 2, 3, \ldots, n$) are constants. Let a_{ij} be an $m \times n$ real matrix and let $\{b_1, b_2, \ldots, b_m\}$ be a set of constants such that

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &\geq b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &\geq b_2 \\ a_{m1}x_1 + a_{m2}x_2 + a_{mn}x_n &\geq b_m \end{aligned} \tag{7.2}$$

and

$$x_j \geq 0, j = 1, 2, \ldots, n \tag{7.3}$$

The problem of determining $(x_1, x_2, \ldots, x_n)$, which makes z minimum or maximum, which satisfies Equations 7.2 and 7.3, is called the general LPP.

The linear function $z = c_1x_1 + c_2x_2 + \ldots + c_nx_n$ is called the objective function of the general linear problem.

Equation 7.2 is called the constraints of the general LPP.

Equation 7.3 is usually known as the set of non-negative restrictions of the general LPP.

Solution: $(x_1, x_2, \ldots, x_n)$ real numbers which satisfy the constraints of the general LPP are called a solution.

Feasible solution: Any solution to a general LPP which satisfies the non-negative restrictions of the problem is called a feasible solution to the general LPP.

Optimum solution: Any feasible solution which optimises (minimises or maximises) the objective function of the general LPP is called the optimum solution.

The goal of the linear programming model is to maximise or minimise the objective function. In the case of maximisation of the objective function, the sign of the constraint function is $\leq$. In the case of minimisation of the objective function, the sign of the constraint function is $\geq$. Constraints may be represented by either inequalities or equalities.

7.2.8 Essential Requirements to Formulate LPP

The following are the basic requirements to apply the technique of LPP to optimise the achievement.

7.2.8.1 Decision Variables

Identify the variables that can be controlled or changed in order to optimally achieve the objective function. These variables have to be defined precisely and completely.

Problems arise when several tasks or activities or products compete for limited resources. These competing products or elements are called decision variables. A decision variable is any activity competing with other activities for limited resources. This is a common problem in the context of industrial production, where several products, activities (departments) and so forth compete for limited resources, capital or raw materials or time. It is necessary to decide the quantities of resources to be analysed for the variables that contribute the maximum towards achieving institutional objectives. The relationship among these variables should be linear.

7.2.8.2 Objective Function

This is represented by a linear mathematical function, with the objective (minimisation or maximisation) precisely defined.

The linear function which is to be optimised is called the objective function. The purpose is expressed along with the constraints in the form of linear equations. Thus, the total process of LPP is to achieve the predetermined objectives, which aim at maximisation or minimisation of the results under study. The objective is expressed in the form of a function or equation; therefore, it is called the objective function. The objective is maximisation in the case of output; profits; productivity of humans, machines or materials; returns; and

so forth. Minimisation is applied in the case of expenses, investment, wastage, loss, cost of production and so forth. To determine the objective function, profit or cost has to be assessed. In stating the objective function, variables are expressed in the form of X_1, X_2 and so forth.

The objective of LPP is optimisation, but it is necessary to define or clearly express whether the objective is to maximise the returns or minimise the costs using the given resources. If Z is a profit maximisation objective, then it can be defined as $Z = 12X_1 + 10X_2 + 8X_3$.

7.2.8.3 Constraint Function

Identify all the constraints. Express them in the form of linear mathematical inequalities. Several problems arise in business or general administration when it is difficult to balance between demand for and supply of resources. For example, if supply is greater than demand, it leads to increased overhead costs and wastage. If demand is greater than supply, it leads to low productivity and indirect losses. Problems will be very critical when essential inputs are scarce and demand is more. This requires special attention of higher authorities to allocate available resources more logically and profitably to optimise results. Such limitations or obstacles are constraints, and the equations that express these limitations are called 'constraint functions'. The objective function will be adjusted according to the given limitation. Hence, the objective function will be followed by the constraint function.

7.2.8.4 Non-Negative Function

This is condition in LPP specifies that the value of the variable being considered in the LPP will never be negative. It will be either zero or greater than zero, but it can never be less than zero. Thus, it is expressed in the form of $x \geq 0$.

7.2.8.5 Alternative Course of Action

One of reason to apply LPP is to select the best course of action. When choices are limited or absent, there no problem will arise. Thus, the presence of different alternatives is a prerequisite to apply LPP.

7.2.8.6 Non-Negative Restriction

All decision variables assume non-negative values.

7.2.8.7 Linearity

For programming problems, it is an essential requirement. Both objectives and constraints must be expressed in terms of linear equations.

Results

1. If an LPP has many (two or more) optimal solutions, it is said to have multiple solutions.
2. If an LPP has only one optimal solution, it is said to have a unique solution.
3. There may be a case where the LPP may not have any feasible solution at all.

4. For some LPPs, the optimum value of of Z may be infinity. In this case, the LPP is said to have an unbounded solution.

7.3 Assumptions of Linear Programming Models

1. Proportionality
2. Additivity
3. Divisibility
4. Certainty

7.3.1 Proportionality

The objective function (Z) is a linear function of the decision variables (xi). For example, the material consumption per unit product remains constant irrespective of the quantities of production, and the total consumption is always proportional to the total production.

7.3.2 Additivity

The concept of linear programming does not consider any synergistic effects among the decision variables, while calculating their total value for the objective function or the constraints that they are associated with.

7.3.3 Divisibility

The decision variables in linear programming models are continuous in nature and can take any non-negative, real, numeric value within the range specified by the constraints. This decision variable is divisible and solves the problems that involve fractional values. For the variables, it does so in the same way in which the problems without any fractional values are solved. The solutions thus obtained are finally rounded off, without making a significant loss of quality in the solution.

7.3.4 Certainty

The LPM assumes that all the constants (Cj, Aij and Bj) have certain values. It assumes that the optimal solution exists for the problem only when the values attributed to the coefficients of variables are constant.

7.4 Graphical Method of Solving an LPP

Here, we will tackle the graphical solution approach. The graphical procedure is useful only when there are two decision variables. When there are more than two variables, it is not possible to plot the solution on a two-dimensional graph.

To find the optimal solution to LPP, we must identify a set, or region of feasible solutions. The technique used to identify the optimal solution is called the graphical solution approach.

Steps

1. **Formulation of the problem**. The problem is expressed in the form of a mathematical model. Here, the objective function and the constraints are written down.
2. **Graphically plotting the constraints**. The constraints are graphically represented. If any constraint is an inequality, with the presumption that it is an equality, it is represented by a straight line. And then, the area indicating the inequality is identified. Change the inequality sign in each constraint by an equality sign and draw all the straight lines on graph paper. Shade the common portion of the graph that satisfies all constraints simultaneously drawn so far. The concluding shaded area is called the feasible region, and any point inside the region is called the feasible solution.
3. **Identification of the region of feasible solutions**. The common region between various constraints and the non-negativity restrictions is identified. This is the region of feasible solutions. The corner points of this region are identified.
4. **Finding the optimal solution**. The values of Z at the various corner points of the region of feasible solutions are calculated. The optimum (maximum or minimum) Z among these values is noted. The corresponding solution is the optimal one.

7.4.1 Infeasible Solution

Infeasible is a condition that arises when no value of the variable satisfies all of the constraints simultaneously. This means there is no unique (single) feasible region. Such a problem arises due to wrong model formulation with conflicting constraints. Infeasible depends solely on the constraints and has nothing to do with the objective function.

Infeasibility is a state that results when there is no solution for an LPP which can satisfy all the constraints (Figure 7.1).

Max. $Z = 3\,x_1 + 2\,x_2$

Subject to the constraints

$x_1 + x_2 \leq 2$

$4\,x_1 + 2\,x_2 \geq 4$

$x_1 \geq 0, x_2 \geq 0$

After drawing the graph for locating the feasible region, it can be seen that there is no feasible region for the problem.

7.4.2 Unbounded Solution

A problem may have an unbounded solution; that is, it may have no limit on the constraints. In such cases, there is no solution (Figure 7.2).

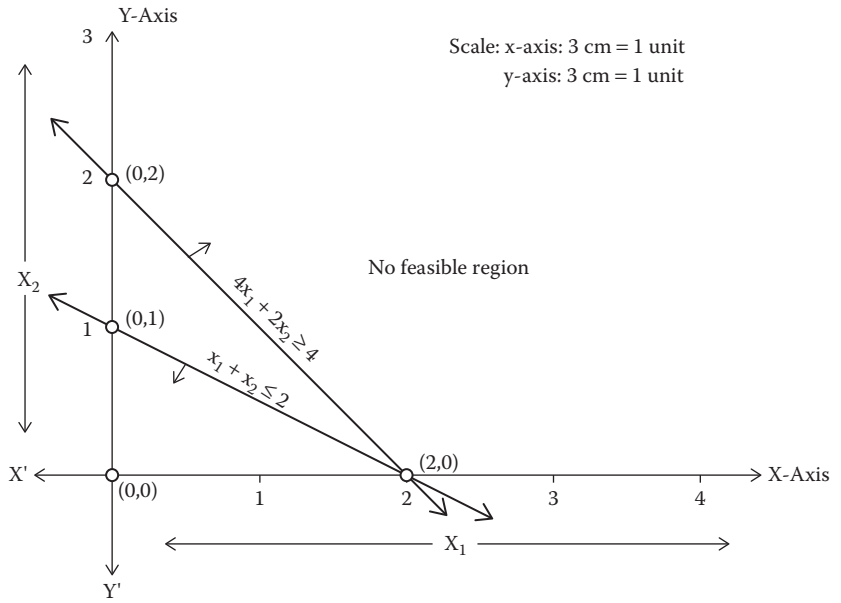


FIGURE 7.1
Graphical method – infeasible solution.

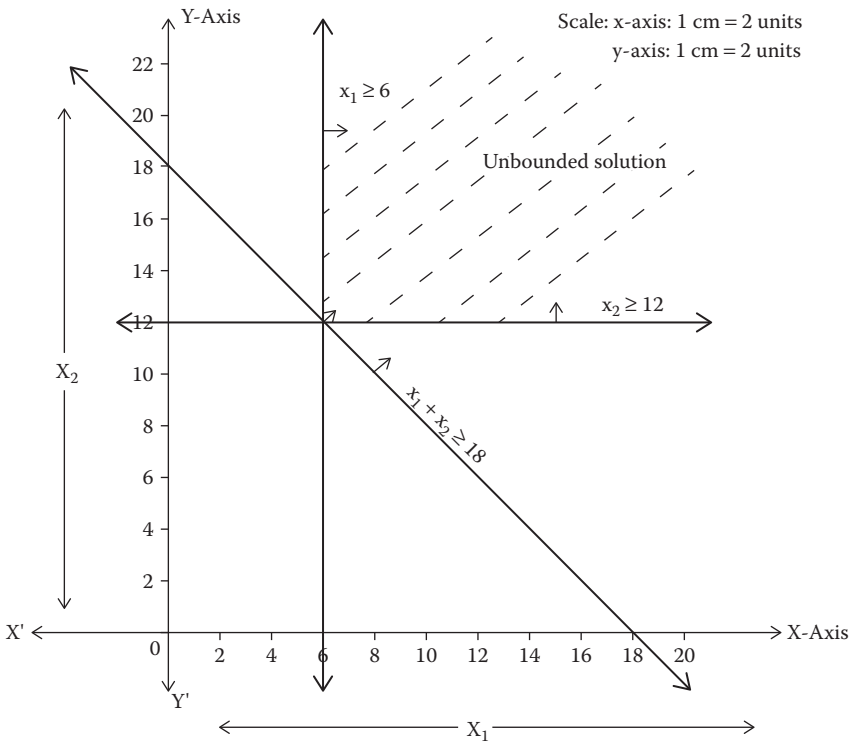


FIGURE 7.2
Graphical method – unbounded solution.

Max. $Z = 15\ x_1 + 17\ x_2$
Subject to the constraints
$x_1 \geq 6$
$x_2 \geq 12$
$x_1 + x_2 \geq 18$
In this case, as the feasible region extends infinitely to the right, the solution is unbounded.

7.4.3 Redundancy

There may be problems with redundant constraints; that is, a constraint may be present without any effect on the feasible region set.

Max. $Z = 3\ x_1 + 5\ x_2$
Subject to the constraints
$x_1 + 2x_2 \leq 16$
$2x_1 + 3x_2 \leq 30$
$x_2 \leq 30$

Here, the third constraint is redundant and is unnecessary to the problem as it has no effect on the feasible region.

7.4.4 Multiple Solutions

Some LPPs may have more than one optimal solution.
This happens when an objective function is parallel to one of the constraints. For example,

Maximise
$Z = 4x_1 + 6x_2$
Subject to the constraints
$4x_1 + 3x_2 \leq 10$
$8x_1 + 12x_2 \leq 19$
$x_1 \leq 7$
$x_1, x_2 \geq 0$

The objective function and the second constraint have the same slope (–3/2) and are parallel to each other. In such cases, the problems will have multiple optimal solutions. Hence, LPP is not always a simple technique and cannot solve each and every problem. However, it can be applied to solve a majority of the problems and hence plays an important role in decision making.

7.5 Duality

Each LPP (the primal problem) has an associated dual problem. For example, a maximisation of profit objective function, subject to resource constraints, has an associated dual problem.

The dual is a minimisation of the total costs of the resources subject to constraints that the value of the resources used in producing one unit of each output is at least as great as the profit received from the sale of that output.

For every linear programming formulation, there exists another unique linear programming formulation called the 'dual', and the original formulation is called the 'primal'. The dual can be considered as the 'inverse' of the primal in every respect.

7.5.1 Primal LPP versus Dual LPP

See Table 7.1 to compare primal LLP with dual LLP.

7.5.2 Conversion of Dual from Primal

Primal	**Dual**
The 'Primal' would be	The 'dual' formulation would be
Minimise $Z = 14x_1 + 16x_2$	Minimise $Z = 12\ y_1 + 10\ y_2$
Subject to the constraints	Subject to the constraints
$3x_1 + 6x_2 \leq 12$	$3\ y_1 + 3\ y_2 \geq 14$
$3x_1 + 2x_2 \leq 10$	$6\ y_1 + 2\ y_2 \geq 16$
$x_1 \geq 0, x_2 \geq 0$	$y_1 \geq 0, y_2 \geq 0$

Exercise

A factory produces three products A, B, and C from four raw materials P, Q, R and S. One unit of A requires three units of P, four units of Q and two units of S. One unit of B requires five units of Q, four units of R and three units of S. One unit of A requires five units of P, four units of R and five units of S. The company has 13 units of material P, 15 units of material Q, 17 units of material R and 19 units of material S. Profits per unit of products A, B and C are Rs. 5, 8 and 11, respectively. Formulate the problem into a linear programming model to maximise the profits.

Solution

1. Identify the decision variables. Let the number of products of Type A be x_1, the number of products of Type B be x_2 and the number of products of Type C be x_3.

TABLE 7.1

Primal LPP vs. Dual LPP

Primal	**Dual**
1. The column coefficients in the primal constraints	1. Become the row coefficients in the dual constraints
2. The coefficients of the primal objective function	2. Become the constants in the right-hand side of the dual constraints
3. The constants of the primal constraints	3. Become the coefficients in the dual objective function
4. The primal constraints have the inequalities of ≤	4. The dual constraints have the inequalities of ≥
5. The primal objective function is a 'maximisation' problem	5. The dual objective function is a 'minimisation' problem

TABLE 7.2

Data Summary Chart

Decision Variables	Products	Type of Raw Material				Profit per Unit (Rs.)
		P	Q	R	S	
x_1	A	3	4	0	5	5
x_2	B	0	5	4	3	8
x_3	C	5	0	4	5	11
Material availability		13	15	17	19	

2. Table 7.2 shows the data summary chart.
3. Define the objective function. The objective is to maximise the profit. Maximise
   $Z = 5x_1 + 8x_2 + 11x_3$.
4. Identify the constraints (availability of material):
   $3x_1 + 5x_3 \leq 13$
   $4x_1 + 5x_2 \leq 15$
   $4x_2 + 4x_3 \leq 17$
   $5x_1 + 3x_2 + 5x_3 \leq 19$
5. The problem can be stated as
   Maximise
   $Z = 5x_1 + 8x_2 + 11x_3$ (objective function)
   subject to
   $3x_1 + 0x_2 + 5x_3 \leq 13$
   $4x_1 + 5x_2 + 0x_3 \leq 15$
   $0x_1 + 4x_2 + 4x_3 \leq 17$
   $5x_1 + 3x_2 + 5x_3 \leq 19$
   $x_1, x_2, x_3 \geq 0$

Exercise

Morarji Textile produces shirts and trousers. It takes 2 hours to make a shirt, while a pair of trousers takes 4 hours. The factory has at most 1000 work-hours per day for the production of garments, and the packing department can pack at most 400 garments per day. If the shirt is sold for Rs. 300 and the trousers are sold at Rs. 500 per piece by the company, how many garments of each type should the factory produce per day to maximise its sales?

Solution

Suppose the company should produce x shirts and y trousers (see data summary chart in Table 7.3).

The problem can be stated as
Maximise
$z = 300x + 300y$
subject to

$$\left.\begin{aligned} x + y &\leq 400 \\ 2x + 4y &\leq 1000 \end{aligned}\right\} \text{Constraints}$$

The number of units produced cannot be –ve.

Therefore, $x \geq 0$, $y \geq 0$.

The line $x + y \leq 400$ meets the x-axis at A (400, 0) and the y-axis at B (0, 400).

TABLE 7.3

Data Summary Chart

	Shirts	Trousers	Max	Constraints
No. of units	$x \geq 0$	$y \geq 0$	400	$x + y + \leq 400$
Production time (hours)	2	4	1000	$2x + 4y + \leq 1000$
Selling price (Rs.)	300	500		$z = \text{Max } 300x + 500y$

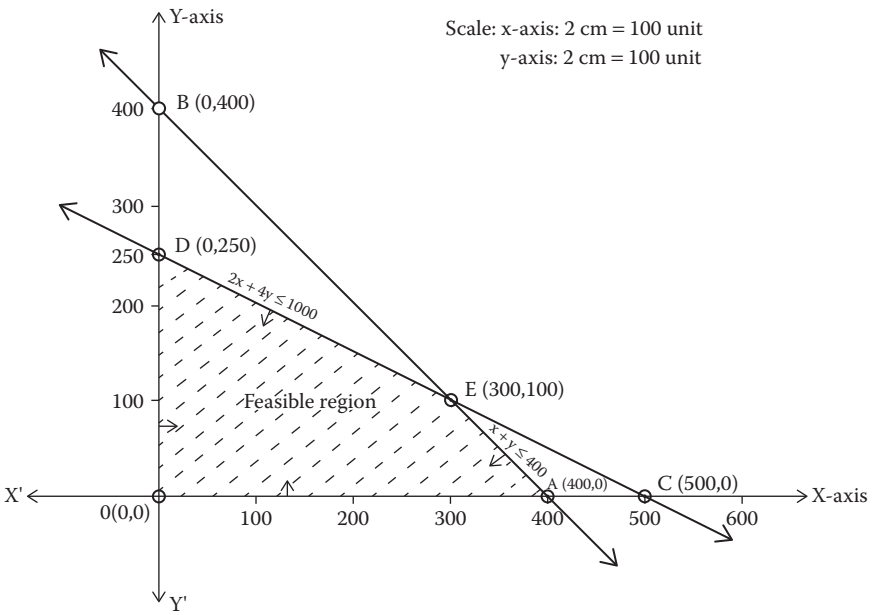


FIGURE 7.3
Graphical representation.

Therefore, $x \geq 0$, $y \geq 0$, $x + y \leq 400$ represent ΔOAB and its interior.

The line $2x + 4y \leq 1000$ meets the x-axis at C (500, 0) and the y-axis at D (0, 250).

Therefore, $x \geq 0$, $y \geq 0$, $2x + 4y \leq 1000$ represent ΔOCD and its interior.

The segments AB and CD intersect at E (300, 100) (Figure 7.3).

The feasible region is $\Delta OAED$ with extreme points O (0, 0), A (400, 0), E (300, 100) and D (0, 250).

The values of the objective function Z are shown in Table 7.4.

Thus, the maximum value of Z occurs at E (300, 100). Hence, the factory should make 300 shirts and 100 pairs of trousers to maximise the sale.

Exercise

A firm makes two products x and y and has a total production capacity of nine units per day. The firm has a permanent contract to supply at least two units of x and at least three units of y per day to another firm. Each unit of x requires 20 machine-hours of production time, and each unit of y requires 50 machine-hours of production time. The daily maximum possible number of machines-hours is 360. All the firm's output can be sold, and the profit made is Rs. 80 per unit of x and Rs. 120 per unit of y. Determine the production schedule for maximum profit and also calculate the profit.

TABLE 7.4

Values of the Objective Function Z

Extreme Points	Value of Z
O (0, 0)	300 (0) + 500 (0) = 0
A (400, 0)	300 (400) + 500 (0) = 120,000
E (300, 100)	300 (300) + 500 (100) = 140,000
D (0, 250)	300 (0) + 500 (250) = 125,000

TABLE 7.5

Data Summary Chart

Decision Variable	Products	Production Capacity (quintal)	Machine-Hours	Profit (Rs.)
x_1	x	1	20	80
x_2	y	1	50	120
Availability		9	360	

Solution

1. Identify the decision variable. Let x_1 be the number of units of product x, and x_2 be the number of units of product y.
2. See Table 7.5 for the data summary chart.
   The problem can be stated as
   Maximise
   (Profits) $Z = 80\, x_1 + 120\, x_2$
   subject to the constraints

$$x_1 + x_2 \leq 9$$
$$20x_1 + 50x_2 \leq 360$$
$$x_1 \geq 2, x_2 \geq 3$$

The line $x_1 + x_2 = a$ meets the x-axis at A (0, 9) and the y-axis at B (9, 0) and represents ΔOAB and its interior.

The line $20x_1 + 50x_2 = 360$ meets the x-axis at C (0, 7.2) and the y-axis at D (18, 0) and represents ΔOCD and its interior.

The line $x_1 = 2$ shows parallel line to y-axis at E (2, 0).

The line $x_2 = 3$ shows a parallel line to the x-axis at F (0, 3) (Figure 7.4).

From the graph, it is observed that the feasible region is $\Delta PQRS$ (Table 7.6).

The maximum profit (value of Z) of Rs. 960 is found at corner point R, that is $x_1 = 3$, $x_2 = 6$. Hence, the firm should produce three units of product x and six units of product y so as to achieve a maximum profit of Rs. 960.

7.6 Summary

The linear programming model is the most effective tool used to allocate the scarce resources effectively. LPP is an optimisation tool that is useful not only in industries and business, but also for a non-profit organisation in making various managerial or

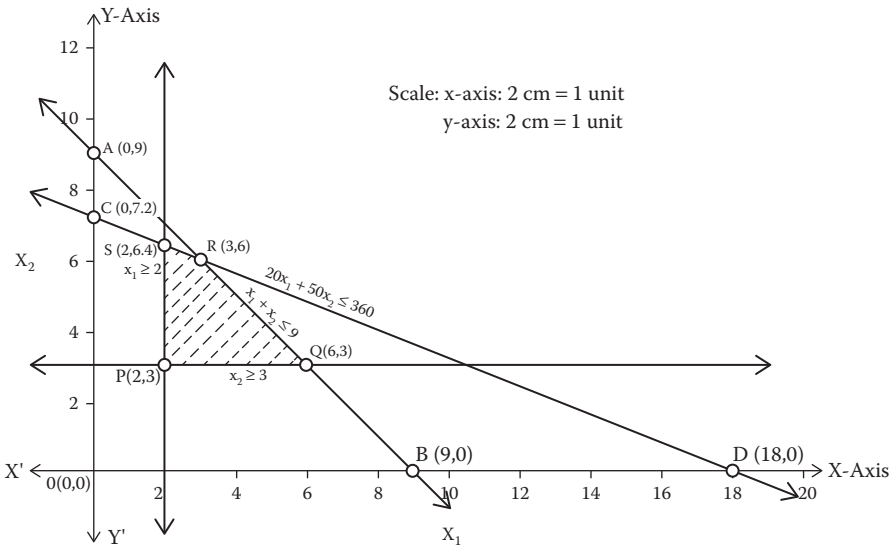


FIGURE 7.4
Graphical representation.

TABLE 7.6

Value of Objective Function Z

Corner Points	Coordinates of Corner Points (x_1, x_2)	Objective Function $Z = 80x_1 + 120x_2$	Value
P	(2, 3)	80 (2) + 120 (3)	520
Q	(6, 3)	80 (6) + 120 (3)	840
R	(3, 6)	80 (3) + 120 (6)	960 (max)
S	(2, 6.4)	80 (2) + 120 (6.4)	928

technological decisions. When there are only two decision variables, it is better to use the graphical method, as it is simple.

The emerging business scenarios coupled with the tremendous growth in competition have necessitated all the production organisations to allocate the scarce resources effectively among competing ends. Linear programming is an effective tool for dealing with such allocation problems. The principles of linear programming were developed to meet the destructive purposes of World War II. However, the principles were later used by industries for constructive purposes. The techniques of linear programming simplify the otherwise complicated procedures of decision making.

The principles of linear programming are based on certain assumptions, like proportionality, additivity, divisibility and certainty. This chapter discussed the methods of formulating a real-world problem as a linear programming model. It also discussed the various methods of solving the LPPs and the other issues in the linear programming concepts.

REVIEW QUESTIONS

1. What is linear programming?
2. What do we mean by optimisation? What is a feasible solution?

3. What are the steps of LPP formulation?
4. What is the graphical method?
5. What is the infeasible solution?
6. What is the unbounded solution?
7. What is the degeneracy condition?
8. Describe the limitations of LPP.

SELF-PRACTICE PROBLEMS

1. Solve the LPP.
   Minimise

$$Z = 2x_1 - 5x_2 + 7x_3$$

   subject to the constraints

$$5x_1 - 2x_2 + 4x_3 \leq 8$$

$$-3x_1 + 5x_2 \leq 14$$

$$-3x_1 + 2x_2 + 6x_3 \leq 12 \text{ and } x_1, x_2, x_3 \geq 0$$

2. A company makes two kinds of belts. Belt A is of high quality, and Belt B is of lower quality. The respective profits are Rs. 10 and 8 per belt. Each belt of Type A requires twice as much time as each belt of Type B, and if all belts were of Type B, the company could make 2000 belts per day. The supply of leather is sufficient for only 900 belts (both A and B combined). Belt A requires a fancy buckle, and only 500 such buckles are available per day. There are only 800 buckles a day available for Type B. Determine the number of belts to be produced for each type so as to maximise profit. Formulate and solve the problem graphically.
3. Use the graphical method to solve the problem.
   Minimise

$$18/4x_1 - 6x_2$$

   subject to the constraints

$$7x_1 - x_2 + x_3 \geq 9$$

$$2x_1 - x_2 + 5x_3 \geq 0 \text{ and } x_1, x_2, x_3 \geq 0$$

4. Solve the LPP.
   Maximise

$$Z = 5x_1 + 6x_2$$

subject to

$$3x_1 + 2x_2 \leq 10$$

$$3x_1 + x_2 \leq 6$$

$$3x_1 - x_2 \leq 10$$

$$x_1, x_2 \geq 0$$

5. A retired person wants to invest up to Rs. 40,000 in fixed income securities. His broker recommends investing in two bonds – Bond A yielding 8% and Bond B yielding 12%. After some consideration, he decides to invest at most Rs. 14,000 in Bond B and at least Rs. 8,000 in Bond A. He also wants the amount invested in Bond A to be at least equal to the amount invested in Bond B. What should the broker recommend if the investor wants to maximise his return on investment? Formulate the problem and solve graphically.

8

Sampling Theory

8.1 Introduction

A sample is not studied for its own sake. The basic objective of its study is to draw an inference about the population. Sampling helps to know the characteristics of the universe or population by examining only a small part of it. The values obtained from the study of a sample such as the average and dispersion are known as a 'statistics'. On the other hand, such values for the population are called 'parameters'. Any characteristic of a sample is called statistics. To study the population characteristic, a manager can go for either complete enumeration (census) or a sampling study. A sample statistic is a numerical summary measure calculated from sample data. The mean, median, mode and standard deviation calculated for sample data are called sample statistics.

8.2 Sample

It is obvious that for any statistical investigation, complete enumeration of the population is rather impracticable. For example, if we want to have an idea of the food expenditure (monthly) of the people in Chennai, then we use the help of sampling. A finite subset of statistical individuals in a population is called a sample.

In sampling some population, a parameter is inferred by studying only a part of the population, which is chosen to study and make the inference and is known as the sample. It is a finite subset of the population, selected from it with the objective of investigating its properties.

8.2.1 Differences between Random Sample and Non-Random Sample

See Table 8.1 for the differences between random and non-random samples.

8.2.2 Differences between Population and Sample

See Table 8.2 for differences between population and sample.

TABLE 8.1

Random Sample vs. Non-Random Sample

	Random Sample	Non-Random Sample
1.	In probability sampling, the entire item in the population has a chance of being chosen in the sample.	In a non-random sample, personal knowledge and opinions are used to identify the item from the population.
2.	Rigorous statistical analysis can be done with a probability sample.	A sample selected by judgement sampling is based on someone's expertise about the population.
3.	Random sampling cannot be done with a judgement sample.	Sampling judgement is sometimes used as pilot or trial sample.
4.	It is more convenient and can be used successfully, but we are unable to measure its validity.	It decides how to take a random sample later.

TABLE 8.2

Population and Sample

	Population	Sample
1.	Collection of items being considered.	Part or portion of the population chosen for study.
2.	Characteristics of population are called parameters.	Characteristic of sample are called statistics.
3.	Population size is represented by N.	Sample size is represented by n.
4.	Population mean is represented by μ.	Sample mean is represented by $\bar{x}$.
5.	Population standard deviation is represented by σ.	Sample standard deviation is represented by S.

8.2.3 Determination of Sample Size

For any sample survey, an important decision that has to be made while adopting a sampling technique is the size of the sample. The number of individuals in a sample is called the sample size. It is the number of units in the sample. Different opinions have been expressed by experts on this point. For example, some have suggested that the sample should be 10% of the size of the population, while others are of the opinion that a sample should be at least 15%.

However, these views are of little use as in practice, appropriate sample size depends on various factors relating to the subject under investigation, like the time aspect, the cost aspect and the degree of accuracy desired. Sampling theory is of little help in arriving at a good estimate of the sample size in any particular situation.

If we know the confidence level and the width of the confidence interval that we want, then we can find the approximate size of the sample.

We know that $E = z\sigma_{\bar{x}}$ is called the maximum error of estimate for population mean μ. As we know, the standard deviation of sample mean is equal to $\sigma/\sqrt{n}$. Therefore, we can write the maximum error of estimate of μ.

$$E = \frac{z\sigma}{\sqrt{n}}$$

So, the sample size

$$n = \frac{z^2\sigma^2}{E^2}$$

We can also find the sample size for estimating the population proportion. The maximum error, E, of the interval estimation of the population proportion, p, is

$$E = z\sigma_{\hat{p}} = z\sqrt{\frac{pq}{n}},\ q = 1 - p$$

So, the sample size $n = \frac{z^2pq}{E^2}$.

Exercise

Determine the sample size for the estimate of μ, when E = 2.4, σ = 23.54 and confidence level = 99%.

Solution

Here, E = 2.4, σ = 23.54 and z = 2.58.

So, sample size is

$$\frac{z^2\sigma^2}{E^2}$$
$$= (2.58)^2\ (23.54)^2/(2.4)^2$$
$$= 640.36$$

Thus, the required sample size is 640.

Exercise

An electronics company has installed a machine that produces a part that is used in transistors. The manager wants to know the proportion of these parts produced by this machine that are defective. According to manager, this estimate has to be within 0.02 of the population proportion for a 99% confidence level. What is the sample size that will limit the maximum error to within 0.02 of the population proportion?

Solution

The company wants the 99% confidence interval to be $\hat{p} \pm 0.02$; hence, E = 0.02.

For the most conservative estimate of sample size, we will generally use p = 0.50 and q = 0.50. The value of z for a 99% confidence level is 2.58.

So, the required sample size is

$$n = \frac{z^2pq}{E^2}$$
$$= (2.58)^2\ (.50)\ (.50)/(0.02)^2$$
$$= 6.66\ (0.25)/0.0004$$
$$= 4160$$

Thus, if the company takes a sample of 4160 parts, there is a 99% chance that the estimate of p will be within 0.02 of the population proportion.

8.3 Sampling

Sampling is the process of inferring something about a large group by studying part of it.

8.3.1 Population

Before defining the sampling, first we will define population. The finite and infinite set of individuals under study is called population or universe. Thus in statistics, population is an aggregate of objects, animate or inanimate, under study. The population may be finite and infinite.

The collection of all elements about which some reference is to be made is called the population.

It is a tool which enables us to draw conclusions about the characteristics of the population after studying only those objects or items that are included in the sample. For example, in an effort to study talcum powder uses in a state, the population could be the sum total of talcum powder users in major cities and towns in the state.

A sampling study consists of

1. Complete enumeration or census
2. Selective enumeration or sample method

8.3.2 Census or Complete Enumeration

In the case of a census or complete enumeration, information relating to characteristics of each and every unit of the population is collected. The unit may be an employee, product or department present in an organisation. The collection of these units under study is called population or universe. For example, when the study is intended to find out the working conditions of workers in the textile industry, the universe of the study will consist of all workers in this industry. Scanning through all the applications for the purpose of recruitment is a good example of complete enumeration.

Advantages

1. Findings of the census method are accurate and reliable
2. Scope for detailed study

Disadvantages

1. Resource constraints
2. Time constraints

8.3.3 Sample or Selective Enumeration

1. The business situation is not required to collect information regarding every unit of the population under study.
2. Some units picked from the population using a sampling technique are called a ‘sample’.
3. This process of selecting a sample and collecting relevant information of the units of the sample is known as sampling enumeration.

8.3.3.1 Characteristics of a Good Sample

While selecting a sample to study the population, a manager should look for the following characteristics:

1. The sample should have all the characteristics of the population from where it is taken.
2. The manager should not be biased in selecting the sample from the population.
3. The findings or decisions made based on the sample study should be applicable to the entire population.

8.4 Sampling Methods

The sampling methods are classified as follows:

1. Purposive or subjective or judgement sampling
2. Probability sampling
3. Mixed sampling

8.4.1 Purposive or Subjective or Judgement Sampling

In judgement sampling, the choice of sample items depends exclusively on the discretion of the investigator. The investigator exercises his judgement in the choice and includes those items in the sample which he thinks is the most typical of the universe with regard to the characteristics under investigation.

Example 9.1: Example for Sample Selection

A sample of 15 students is to be selected from a class of 90 for analysing the spending habits of the students; the investigator would select 15 students who, in his opinion, are the representative of the class.

In this method, a desired number of sample units is selected deliberately or purposely depending on the object of the inquiry, so that only the important items representing the true characteristics of the population are included in the sample.

Drawbacks

It is highly subjective in nature since the selection of the sample depends entirely on the personal convenience, beliefs, biases and prejudices of the investigator.

Example 9.2: Example for Sample Selection

If in a socio-economic survey it is desired to study the standard of living of the people in City A, and if the investigator wants to show that the standard has gone down, then he may include individuals in the samples only from the low-income stratum of the society and include the people from the path colonies, like A_1, A_2, A_3, A_4, A_5 and A_6, and so this method cannot be worked out for large samples and is expected to give good results in small samples only, provided the selection of the sample is representative.

8.4.2 Probability Sampling

This provides a scientific technique of drawing samples from the population according to some laws of chance in which the universe has some definite preassigned probability of being selected in the sample.

Different types of sampling are where

1. Each sample unit has an equal chance of being selected.
2. Sampling units have varying probabilities of being selected.
3. The probability of selection of a unit is proportional to the sample size.

8.4.3 Mixed Sampling

Sampling design in which the sample units are selected partly according to some probability laws and partly according to a fixed sampling rule (no use of chance) is called mixed sampling.

When a sample is required to be selected from a population, it is necessary to decide which method is to be applied. The choice would depend on the nature of the data and the purpose of the inquiry. The various methods available for sampling are

1. Simple random sampling
2. Stratified sampling
3. Systematic random sampling or quasi-random sampling
4. Cluster sampling
5. Multi-stage sampling
6. Area sampling
7. Quota sampling

8.5 Simple Random Sampling

In this technique, a sample is drawn so that each and every unit in the population has an equal and independent chance of being included in the sample.

If the unit selected in any draw is not replaced in the population before making the next draw, then it is called simple random sampling without replacement (SRSWOR).

And if it is replaced before making the next draw, then the sampling plan is called simple random sampling with replacement (SRWSR). An important feature of SRSWOR is that the probability of selecting a specified unit of a population at any given draw is equal to the probability of its being selected at the first draw.

In SRSWOR from a population of size N, the probability remains constant throughout the drawing.

8.5.1 Mathematically

If E_r is the event that any specified unit is selected at the rth draw, then $P(E_r) = 1/N = P(E_1)$; that is, the chance of selection of any specified items is the same at any draw as it was in the first draw, that is 1/N.

8.5.2 Selection of SR Sample

A random sample may be selected by

1. Lottery method
2. Use of table of random numbers

8.5.2.1 Lottery Method

This consists of identifying each and every number or unit of the population with a distinct number which is recorded on a slip or card. The slips should be as homogeneous as possible in shape, size, colour and so forth to avoid the human bias.

If the population is small, then the slips are put in a bag and thoroughly shuffled after each draw. The sampling units corresponding to the numbers on the selected slips will constitute a random sample. For example, we want to draw a random sample of 100 individuals from a population of 1000 individuals. We assign the numbers 1–1000 to identical slips bearing the numbers1–1000. These slips are then placed in a big container and shuffled thoroughly. Finally, a sample of 100 slips is drawn out one by one. The individuals bearing the numbers on these selected slips will constitute the desired sample.

If the population to be sampled is fairly large, then we may adopt the lottery method in which all the slips or cards are placed in a metal cylinder, which is thrown into a large, rotating drum working under a mechanical system. The rotation of the drum results in a sample of desired size n being drawn out of the container mechanically, and the corresponding n sample unit constitutes the desired random sample.

8.5.2.2 Use of Table of Random Numbers

In the table, random numbers, each of the digits 0, 1, 2, 3, 4, …, 9, appear with approximately the same frequency and independently of each other.

If we have to select a sample from a population of size N (≤99), then the numbers can be combined two by two to give pairs from 00 to 99. Similarly, if $N \leq 999$ or $N \leq 9999$ and so on, then combining the digits three by three (or four by four and so on), we get numbers from 000 to 999 or 0000 to 9999 and so on. Since each of the digits 0, 1, 2, …, 9 occurs with approximately the same frequency and independently of each other, so does each of the pairs 00–99, triplets 000–999, quadruplets 0000–9999 and so on. The method of drawing a random sample comprises the following steps:

1. Identify N units in the population with the number 1–N.
2. Select at random only a page of the random number table and pick up the numbers in any row, column or diagonal random.
3. The population units corresponding to the numbers selected in the above steps constitute the random samples.

Table 8.3 shows random numbers.

Exercise

Draw a random sample (without replacement) of 15 students from a class of 450 students.

TABLE 8.3

Random Numbers

2952	6641	3392	9792	7979	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2370	7483	3408	2762	3563	1089	6913	7691
0560	5246	1112	6107	6008	8126	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

TABLE 8.4

Grouping of Numbers in Threes

295	266	413	992	979	279	795	911	317	056	244
167	952	415	451	396	720	353	561	300	269	323
707	483	340								

TABLE 8.5

Students Corresponding to the Numbers

295	266	413	279	317	56	244	167
415	391	353	300	269	323	340	

Solution

First, we identity the 450 students of the college with numbers from 1 to 450. Starting with the first number in the above extract from Tippet's random number tables and moving row-wise. we pick out one by one the 3-digit numbers less than or equal to 450, until the 15 numbers ≤450 are discarded and the repeated numbers, if any, are taken only once. The above numbers grouped in threes are as shown in Table 8.4.

Thus, the students corresponding to the numbers are shown in Table 8.5.

Constitute the desired random sample of size 15 merits.

1. Since it is a probability sampling, it eliminates bias due to the personal judgements or discretion of the investigator.
2. Because of its random character, it is possible to ascertain the efficiency of the estimates by considering the standard errors of their sampling distributions.
3. The theory of random sampling is highly developed so that it enables us to obtain the most valuable and maximum information at the least cost and results in savings in time, money and labour.

8.5.2.2.1 Methods of Drawing Random Sample from Random Number Table

1. Identify each unit of the population by numbers in a serial manner.
2. Take up any page from the random number table and pick up the required number of 'numbers' in an orderly manner (row-wise or column-wise) until a value of the maximum serial number of the population is reached.
3. Select those units of the universe which bear the number thus picked up from the table. These units will constitute the desired sample.

8.5.2.2.2 Special Steps in Case of Small-Sized Universe

1. If the universe size is within three digits only: In such cases, it is advisable to convert the four-digit number of the table into three-digit numbers by taking the first three

digits of the first number, the fourth digit of the first number with the first two digits of the second number and so on. The four-digit numbers thus extracted from this Tippet's table can be converted into three-digit numbers row-wise as follows:

295, 266, 413, 992, 979, 279, 695, 911, 317, 056, 244 and so on

This conversion should be continued until the required number of random numbers is obtained for the sample. After this, the next two steps cited under 2 and 3 above (Section 8.5.2.2.2) are to be followed to get the sample.

If the universe size is within two digits only: In such cases, it is advisable to convert the four-digit numbers in the tables into two-digit numbers by breaking each random number into two equal parts. Thus, the random numbers given by Tippet, shown as above, can be converted into two-digit numbers row-wise as

29, 52, 66, 41, 39, 92, 97, 92, 79, 69 59, 11, 31, 70, 56, 24, 41, 67, 95, 24 and so on

As cited above, this conversion should be continued until the required number of random numbers is obtained for the sample. After this, the sample should be selected as per steps 2 and 3 mentioned above (Section 8.5.2.2.2).

If the universe size is too small, say $\leq$20, each random number should be split into two parts as above and each unit of the universe should be assigned different numbers out of certain serial numbers in a logical order, as shown below relating to a universe of 15 units.

The unit serial numbers assigned are

1. [1 + (1 × 15)], that is 16; [1 + (2 × 15)], that is 31; [1 + (3 × 15)], that is 46; and so on
2. [2 + (1 × 15)], that is 17; [2 + (2 × 15)], that is 32; [2 + (3 × 15)], that is 47; and so on

15 (last unit): 0, [0 + (1 × 15)], that is 15; [0 + (2 × 15)], that is 30; [0 + (3 × 15)], that is 45; and so on.

After this, the required number of units for the sample will be selected with reference to the converted two-digit numbers that correspond to the serial numbers thus assigned to the units of the population.

Exercise

Select a sample of 10 persons out of 5000 through the random numbers of Tippet reproduced earlier.

Solution

After identifying each of the 5000 persons by a serial number, the first 10 random numbers coming within a value of 5001 are picked up row-wise from the given table as

2952, 3992, 3170, 4167, 1545, 1396, 1300, 2693, 2370 and 3408

The persons corresponding to these numbers will constitute the desired sample.

Merits of simple random sampling

1. It works well for small populations.
2. Each element of the population has an equal probability of being included in the sample.

3. It is the easiest method of sampling.
4. It is the most commonly used method of sampling.
5. It does not require any additional information on the frame other than the complete list of members of the survey population, along with information for contact.
6. The theory behind it is that well-established standard formulae exist to determine the sample size; these formulae are easy to use.

Demerits of simple random sampling

1. Simple random sampling requires a complete and up-to-date list of the population units to be sampled. In practice, since this is not readily available in many inquiries, it restricts the use of this sampling design. This is hard to achieve in practice.
2. In field surveys, if the area of the coverage is fairly large, then the units selected in the random sample are expected to be scattered widely geographically, and thus it may be quite time-consuming and costly to collect the required information or data.
3. If the sample is not sufficiently large, then it may not be representative of the population, and thus may not reflect the true characteristics of the population.
4. The numbering of the population units and the preparation of the slips is quite time-consuming and uneconomical, particularly if the population is large.
5. It is expensive to conduct, as those sampled may be scattered over a wide area.

Note: The most widely used type of sampling is a simple random sampling.

1. In sampling with replacement (SRSWR), each card drawn is replaced back in the container before making the next draw.
2. In sampling without replacement (SRSWOR), with cards drawn and not returned, since cards are drawn one by one, a through mixing is required before the next draw.

8.6 Stratified Random Sampling

The population is divided into subgroups called subgroup strata and a sample is randomly selected from each stratum. When a population can be clearly divided into groups based on the same characteristics, we may use stratified random sampling. It guarantees each group is represented in the sample. The groups are also called strata. For example, college students can be grouped as full-time or part-time, male or female or traditional or non-traditional.

Once the strata are defined, we can apply simple random sampling with each group or stratum to collect the sample. It is used when the population is homogeneous. Each stratum is called a subpopulation. The entire subsample combines together to form the stratified sample. The process of obtaining and examining a stratified sample with a view to estimating the characteristics of the population is known as stratified sampling.

The different numbers of samples are drawn at random from different strata or divisions of the universe. For this, the entire universe is first divided into certain numbers of strata on the basis of certain criteria known as stratifying factors, such as age, sex, income,

education, status, geographical area, economic condition and sociological character. While dividing and subdividing a universe into certain strata, care must be taken to see that

1. There is a remarkable heterogeneity between the various strata.
2. There is a remarkable homogeneity between the different units of each stratum.
3. There is no overlapping between any strata, which means that no unit of the universe finds a place in more than one stratum or subdivision.
4. The number of the strata or subdivisions is not too large and remains preferably within nine, or else it may give rise to various complications.

8.6.1 Why Strata Are Created

1. It can make the strategy more efficient.
2. A larger sample is needed to get a more accurate estimation of a characteristic that varies greatly compared with a characteristic that does not. For example, if every person in a population had the same salary, then a sample of one individual would be enough to get a precise estimate of the average salary.

For example,

1. A retailer can use stratified sampling as explained below. She would analyse the bill copies according to the item purchased: TV buyers, stereo buyers and VCR buyers. For each stratum, random sampling would be done.
2. Suppose we want to select a sample of students from the university for the study of income level of newly joined students. In this case, instead of selecting a simple random sample or systematic random sample, we may consider applying a different method for selection of sample. First, we divide all students of the university into different groups based on income levels. (Auxiliary information related to the character under study may be used to divide the population into various groups.)
3. We may form three groups of low-, medium- and high-income students. We will now have three subpopulations, which are usually called strata. We then select samples from each stratum.
4. Suppose you want to estimate how many high school students have part-time jobs at the national level and also in each province. If you were to select a simple random sample of 25,000 people from a list of all high school students in Country A (assuming such a list was available for selection), you would end up, on average, with just a little more than 100 people from City P, since they account for less than half of a percent of the whole of Country A's population. This sample would probably not be large enough for the kind of detailed analysis you had in mind, stratified by province. Thus, in order to get good representation of City P, you would use a larger sample than the one allotted to it by the simple random sampling approach.

Stratification is most useful when the stratifying variables are

1. Simple to work with
2. Easy to observe
3. Closely related to the topic of the survey

8.6.2 Size of the Sample

After the universe is divided into a large number of strata, the next thing to do is to fix the size of the sample for the entire universe in a reasonable manner. After this, the next important thing to be done is to allocate the sample size to various strata sizes from which the data are to be collected. Such allocation can be done in any of the following manners:

1. Proportionate
2. Disproportionate
3. Optimum

8.6.2.1 Proportionate Manner

Under this manner, a sample size of the different strata is fixed at a certain percentage or ratio or in respective proportion to the size of the population, that is

$$\frac{n_1}{N}, \frac{n_2}{N}, \frac{n_3}{N}$$

where n_1, n_2 and n_3 represent the numbers of units in the first, second and third strata, respectively, and N is the numbers of units in the entire population. For example, if a population of 5000 units is divided into three strata containing 2500, 1500 and 1000 units, respectively, then the size of the samples of these three strata would be 5:3:2, which is the ratio of the size of the strata in the size of the population.

8.6.2.2 Disproportionate Manner

No proportion or ratio is maintained between the samples of the different strata. They are fixed arbitrarily without any regard for the sizes of the different strata. It may so happen that the size of all the strata may be fixed at an equal number not withstanding the different sizes of the strata.

8.6.2.3 Optimum Manner

The size of the samples of the different strata is fixed in the light of this optimisation principle, that is the principle of obtaining maximum benefits at the minimum cost. Thus, for a stratum which is expected to yield maximum information at a minimum cost, the size of the sample may be fixed at a large number, whereas for a stratum which is expected to yield less information and cause more expenditure, the size of the sample may be fixed at a smaller number.

Exercise

From the following data relating to the strength of the students in various classes of three faculties of a college, determine the number of arts, science and commerce students to be selected from each class in a sample of 5% if a proportionate stratified sample method is used. Also, determine their respective numbers to be selected for the sample if the size of the sample retains 5% of the universe. The data strength of the students in college is shown in Table 8.6.

TABLE 8.6

Strength of Students in Various Classes

	Faculties			
Class	**Arts**	**Science**	**Commerce**	**Total**
1st year	300	250	200	750
2nd year	200	150	100	450
3rd year	150	100	50	300
Total	650	500	350	1500

TABLE 8.7

Sample Size of the Different Strata

	Faculties			
Class	**Arts**	**Science**	**Commerce**	**Total**
1st year	15	13	10	38
2nd year	10	7	5	22
3rd year	8	5	2	15
Total	33	25	17	75

Solution

1. The size of the sample for the universe, that is the college, would be 75 (i.e. 5% of 1500, the total number of students in the college). Since the proportionate stratified method used is at 5%, the sample size of the nine different strata would be at 5% of their respective size, as shown in Table 8.7.

   Stratification means division into layers or groups. Stratified random sampling involves the following steps or subpopulation, called strata, such that

   a. The units written for each stratum (subgroup) are as homogenous as possible.
   b. The differences between various strata are as marked as possible; that is, the stratum means differ as widely as possible.
   c. Various strata are non-overlapping. This means each and every unit in the population belongs to one and only one stratum.

   Some of the commonly used stratifying factors are age, sex, income, occupying educational level, geographic area and economic status.

   Stratification will be effective only if it possesses the three characteristics, a, b and c, enumerated above.

   Thus, in stratified sampling the given population of size N is divided into say k relatively homogeneous strata of sizes N_1, N_2, …, N_k, respectively, such that

$$N=\sum_{i=1}^{k} N_i$$

2. Draw simple random samples (without replacement) from each of the k strata (i = 1, 2, 3, …, k) such that

$$\sum_{i=1}^{k} n_i = n,$$

where n is the total sample size from a population of size N. The sample of

$$\sum_{i=1}^{k} n_i = n,$$

units is called a stratified random sample (without replacement). The technique of drawing such a sample is called stratified random sampling.

Merits of stratified random sampling

1. More representatives. It provides more representative samples for a universe. Since the population is first divided into various strata and then the samples are drawn at random from each of the strata, it gives an adequate representation to each stratum or division of the population, and thereby overrules the possibility of any essential group of the population being left out completely. The sample itself is free.
2. More beneficial. It overcomes the drawbacks of both purposeful and random sampling, and all the same, it enjoys the advantages of both methods by dividing a heterogeneous universe into a number of homogeneous strata with respect to purposeful characteristics and then uses the technique of random sampling in the drawing of samples from each stratum. This type of sampling balances the uncertainty of random sampling against the bias of deliberate selection.
3. Greater precision. There is greater precision in the estimates of the various parameters (measures of the population) provided by this sampling, as variability in each stratum is reduced to a considerable extent.
4. The aim of stratification is to increase the efficiency of sampling by dividing a heterogeneous universe in such a way that (a) there is maximum homogeneity within each stratum and (b) a marked difference is possible between the strata.
5. Less expensive. It reduces the time and expenses of interviewing by concentrating the units of the different strata in a geographical manner.
6. Administrative convenience. It ensures administrative convenience by dividing the population into certain homogeneous strata and substrata. The various problems can be tackled effectively through stratification of the universe by considering each stratum as different and approaching them independently during sampling.
7. In stratified random sampling, the sampling is designed so that a designated number of items are chosen from each stratum. In simple random sampling, the sample items are chosen at random from the entire universe.

Demerits of stratified random sampling

1. Difficulty in weighting. It encounters difficulty in weighting the different strata. If the weights are not properly assigned, conclusions drawn through such sampling prove to be misleading.
2. Difficulty in stratification. It requires proper stratification of a heterogeneous universe into a number of homogeneous groups or strata, which is very difficult to do. If stratification were faulty, the results would be defective.
3. Difficulty in determination of the size of the strata. It needs proper determination of the sample size of the different strata, which involves several difficulties.

4. More expensive. Since the stratified samples are likely to be more widely distributed geographically, this method may prove to be more expensive in the matter of collection.
5. Overlapping of the strata. There is a likelihood of the different strata overlapping with each other. In such cases, the sampling will be difficult and its results disastrous.

Note:

1. Each stratum should be constructed in a way which will minimize differences among sampling units within strata, and maximize difference among strata.
2. The higher the homogeneity, the higher is the efficiency.
3. Members of a stratum are similar to each other and are different from the members of another stratum.

8.7 Systematic Random Sampling or Quasi-Random Sampling or Interval Sampling

Systematic sampling means there is a gap, or interval, between each selected unit in the sample. A systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced intervals until the sample has been formed.

The initial unit of the sample is selected at random from the initial stratum of the universe, and the other units are selected at a certain space interval from the arranged universe in a systematic order, like numerical, alphabetical or geographical. For this, the size of the sample is determined first on some basis, and then the following model determines the space interval of the universe.

$$K = \frac{N}{n},$$

where

K = space interval of the universe from which samples are to be taken
N = size of the universe
n = size of the sample

Thus, if there are 100 students from which only 10 are to be selected for the sample, all 100 students will first be arranged in serial order from 1 to 100, and then the space interval will be determined by

$$K = \frac{N}{n}$$

$$= \frac{100}{10}$$

$$= 10$$

After this, 1 student will be selected at random from the first 10 students. If this serial number comes out to be 9, then the serial numbers of the other nine students to be selected for the sample would be 19, 29, 39, 49, 59, 69, 79, 89 and 99. The sample would then consist of 10 students bearing the serial numbers 9, 19, 29, 39, 49, 59, 69, 79, 89 and 99.

Note: If the values of K come out in a fraction, they should be approximated to the nearest minimum integer, as shown below.

$$K = \frac{100}{9} = 11.11 \text{ approximated to } 11$$

$$K = \frac{100}{15} = 6.67 \text{ approximated to } 6$$

$$K = \frac{110}{20} = 5.5 \text{ approximated to } 5.$$

Exercise

There are 128 students in a BCom class bearing roll numbers from 501 to 628. Five students are to be selected from them for a study tour. Using the systematic random sampling method, determine the roll number of students who will be selected for the purpose.

Solution

Here, N = 128, n = 5 and

$$K = \frac{128}{5}$$

$$= 25.6$$

$$\cong 25$$

Let us select the students bearing roll number 525 from the first space interval at random as the initial unit of the sample. Then, the roll number of the other students to be selected would be at the every 25th position in the rolls of the class, counting from roll number 525. Thus, the roll numbers of the five students selected for the study tour would be 525, 550, 575, 600 and 625.

This method of sampling is also called quasi-random sampling. In this method, the initial unit of the sample is selected at random and the other units to be selected are predetermined by their space interval. This technique of sample selection is usually recommended where a complete and up-to-date list of the population is available and arranged in the some order.

Merits

1. It is very simple to understand and easy to operate. It is much easier than simple random sampling.
2. It provides satisfactory results. It is more efficient when elements are grouped together.
3. The time and work involved in sampling by this method are relatively less. It saves a lot of time and labour in comparison to simple random and stratified random sampling.

4. It enjoys most of the same advantages as proportionate stratified random sampling, as it is conducted more or less in that manner by dividing the universe into certain space intervals and selecting one unit of sample from each of them.
5. If the populations are sufficiently large, systematic sampling can often be expected to yield results similar to those obtained by proportional stratified sampling.

Demerits

1. It needs a complete and up-to-date frame of the population, which is not available in most cases.
2. It is less representative if we are dealing with the population having 'hidden periodicities'.
3. It selects unrepresentative items when there are cyclic features in the population and the space interval K is equal to or a multiple of the cycle. For example, if every fifth boy in a class happens to be a spendthrift and every such boy is selected for the sample, the same sample would not be representative of the class.

8.7.1 Application of Systematic Sampling

1. It can be used for estimation of the precision of results.
2. Systematic samples are treated as simple random samples.
3. The simple random sampling method for the selection of samples will not be practical if the size of the population is large.

For example, if we want to select 300 households from a list of 90,000, it is very time-consuming to take a simple random sampling. In such cases, it is better to use systematic random sampling.

In systematic random sampling, we first randomly select one member from the first k units (where k = N/n, N is population size and n is sample size). Then every kth member, starting with the first selected member, is included in the sample.

For the selection of samples in the above example, we would arrange all 90,000 households on any characteristic. Since sample size should be equal to 300, the ratio of population and sample size is 90,000/300 = 300. So k is 300. We randomly select one household from the first 300 households in the arranged list. Suppose we select the 105th household. This is the first unit of the sample. We then select every 300th from first selected household. In other words, our sample contains the households with numbers 105, 405, 705, 1005, 1305, 1605 and so on.

Note: Systematic random sampling does not give a simple random sample because every member of the population does not have the same probability of being selected.

8.8 Cluster Sampling

In cluster sampling, the population is divided into well-defined groups or clusters. Then, a few of these clusters are selected based on the assumption that they represent the entire universe. All the units of the selected cluster are studied to arrive at a conclusion. The

selection of these clusters is done by using any one of the sampling methods. All the units belonging to these selected clusters constitute the sample desired.

For example,

1. Suppose that the Department of Agriculture wishes to investigate the use of pesticides by farmers in Country U. A cluster sample could be taken by identifying the different cities in Country A as clusters. A sample of these cities (clusters) would then be chosen at random, so all farmers in those cities selected would be included in the sample. It can be seen here that it is easier to visit several farmers in the same cities than it is to travel to each farm in a random sample to observe the use of pesticides.
2. If we are interested in obtaining the income or opinion data in a city, the whole city may be divided into N different blocks or localities (which determine the cluster) and a simple random sample of n blocks drawn. The individuals in the selected blocks determine the cluster sample.

Note: Cluster sampling is an example of two-stage sampling or multi-stage sampling. In the first stage, a sample of area is chosen; in the second stage, a sample of respondents within that area is selected.

8.8.1 Importance of Cluster Sampling

1. It allows for great economies in data collection costs since the travel-related costs etc. are smaller.
2. The data collection for nearby elements (units) is easier, faster, cheaper and more convenient than observing units scattered over a geographical region.
3. In practical situations where the sampling efficiency is of less importance but the cost and time is of greater significance, the cluster sampling procedure is extensively used.
4. The population is divided into defined groups or cluster. All units of selected cluster are study. Selection of cluster is done by any sampling method. It is heterogeneous.

8.8.2 Application

1. In a prepoll survey, the entire voting population is divided into clusters. Some clusters are selected as samples, and every element of this cluster is studied.
2. The Department of Agriculture wishes to investigate the use of pesticides by farmers in Country U. A sample of cities or clusters is chosen at random.

Merits

1. This allows for great economic savings in data collection, as travel-related costs are smaller. It is simple to understand and easy to operate.
2. It provides an unbiased estimate of the population parameter if properly done.
3. Economically, it is more efficient than simple random sampling. It is less expensive and less time-consuming.

4. It produces satisfactory results when the clusters are of equal and small size.
5. The cost of the first sample is lower, especially with a geographic cluster.
6. It is easy to conduct without a population list.
7. It is often used in marketing research.
8. If the cost of data collection is say Rs. 40 under simple random sampling, it is only Rs. 10 under cluster sampling (i.e. one-fourth that of simple random sampling).

Demerits

1. It is rarely used in single-stage sampling plan.
2. There is a higher rate of sampling errors, which is difficult to measure.
3. It has greater economy than simple random sampling (as a geographically dispersed population can be expensive to survey – area sampling or geographical cluster sampling).
4. It may produce fallacious results if the cluster selected consists of a group of biased informants.
5. It cannot be recommended for an area with a large number of persons or households.

8.9 Multi-Stage Random Sampling

This is carried out in multiple stages, say two, three or four.

First, the universe is divided into some clusters from which certain clusters are selected at random as the first-stage samples.

Second, the selected first-stage samples are again subdivided into some clusters from which again certain clusters are selected at random as the second-stage samples.

Third, the selected second-stage samples are again subdivided into some clusters from which certain clusters are again selected at random as third-stage samples.

In this, the process of division and subdivision of the clusters and the selection of the multi-stage samples are carried out until the sample size is reduced to a reasonable extent. For example, in an urban inquiry of a sample of towns, a subsample of households may be selected and then, if necessary, from each of the selected households a third-stage sample of individuals may be taken.

Merits

1. It is very flexible compared with other methods of sampling.
2. In this method, the subsequent stages of sampling are needed only for a limited number of units, that is for those which were only selected in the preceding stages. As such, it saves a lot of time, energy and cost.
3. It leads to administrative efficiency by permitting the fieldwork to be concentrated and yet cover a large area.

4. It is of great utility in surveys of underdeveloped areas where no up-to-date and accurate frame is available for the subdivision of the materials into reasonably small sample units.

Demerits

1. It is likely to cause a large number of errors, as it involves a process of divisions and subdivisions of the various strata or clusters in different stages.
2. It leads to greater variability of the estimates than any other methods of sampling. In general, it is less efficient than a suitable single-stage random sampling.

8.10 Area Sampling

The samples are collected at random on an area basis. For this, the entire universe is first divided into certain parts on a geographical basis, and then from each such part the required number of items are selected at random, which constitutes the desired sample. This method is suitable for conducting an opinion poll with regard to any common problems of the population.

Merits

1. It ensures proportional representation of each of the segments of the population.
2. It enjoys the advantages of the stratified sampling method discussed earlier.

Demerits

1. It requires the arrangement of the items in geographical order, which is a difficult task.
2. It gives more weightage to those areas in which there is a greater concentration of the items, and thus each of the geographical areas is not equally represented.

8.11 Quota Sampling

This is a type of judgement sampling. In a quota sample, quotas are set up according to some specified characteristics, such as so many in each of several income groups, so many in each age, so many with certain political or religious affiliations and so on, but within the quotas, the selection of sample items depends on personal judgement. For example, in a radio listening survey, the interviewers may be told to interview 5000 people living in a certain area, and that out of every 1000 persons interviewed, 60 are to be housewives, 225 farmers and 155 children under the age of 13. Within these quotas, the interviewer is free to select the people interviewed. The cost per person interviewed may be relatively small for a quota sample, but there are numerous opportunities for biases which may invalidate the results.

8.12 Non-Random/Non-Probability Sampling and Judgement Sampling

In judgement sampling, personal knowledge and opinion are used to identify the items from the population.

There are three types of sampling selection of the sample population:

1. Judgement sampling
2. Convenience sampling
3. Sequential sampling

8.12.1 Judgement Sampling or Purpose Sampling or Deliberated Sampling

This approach is used when a sample is taken based on certain judgements about the overall population. It is subject to the researcher bases and is perhaps even more biased than haphazard sampling. It should be carried out by an expert in the field of judgement, which influences the final outcome. For example, statisticians use this method in a laboratory setting where there is a choice of experimental subjects (i.e. animals, humans, fruits and vegetables).

8.12.2 Convenience Sampling or Haphazard or Accidental Sampling or Chunk Sampling

Convenience sampling does not produce the representative sample of the population because people or items are only selected for a sample if they can be easily and conveniently accessed. A convenience sample is obtained by selecting convenient population units.

It is based on the convenience of the researcher. It uses the sources available to him or her. He or she may use telephone directory opinion poll list of employees of an organisation.

It is not normally representative of the target population because sample units are only selected if they can be accessed easily and conveniently.

The method of convenience sampling is also called the chunk. A chunk refers to that fraction of the population being investigated which is selected neither by probability nor judgement but convenience. For example, a sample is obtained from readily available lists, such as automobile registrations or telephone directories.

Examples of convenient sample include selecting

1. The first 20 cars to enter a car park
2. The first 20 people to walk to a turnstile at a sporting event
3. The females in the first row of a concert
4. The first 200 customers to enter a department store
5. The female moviegoers sitting in the first row of a movie theatre
6. The first four callers in a radio contest

Advantages of convenience sampling

1. The method is simple and easy to use.
2. Accurate results in the population are homogeneous.

Disadvantages of convenience sampling

1. Convenience sampling does not produce a representative sampling of the population.
2. It a greatly offset by the sample being biased.

8.12.3 Sequential Sampling

Here, the size of the sample is not fixed. The sampling process takes place depending on the results of the first sample. For example, a manager draws a lot from his inventory and tests for acceptability. If it is acceptable, there will be further sampling required, but if it is found unacceptable, the entire stock will be rejected. So, when the result of the first sample falls near to the acceptable standard, the manager take another sample before deciding on the quality of the inventory.

8.12.3.1 Application

It used for statistical control. A manager draws a lot from the inventory and tests it for acceptability. If it is acceptable, there will be no further sample required. But if it is found unacceptable, the entire lot is rejected.

8.13 Error

This is the difference between the true value and the estimated or approximated value. The errors in statistics arise due to a number of factors, such as the following:

1. Approximation in measurements. For example,
   a. The heights of individuals may be approximated to 1/10 of centimetre.
   b. Age may be measured correctly to the nearest month.
   c. Weight may be measured correctly to 1/10 of a kilogram.
   d. Distance may be measured correctly to the nearest metre.

   In all such measurements, there is bound to be a difference between the observed value and the true one.
2. Approximation in rounding of the figures to the nearest hundredths, thousandths, millionths and so forth, or the rounding of the decimals
3. The biases due to faulty collection and the analysis of the data and biases in the presentation and interpretation of the results
4. Personal biases of the investigators and so on

8.13.1 Sampling Error

The error involved in approximations of the population characteristics on the basis of the sample is known as sampling error. The error arising due to drawing inferences about the

population on the basis of few observations (sampling) are known as sampling errors (or sampling fluctuations).

8.13.1.1 Reasons for Sampling Errors

1. **Facility selection of the sample.** By the use of a defective sampling technique for the selection of a sample, some bias is introduced.
2. **Substitution.** If difficulties arise in enumerating a particular sampling unit included in the random sample, the investigators usually substitute a convenient member of the population. This leads to some bias since the characteristics possessed by the substitute unit will usually be different from those possessed by the unit originally included in the sample.
3. **Faulty demarcation of sampling units.** Bias due to defective demarcation of the sampling units is particularly significant in area surveys, such as agricultural experiments in the field or crop-cutting surveys.
4. **Error due to bias in the estimation method.** Improper choice of the estimation techniques might introduce error. For example, if, in place of simple random sampling, deliberate sampling has been used in a particular case, some bias is introduced in the result and hence such errors are known as biased sampling errors.
5. **Variability of the population.** Sampling error also depends on the variability or heterogeneity of the population to be sampled.

8.13.2 Non-Sampling Error

Non-sampling errors are a consequence of certain factors which are within human control. They are due to certain causes which can be traced and may arise at any stage of the inquiry, planning and execution of the survey and collection, processing and analysing of the data.

8.13.2.1 Important Factors Responsible for Non-Sampling Errors in Any Survey

1. Faulty planning, including vague and faulty definitions of the population or the statistical units to be used; incomplete list of population members.
2. Vague and imperfect questionnaire, which might result in incomplete or wrong information.
3. Defective methods of interviewing and asking questions.
4. Vagueness about the type of data to be collected.
5. Exaggerated or wrong answers to the questions, which appeal to the pride, prestige or self-interest of the respondents. For example, a person may overstate his education or income or understate his age, or he may give a wrong statement to safeguard his self-interest.
6. Personal bias of the investigator.
7. Lack of trained and qualified investigator and lack of supervisor staff.
8. Failure of respondent's memory to recall the events or happenings in the past.

9. Non-response and inadequate or incomplete response. Bias due to non-response results; if in a house-to-house survey, the respondent is not available or refuses to furnish the information.

   Incomplete response error is introduced if the respondent is unable to furnish information on a certain question or if he or she is unwilling or even refuses to answer certain questions.
10. Improper coverage. If the objectives of the survey are not precisely stated in clear-cut terms, this may result in the inclusion in the survey of certain units, which were not to be included in the survey under the objectives. For example, in a census to determine the number of individuals in the age group, say 45 years, more or less serious errors may occur in deciding whom to enumerate unless a particular community or area is not specified and also the time at which the age is to be specified.
11. Compiling errors. Wrong calculations or entries made during the processing and analysing of the data. Various operations of data processing; editing and coding of the responses, punching of cards and tabulating and summarising the original observations made in the survey are potential sources of error. Compilation errors are subject to control through verification, consistency, checks and so forth.
12. Publication errors. The errors committed during presentation and printing of tabulated results are basically due to two sources:
    a. Mechanics of publication – proofing error
    b. Failure of the survey organisation

8.13.2.2 Biased Errors or Cumulative Errors

1. Bias on the part of the enumerator or investigator, whose personal beliefs and prejudices are likely to affect the results of the inquiry.
2. Bias in the measuring instrument or the equipment used for recording the observations.
3. Bias due to faulty collections of the data and in the statistical techniques and formulae used for the analysis of the data.
4. Respondent's bias. An appeal to the pride or prestige of an individual introduces a bias called prestige bias, by virtue of which he or she may upgrade his or her education, occupation, income and so forth, or understate his or her age, thus resulting in wrong information to safeguard his or her personal interests. For example, for income tax purposes a person may give an understatement of his salary or income or assets.
5. Bias in the technique of approximations. If, while rounding off, each individual value is approximated to either the next highest or lowest number so that the entire errors move in the same direction, there is bias for overstatement or understatement, respectively. For example, if the figures are to be rounded off to the next highest or lowest hundred, then each of the values 405 and 496 will be recorded as 500 and 400, respectively.

Note: Biased errors have a tendency to grow in magnitude with an increase in the number of the observations, and hence are also known as cumulative errors.

8.13.2.3 Unbiased Errors (Compensatory Error)

Unbiased errors are when the chances of making an overestimate are almost same as the chances of making an underestimate. Since these errors move in both directions, the errors in one disruption are more or less neutralised by the errors in the opposite direction, and consequently, the ultimate result is not much affected. For example, if the individual values, say 385, 415, 355 and 445, are rounded to the nearest complete unit, that is hundred, each one of them would be recorded as 400. In this case, the values 385 and 355 give overestimating errors of magnitudes 15 and 45, respectively, while the values 415 and 445 give underestimating errors of magnitudes 15 and 45, respectively, and in the ultimate result (approximation) they get neutralised.

Note: If the number of observations is quite large, unbiased errors will not affect the final result much. Since the errors in one direction compensate for the errors in the other direction, unbiased errors are termed 'compensatory errors'.

8.14 Summary

Sampling theory is a study of relationships existing between a population and samples drawn from the population. This chapter discussed enumeration methods like complete enumeration and sampling methods.

There are several techniques that can be used to obtain a representative sample. There are two methods of selecting samples from the universe – (1) non-random or judgement sampling and (2) random or probability sampling. This chapter discussed random sampling methods like simple random sampling, stratified sampling, systematic sampling and cluster sampling. Some non-random sampling methods like judgement sampling, convenience sampling and sequential sampling were also discussed.

REVIEW QUESTIONS

1. You are requested to plan a survey with a view to control and abolish street begging in City X. Outline the main steps you would take and draft a suitable schedule to collect the necessary data.
2. 'Sampling is necessary under certain conditions'. Explain this with illustrative examples. Point out the importance of sampling in solving business and economic problems. What are the principles on which sampling methods rest?
3. Explain the terms *random sample*, *stratified random sample and purposive sample*. Explain the importance of sampling theory in economics.
4. State the advantages of adopting sampling procedures in carrying out large-scale surveys.
5. a. What is systematic random sampling? How does it differ from purposive sampling?
   b. What are the advantages of random sampling?
   c. How does the size of a sample affect 'sampling errors'?
6. Distinguish between the 'census' and 'sampling' methods of collecting data and compare their merits and defects.

7. Give a comparative account of the various methods of selecting a sample.
8. Point out the significance of sampling. Distinguish between random sampling and deliberate sampling.
9. What is the utility of sampling in statistics? Briefly explain
   a. Random sample
   b. Biased sample
   c. Stratified sample
   d. Population or universe
   e. Quota sampling
10. Explain the importance of sampling. What are the well-known methods of sampling in use? An industry is composed of 100 independent organisations. A sample of 30 was selected composed of 10 small, 10 middle-sized and 10 large units. Is this sample a satisfactory one? Give reasons.
11. What do you understand by *sampling*? In order to determine a new cost of living index, it is proposed to make a survey of the income and expenditure of 1000 households in a large city. Describe carefully two methods which might be used to select the sample households.
12. Distinguish between the census and sampling methods of data collection and compare their merits and demerits. Why is the sampling method unavoidable in certain situations?
13. Explain the terms *population* and *sample*. Explain why it is sometimes necessary and often desirable to collect information about the population by conducting a sample survey instead of complete enumeration.

9

Hypothesis Testing

9.1 Introduction

In the test of hypotheses, we always begin with an assumption, the null hypothesis. The null hypothesis asserts that there is no significant difference between the statistic and the population parameter, and whatever observed difference is there, it is merely due to chance fluctuation in sampling from the same population.

9.2 Some Basic Concepts

9.2.1 Null Hypothesis

The null hypothesis is usually denoted by the symbol H_0. The 'no difference' attitude on the part of a statistician before drawing any sample is the basis of the null hypothesis. R A Fisher (1936) defined null hypothesis as the hypothesis which is tested for possible rejection, under the assumption that it is true.

9.2.2 Alternative Hypothesis

Any hypothesis which contradicts the null hypothesis H_0 is called an alternative hypothesis and is denoted by H_1. For example, if we have to test whether the population mean μ has a specified value μ_0, then

Null hypothesis

$$H_0 : \mu = \mu_0$$

Alternate hypothesis

$$H_1 : \mu \neq \mu_0 (\mu > \mu_0 \text{ or } \mu < \mu_0)$$

or

$$H_1 : \mu < \mu_0$$

or

$$H_1 : \mu > \mu_0$$

9.2.3 Hypothesis Testing

One objective of sampling theory is hypothesis testing. Hypothesis testing begins by making an assumption about the population parameter. Then, we gather sample data and determine the sample statistic.

To test the validity of our hypothesis, the difference between the hypothesised value and the actual value of the sample statistic is determined.

If the difference between the hypothesised population parameter and the actual value is large, then we automatically reject our hypothesis. If it is small, we accept it.

9.2.4 Power

$(1 - \beta)$ indicates how powerful the test is. A high $(1 - \beta)$ (i.e. close to 1) implies that the test is doing exactly what is should be doing: rejecting H_0 when it is false.

A low $(1 - \beta)$ indicates poor performance.

We plot $(1 - \beta)$ against each value of possible μ. The graph (the power curve) will indicate the value of μ ($\neq \mu_0$), where the test is performing well.

Note that as μ is approaches μ_0, $(1 - \beta)$ approaches α and as μ moves away from μ_0, $(1 - \beta)$ approaches 1. At $\mu = \mu_0$, $(1 - \beta)$ is not defined.

9.2.5 Critical Region or Region of Rejection

The critical region is the area under the sampling distribution in which the test statistic value has to fall for the null hypothesis to be rejected.

9.2.6 Region of Acceptance

The set of Z-scores inside the range –1.96 to 1.96 is called the region of acceptance of the hypothesis.

9.2.7 Critical Values

The values –1.96 and 1.96 are called the critical values at the 5% level of significance.

9.2.8 Z-Score

Suppose that under a given hypothesis the sampling distribution of a statistic θ is approximately a normal distribution with mean E (θ) and standard deviation (SD) (standard error [SE]) σ_θ.

Then,

$$Z = \frac{\text{Observed value} - \text{Expected value}}{\text{SE of } \theta}$$

is called the standardised normal variable or Z-score, and its distribution is the standardised normal distribution with mean 0 and standard deviation 1.

9.2.9 Inferential Statistics

Concerned with populations, use sample data to make an inference about the population or to test the hypothesis considered at the beginning of the research study.

Help the researcher to determine if the difference found between two or more groups, such as an experimental and a control group, is a real difference or only a chance difference that occurred because an unrepresentative sample was chosen from the population.

9.2.10 Types of Errors

A Type I error is rejecting H_0 when it is actually true and is designated by the Greek letter β.

A Type II error is accepting H_0 when it is actually false and is designated by the Greek letter α.

9.2.11 Level of Significance

1. It is the probability of making a Type I error.
2. It is represented by α.
3. It is the probability of rejecting the null hypothesis when it is true.
4. In health science, we consider α to be either 1% (0.01) or 5% (0.05).
5. Five percent means the researcher is willing to take risk being wrong 5 times out of 100 when rejecting the H_0.
6. One percent means the researcher is willing to take risk being wrong 1 time out of 100 when rejecting the H_0.

9.2.12 Confidence Interval

This is the range of values which, with a specified degree of probability, is thought to contain the population value. It contains a lower and an upper limit. For example, Vedanta conducted a study on 100 households to assess the per capita income of households and found that the mean per capita income is Rs. 150 with the standard deviation of Rs. 3. In this case, the researcher may calculate the confidence interval.

9.2.13 Degrees of Freedom

1. The interpretation of a statistical test depends on the degrees of freedom (d.f.).
2. It is the number of classes to which the values are assigned arbitrarily or at will without violating the restrictions.
3. It indicates the number of values that are free to vary.
4. The procedure to calculate the degrees of freedom varies from test to test.

9.2.14 Test of Significance

Parametric test: T-test, z-test, analysis of variance (ANOVA)

Non-parametric test: Chi-square, median test, McNemar, Mann–Whitney, Wilcoxon, Fisher's exact

9.2.15 Parametric Test

1. It is known as a normal distribution statistical test.
2. Statistical methods of inference make certain assumptions about the populations from which the samples are drawn. For example, populations are normally distributed, have the same variance.

9.2.16 Non-Parametric Tests

1. A normality assumption is not required.
2. Ordinal or interval scale data is used.
3. It can be applied for small samples.
4. Samples use the sign test.
5. Independent samples use Mann–Whitney U-statistics.
6. Randomness uses run tests.
7. Several independent samples use the Kruskal–Wallis test.

9.3 Probability Distributions

9.3.1 Binomial Distribution

It was discovered by Jacob Bernoulli, a Swiss mathematician of the seventeenth century.

9.3.1.1 Assumption

The probability of the outcome remains constant over time.

9.3.1.2 Bernoulli Variable

This is a random variable x which assumes the values of 1 and 0 with respective probabilities p and $q = 1 - p$.

The Bernoulli distribution is

x	1	0
p(x)	p	q

9.3.1.3 Random Variable

This is a quantity obtained from an experiment that may by chance result in different values. A random variable is a function which has its domain confined to a sample space and its range confined to a real space; such restricted functions are called random variables. It is a variable which assumes different numerical values as a result of a random experiment or random occurrences.

Note:

1. Every function has a domain and range.
2. The domain must be a real number.

3. A random variable must assume numerical values. For example, a count of number of accidents during a week might be 10 or 11 or 12 or some other number. It can be written as

$$p(x) = p^x \ q^{x-1}, \ x = 0,1$$

Here, the occurrence of the value 0 may be termed a 'failure':

$$\therefore p\,[\text{Success}] = p$$

and

$$p\,[\text{Failure}] = q = 1 - p$$

9.3.1.4 Characteristics of Bernoulli Process

1. Each trial has only two possible outcomes. For example, a newborn baby is male or female.
2. The probability of the outcome of any trial remains fixed over time. For example, the probability of the baby being male or female remains fixed throughout.
3. The trials are statistically independent. For example, the outcome of the baby being male or female does not affect the outcome of any other baby being so.

Exercise

Find the probability of getting exactly three heads in four tosses of a biased coin, where P (H) = ¾ and P (T) = ¼.

Solution

Given
n = 4 (total number of trials)
x = 3 (certain number of successes)
p = ¾ = 0.75 (probability of success)
q = ¼ = 0.25 (probability of failure)

$$p(X = x) = (n_{Cx})\, p^x \ q^{n-x}, x = 0,1,2,\ldots,n$$

where:

$$(n_{Cx}) = \frac{n!}{x! \ (n-x)!}$$

$$p(X = 3) = (4_{C_3})(0.75)^3 \ (0.25)^{4-3}$$

$$= 0.421875$$

It can be shown for the Binomial distribution that

$$\mu = E(x) = np$$

$$\sigma^2 = V(x) = npq$$

Exercise

A fair coin is tossed 10 times. If getting heads is defined as success, find out the probability of getting four successes in the 10 trials.

Solution

p = probability of getting heads = 0.5
q = probability of getting heads = 1 – p = 0.5
r = number of successes = 4
n = number of trials = 10
The probability of getting four successes in 10 trials is

$$p(X=4)=\frac{10!}{4!\ (10-4)!}(0.5)^4\ \ (0.5)^{10-4}=0.2051$$

Exercise

A binomial experiment is repeated nine times. If the probability of success is 0.6, find the probability of getting four successes.

Solution

n = 9, p = 0.6, q = 0.4, r = 4

$$p(X=4)=\frac{9!}{4!\ (9-4)!}(0.6)^4\ \ (0.4)^{9-4}=0.167$$

Exercise

If boys and girls are equally likely, find the number of families out of 600, consisting of three children, with each having

1. Two boys and one girl
2. At most one boy
3. All girls

Solution

Let
X = number of boys out of 3 children in a family
P = probability that child is a boy
= 1/2
n = 3
$\therefore X \sim B(3, 1/2)$

1. Two boys and one girl

$$P(X=2)=C_2^3(1/2)^2(1/2)^1$$

$$=3\times\frac{1}{8}=\frac{3}{8}$$

Therefore, the number of families having two boys and one girl is

$$600 \times \frac{3}{8} = \frac{1800}{8} = 225$$

2. At most one boy

$$\begin{aligned} P(X \le 1) &= P(X = 0) + P(X = 1) \\ &= C_0^3 (1/2)^0 (1/2)^3 + C_1^3 (1/2)^1 (1/2)^2 \\ &= 1 \times \frac{1}{8} + 3 \times \frac{1}{8} = \frac{4}{8} = \frac{1}{2} \end{aligned}$$

Therefore, the number of families is $600 \times \frac{1}{2} = 300.$

3. All girls means no boy

$$\therefore P(X = 0) = C_0^3 (1/2)^0 (1/2)^3 = \frac{1}{8}$$

$$\therefore \text{Number of families } = 600 \times \frac{1}{8} = 75$$

9.3.2 Poisson Distribution

A probability distribution, which has the following probability mass function (pmf), is called a Poisson distribution.

$$P(X = x) = p(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \ldots, \lambda > 0$$

where:

X = discrete random variable (Poisson variate) (number of occurrences [successes])
x = specific value X can take
λ = number of occurrences per interval of the time (parameter of Poisson distribution) or the arithmetic mean number of occurrences (successes) in a particular interval of time
e = the constant 2.71828 (base of the Naperian log arithmetic system)

9.3.2.1 Definition

A distribution often used to approximate binomial probabilities. When n is large and p is small, what is considered 'large' or 'small' is not precisely defined, but a general rule is that n should be ≥20 and p ≤ 0.05.

9.3.2.2 History

The distribution is named after Simeon Poisson, who described it in 1837.

The Poisson distribution is formed by counting something; hence, it is a discrete probability distribution.

9.3.2.3 Need for Poisson Probability Distribution

The binomial probability distribution for probabilities of success ($p < 0.05$) could be computed. But, the calculation would be quite time-consuming (especially for a large n of say 100 or more). The distribution of probabilities would become more and more skewed as the probability of success became smaller.

The limiting form of the binomial distribution, where the probability of success is very small and n is large, is called the Poisson probability distribution.

9.3.2.4 Applications of the Poisson Distribution

It is used as a model to describe

1. Such phenomena as the distribution errors in data entry
2. The number of scratches and other imperfections in newly painted panels
3. The number of defective parts in outgoing shipments
4. The number of customers waiting to be served at a restaurant
5. The number of customers waiting to get into an attraction at Disney World
6. The distribution of telephone calls going through a switchboard system
7. The arrivals of trucks at a toll booth
8. The average number of radioactive particles passing through a counter, during 1 millisecond in a laboratory experiment
9. The number of deaths occurring in a city in a day
10. The number of road accidents occurring in a city in a day
11. The number of incoming telephone calls at an exchange in 1 minute
12. The number of vehicles crossing a junction in 1 minute

9.3.2.5 Properties of the Poisson Distribution

1. The mean and variance of a Poisson distribution is $\lambda = np$.
2. The binomial distribution tends to be a Poisson distribution, under the following conditions:
   a. P is very small ($p \to 0$).
   b. n is very large ($n \to \infty$).
   c. $np = \lambda$ is fixed.
3. The probability that a single event will occur during a very short time interval is proportional to the length of the time interval.
4. The probability that more than one event will occur in such a short time interval is negligible.
5. The number of events occurring in a fixed time interval of a random variable X has a Poisson distribution.
6. Each outcome is classified into one of two mutually exclusive categories.
7. The probability of a success remains the same from trial to trial.

8. Each trial is independent.
9. The distribution results from a count of the number of successes in a fixed number of trials.

Example 9.1: Use of Poisson Distribution

Assume that billing clerks rarely make errors in data entry on the billing statements. Many statements of clerks have no mistakes. Some have one, a very few have two and rarely a statement will have three. A random sample of 1000 statements revealed 300 errors.

$$\text{AM}(\bar{x}) = \frac{\text{No. of mistakes}}{\text{No. of billing statements}}$$

$$\lambda = \frac{300}{1000} = 0.3$$

This is a sample mean $\bar{x}$, which is used to estimate the population mean, λ, for a model (Poisson) of the process. The probability of no (0) mistakes appearing in a statement is computed by

$$p(x) = \frac{e^{-\lambda}\,\lambda^{x}}{x!}$$

Here

$$e = 2.71828,\ \lambda = 0.3,\ x = 0$$

$$p(0) = \frac{(2.71)^{-0.3}(0.3)^{0}}{0!} = 0.7408$$

Exercise

The number of persons joining a cinema queue in a minute has a Poisson distribution with a parameter of 5.8. Find the probability that

1. No one joins the queue in a particular minute
2. Two or more persons join the queue in the minute

Solution

Let X be the number of persons joining the queue in the minute. Then,

$$X \sim P\ (\lambda = 5.8)$$

Then, the pmf is

$$P(X = x) = p(x) = \frac{e^{-5.8}(5.8)^{x}}{x!},\quad x = 0, 1, 2, \ldots$$

1. P (no one joins the queue) $= P[X = 0]$

$$= p(0)$$
$$= \frac{e^{-5.8}(5.8)^0}{0!}$$
$$= 0.003$$

2. P (two or more join) $= P[X \geq 2]$

$$= 1 - P[X < 2]$$
$$= 1 - \{p(0) + p(1)\}$$
$$= 1 - \left\{\frac{e^{-5.8}(5.8)^0}{0!} + \frac{e^{-5.8}(5.8)^1}{1!}\right\}$$
$$= 1 - e^{-5.8}(1 + 5.8)$$
$$= 1 - (0.003)(6.8)$$
$$= 0.9796$$

9.3.3 Normal Probability Distribution

This distribution is also known as the Gaussian distribution. This is a continuous probability distribution and is an important distribution in statistics. For many variables, we observe that they follow the normal probability distribution. The probability rule for a normal distribution is given as

$$P(X \leq r) = \int_{-\infty}^{r} \frac{1}{\sigma \times \sqrt{2\pi}} \times e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \times dx$$

where:
- μ = mean of the distribution
- σ = standard deviation of the distribution
- m and σ = parameters of the distribution

The notation we use is

$$X \sim N(\mu, \sigma)$$

9.3.3.1 Discrete Random Variable

If the random variable can assume only a limited number of values, then it is called a discrete random variable and can only take whole numbers as values.

9.3.3.2 Continuous Random Variable

It assumes an infinite number of possible values, within a specified range. It usually results from measuring something because wide variety of values can be measured within a given range, such as

1. Weight of an individual (the weight might be 111.0, 111.1 or 111.12 kg and so on) depending on the accuracy of the scale
2. Life expectancy of alkaline batteries
3. Volume of a shipping container
4. Weight of impurities in a steel ingot
5. Mean temperature of each day for a city
6. Measurement of rainfall in centimetres on each day of monsoon season

9.3.3.3 Characteristics of Normal Distribution

1. The normal curve is bell-shaped and has a single V peak at the exact centre of the distribution; thus, it is unimodal.
2. The two tails of the normal probability distribution extend indefinitely and never touch the horizontal axis.
3. The curve is symmetric about mean m.
4. The mean of a normally distributed population lies at the centre of its normal curve. For a normal distribution, mean = median = mode
5. The area under the curve gives probability.
   a. $P(-\infty < X < \infty) = 1$
   b. Since the curve is symmetric about the mean, we get

$$P(X < m) = P(X > m) = 0.5$$

6. A normal distribution with mean 0 and standard deviation 1 is called the standard normal distribution. This is denoted by Z. Z is called the standard normal variate (SNV). Thus Z ~ N (0, 1). The figure for SNV is as follows:
   a. For SNV we get
      i. $P(Z < 0) = 0.5$
      ii. $P(Z > 0) = 0.5$
   b. We have a probability table for finding probability (area under the curve for different values of Z). Generally, the table gives the area under the curve when Z lies between 0 and a certain value Z. This may be denoted by Φ (Z) or A (Z lies between 0 and Z).
7. Any normal distribution can be converted to Z with the formula

$$Z = \frac{X - \mu}{\sigma}$$

Thus, if $X \sim N(\mu, \sigma)$ and we define

$$Z = \frac{X - \mu}{\sigma}$$

then

$$Z \sim N(0,1).$$

Because of this property, use of the normal distribution and finding probability is easier.

If σ is the standard deviation of the normal distribution, 80% of the observation will be in the interval

$$\mu - 1.28\sigma \text{ to } \mu + 1.28\sigma.$$

$$95\% = \mu - 1.96\sigma \text{ to } \mu + 1.96\sigma$$

$$98\% = \mu - 2.33\sigma \text{ to } \mu + 2.33\sigma$$

9.3.3.4 Gaussian or Normal Curve

If large values are collected for any character, and a frequency table is prepared with a small class interval, the frequency curve of this data will give a bell-shaped symmetrical curve, which is known as a Gaussian or normal curve.

1. The shape of this curve depends on the mean and standard deviation of data.
2. If the standard deviation (variation) is very high, the width of the curve is also more.
3. If we move both sides from the mid-point ($\bar{x}$ = mean), the height of the curve decreases.
4. If the mean and standard deviation are μ and σ, respectively, then the normal curve is called the standard normal curve.
5. If the total area under the curve is considered unity, the normal curve is called the normal probability curve.

9.3.3.5 Properties of the Normal Probability Curve

1. The normal curve is bell-shaped and symmetrical.
2. The height of the normal curve is at its maximum at the mean.
3. The curve approaches nearer and nearer to the base, but it never touches it. That is, the curve is asymptotic to the base on either side. Hence, its range is unlimited or infinite in both directions.
4. Since there is only one maximum point, the normal curve is unimodal; that is, it has only one mode.
5. The variable distributed according to the normal curve is a continuous one.

   The area under the normal curve is distributed as follows:

a. Mean ± 1σ covers 68.27% of the area (34.135% of the area will lie on either side of the mean).
b. Mean ± 2σ covers 95.45% of the area.
c. Mean ± 3σ covers 99.73% of the area.

9.3.3.6 Importance of Normal Probability Curve

1. Data obtained from biological measurements approximately follows a normal distribution.
2. The binomial and Poisson distributions can be approximated to a normal distribution.
3. For a large sample, any statistic (i.e. sample mean or sample standard deviation) approximately follows the normal distribution, and as such, it can be studied with the help of the normal curve.
4. The normal curve is used to find the confidence limits of the population parameters.
5. The normal distribution also forms the basis for the test of significance (Figure 9.1).

9.3.3.7 Finding Probability for Different Values of Z (Using Table)

1. Find P (0 < Z < 1.57)
   This probability given in the table as $A(0-1.57)=0.4418$.

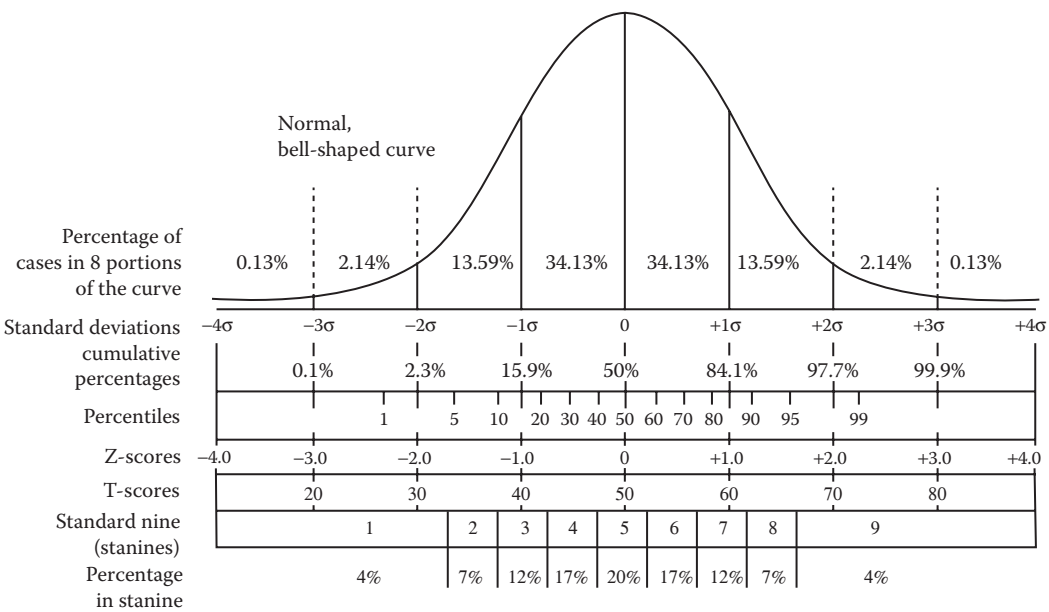


FIGURE 9.1
Normal probability (bell-shaped) curve.

2. P (–1.23 < Z < 0)
   This area is not given in the table. Using symmetry, we get

$$A(-1.23 \text{ to } 0) = A(0 - 1.23)$$
$$= 0.3906$$

3. P (–0.62 < Z < 1.35)
   This probability is

$$A(0 - 1.35) + A(0.62 - 0)$$
$$= A(0 - 1.35) + A(0 - 0.62)$$
$$= 0.4115 + 0.02324$$
$$= 0.43474$$

4. P (Z > 1.23)
   This is a one-tailed probability, so we know

$$P(Z > 0) = 0.5$$
$$\therefore p(Z > 1.35) = 0.5 - A(0 - 1.23)$$
$$= 0.5 - 0.3906$$
$$= 0.1094$$

5. P (0.58 < Z < 2.37)

$$= A(0 - 2.37) - A(0 - 0.58)$$
$$= 0.4911 - 0.2190$$
$$= 0.2721$$

Exercise

In a sample of 120 tablets in a factory, the mean and standard deviation of active ingredient contents were 11.35 and 3.03, respectively. Find the percentage of tablets having an active ingredient content between 9 and 17 in the whole sample, assuming that the active ingredient content is normally distributed.

Solution

Let X be the active ingredient content in the tablet.

$$\therefore X \sim N(11.35, 3.03)$$

Mean = m = 11.35
SD = σ = 3.03

We require

$$P(q < X < 17) = P\left(\frac{9-11.35}{3.03} \le \frac{X-11.35}{3.03} \le \frac{17-11.35}{3.03}\right)$$

$$= P(-0.78 \le X \le 1.86)$$

$$= A(0-0.78) + A(0-1.86)$$

$$= 0.2823 + 0.4686 = 0.7509$$

The percentage of tablets containing active ingredient between 9 and 17 is

$$100 \times 0.7509 = 75.09\%$$

Exercise

The mean and standard deviation of a group of 600 students on a certain test are 58.5 and 10, respectively. The group is divided into five subgroups of 600 students as A (Very good), B (good), C (Average), D (Below Average) and E (Poor), which contain 25%, 20%, 30%, 10% and 15% of the students, respectively.

If the scores on the test are distributed normally, find the limit of scores of three subgroups to the nearest whole number.

Solution

Let X be the scores of students.

$$X \sim N(58.5, 10)$$

1. Let Y be the minimum score for Group A.
   $\therefore P(X > Y) = 0.25$ ($\because$ 25% students in Group A)

$$\therefore P\left(\frac{X-58.5}{10} > \frac{Y-58.5}{10}\right) = 0.25$$

$$\therefore P\left(Z > \frac{Y-58.5}{10}\right) = 0.25$$

$$\therefore P\left(0 < Z < \frac{Y-58.5}{10}\right) = 0.25 = 0.5 - 0.25$$

From the table,

$$P(0 < Z < 0.67) = 0.25$$

$$\therefore \frac{Y-58.5}{10} = 0.67$$

$$\therefore Y = 65.2$$

Therefore, the limits for Group A are above a score of 65.2, that is above 65.

2. Let Y be the minimum score for Group B.
   $\therefore$ P (X > Y) = 0.45 (% students in Groups A and B)

$$\therefore P\left(\frac{X-58.5}{10} > \frac{Y-58.5}{10}\right) = 0.45$$

$$\therefore P\left(Z > \frac{Y-58.5}{10}\right) = 0.45$$

$$\therefore P\left(0 < Z < \frac{Y-58.5}{10}\right) = 0.05 = 0.5 - 0.45$$

From the table,

$$P(0 < Z < 0.125) = 0.05$$

$$\therefore \frac{Y-58.5}{10} = 0.125$$

$$\therefore Y = 59.75$$

Therefore, the limits for Group B are 59.75–65.2, that is 60–65.

3. Let Y be the minimum score for Group C.

$$\therefore P\ (X > Y) = 0.75$$

$$\therefore P\left(\frac{X-58.5}{10} > \frac{Y-58.5}{10}\right) = 0.75$$

$$\therefore P\left(Z > \frac{Y-58.5}{10}\right) = 0.75$$

$$\therefore P\left(0 < Z < \frac{Y-58.5}{10}\right) = 0.7 - 0.5 = 0.25$$

From the table,

$$P(0 < Z < 0.67) = 0.25$$

$$\therefore \frac{Y-58.5}{10} = 0.67$$

But,

$$\frac{Y-58.5}{10}$$

lies on the left-hand side of O. Therefore, it is negative.

$$\therefore \frac{Y - 58.5}{10} = -0.67$$

$$\therefore Y = 51.8$$

Therefore, the limits for Group C are 51.8–59.75, that is 52–60

Exercise

For a standard normal variate X, find k such that

1. $P(-k < X < k) = 0.95$
2. $P(|X| \geq k) = 0.01$

Solution

1. Let m be the mean and σ be the standard deviation.

$$\therefore P(-k < X < k) = 0.95$$

$$P\left(\frac{-k-m}{\sigma} < \frac{X-m}{\sigma} < \frac{k-m}{\sigma}\right) = 0.95$$

$$P\left(\frac{-k-m}{\sigma} < Z < \frac{k-m}{\sigma}\right) = 0.95$$

From the table,

$$P(-1.96 < Z < 1.96) = 0.95$$

$$\therefore \frac{k-m}{\sigma} = 1.96$$

$$\therefore k = m \pm 1.96\sigma$$

2. $P(|X| \geq k) = 0.01$

$$\therefore P(-k < X < k) = 1 - 0.01 = 0.99$$

$$P\left(\frac{-k-m}{\sigma} < \frac{X-m}{\sigma} < \frac{k-m}{\sigma}\right) = 0.99$$

$$P\left(\frac{-k-m}{\sigma} < Z < \frac{k-m}{\sigma}\right) = 0.99$$

From the table,

$$P(-2.58 < Z < 2.58) = 0.99$$

$$\therefore \frac{k-m}{\sigma} = 2.58$$

$$\therefore k = m + 2.58\sigma$$

Exercise

The particle size analysis of powder shows a normal distribution. Particle size analysis data of 10,000 particles shows the mean size as 750 μm and standard deviation as 50. Find

1. The number of particles with a size less than 700 μm
2. The number of particles with a size between 700 and 800 μm

Solution

Let X be the size of the particle.

$$\therefore X \sim N(750, 50)$$

1. We require

$$P(X < 700) = P\left(\frac{X - 750}{50} < \frac{700 - 750}{50}\right)$$
$$= P(Z < -1)$$
$$= 0.5 - A(0 \text{ to } 1)$$
$$= 0.5 - 0.3413 = 0.1587$$

$$\therefore \text{Number of particles} = 10,000 \times 0.1587 = 1587$$

2. Now,

$$P(700 < X < 800 = P\left(\frac{700 - 750}{50} < \frac{X - 750}{50} < \frac{800 - 750}{50}\right)$$
$$= P(-1 < Z < 1)$$
$$= A(0 \text{ to } 1) + A(0 \text{ to } 1)$$
$$= 0.3413 + 0.3413$$
$$= 0.6826$$

$$\therefore \text{Number of particles} = 10,000 \times 0.6826 = 6826$$

Exercise

A random variable X has the following probability distribution:

X	0	1	2	3	4	5	6
P (X)	k	3k	5k	7k	9k	11k	13k

Find

1. k
2. P ($X \geq 2$)
3. P ($0 < X < 5$)

Solution

1. Sum of probability = 1

$$\therefore 49k = 1$$

$$\therefore k = \frac{1}{49}$$

2. $P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$

$$= 45k$$

$$= \frac{45}{49}$$

3. $P(0 < X < 5) = P(X = 1) + P(X = 2) + P(X = 4)$

$$= 24k$$

$$= \frac{24}{49}$$

Exercise

The burning time of an experimental rocket is a random variable which has a normal distribution with $\mu = 36$ seconds and $\sigma = 0.04$ seconds. What are the probabilities that this kind of rocket will burn for the following?

1. Less than 4.25 seconds
2. More than 4.40 seconds
3. 4.30–4.42 seconds

Solution

Let X be the burning time of the rocket:

$$X \sim N(3.6, 0.04)$$

1. $$P(X < 4.25) = P\left(\frac{X - 3.6}{0.04} < \frac{4.25 - 3.6}{0.04}\right)$$

$$= P(Z < 16.25)$$

$$= 1$$

2. More than 4.40 seconds

$$P(X < 4.40) = P\left(\frac{X - 3.6}{0.04} < \frac{4.40 - 3.6}{0.04}\right)$$

$$= P(Z > 20)$$

$$= 0$$

3. P (4.40 < X < 4.42)

$$= P(17.5 < Z < 20.5)$$

$$= 0$$

9.3.3.8 Standard Normal Distribution

1. It is a normal distribution with a mean $\mu = 0$ and a standard deviation $\sigma = 1$.
2. The observation values in a standard normal distribution are denoted by Z.

Exercise

What is the probability that an observation from a standard normal distribution will lie in the interval –1.28 to 1.28?

Solution

We know that for a normal distribution, 80% of the observations lie between $\mu - 1.28\sigma$ and $\mu + 1.28\sigma$, and for a standard normal distribution, $\mu = 0$ and $\sigma = 1$.
Therefore, 80% of the observations will lie between –1.28 and +1.28.
Similarly,

95% = –1.96 and 1.96

98% = –2.33 and +2.33

9.3.3.9 Standard Normal Variables

Suppose we have a normal population. We can represent it by a normal variable X. We can convert any value of X into a corresponding value of Z of the standard normal variable by using the formula

$$Z = \frac{X - \mu}{\sigma}$$

where:
- X = value of any random variable
- μ = mean of the distribution random variable
- σ = standard deviation of the distribution
- Z = number of standard distribution from X to the mean of the distribution and is known as the Z-score or standard score.

Exercise

A normal variable X has a mean of 56 and a standard deviation of 12. Find the Z-value corresponding to the X-value of –5.

Solution

$$Z = \frac{X - \mu}{\sigma} = \frac{-5 - 56}{12} = -5.08$$

Exercise

A normal variable has a mean of 10 and standard deviation of 5. What is the probability that the normal variable will take a value in the interval 0.2–19.8?

Solution

$$\text{Probability}(0.2 < X < 19.8) = \text{probability}\left(\frac{0.2-10}{5} < Z < \frac{19.8-10}{5}\right)$$

$$= \text{probability}(-1.96 < Z < 1.96)$$

$$= 95\%$$

because 95% of the area under the standard normal curve lies in the interval –1.96 to 1.96

Steps

1. The area under the standard normal curve between 0 and +1.96 is 0.4750.
2. Due to symmetry of the standard normal distribution, the area under the curve between –1.96 and +1.96 is twice the area under the curve between 0 and +1.96.

$$\therefore \text{Probability}(-19.6 < Z < +1.96) = 0.475 \times 2$$

$$= 0.950 \text{ or } 95\%$$

Note: Any normal variable can be converted into a standard normal variable.

Therefore, we can use a standard normal distribution table to find the probability that the variable will take a value within any given intervals.

9.4 t-Test

It is used to find the significant difference between two means. It is applied for

1. Randomly selected homogenous samples
2. Quantitative data (numerical data, not the frequency distribution)
3. Normally distributed variables
4. Sample sizes of <30

9.4.1 Types of t-Tests

Unpaired t-test: Applied when samples are drawn from two different populations

Paired t-test: Commonly used in nursing research and applied on paired data of independent observations made on the same sample before and after the intervention

9.4.2 Assumptions for the t-Test Application

1. Observations are independent.
2. Observations are drawn at random.
3. The sample size is <30.
4. The population standard deviation is unknown.
5. The population must be normal.

9.4.3 Characteristics of Student's t or the t-Distribution

The t-distribution was developed by William S Grossett in 1908. This distribution is flatter, more 'spread out' than the normal Z-distribution, and has the following characteristics:

1. Continuous distribution
2. Bell-shaped and symmetrical
3. Family of t-distributions
4. More spread out
5. Flatter at centre than the standard normal distribution

9.4.4 t-Distribution with (n – 1) Degrees of Freedom

The t-distribution with (n – 1) degrees of freedom is defined as shown in Figure 9.2.

9.4.5 Uses of t-Distribution

The best use of the t-distribution is when the degrees of freedom is <30. It is used instead of the normal distribution whenever the standard deviation is estimated.

9.4.6 Test for the Population Mean (Single)

Steps

1. State the null and alternate hypotheses.
2. Select the level of significance. The 0.01 level is used.
3. Give the test statistic.

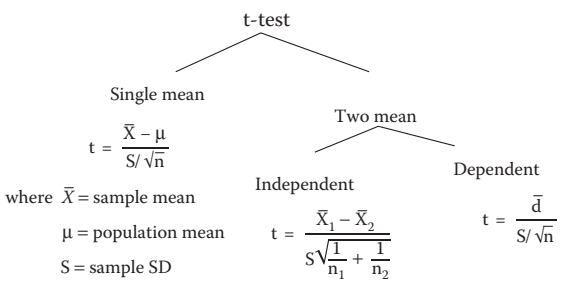


FIGURE 9.2
t-Distribution.

The formula for t is

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad \text{or} \quad t = \frac{\bar{X} - \mu}{S/\sqrt{n-1}}$$

4. Formulate the decision rule d.f. = n – 1, a left-tailed test, and the level is 0.01.
5. Compute t and arrive at a decision.

Exercise

Table 9.1 represents the number of men and the number of women who choose each of five possible answers to an item in attitude scale.

Does their data indicate a significant difference in attitude towards this question? (Note: Test the independence [null hypothesis]).

Solution

Let us take the hypothesis as

H_0: There is no significant difference in attitude towards this question.
H_1: There is a significant difference in attitude towards this question.
Apply the t-test (Table 9.2).

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$\bar{x_1} = \frac{\Sigma x_1}{n_1} = \frac{100}{5} = 20 \qquad \bar{x_2} = \frac{\Sigma x_2}{n_2} = \frac{60}{5} = 12$$

$$\therefore S = \sqrt{\frac{\Sigma\left(x_1 - \bar{x_1}\right)^2 + \Sigma\left(x_2 - \bar{x_2}\right)^2}{n_1 + n_2 - 2}}$$

TABLE 9.1

Number of Men's and Women's Choices for an Item

	Approve Strongly	Approve	Indifferent	Disapprove	Strongly Disapprove	Total
Men	25	30	10	25	10	100
Women	10	15	5	15	15	60

TABLE 9.2

Testing Independence by Applying the t-Test

Men x_1	$(x_1 - \bar{x}_1)$ $\bar{x}_1 = 40$	$(x_1 - \bar{x}_1)^2$	Women x_2	$(x_2 - \bar{x}_2)$ $\bar{\bar{x}}_2 = 12$	$(x_2 - \bar{\bar{x}}_2)^2$
25	5	25	10	–2	4
30	10	100	15	3	9
10	–10	100	5	–7	49
25	5	25	15	3	9
10	–10	100	15	3	9
$\sum x_1 = 100$	$\sum (x_1 - \bar{x}_1) = 0$	$\sum (x_1 - \bar{x}_1)^2 = 350$	$\sum x_2 = 60$	$\sum (x_2 - \bar{x}_2) = 0$	$\sum (x_2 - \bar{x}_2)^2 = 80$

$$S = \sqrt{\frac{350 + 80}{5 + 5 - 2}}$$

$$S = 7.33$$

$$\therefore t = \frac{20 - 12}{7.33} \quad \sqrt{\frac{5 \times 5}{5 + 5}}$$

$$t_{calculated} = 1.72$$

Degrees of freedom $(\nu) = n_1 + n_2 - 2 = 5 + 5 - 2 = 8$

$$\therefore t_{0.05 \text{ at } 8 \text{ d.f.}} = 2.306$$

The calculated value of t is less than the table value of t at the 0.05 level of significance, and hence the null hypothesis (H_0) is accepted and alternate hypothesis (H_1) is rejected. We therefore conclude that there is no significant difference in attitude towards this question.

9.4.7 Hypothesis Tests of Mean When Population Standard Deviation Is Known and Unknown for Large Samples (p-Value Approach)

The hypothesis test of the mean for a large sample test can be done by using the p-value. This is known as the probability value approach or p-value approach.

In this method, first we estimate the p-value for the test, which is the smallest level of significance at which the given null hypothesis is rejected. Using this p-value, make the decision.

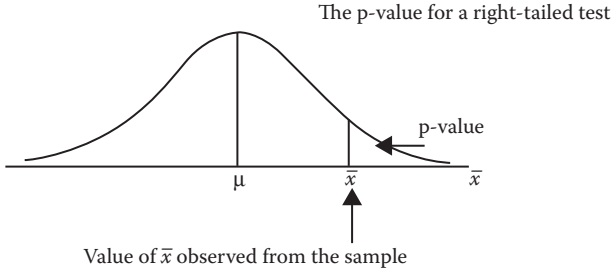
We reject the null hypothesis if

$$\text{p-value} < \alpha \text{ or } \alpha > \text{p-value}$$

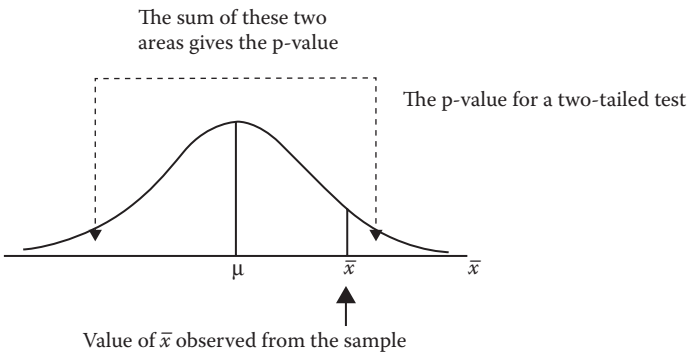
We do not reject the null hypothesis if

$$\text{p-value} \geq \alpha \text{ or } \alpha \leq \text{p-value}$$

The following figure shows the p-value for a right-tailed (one-tailed) test for μ.



For a two-tailed test, the p-value is twice the area in the tail of the sampling distribution curve beyond the observed value of the sample statistic.



For a large sample, the sampling distribution of the sample mean ($\bar{x}$) is approximately normal. We can use the normal distribution to find the p-value for a test of hypothesis when the sample size is large. To find the area under the normal distribution curve beyond the sample mean $\bar{x}$, first we calculate the z-value for $\bar{x}$.

For a large sample, the value of z for $\bar{x}$ for a test of hypothesis for μ is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \text{ if } \sigma \text{ (standard deviation) is known}$$

$$z = \frac{\bar{x} - \mu}{s_{\bar{x}}} \text{ if } \sigma \text{ is not known}$$

where

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and

$s_{\bar{x}} = \frac{s}{\sqrt{n}}$.

This is the observed value of z.

9.4.8 Test for Equality of Means for Small and Independent Samples

To do this, one must make a confidence interval and test a hypothesis about the difference between two population means when samples are small and independent.

9.4.8.1 Assumption

The two populations from which the two samples are drawn are approximately normal.

Case I: When population standard deviations are known. If the above assumption is true and the population standard deviations are known, then we can use the normal distribution for inference about the difference of population means.

Case II: When population standard deviations are unknown. If the standard deviations of the populations are not known, then the normal distribution is replaced by the t-distribution to make inferences about the difference of population means.

Here, we will make one more assumption, that the standard deviations of both populations are equal ($\sigma_1 = \sigma_2 = \sigma$).

When the standard deviations of two populations are equal, then we can use a common standard deviation for both (σ for both σ_1 and σ_2). Because σ is unknown, we replace it by its point estimator, s_p, which is called the pooled sample standard deviation.

The pooled standard deviation for two samples is calculated by using the following formula:

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}}$$

where:

$s_1^{\,2}$ and s_2^2 = variances of the two samples
$n_1 - 1$ = degrees of freedom for Sample I
$n_2 - 1$ = degrees of freedom for Sample II
$(n_1 + n_2 - 2)$ = degrees of freedom for the two samples taken together

The standard deviation $\sigma_{\bar{x}_1 - \bar{x}_2}$ of $\bar{x}_1 - \bar{x}_2$ is estimated by the standard deviation of $(\bar{x}_1 - \bar{x}_2)$. The standard deviation of $(\bar{x}_1 - \bar{x}_2)$ $s_{\bar{x}_1 - \bar{x}_2}$ is

$$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

9.4.8.2 Confidence Interval

The confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t\, s_{\bar{x}_1 - \bar{x}_2}$$

9.4.8.3 t-Distribution Value

Here, the value of t is obtained from the t-distribution table for a given confidence level and $(n_1 + n_2 - 2)$ degrees of freedom.

9.4.8.4 Hypothesis Testing

Hypothesis testing for $\mu_1 - \mu_2$, the value of the test statistic t for $\bar{x}_1 - \bar{x}_2$, is calculated by using the following formula:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

The value of $\mu_1 - \mu_2$ in the above formula is substituted from the null hypothesis.

Exercise

An insurance company wants to know if the average speed at which urban people drive cars is greater than that of rural drivers. The company took a sample of 40 cars driven

by urban people and another sample of 30 cars driven by rural people and found the mean speeds to be 72 and 68 kph, respectively. If standard deviations for both are 2.2 and 2.5 kph, respectively, and assuming that the speeds at which all urban and all rural people drive cars on the highway are normally distributed and the populations have the same standard deviations, then construct a 95% confidence interval for the difference between the mean speeds of both drivers. Test a hypothesis at the 1% significance level of whether the mean speed of cars driven by all urban people on the highway is greater than that of cars driven by rural persons.

Solution

Let μ_1 and μ_2 be the mean speeds of the urban population and rural population, respectively. Here, $n_1 = 28$, $n_2 = 20$, $\bar{x}_1 = 72$, $\bar{x}_2 = 68$, $s_1 = 2.2$ and $s_2 = 2.5$.

The confidence level is $1 - \alpha = 0.95$.

So, the standard deviation of $(\bar{x}_1 - \bar{x}_2)$ is

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(28 - 1)2.2^2 + (20 - 1)2.5^2}{28 + 20 - 2}},$$

$$= 8.043696$$

So,

$$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (8.043696)\sqrt{\frac{1}{28} + \frac{1}{20}} = 1.942738$$

The area in each tail is

$$\alpha/2 = 0.5 - (0.95/2) = 0.025$$

The degrees of freedom is

$$n_1 + n_2 - 2 = 28 + 20 - 2 = 46$$

The t-value at df = 46 and area 0.025 of the t-distribution curve is 2.013.

So, the 95% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t\, s_{\bar{x}_1 - \bar{x}_2}$$

$$72 - 68 \pm 2.013(1.942738)$$

$$(0.089268 \text{ to } 7.910732)$$

So, with 95% confidence we can state that based on these two sample results, the difference in the mean speeds of the urban and rural population lies between 0.089 and 7.911.

Testing of Hypothesis

The significance level is $\alpha = 0.01$.

The null and alternative hypotheses are

H_0: $\mu_1 - \mu_2 = 0$ (mean speeds are not different)
H_1: $\mu_1 - \mu_2 > 0$ (mean speeds of urban people are greater than those of rural people)

The > sign in the alternative hypothesis indicates that the test is right-tailed. The significance level is 0.01. So, the area in the right tail of the t-distribution is $\alpha = 0.01$.

$$df = 28 + 20 - 2 = 46$$

From the t-distribution table, the value of t for df = 46 and the 0.01 area in the right tail of the t-distribution is 2.410.

From the above, we know that $s_{\bar{x}_1 - \bar{x}_2} = 1.942738$, so the calculated value of the t-statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

$$= \frac{(72 - 68) - 0}{1.942738}, (\mu_1 - \mu_2 = 0 \text{ null hypothesis})$$

$$= 2.05895$$

Because the value of the test statistic t = 2.059 for $\bar{x}_1 - \bar{x}_2$ falls in the non-rejection region ($t_{calculated} < t_{tabulated}$), we accept the null hypothesis. Hence, we conclude that the mean speeds for both urban and rural people are the same.

9.4.9 Equality of Means for Dependent Samples

Two samples are said to be dependent when for each data value collected from one sample there is a corresponding data value collected from the second sample, and both data values are collected from the same source.

9.4.10 Paired t-Test

This is a test in which the samples are dependent and are paired so that each observation in one sample is associated with some particular observation in the second sample.

9.4.11 Paired Difference

In dependent samples, the difference between the two data values for each element of the two samples is denoted by d. This value is also called the paired difference.

Note: Each sample contains the same number of values because each source gives a pair of values; that is, both samples are the same size. So we denote the common sample size of n, and the degrees of freedom for the paired sample is n – 1.

The mean and standard deviation of the paired differences for the two samples are denoted by $\bar{d}$ and s_d, respectively.

$$\bar{d} = \frac{\Sigma d}{n}$$

$$S_d = \frac{\Sigma d^2 - (\Sigma d)^2 / n}{n-1}$$

If the number of paired values is large, then because of the central limit theorem, the sampling distribution of $\bar{d}$ is approximately normal with its mean $\mu_{\bar{d}}$ and standard deviation $\sigma_{\bar{d}}$.

$$\mu_{\bar{d}} = \mu_d \text{ and } \sigma_{\bar{d}} = \sigma_d / \sqrt{n}$$

If the number of paired values is small and σ_d is not known, then the t-distribution is used to make an inference about μ_d. The standard deviation $\sigma_{\bar{d}}$ of $\bar{d}$ is estimated by $s_{\bar{d}}$. Then, the estimate of the standard deviation of the paired differences is $s_{\bar{d}} = s_d / \sqrt{n}$.

9.4.11.1 Confidence Interval

The confidence interval for μ_d is $\bar{d} \pm ts_{\bar{d}}$, the where the t-value is obtained from the t-distribution table for a given confidence level and n – 1 degrees of freedom.

9.4.11.2 Hypothesis Testing

Hypothesis testing for μ_d, the value of the test statistic t for $\bar{d}$, is computed as

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}}$$

Exercise

A company sent 10 of its salespersons to attend the course 'How to Be a Successful Salesperson'. Table 9.3 gives the 1-week sales of these salespersons before and after they attended this course. Using the 5% level, conclude the result of attending this course.

Solution

Let d be the difference between weekly sales before attending the course and after attending the course (Table 9.4).
Paired difference mean

$$\bar{d} = \frac{\Sigma d}{n}$$

$$= \frac{-32}{10} = -3.2$$

TABLE 9.3

Data of Salesperson before and after Attending the Course

Sales before attending the course	11	21	15	25	24	26	12	18	22	20
Sales after attending the course	15	24	13	31	28	25	12	24	28	25

TABLE 9.4

Calculation of t-Test

Sales Before	Sales After	Difference d	d^2
11	15	–4	16
21	24	–3	9
15	13	2	4
25	31	–6	36
24	28	–4	16
26	25	1	1
12	12	–1	1
18	24	–6	36
22	28	–6	36
20	25	–5	25
		$\sum d = -32$	$\sum d^2 = 180$

Standard deviation of the paired differences

$$S_d = \frac{\Sigma d^2 - (\Sigma d)^2 / n}{n-1}$$

$$= \frac{\sqrt{180 - 102.4}}{10} = 7.76$$

The standard deviation of $\bar{d}$ is

$$s_{\bar{d}} = \frac{S_d}{\sqrt{n}}$$

$$= \frac{7.76}{\sqrt{10}} = 2.46$$

Let μ_1 and μ_2 be the mean sales before and after the course. Then $\mu_d = \mu_1 - \mu_2$. Then our null and alternative hypotheses are

$$H_0 : \mu_d = 0 (\text{mean weekly sales do not increase})$$

$$H_1 : \mu_d < 0 (\text{mean weekly sales do increase})$$

For the determination of the rejection and non-rejection regions, the < sign in the alternative hypothesis shows that the test is left-tailed. The significance level is 0.05. Hence, the area of the left tail is

$$\alpha = 0.05$$

and

$$\text{d.f.} = n - 1 = 10 - 1 = 9$$

So, the critical value of t at d.f. = 9 is $t_{\text{Tabulated}} = -1.833$.

Next, the calculated value of the test statistic is

$$t_{Calculated} = \frac{\bar{d} - \mu_d}{s_{\bar{d}}}$$

$$= \frac{-3.2 - 0}{2.46} = -1.30$$

$\mu_d = 0$ by the null hypothesis.

Conclusion

The tabulated value of the test statistic is –1.833, and the calculated test statistic is –1.30, so we reject the null hypothesis. Consequently, we conclude that the mean weekly sales for all salespersons do increase.

Exercise

A researcher wanted to analyse whether the mean costs of land are the same regardless of the three metropolitan cities where they are located. Random samples of the prices at which the lands are being purchased in all three areas are given in Table 9.5.
Calculate H and report your results with a p-value of 0.05.

Solution

Let us take the hypothesis as

Null hypothesis H_0: There is no significant difference in the mean cost of land.
Alternate hypothesis H_1: There is a significant difference in the mean cost of land.

Apply the ANOVA technique (Table 9.6).

$$\therefore \bar{\bar{x}} = \frac{\bar{x_1} + \bar{x_2} + \bar{x_3}}{3} = \frac{91.2 + 79 + 55.4}{3} = 75.2$$

The variance between samples is shown in Table 9.7.
Therefore, the sum of squares between samples (SSB) is 1280 + 0.2 + 1960.2 = 3240.4.
The sum of squares within samples (SSW) is 6170.8 + 1482 + 729.2 = 8382 (Table 9.8).
Table 9.9 shows the ANOVA.

$$F_{cal} = \frac{1620.2}{698.5} = 2.31$$

TABLE 9.5

Mean Cost of Land (Thousands of Dollars)

Observation	Low	Moderate	High
1	150	61	40
2	88	69	55
3	40	110	73
4	95	75	45
5	83	80	64

TABLE 9.6

Calculation of ANOVA

	x_1	x_2	x_3
	150	61	40
	88	69	55
	40	110	73
	95	75	45
	83	80	64
Total	$\Sigma x_1 = 456$	$\Sigma x_2 = 395$	$\Sigma x_3 = 277$
$\bar{x}$	$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = 91.2$	$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = 75$	$\bar{x}_3 = \frac{\Sigma x_3}{n_3} = 55.4$

TABLE 9.7

Calculation of Variance between Samples

	$\left(\bar{x}_1 - \bar{\bar{x}}\right)^2$	$\left(\bar{x}_2 - \bar{\bar{x}}\right)^2$	$\left(\bar{x}_3 - \bar{\bar{x}}\right)^2$
	256	0.04	392.04
	256	0.04	392.04
	256	0.04	392.04
	256	0.04	392.04
	256	0.04	392.04
Total	1280	0.2	1960.2

TABLE 9.8

Calculation of Variance within Samples

	$\left(x_1 - \bar{x}_1\right)^2$		$\left(x_2 - \bar{x}_2\right)^2$		$\left(x_3 - \bar{x}_3\right)^2$
x_1	$\bar{x}_1 = \mathbf{91.2}$	x_2	$\bar{x}_2 = \mathbf{75}$	x_3	$\bar{x}_3 = \mathbf{55.4}$
150	3457.44	61	196	40	237.16
88	10.24	69	36	55	0.16
40	2621.44	110	1225	73	309.76
95	14.44	75	0	45	108.16
83	67.24	80	25	64	73.96
Total	6170.8		1482		729.2

TABLE 9.9

Analysis of Variance

Sources of Variation	Sum of Squares	Degrees of Freedom (ν)	Mean Sum of Squares
Between	SSB = 3240.4	c – 1 = 2	MSB = 1620.2
Within	SSW = 8382	n – c = 15 – 3 = 12	MSW = 698.5

For $v_1 = 2$ and $v_2 = 12$,

$$F_{\text{Tabulated at 0.05 level of significance d.f. 2 and 12}} = 3.88$$

We find that the calculated value of F is less than the tabulated value, that is 2.31 < 3.88.

Therefore, the null hypothesis H_0 is accepted and the alternate hypothesis H_1 is rejected. Hence, there is no significant difference in the mean cost of land.

9.5 Summary

Many problems arise in the daily life of a manager where a decision has to be made on the basis of sample data. The decision made by the managers always has an inherent risk of being wrong. Hence, managers need to verify whether the parameters selected for making decisions are being influenced by sampling fluctuations alone or by some other factors also. Hypothesis testing is a handy tool for this purpose.

Hypothesis testing is a process of testing the significance regarding a parameter of the population on the basis of a sample. The hypothesis being tested is called the null hypothesis and generally specifies an exact parameter value. Any hypothesis complementary to the null hypothesis is called an alternative hypothesis. There are two-tailed tests and one-tailed tests of hypotheses. A two-tailed test of a hypothesis will reject the null if the sample mean is significantly higher or lower than the hypothesised population mean value. The one-tailed tests are of two types: left-tailed (lower-tailed) test and right-tailed (upper-tailed) test. A left-tailed test of a hypothesis will reject the null hypothesis in favour of the alternative hypothesis if the sample mean is significantly below the hypothesised population mean. A right-tailed test of a hypothesis will reject the null hypothesis in favour of the alternative hypothesis if the sample mean is significantly higher the hypothesised population mean.

This chapter discussed the steps involved in testing the significance of a hypothesis and the types of possible errors in that process.

The methods of testing the hypothesis differ with the size of the sample. The chapter also discussed the test of significance for a large sample (i.e. a sample with a minimum size of 30) and for a small sample.

REVIEW QUESTIONS

1. Discuss the main principles of large sample theory with special reference to the sampling of attributes.
2. Why should there be different formulae for testing the significance of the difference between means when the samples are (a) small and (b) large?
3. What is sampling distribution? Explain the role of standard errors in large sample tests.
4. How would you test the significance of the difference of two variances when the samples are (a) large and (b) small? Mention the assumptions involved and the statistics used in each case.
5. Define Student's t and write down without proof its sampling distribution.
6. a. Explain the terms *null hypothesis* and *level of significance*.
   b. Explain the concept of null hypothesis.
7. What do you understand by t-test and Fisher's F-test? Indicate some practical applications of these tests.

8. Describe the different uses of the F-statistic, stating clearly the assumption involved.
9. Explain how the Student's t-test is a landmark in the development of statistical methods.
10. Explain the uses of the t-test and F-test, indicating in each case what exactly is sought to be tested and under what assumptions the test will be valid.
11. Explain clearly what is meant by the sampling distributions of a statistic. Develop your answer with reference to one statistic.

SELF-PRACTICE PROBLEMS

1. The incomes of a random sample of engineers in Industry I are Rs. 630, 650, 680, 690, 710 and 720 per month. The incomes of a similar sample from Industry II are Rs. 610, 620, 650, 660, 690, 700, 710, 720 and 730 per month. Discuss the validity of the suggestion that Industry I pays its engineers much better than Industry II.

   [Answer: $t = 0.1$, $v = 14$; for $v = 14$, $t_{0.05} = 2.15$. The suggestion is valid.]
2. The following table gives the retail prices of a commodity in some shops selected at random in four cities:

City	Prices (Rs. per lb)			
A	22	24	27	23
B	20	19	23	—
C	19	17	21	18
D	24	25	29	26

   Carry out the ANOVA to test the significance of the difference between the prices of the commodity in four cities.
3. Four machines produce steel wire; the following data gives the diameters at 10 positions along the wire for each machine. Examine, by performing the ANOVA whether the machine means can be regarded as constant.

Machine	Diameters (thousandth of an inch)									
A	12	13	13	16	16	14	15	15	16	17
B	12	14	14	16	16	18	17	19	20	18
C	14	21	17	14	19	18	17	17	16	15
D	23	27	25	21	26	24	57	24	20	21

10

The Chi-Square Tests

10.1 Introduction

A test of significance might be based on the assumption that the sample values were drawn from universes having the same variance. The testing procedures also assumed that the unknown values of the parameters, about which statistical interferences were to be made, could be estimated from statistics obtained from random samples. This approach to inferential statistics is called parametric methods, since the concern is with the value of a parameter.

There are many situations in which it is not possible for the statistician to make a rigid assumption about the shape of the population from which samples are being drawn. This limitation has led to the development of a group of alternate techniques known as non-parametric or distribution-free methods.

A non-parametric method may be defined as a statistical test in which no hypothesis is made about specific values of parameters. Distribution-free tests may be defined as methods for testing a hypothesis that does not depend on assumptions concerning the form of the underlying distribution.

10.2 Chi-Square, χ^2

The χ^2-test (pronounced 'chi-square test') is one of the simplest and most widely used non-parametric tests in statistical work. The χ^2-distribution has very many applications in situations that involve the testing of hypotheses concerning discrete or qualitative data. The Greek letter χ was first used to describe this statistic by Karl Pearson in 1900. The quantity χ^2 describes the magnitude of discrepancy between theory and observation; that is, with the help of the χ^2-test, we are in a position to know whether a given discrepancy between theory and observation may be attributed to chance or whether it results from the inadequacy of the theory to fit the observed facts. If χ^2 is zero, it means that the observed and expected frequencies completely coincide. The greater the discrepancy between the observed and expected frequencies, the greater is the value of χ^2.

The square of a standard normal variable is called a chi-square variate with 1 degree of freedom (d.f.). For example, if X is a random variable following a normal distribution with a mean μ and standard deviation σ, then $(X-\mu)/\sigma$ is a standard normal variate and $((X-\mu)/\sigma)^2$ is a chi-square variate with 1 degree of freedom.

10.2.1 Need for the χ^2-Test

To test hypotheses about a population mean or two or more population means, it was assumed that the population was normal. A test of hypothesis can be made for interval data.

But for the χ^2-distribution, assumptions are not necessary for shaping the parent population, and tests can be conducted if the data is not interval scale, but is nominal or ordinal.

10.2.2 Conditions for the Validity of χ^2

10.2.2.1 Assumptions

1. N, the total frequency, that is sample size, should be large enough, $N > 50$.
2. The expected frequency should be greater than 5, that is $f_e > 5$.
3. The data should be given in original numbers, that is natural numbers.
4. The sample observation should be random and independent.

10.2.2.2 Interval Scale

It deals with data that was at least an interval scale such as weights, incomes and wages.

10.2.2.3 Nominal-Level or Nominal-Scale Data

This type of data can only be classified into categories, such as male and female, literate and illiterate.

10.2.2.4 Ordinal-Level Data

The data measurement assumes that one category is ranked higher than the next one. For example, a ranking of outstanding is higher than good, good is higher than fair, brighter is higher than lighter and so on.

10.2.3 Degrees of Freedom

1. While comparing the table value of χ^2 with the calculated value, we have to determine the degrees of freedom.
2. Degrees of freedom is the number of classes to which the values are assigned arbitrarily or at will without violating the restrictions.
3. For example, we choose any four numbers whose total is 40. Here, we have a choice to select any three numbers, say 10, 5 and 20, and the fourth number is 5, that is [40 – (10 + 5 + 20)].

   Thus, our choice of freedom is reduced by 1, on the condition that the total is 40. Therefore, the restriction placed on the freedom is 1 and the degrees of freedom is 3. As the restriction increases, the degrees of freedom is reduced.

$$\upsilon = n - k$$

$$= 4 - 1$$

$$= 3$$

where:

υ = degrees of freedom

k = no. of independent constraints

n = no. of frequency classes

4. If more restrictions are placed, our freedom of choice will be still curtailed.

   For example, if there are 10 classes and we want our frequencies to be distributed in such a manner that the number of cases, the mean and the standard deviation agree with the original distribution, now we have three constraints:

$$\upsilon = n - k$$
$$= 10 - 3$$
$$= 7$$

   Thus, the number of degrees of freedom is obtained by subtracting the number of constraints from the number of frequency classes.

10.2.3.1 In Binomial Distribution

The number of degrees of freedom is one less than the number of classes:

$$\upsilon = n - 1$$

10.2.3.2 In Poisson Distribution

The number of degrees of freedom is two less than the number of classes. (We use total frequency and arithmetic mean.)

$$\upsilon = n - 2$$

10.2.3.3 In Normal Distribution

The number of degrees of freedom is three less than the number of classes. (We use total frequency, mean and standard deviation.)

$$\upsilon = n - 3$$

10.2.3.4 For a Contingency Table

$$\upsilon = (c - 1)(r - 1)$$

where:

c = number of columns
r = number of rows

10.2.3.5 Important Characteristics of Degrees of Freedom (ν)

1. The distribution of χ^2 depends on the degrees of freedom.
2. There is a different χ^2-distribution for each number of degrees of freedom.
3. The distribution is much skewed to the right for small degrees of freedom.
4. As the degrees of freedom increases, the curve becomes more and more symmetric.
5. The mean of the χ^2-distribution is equal to the number of degrees of freedom.
6. The variance is equal to 2 degrees of freedom.

10.2.4 General Aspects of χ^2

1. Tests can be performed on the actual numbers but not on the percentages and proportions. If the data is in percent or proportion, then it needs to be converted into absolute numbers before performing the χ^2-test.
2. No theoretical cell frequency should be small. Here, again it is hard to say what constitutes smallness, but 5 should be regarded as the very minimum and 10 is better. In practice, data not infrequently contain cell-frequencies below these limits. As a rule, the difficulty may be met by amalgamating such a cell into a single cell entitled '10 or greater than 10'.
3. The chi-square test works only when the sample size is large enough; usually, the sample size needs to be >50.
4. Observations drawn need to be random and independent.

10.2.5 Characteristics of the Chi-Square Distribution

1. χ^2 is always positive. The computed value of χ^2 is always positive because the differences between E_{ij} and O_{ij} are squared, that is $(E_{ij} - O_{ij})^2$.
2. The shape of the χ^2-distribution depends on number of cells. That is, the number of degrees of freedom is determined by (n – 1), where n is the number of samples (categories).

   Therefore, the shape of the χ^2-distribution does not depend on the size of the sample.

   For example, if 300 employees of an airline were classified into one of three categories, that is flight personnel, ground support and administrative personnel, there would be $n - 1 = 3 - 1 = 2$ degrees of freedom.
3. The χ^2-distribution is truly skewed. However, as the number of degrees of freedom increases, the distribution begins to approximate the normal distribution.
4. The greater the discrepancy between the observed frequency, the greater is the value of χ^2.
5. Large values of χ^2 indicate disagreement between the observed frequency (O_i) and the expected frequency (E_i) under the null hypothesis.
6. The critical value regions will lie towards the extreme right tail of the χ^2-distribution. Therefore, it is called a right-tailed test or positively skewed.

7. It depends only on the set of observed and expected frequencies and on degrees of freedom.
8. It does not make any assumptions regarding the parent population.
9. It does not involve any population parameters (statistics) and is therefore known as a non-parametric test.
10. It is a distribution-free test because there are no assumptions.
11. If χ^2 is 0, this means that the observed and expected frequencies completely coincide.
12. Uses for quantitative data. Other distributions cannot be used for qualitative data. Generally, for qualitative data a non-parametric test is preferred, and in parametric tests we have only χ^2 for qualitative tests.
13. Goodness of fit, validity, checking or confirmation, quality checking or fitting only with the help of the χ^2-test. In goodness of fit, the null hypothesis, H_0, is proposed for acceptance.

10.2.6 Application of Chi-Square

1. It is applicable for testing the significance of the concerned discrete or qualitative data that involves testing of a hypothesis.
2. It describes the magnitude of discrepancy between an observation (experiment) and theory, and may be distributed by chance (fluctuation of sampling) or result from the inadequacy of the theory.
3. It can evaluate the relationship between two or more variables.
4. We will not always be interested in means and proportions. There are many managerial situations where we will be concerned with the variability in a population, instead of means and proportions.
5. The χ^2-test enables us to test whether more than two population proportions can be considered equal.
6. We can use the χ^2-test to determine if the two attributes are independent of each other. The χ^2-test is a test of independence that determines whether the difference between the proportions representing more than two samples is significant.
7. It determines whether the two attributes according to which a population is categorised are independent of each other, and it also serves as a test for goodness of fit.
8. It is useful when analysing cross-tabulation.
9. It helps to determine whether there is any significant association between the variables involved in the research problem.

10.2.7 Limitations of Chi-Square

1. If there are only two cells, the expected frequency (E_i) in each cell should be five or more.
2. For more than two cells, χ^2 should not be applied if more than 20% of E_i cells have less than five frequencies.

10.3 Chi-Square Test of Goodness of Fit

This is a non-parametric test developed by Karl Pearson in the early 1900s. χ^2 is a test statistic used to test a hypothesis that provides a set of theoretical frequencies with which observed frequencies are compared; goodness means +ve. It is used to determine how well an observed set of data fits an expected set, by using the formula

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

where:

n = number of categories
O_i = observed frequency in a particular category
E_i = expected frequency in a particular category

10.3.1 Procedure for χ^2-Test of Goodness of Fit – Steps

1. Formulate H_0 and H_1.

   Null hypothesis H_0: There is no difference between the set of observed frequencies and the set of expected frequencies; that is, any difference can be attributed to chance.

   Alternate hypothesis H_1: There is a difference between the two set of frequencies.
2. Calculate the expected frequency.
3. The test statistic. Calculate χ^2 by using the formula.
4. Decide on the test level of significance. Alpha (α) = 5%, a level of significance which is the same as the Type I error of probability. Thus, the probability is 0.05 that a true null hypothesis will be rejected, and the degrees of freedom = (n – 1), where n = number of samples.
5. The decision rule and the critical value. Determine the critical (tabulated) value of χ^2 and then compare it with the calculated value.

10.3.2 Critical Value

The decision rule in hypothesis testing requires finding a number that separates the region, where we do not reject H_0 from the region of rejection. This number is called the critical value.

6. Deduce the conclusion.

10.3.3 Decision Rules

1. The decision rule indicates that if there are large differences between the expected and observed frequencies, resulting in a computed χ^2 of more than the table

value, the null hypothesis should be rejected. If the calculated value of chi-square ($\chi^2_{calculated}$) is greater than the tabulated value of chi-square ($\chi^2_{tabulated}$), we reject H_0 and accept the alternate hypothesis, H_1.

2. If the difference between the observed and expected frequencies is small, then H_0 should not be rejected, that is accepted. Clearly, if the differences between E_{ij} and Oij are small, the computed χ^2-value will be less and the null hypothesis should not be rejected, because such small differences between the expected and observed frequencies are probably due to chance. Do not reject the null hypothesis, H_0, if the calculated value of χ^2 is less than or equal to the χ^2 table value.

10.4 Chi-Square Test – Test of Independence

10.4.1 Characteristics

1. It is used to evaluate the relationship between two or more variables.
2. It is useful when analysing cross-tabulation.
3. It helps in determining whether there is any significant association between the variables involved in the research problem.

   For example, suppose N number of observations are classified according to criteria. We may ask whether the criteria are relative or independent, for example

   1. Whether a particular drug is effective in controlling fever
   2. Whether there is any association between petrol consumption and air pollution

10.4.2 Procedure for χ^2-Test of Independence – Steps

1. Formulate the null (H_0) and alternate (H_1) hypotheses.
2. Calculate the expected frequency.

$$E_{ij} = \frac{n_i \times n_j}{n}$$

where:

E_{ij} = expected frequency of a cell corresponding to a particular row and column
n_i = row total
n_j = column total
n = total sample size

3. Calculate χ^2 by using the formula.
4. Decide on the test level of significance and degrees of freedom.

   Degrees of freedom = [Number of rows – 1] [Number of columns – 1]
5. Determine the critical value of χ^2 and then compare it with the calculated value.

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}\left[\frac{(E_{ij} - O_{ij})^2}{E_{ij}}\right]$$

where:
O_{ij} = observed frequency in ith row and jth column
E_{ij} = expected frequency in ith row and jth column
n = number of rows
k = number of columns

6. Deduce the conclusion.

For example,

H_0: Given data follows a Poisson distribution.

H_0: X batch is a successful batch.

Exercise

Table 10.1 shows the data obtained during an epidemic of dengue.
Test the efficiency of inoculation in preventing the attack of dengue.

Solution

1. Use the formulae for H_0 and H_1.
   H_0: Inoculation is not effective in preventing the attack of dengue.
   H_1: Inoculation is effective in preventing the attack of dengue.
2. Calculate the expected frequency.

$$E_{ij} = \frac{n_i \times n_j}{n}$$

where:
E_{ij} = expected frequency of a cell corresponding to a particular row and column
n_i = row total
n_j = column total
n = total sample size

The observed frequencies and corresponding expected frequencies are shown in parentheses in Table 10.2.

3. Calculate χ^2 by using the formula

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}\left[\frac{(E_{ij} - O_{ij})^2}{E_{ij}}\right]$$

where:
O_{ij} = observed frequency in ith row and jth column
E_{ij} = expected frequency in ith row and jth column
n = number of rows
k = number of columns

TABLE 10.1

Data of an Epidemic of Dengue

	Attacked	Not Attacked	Total
Inoculated	31	469	500
Not inoculated	185	1,315	1500
Total	216	1,784	2000

TABLE 10.2

Observed and Expected Frequencies

	Attacked	Not Attacked	Total
Inoculated	$\frac{316\times500}{2000}$ = (54) 31	$\frac{1784\times500}{2000}$ = (446) 469	500
Not inoculated	$\frac{316\times500}{2000}$ =(162) 185	$\frac{1784\times1500}{2000}$ =(1338) 1315	1500
Total	216	1784	2000

TABLE 10.3

Calculation of χ^2

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
31	54	23	529	9.80
185	162	23	529	3.26
469	446	23	529	1.18
1315	1338	23	529	0.39
$\sum O_i = 2000$	$\sum E_i = 2000$			$\sum(O_i - E_i)^2/E_i = 14.63$

The calculation is shown in Table 10.3.

4. Decide on the test level of significance and degrees of freedom.
   Degrees of freedom = [Number of rows – 1] [Number of columns – 1]
   = [2 – 1] [2 – 1]
   = 1
   The level of significance is 5%.
5. Determine the critical value of χ^2 and then compare it with the calculated value.
   From the χ^2-distribution table,

$$\chi^2_{\text{Tabulated}} \text{ (1 d.f. at 5\%)} = 3.84$$

$$\chi^2_{\text{Calculated}} = 14.63$$

$$\chi^2_{\text{Tabulated}} (3.84) > \chi^2_{\text{calculated}} (14.63)$$

See the graph in Figure 10.1.

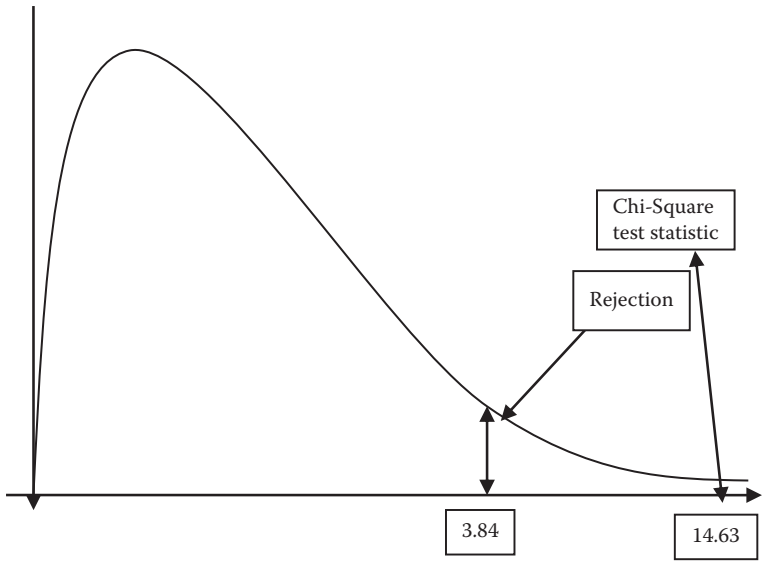


FIGURE 10.1
Chi-square test of independence.

Conclusion: There is evidence to reject the null hypothesis that

$$\chi^2_{\text{Calulated}} > \chi^2_{\text{Tabulated}}$$

Therefore, we conclude that the calculation is effective in preventing the attack of dengue.

Exercise

Channel viewership is segregated according to age group. Evaluate the statistical significance of association among the variables involved in the cross-tabulation.

Table 10.4 shows the channel viewership distribution according to age group.

Solution

Observation: The 15–35 age group respondents prefer Channel A compared with the other channels. But, it is not certain whether this observation is representative of the entire population or whether it is due to sampling error. This dilemma can be resolved by subjecting the data to a χ^2-test.

TABLE 10.4

Channel Viewership Distribution according to Age Group

Channel →, Age Group ↓	**A**	**B**	**C**	**Total**
10–15	20	30	30	80
15–35	80	70	50	200
≥35	60	40	20	120
Total	160	140	100	400

Steps

1. Formulate H_0 and H_1.
   H_0: There is no significant association between age groups and channel viewership.
   H_1: There is a significant association between age groups and channel viewership.
2. Calculate the expected frequency.
   The expected frequency value for each category (age group) can be calculated by using the formula

$$E_{ij} = \frac{n_i \times n_j}{n}$$

where:
- E_{ij} = expected frequency of a cell corresponding to a particular age group and a particular channel
- n_i = row total
- n_j = column total
- n = total sample size

Table 10.5 shows the observed frequencies and corresponding expected frequencies in parentheses.

3. Calculate $\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}\left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right]$

where
- O_{ij} = observed frequency in ith row and jth column
- E_{ij} = expected frequency in ith row and jth column

The calculation of χ^2 is shown in Table 10.6.

4. Decide on the level of significance and degrees of freedom.
   Let us set the level of significance at 1%.
   Degrees of freedom $\nu = (n - 1)(k - 1)$,

where
- ν = degrees of freedom
- n = number of rows
- k = number of columns

TABLE 10.5

Observed and Expected Frequencies

Channel →, Age Group ↓	**A**	**B**	**C**	**Total**
10–15	$\frac{80\times60}{400} = (32)20$	$\frac{80\times140}{400} = (28)30$	$\frac{80\times100}{400} = (20)30$	80
15–35	$\frac{200\times160}{400} = (80)80$	$\frac{200\times140}{400} = (70)70$	$\frac{200\times100}{400} = (50)50$	200
≥35	$\frac{120\times160}{400} = (48)60$	$\frac{120\times140}{400} = (42)\ 40$	$\frac{100\times120}{400} = (30)20$	120
Total	160	140	100	400

TABLE 10.6

Calculation of χ^2

O_{ij}	E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
20	32	–12	144	4.5
30	28	2	4	0.1428571
30	20	10	100	5.0
80	80	0	0	0
70	70	0	0	0
50	50	0	0	0
60	48	12	144	3
40	42	–2	4	0.095238
20	30	–10	100	3.33333
$\Sigma O_{ij} = 400$	$\Sigma E_{ij} = 400$			$\chi^2 \cong 16.08$

Here, n = 3, k = 3 and
$\nu = (3 - 1)(3 - 1) = 2 \times 2 = 4$

5. Determine the critical value and compare it with the calculated value.
   From the χ^2-distribution table,
   $\chi^2_{\text{Tabulated}}$ for 4 d.f. at the 1% level of significance = 13.28
   $\chi^2_{\text{Calculated}} = 16.08$

$$\chi^2_{\text{Tabulated}}\ (13.28) < \chi^2_{\text{Calculated}}\ (16.08)$$

Graph the χ^2-test, for 4 degrees of freedom, at the 1% level of significance showing the region of rejection (Figure 10.2).

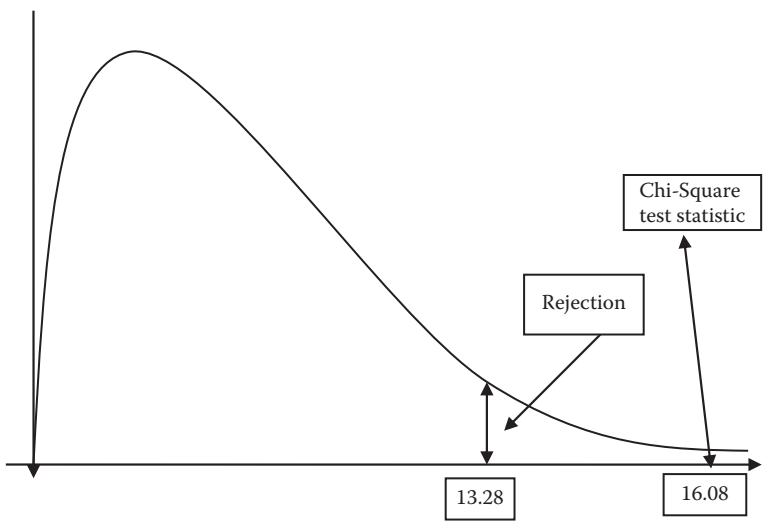


FIGURE 10.2
Chi-square test of independence.

6. Deduce the conclusion: evidence to reject the null hypothesis (H_0).

$$\chi^2_{Calculated}\ (16.08) > \chi^2_{tabulated}\ (13.28)$$

Therefore, the null hypothesis that there is no significant association between the age group and the channel viewership is rejected. This implies that the channel viewership is significantly (1% level) dependent on the age group.

Conclusion

The channel viewership is significant (1% level) dependent on the age group.

10.5 Strength of Association

1. The test of independence will only enable the researcher to identify whether there is an association between the two variables.
2. The test will not describe the strength or magnitude of the association.
3. The strength of the association can be evaluated using two key techniques:
   a. Phi-coefficient
   b. Coefficient of contingency

10.6 Phi-Coefficient

In order to measure the strength, the phi-coefficient is used. It is denoted by Φ. The phi-coefficient measures the strength of association between only two variables (i.e. with two rows and two columns). It fails to measure the strength of association between more than two variables.

A formula for measuring the strength of association between two variables is

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

where

χ^2 = calculated chi-square value
n = sample size

A limitation is that Φ measure is suitable for only 2×2 tables.

10.7 Coefficient of Contingency

This technique can measure the strength of association between more than two variables. It can be calculated for tables of any size. The coefficient of contingency (C) can be calculated by using the given formula

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

where:

χ^2 = calculated chi-square value
n = sample size

1. The coefficient varies from 0 to 1.
2. Value 0 indicates there is no association between the variables.
3. Value 1 indicates the maximum strength.

For example, the calculated χ^2-value is 16.08, and the sample size is 200. The coefficient of contingency is given by

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

$$= \sqrt{\frac{16.08}{16.08 + 200}}$$

$$= 0.2727945796$$

10.8 Summary

In this chapter, we have looked at some situations where we can develop tests based on the chi-square distribution. We started by testing the variance of a normal population where the test statistic used was $(n-1)s^2/\sigma^2$ since the distribution of the sample variance s^2 was not known directly. We found that such tests could be one-tailed depending on our null and alternative hypotheses.

We then describe a multinational experiment and found that if we have data that classify observations into k different categories, and if the conditions for the multinomial experiment are satisfied, then a test statistic called the chi-square statistic, defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

will have a chi-square distribution with specified degrees of freedom. Here, O_i refers to the observed frequency of the ith category and E_i to the expected frequency of the ith category, and the degrees of freedom is equal to the number of categories minus 1, minus the number of independent parameters estimated from the data to calculate the E_i values. This concept was used to develop tests concerning the goodness of fit of the observed data to any hypothesised distribution, and also to test whether two criteria for classification are independent.

The chapter also discussed the concepts of the chi-square test and its uses and testing for goodness of fit. The various tests of hypothesis in the earlier sections are based on the assumption that the sampling distribution follows a normal distribution curve. However, it is not always possible to assume the distribution pattern from which the samples are drawn. To overcome this difficulty, many managers use chi-square tests. Chi-square is a test statistic used to test a hypothesis that provides a set of theoretical frequencies with which observed frequencies are compared.

REVIEW QUESTIONS

1. What is the χ^2-test of goodness of fit? What precautions are necessary in using this test?
2. Describe the χ^2-test of significance and state the various uses to which it can be put.
3. Illustrate with examples the usefulness of the χ^2-test as a test for independence.
4. Describe the use of the χ^2-test in testing the independence of attributes in a 2×2 contingency table.
5. Explain how the χ^2-distribution can be used to (a) test the goodness of fit and (b) test the independence of the cell frequencies in a 2×2 contingency table.

SELF-PRACTICE PROBLEMS

1. The following data relates to the sales in a time of trade depression of a certain proprietary article in wide demand. Does the data suggest that the sales are significantly affected by the depression?

Districts Where Sales Are	Districts Not Hit by Depression	Districts Hit by Depression	Total
Satisfactory	250	80	330
Not satisfactory	140	30	170
Total	390	110	500

[Answer: $\chi^2 = 0.84$; $\chi^2_{0.05} = 3.84$. The hypothesis holds well.]

2. Two sample polls of votes for two candidates A and B for a public office are taken, one from among residents of urban areas and the other from residents of rural areas. The results are given below. Examine whether the nature of the area is related to voting preferences in the election.

Votes for Area	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

[Answer: $\chi^2 = 10.6$, the table value of χ^2 at the 5% level for $v = 1 = 3.84$. The hypothesis does not hold well.]

11

Business Forecasting

11.1 Introduction

The business decisions analysis of an industry gives valued information that is useful for determining business forecasting. Unfavourable business forecasting results are useful for either improving the business performance of an industry or liquidating it at the earliest. This vital contribution of business decision analysis to an industry has to be acknowledged by the management from the viewpoint of its business interests in the future.

Data on demands of the market may be needed for a number of purposes to assist an organisation in its long-, medium- and short-term decisions.

11.2 Forecasting

Growing competition, rapidly changing circumstances and the trend towards automation demand that decisions in business are not based purely on guesses and hunches but, rather, on a careful analysis of data concerning the future course of events.

When estimates of future conditions are made on a systematic basis, the process is referred to as *forecasting*, and the figure obtained is known as a *forecast*.

In a world where the future is not known with certainty, virtually every business and economic decision rests on a forecast of future conditions. When someone assumes the responsibility for running a business, they automatically take on the responsibility for forecasting the future and the successful or unsuccessful course of events.

Forecasting aims at reducing the area of uncertainty with respect to costs, profit, sales, production, pricing, capital investment and so on.

If the future were known with certainty, forecasting would be unnecessary. Decisions could be made and plans formulated on a once-and-for-all basis, without the need for subsequent revision. But uncertainty does exist, future outcomes are rarely assured, and therefore, an organised system of forecasting is necessary rather than the establishment of predictions based on hunches, intuition or guesses.

Forecasting is essential for a number of planning decisions and often provides a valuable input on which future operations of the business enterprise depend.

Areas where forecasts of the future product demand would be useful are

1. Specification of production
2. Planning of equipment, manpower and procurement

3. Budget allocation depending on production and sales
4. Determination of inventory policy
5. Decision on expansion and changes in production process
6. Future trends of product development, diversification, scrapping and so on
7. Design of pricing policy
8. Planning of distribution and sales promotion

11.3 Future Uncertainty

The future is uncertain, and any forecast at best is an educated guess with no guarantee of coming true.

Forecasting generally refers to the scientific methodology that often uses past data along with some assumptions to come up with a forecast of future demand.

A *prediction* is an estimate made by an individual using an intuitive 'hunch', which may, in fact, turn out to be true.

The demand for a particular product (say, a raincoat) would depend on competitors' prices, advertising, weather conditions, population, and a number of factors that might even be difficult to identify. In spite of these complexities, a forecast has to be made so that the manufacturers of raincoats (a product that exhibits a seasonal demand) can plan for the next season.

11.4 Forecasting for Planning Decisions

The primary purpose of forecasting is to provide valuable information for planning the design and operation of the enterprise. Planning decisions may be classified as long term, medium term and short term.

Long-term decisions include decisions such as plant expansion or new product introduction, which may require new technologies. Such decisions are characterised by lack of quantitative information and absence of historical data on which to base the forecast of future events. Intuition and the collected opinion of experts in the field play a significant role in developing forecasts.

Medium-term decisions involve

1. Planning the production levels in a manufacturing plant over the next year
2. Determining manpower requirements or inventory policy

Short-term decisions involve

1. Daily production planning
2. Scheduling decisions

11.5 Steps in Forecasting

Forecasting business change involves more than analysis of such things as secular trend, seasonal variations, cyclical variations and a consideration of cause and effect.

Forecasting of business fluctuations consists of the following steps:

1. Understanding the occurrence of changes in the past

   If an attempt is made to forecast business fluctuations without understanding why past changes have taken place, the forecast will be purely mechanical, based solely on the application of mathematical formulae and subject to serious error.

2. Determining phases of business activity to be measured

   It is necessary to measure certain phases of business activity to predict what changes will probably follow the present level of activity.

3. Selecting and compiling data to be used as measuring devices

   Statistical data cannot be selected and compiled in an intelligent manner unless there is a sufficient understanding of business fluctuations; it is important that reasons for business fluctuations be stated in such a manner that it is possible to secure data related to the reasons.

4. Analysing the data

   The data is analysed in the light of one's understanding of the reason why change occurs.

If it is reasoned that a certain combination of forces will result in a given change, the statistical part of the problem is to measure these forces, and to draw conclusions on the future course of action from the data available. The methods of drawing conclusions may be called *forecasting techniques*, which represent a large number of analytical devices for summarising data and drawing inferences from the summaries.

11.6 Methods of Forecasting

The rule-of-thumb method has been practiced in business. It consists in deciding about the future in terms of past experience. This method is very widely used in business. However, it can lead to absurd conclusions if employed by the inexperienced.

The techniques of forecasting are applicable to every sphere of business activity. Forecasting business change involves more than an analysis of statistical data – it also embodies the prediction of economic change, such as

1. Secular trend
2. Seasonal variation or cyclical variation
3. Consideration of cause and effect

Methods of forecasting are

1. Business barometers
2. Extrapolation
3. Regression analysis
4. Econometric models
5. Forecasting by the use of time series analysis
6. Opinion polling
7. Causal models

A forecast is usually a combination of several techniques.

11.6.1 Business Barometers

Practical forecasting is a series used as an 'index' or 'indicator' of basic conditions related to the industry. The term *barometer* is used in business statistics to mean an indicator of the present economic situation and to designate an indicator of future conditions.

Series that aid in business forecasting include

1. Gross national product
2. Employment
3. Wholesale prices
4. Consumer prices
5. Industrial production
6. Volume of bank deposits and currency outstanding
7. Consumer credit
8. Disposable personal income
9. Departmental store sales
10. Stock prices
11. Bond yields

Index numbers relating to different activities in the field of production, trade, finance and so on may be combined into a general index of business activity. *General index* refers to the general conditions of trade and industry. But the behaviour of individual industries and trades might show a different trend from that of the composite business activity index. The trends indicated by barometers will guide the business person as to whether the stocks of goods should be increased or released, whether to increase investment or not, and so on.

11.6.2 Extrapolation

Extrapolation relies on the relative constancy in the pattern of past movements in some time series. Extrapolation is used for sales forecasts and for other estimates when 'better' forecasting methods may not be justified.

Since extrapolation assumes that the variable will follow its established pattern of growth, the problem is to determine the trend curve and the values of its parameters.

Trend curves suitable for business forecasting application are

1. Arithmetic trend: The straight-line arithmetic trend assumes that growth will be by a constant absolute amount each year.
2. Semi-log trend: The semi-logarithmic trend assumes a constant percentage increase each year. Since the annual increment is constant in logarithms, this line translates into a straight line when drawn on paper in a logarithmic vertical scale.
3. Modified exponential trend: This curve assumes that each increment of growth will be a constant percentage (less than 100%) of the previous one. The line trends generally to approach, but never quite reach, a constant asymptote, which may be thought of as an upper limit.
4. Logistic curve: The logistic curve has both an upper asymptote and a lower asymptote. It assumes a 'law of growth' involving increasing increments from an initial low value and then gradual slowing down of growth as 'maturity' is approached.
5. The Gompertz curve: The Gompertz curve is often used to describe growth of industrial output.

Selection of an appropriate growth curve can be guided by empirical and theoretical considerations. Empirically, it is a question of selecting the curve that best fits the past movement of the data. Theoretical matters that intervene in this logic may support a particular growth pattern.

For example, population growth when there are no resource or choice restraints implies a geometric pattern of growth. With limited resources, population is sometimes thought to grow along a logistic curve.

11.6.3 Regression Analysis

The regression approach offers many valuable contributions to the solution of the forecasting possible or theoretically suggested relationships between variables in a complex economy those which will be useful for forecasting. If two variables are functionally related, knowledge of one will make possible an estimate of the other.

For example, if we know that advertising expenditure and sales are correlated, we can find out the probable increase in sales for a given advertising expenditure, or vice versa.

Regression relationships may involve only one predicted, or dependent, and one independent variable (simple regression), or they may involve relationships between the variable to be forecast and several independent variables (multiple regression).

Statistical techniques to estimate the regression equations are often complex and time-consuming, but there are many computer programs now available that estimate simple and multiple regressions cheaply and quickly.

The dangers in using regression analysis for forecasting are

1. There is a possibility of a mechanistic approach, accepting with little question the relationship that the calculations reveal – perhaps that with the highest r^2 – and applying it to forecasting. There are many possibilities for spurious correlation among time series, as many series move together over time even where there is no conceivable connection between them.

2. There is a risk that the estimated regression is false. The forecaster must always use judgement and knowledge of the facts and of the underlying theory.

11.6.4 Econometric Models

Econometric techniques, which originated in the early 1800s, have recently gained popularity for forecasting. Much of the revival of econometrics is attributed to the growth of computer technology.

The term *econometrics* refers to the application of mathematical economic theory and statistical procedures to economic data to verify economic theorems and to establish quantitative results in economics. An econometrician is, therefore, an economist, a statistician and a mathematician, all in one.

Econometric models take the form of a set of simultaneous equations. The values of the constants in such equations are supplied by a study of statistical time series, and a large number of equations may be necessary to produce an adequate model. The work of computation is greatly facilitated by electronic data-processing equipment such as computers.

At the present time, most short-term forecasting uses only statistical methods with little qualitative information. However, in the years to come, when most large companies develop and refine econometric models of their major businesses, this forecasting tool will become more popular.

The development of an econometric model requires sufficient data that the correct relationships can be established. When data is scarce – for example, when a product is first introduced into a market – this method cannot be profitably employed.

The econometric model is most formal, since the forecast is based on an explicit mathematical model. The model states in detail and in quantitative terms the ways in which the various aspects of the economy are interrelated. Theoretically, the model makes possible a wholly mechanical forecast once values have been estimated for the exogenous variables. But, in practice, qualitative and quantitative forecasters have tended to come together.

The 'artist' forecaster has become fully aware of the need for quantitative relationships, while the econometric forecaster has learned that in some instances, quantitative relationships have to be modified by qualitative factors.

The econometric model provides the forecaster with a record of the prediction with a clear statement of the assumptions concerning exogenous variables and the solution of the model – it is often possible, or at least it is made easier, to trace and reproduce the causes for success as well as failures.

One can learn just where errors were made and where improvements can be made. Thus, discredited hypothesis may be dropped and new ones can be substituted, which ultimately will lead to better understanding of the economic system and business fluctuations.

Econometric models are not very popular in practice, because it is probably neither necessary nor feasible for every business forecaster to construct their own model of the economy. The effort and cost involved in a fully developed econometric model are well beyond most forecasting operations. Most forecasters will probably rely on the basis aggregate models developed at research institutes or universities. These models may be used to make predictions and to test out alternative assumptions about government policy or the other exogenous aspects of the economy. With the help of the various forecasting methods and of their interrelationships, forecasts will be improved.

11.6.5 Forecasting by the Use of Time Series Analysis

Time series analysis helps to identify and explain

1. Any regular or systematic variation in the series of data that is due to seasonality – the *seasonals*
2. Cyclical patterns that repeat every 2 or 3 years or more
3. Trends in the data
4. Growth rates of these trends

Most existing methods identify only the seasonal, the combined effect of trends and cycles, and the irregular, or chance, component. That is, they do not separate trends from cycles.

This is not to say that those other effects are not to some degree manageable. The suggestion is that the analysis of trend and seasonal effects and the projection of these two sets of forces should be understood to be the first step in the forecast, and that, taking into account such cyclical and residual forces as may be manageable, further refinements should then be made.

Many statisticians consider time series analysis a somewhat useless tool. One critic, M J Moroney, said that 'Economic forecasting like weather forecasting in England is only valid for the next six hours or so. Beyond that it is sheer guesswork'.

In any event, although the limitations of time series analysis must be understood, the importance and usefulness of these procedures should not be underestimated.

Purpose of analysis:

1. It provides an initial approximation forecast that takes into account those empirical regularities that can, with reasonable assurance, be expected to persist.
2. After the trend and seasonal effects have been identified and measured, the original data may be adjusted for these influences, yielding a new historical time series consisting of the trend and seasonally adjusted data. This new time series may be helpful in the analysis and interpretation of cyclical and residual influences.

This method of forecasting can be used only when several years' data for a product or product line are available, and when relationships and trends are both clear and relatively stable.

11.6.6 Opinion Polling

Opinion polling is a basis for forecasting. The Survey Research Center of the University of Michigan conducts an annual poll regarding the future plans of consumers. The answers to many questions are translated into short-run demand for colour television sets, automobiles and other consumer products.

11.6.7 Causal Models

A causal model is the most sophisticated kind of forecasting tool. It expresses mathematically the relevant causal relationships and may include pipeline considerations (i.e. inventories) and market survey information. It may also directly incorporate the results of a

time series analysis. The causal model takes into account everything known of the dynamics of the flow system and uses predictions of related events such as competitive action, strikes and promotions.

If the data is available, the model generally includes factors for each location in the flow chart and connects these by questions to describe overall product flow. If certain kinds of data are lacking, initially, it may be necessary to make assumptions about some of the relationships and then track what is happening to determine whether the assumptions are true. Typically, a causal model is continually revised as more knowledge about the system becomes available.

11.7 Choice of a Method of Forecasting

Choice of a method of forecasting depends on

1. Context of the forecast
2. Relevance and availability of historical data
3. Degree of accuracy desirable
4. Time period to be forecast
5. Cost benefit or value of the forecast to the company
6. Time available for making the analysis

These factors must be weighed constantly and on a variety of levels. In general, for example, the forecaster should choose a technique that makes the best use of available data. If they can readily apply one technique of acceptable accuracy, they should not try to 'gold plate' by using a more advanced technique that offers potentially greater accuracy but that requires non-existent information or information that is costly to obtain.

Furthermore, where a company wishes to forecast with reference to a particular product, it must consider the stage of the product's life cycle for which it is making the forecast.

11.8 Theories of Business Forecasting

1. Sequence or time-lag theory
2. Action and reaction theory
3. Economic rhythm theory
4. Specific historical analogy
5. Cross-section analysis

11.8.1 Sequence or Time-Lag Theory

Time-lag theory is based on the assumption that most business data have lag and lead relationships; that is, changes in business are successive and not simultaneous. There is a time lag between different movements.

For example, expenditure on advertisement may not at once lead to an increase in sales. Similarly, when government makes use of deficit financing, it leads to inflationary pressures: the purchasing power of people goes up; the wholesale prices and retail prices start rising. With the rise in retail prices, the cost of living goes up, and with it there is a demand for increased wages. Thus, one factor, that is, more money in circulation, has affected various fields of economic activity not simultaneously but successively. Similarly, when the excise duties are increased by the government, they result in increase in prices, which would lead to higher demand for wages.

The reliability of the forecast depends in this case on the accuracy with which time lag is estimated. Also, forecasting should not be done mechanically; due allowance should be made for the effects of the current economic conditions and other special factors operating at the time, and the forecast should be modified in the light of these special factors.

11.8.2 Action and Reaction Theory

This theory is based on the following assumptions:

1. Every action has a reaction after some time.
2. The magnitude of reaction is based on the magnitude of the original action.

Thus, if the price of rice has gone up above a certain level in a certain period, there is a likelihood that after some time it will go down below the normal level.

Thus, according to this theory, a certain level of business activity is normal – subnormal or abnormal conditions cannot remain so for ever; there is bound to be a reaction to them. Thus, we find four phases of a business cycle:

1. Prosperity
2. Decline
3. Depression
4. Improvement

Since the theory regards a certain level of business activity as normal, the normal level must be very carefully estimated at the time of making forecasts. However, in practice, it is really difficult to decide precisely what constitutes 'normality'.

11.8.3 Economic Rhythm Theory

The basic assumption of this theory is that 'history repeats itself', and hence, the exponents of this theory believe that economic phenomena behave in a rhythmic order.

Cycles of nearly the same intensity and duration tend to recur. Thus, the available historical data has to be analysed into its component parts, and different types of fluctuations influencing it have to be segregated. A trend is then obtained, which will represent a long-term tendency to growth or decline. This trend line is projected a number of years into the future, either by the freehand method or by the mathematical method. This is done on the assumption that the trend line represents the normal growth or decline of the series.

Limitations of economic rhythm theory:

1. Business cycles are not strictly periodic, and the statistical extrapolation of cycles is not very satisfactory.

2. An error can be committed by an increase in either amplitude or duration, whereas the business person is primarily interested in predicting the turning points of a cycle.

11.8.4 Specific Historical Analogy

This theory is based on the assumption that not all business cycles are uniform in amplitude or duration, and as such, the use of history is made not by projecting any fancied economic rhythm into the future, but by selecting some specific previous situation that has many of the earmarks of the present and concluding that what happened in that previous situation will happen in the present one also.

A time series relating to the data in question is thoroughly scrutinised, and from it a period is selected in which conditions were similar to those prevailing at the time of making the forecasts. The course which events took in the past under similar circumstances is then studied, which gives an idea of the likely course that the phenomenon in question will follow.

For example, after World War II, many persons forecast a depression, because World War I had been followed by a depression.

11.8.5 Cross-Section Analysis

This theory is based on the knowledge and interpretation of current forces rather than projection of past trends. The theory assumes that no two cycles are alike, but that like causes always produce like results.

All the factors bearing on a given situation are assembled, and relying on knowledge of economic processes, the forecaster concludes whether the situation is favourable or not. Immediate recognition is given to the fact that business conditions are shaped by simultaneous inflationary and deflationary forces. Predominance of inflationary forces results in booms, whereas predominance of deflationary forces leads to depression.

The forecaster who uses this method

1. Itemises inflationary forces
2. Enumerates stable forces
3. Itemises deflationary forces

Weights are assigned to these forces on the basis of judgement.

The dominant forces change from time to time. Factors that may be considered include

1. Technological development
2. Supply–demand relationship
3. Governmental policies
4. Business people's expectations

Several organisations regularly conduct surveys of executive opinions concerning future trends of general business conditions and selected series of business data.

11.9 Forecasting Agencies

Business forecasting has become a specialised job, and in many advanced countries, such as the United States and the United Kingdom, there are forecasting agencies, which employ expert statisticians to analyse and interpret statistical material and publish results.

However, in India, such agencies, though badly needed, have not yet come into being. The important forecasting agencies of the United States and the United Kingdom are

- United States: Harvard Committee of Economic Research
  - Brookmire Economic Services
- United Kingdom: London and Cambridge Economic Services
  - Economists' Organisation

Most business people, even today, depend on their intuition and judgement rather than on scientific analysis of facts for deciding on a future course of action, and, thus, scientific forecasting is practically absent. However, it is gratifying to note that a large number of governmental and non-governmental agencies are engaged in the task of collection, analysis and interpretation of data affecting the various aspects of business and the economy. If the business person supplements their judgement with these important indicators, a much better forecast will be possible.

11.10 Caution While Using Forecasting Techniques

Forecasting business conditions is a complex task, which cannot be accomplished with exactness. The economic, social and political forces that shape the future are many and varied; their relative importance changes almost constantly.

It is obvious, therefore, that statistical methods cannot claim to be able to make the uncertain future certain. It does not follow from this disclaimer that statistical methods have nothing to contribute to business forecasting. The choice is not between forecasting and not forecasting, because the lack of a forecast implies a dangerous type of forecast; the mere warning of a possibility of a change is better than no warning at all. As is widely said, 'Forewarned is forearmed'.

Also, it should be remembered that forecasts are not made just for the sake of forecasting; that is, they are not ends in themselves. Forecasts are made to assist management in determining a strategy and alternative strategies.

11.11 Advantages of Forecasting

1. Help to predict the future

   Forecasting does not provide you with a crystal ball to see exactly what will happen to the market and your company over the coming years, but it will help to

give you a general idea. This will provide you with a sense of direction, which will allow your company to get the most out of the marketplace.

2. Learn from the past

   Looking at what has happened in the past can help companies predict what will happen in the future, thus making the company stronger and more profitable.

3. Save on staffing costs

   Forecasting allows companies to predict how much product will need to be produced to meet customer demand. Companies can use this data to accurately determine how many employees they will need to have on hand to meet the required level of production.

4. Remain competitive

   A business that does not use forecasting techniques will likely succumb to its competition in a short time. Having a general idea of what sales to expect in the following period is very important. This will help a company prepare to meet customer demand; otherwise, customers will look to fulfil their need elsewhere.

5. Prepare for new business

   By forecasting demand, a company can see whether an increase in sales is likely imminent. This will allow the company to prepare for this increase in business by providing extra staff or production facilities to meet this new level of demand.

11.12 Disadvantages of Forecasting

1. Basis of forecasting

   The most serious limitations of forecasting arise out of the basis used for making forecasts. Top executives should always bear in mind that the bases of forecasting are assumptions, approximations and average conditions. Management may become so concerned with the mechanism of the forecasting system that it fails to question its logic.

2. Reliability of past data

   Forecasting is made on the basis of past data and current events. Although past events/data are analysed as a guide to the future, a question is raised as to the accuracy as well as the usefulness of these recorded events.

3. Time and cost factor

   These factors suggest the degree to which an organisation will go for formal forecasting. The information and data required for forecasting may be in a highly disorganised form; some may be in qualitative form. The collection of information and conversion of qualitative data into quantitative data involve a lot of time and money.

11.13 Summary

The chapter has emphasised the importance of forecasting in all planning decisions. There are methods, such as moving averages or exponential smoothing, that are based on averaging past data. Regression is also used in the estimation of parameters of causal or econometric models.

REVIEW QUESTIONS

1. What do you mean by forecasting?
2. State the methods of forecasting.
3. Why is forecasting so important in business? Identify applications of forecasting for
   i. Long-term decisions
   ii. Medium-term decisions
   iii. Short-term decisions
4. How would you conduct an opinion poll to determine student reading habits and preferences towards daily newspapers and weekly magazines?
5. What is meant by business forecasting? Give a critical estimate of the methods used in business forecasting.
6. Distinguish between the 'historical analysis of past conditions' and the 'cross-section analysis of current events' as methods of business forecasting.
7. What is business forecasting? What are the assumptions on which business forecasts are made? Describe the techniques of forecasting that are commonly employed by big business houses.

12

Correlation Analysis

12.1 Introduction

We study one variable at a time with certain characteristics of the given data. These characteristics are measures of central tendency, measures of dispersion, skewness, kurtosis and so on. When we study the blood pressure of a patient, it may depend on the age of the patient, or it may depend on the patient's particular disease. The sales of a cosmetic may depend on advertisements.

Examples of correlation problems are found in the study of the relationship between IQ and aggregate percentage marks obtained by a person in the Staff Selection Commission (SSC) examination, between blood pressure and metabolism, or between the height and weight of individuals. In these examples, both variables are observed as they naturally occur, since neither variable is fixed at a predetermined level.

12.2 Correlation

So far, we have studied problems relating to one variable only. In practice, we come across a large number of problems involving the use of two or more than two variables.

If two quantities vary in such a way that movements in one are accompanied by the movements in the other, these quantities are correlated. When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.

For example, there exists some relationship between age of husband and age of wife, and between price of commodity and amount demanded. In such cases, we require to study two or more variables at a time and the relationship between these two or more variables.

The statistical tool with the help of which these relationships between two or more than two variables are studied is called *correlation*. The correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables.

A relationship exists between two variables when one depends on the other.

In some correlations, the two variables depend on each other and affect each other. A high degree of correlation between two variables may be due to a third variable acting on both.

12.2.1 Correlation Coefficient

The measure of correlation called the *correlation coefficient* summarises in one number the direction and degree of correlation.

12.2.2 Correlation Analysis

Correlation analysis is a technique that measures the nature, degree and extent of relationship existing between two or more variables.

The correlation coefficient may also be said to be a measure of *covariance* between two series.

12.2.3 Bi-Variate Correlation

The study of correlation between two variables is *bi-variate* correlation. The study of correlation between three or more variables is *multivariate* correlation.

12.2.3.1 Bi-Variate Data

To study the correlation between two variables, we require pairs of observations. Such data is called *bi-variate data*.

The detection and analysis of correlation between two statistical variables requires a relationship of some sort which associates the observations in pairs, one of each pair being a value of each of the two variables.

12.2.4 Correlation: Cause and Effect Relation

1. A high degree of correlation between two variables may be proved mathematically even though they are not related to each other. Therefore, the correlation between the two may be due to chance or coincidence (Table 12.1).
2. A relationship exists between two variables when one depends on the other. In this case, one variable is dependent, and the other may be totally independent, that is, not influenced by the other.

   For example, during holidays, there will be a heavy rush to pilgrim centres. But pilgrim centres cannot cause holidays.
3. In some correlations, the two variables depend on each other they affect each other; for example, income and expenditure.
4. A high degree of correlation between two variables may be due to a third variable acting on both.

TABLE 12.1

Cause and Effect Relation of Correlation

Correlation between	Effect of
x and y	Chance or coincidence
x and y	Because of x but not y
x and y	Both x and y
x and y	p or q

For example, during holidays, there is more rush to tourist centres due to the travel concessions declared by the government as part of the tourism development programme.

12.2.5 Significance of Correlation

The coefficient of correlation is one of the most widely used and also one of the most widely abused statistical measures. It is abused in the sense that one sometimes overlooks the fact that correlation measures nothing but the strength of a linear relationship and that it does not necessarily imply a cause–effect relationship. In business and industry, we come across several situations in which the relation between two variables will have a decisive influence on policy formers.

1. Increasing the supply of a product depends on the market demand for the same product.
2. The amount to be spent on advertising is determined on the basis of past sales and estimations of future sales.
3. Correlation techniques help in reducing mass data, which is in the form of pairs of variables, into a simple figure that facilitates a comfortable interpretation of critical analysis of the total data. For this reason, correlation is widely used in business and economic analysis.
4. For socio-economic studies and for analysis of economic problems, correlation is extensively applied. The study of the increasing number of unemployed people and increasing number of crimes, and so on, helps in understanding socio-economic problems.
5. The nature of relationship between two variables gives an idea about their behaviour and the extent of the influence of one variable on the behaviour of the other. To determine price policy, it is necessary to understand market behaviour and consumers' purchasing power.
6. With the help of correlation analysis, we can measure in one number the degree of relationship existing between the variables.
7. In business, correlation analysis enables the executive to estimate costs, sales, prices and other variables on the basis of some other series with which these costs, sales or prices may be functionally related.
8. Progressive development in the method of science and philosophy has been characterised by an increase in the knowledge of relationships or correlations.

12.2.6 Limitations of Correlation

1. In the absence of a relationship between variables, the measure becomes useless, and it will be a nonsense correlation.

   For example, correlation between number of failures in the common entrance examination and number of trees in the area will be an illogical correlation.
2. Coefficient of correlation is abused in the sense that one sometimes overlooks the fact that correlation measures nothing but the strength of linear relationships and that it does not necessarily imply a cause-effect relationship.
3. Positive correlation may exist in a small sample, but this may not be found universally.

12.2.7 Properties of Correlation

1. The degree of relationship is expressed by a coefficient, which ranges from –1 to +1.
2. The direction of change is indicated by a negative or positive sign.

12.3 Types of Relationships

Figure 12.1 shows additional relations that may exist between two variables:

1. Positive or negative
2. Simple, partial and multiple
3. Linear and non-linear

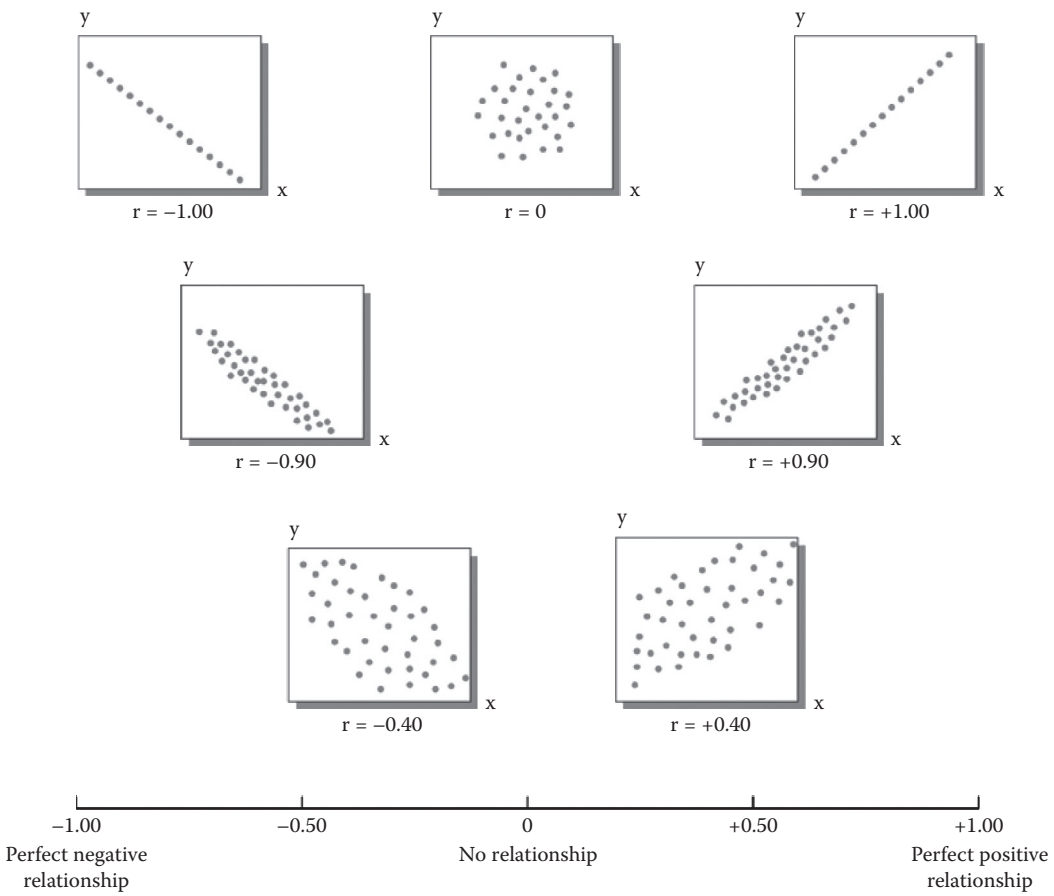


FIGURE 12.1
Types of relationships.

12.3.1 Positive or Negative

If as one variable is increasing, the other, on average, is also increasing, or if as one variable is decreasing, the other, on average, is also decreasing, the correlation is said to be positive; otherwise, it is negative.

12.3.2 Simple, Partial and Multiple

When only two variables are studied, it is a problem of simple correlation. When three or more variables are studied, it is a problem of either multiple or partial correlation. In multiple correlation, three or more variables are studied simultaneously.

1. Linear and non-linear

   If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, the correlation is said to be linear; otherwise, it is non-linear or curvilinear.
2. Simple correlation

   The relationship is confined to two variables; for example, yield of a crop and use of chemical fertiliser.
3. Multiple correlation

   The relationship is between more than two variables; for example, yield of a crop, use of irrigation facilities and amount of rainfall.
4. Partial correlation

   Here, an analyst recognises more than two variables but considers only two variables, keeping the others constant; for example, the relation between the yield of a crop and the use of irrigation facilities, keeping the variable rainfall constant.
5. Total correlation

   Total correlation is based on all the relevant variables, which, however, is normally not feasible; for example, considering all three variables – yield of a crop, use of chemical fertilisers and amount of rainfall – and finding the correlation.

12.3.3 Linear and Non-Linear or Curvilinear Correlation

1. Linear correlation

   Here, the amount of change in one variable tends to bear a constant ratio to the amount of change in the other. The graph of the variables having a linear relationship will form a straight line.
2. Non-linear correlation

   Here, the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

   For example, if the rainfall is doubled, it is not necessary that the yield also doubles.
3. Disadvantage

   The techniques of analysis for measuring non-linear relationships are more difficult and complicated.
4. Inverse linear relation

The nature of the relationship is linear. In this case, the line slopes downward. Therefore, similar values of Y are associated with larger values of X. The relationship is called *inverse linear relation*.

5. Inverse non-linear relation

   The nature of the relationship is not linear. It is a curvilinear relation. Therefore, smaller values of Y are associated with larger values of X. This relationship is inverse *non-linear relation*.

6. Direct non-linear relation

   This is a curvilinear relation. Therefore, values of Y are associated with larger values of X. Therefore, this relation is direct and curvilinear.

   In most of the practical cases, we find a non-linear relationship between the variables. However, since techniques of analysis for measuring non-linear correlation are far more complicated than those for linear correlation, we generally make an assumption that the relationship between the variables is of the linear type.

7. Perfectly positive correlation

   If the trend of the points or dots lies on a straight line, moving from the bottom left-hand corner to the top right-hand corner of the graph, the correlation is said to be *perfectly positive*.

8. Negative or inverse correlation

   An increase or decrease in the value of one variable leads to a decrease or increase in the value of the other variable; that is, the movement of the variable is in the opposite direction; for example, increase in price of a commodity and decrease in amount demanded.

9. Positive or direct correlation

   When the increase or decrease in the value of one variable leads to an increase or decrease in the value of the other variable; that is, the movement of the variable is in the same direction; for example, supply and demand of a commodity.

10. Low degree of positive correlation

    If the points are widely scattered and rising from the bottom left-hand corner to the top right-hand corner, the correlation is of lower degree but positive.

11. Low degree of negative correlation

    If the points are scattered, declining from the top left-hand corner to the bottom right-hand corner, the correlation is of lower degree but negative.

12.4 Difference between Positive and Negative Correlation

The difference between positive and negative correlation is summarised in Table 12.2.

12.5 Distinction between Simple, Partial and Multiple Correlation

The distinction between simple, partial and multiple correlation is based on the number of variables studied.

TABLE 12.2

Positive vs. Negative Correlation

Positive Correlation	Negative Correlation
Both the variables are varying in the same direction.	The variables are varying in opposite directions.
As one variable is increasing, the other, on average, is also increasing.	As one variable is increasing, the other is decreasing.
As one variable is decreasing, the other, on average, is also decreasing.	As one variable is decreasing, the other is increasing.

1. When only two variables are studied, it is a simple correlation.
2. In multiple correlations, three or more variables are studied simultaneously.
3. In partial correlation, we recognise more than two variables but consider only two variables to be influencing each other, the effect of other influencing variables being kept constant.

12.5.1 No Correlation

If the plotted points lie on a straight line parallel to the X-axis, or in a haphazard manner, this shows the absence of any relationship between the variables. The graph shows no correlation between variables X and Y.

12.6 Lag and Lead in Correlation

The study of lag and lead is of special significance while studying economic and business series. In the correlation of time series, the investigator may find that there is a time gap before a cause and effect relationship is established.

For example, the supply of a commodity may increase today, but this may not have an immediate effect on prices. It may take a few days or even months for prices to adjust to the increased supply.

The difference in the period before a cause and effect relationship is established is called *lag*.

While computing correlation, this time gap must be considered; otherwise, fallacious conclusions may be drawn. The pairing of items is adjusted according to the time lag.

12.7 Methods of Studying Correlation

1. Scatter diagram
2. Karl Pearson's coefficient of correlation
3. Spearman's rank correlation coefficient
4. Method of least squares

12.7.1 Scatter Diagram Method or Dotogram or Scatter Gram or Dot Chart

The simplest device for ascertaining whether two variables are related is to prepare a dot chart called a scatter diagram. The method is so called because it indicates the scatter of the various points. When this method is used, the given data are plotted on graph paper in the form of dots, that is, for each pair of X and Y values, we put a dot and thus obtain as many points as the number of observations. By looking at the scatter of the various points, we can form an idea as to whether the variables are related or not. The more the plotted points "scatter" over a chart, the less relationship there is between the two variables.

A close observation of these points shows the trend of correlation between variables. If the points show an upward trend, it is positive correlation.

If the tendency is downward, it is negative correlation. The more the plotted points 'scatter' over a chart, the less relationship there is between the two variables.

1. If all the points lie on a line from the lower left-hand corner to the upper right-hand corner, correlation is said to be positive, that is, $r = +1$.
2. If all the points are lying on a straight line from the upper left-hand corner to the lower right-hand corner of the diagram, correlation is perfectly negative ($r = -1$).
3. If points fall in a narrow band, there is a high degree of correlation between the variables.

For example, suppose we have two variables, x and y, and there are n pairs of values (x_1, y_1), (x_2, y_2), …, (x_n, y_n). Then, these values are plotted on the x-axis and y-axis in the xy plane.

Generally, the independent variable is plotted along the horizontal or x-axis, and the dependent variable along the vertical or y-axis. The picture obtained is called a *scatter diagram*.

1. Interpretation of scatter diagram
   i. If the pattern of points (or dots) on a scatter diagram is more scattered, the degree of relationship between the two variables is lower.
   ii. The nearer the points come to the line, the higher the degree of relationship.
2. Uses
   i. It is a simple and non-mathematical method of studying correlation between the variables.
   ii. As such, it can be easily understood, and a rough idea can very quickly be formed as to whether or not the variables are related.
   iii. It gives an idea about direction and closeness of relation.
   iv. It shows whether the relation is linear or non-linear.
   v. It is not influenced by the size of extreme items.
   vi. It is the first step in investigating the relationship between two variables.
   vii. It enables correlation to be measured diagrammatically.
3. Limitations
   i. Fails to give accurate measures of correlation. It gives only a vague idea. Therefore, it cannot be used for further analysis or to make decisions.

ii. Gives an idea about the direction of correlation and also whether it is high or low, but we cannot establish the exact degree of correlation between the variables. Therefore, it cannot be used for further analysis or to make decisions.

12.7.2 Karl Pearson's Coefficient of Correlation or Pearsonian Coefficient of Correlation

Karl Pearson (1867–1936), a British biometrician, developed a mathematical method to calculate the coefficient of correlation. The coefficient of correlation is denoted by *r*, which is a measure of the degree of linear relationship or correlation between two variables, say, x and y, one of which happens to be an independent variable and the other a dependent variable. It is universally used for describing the degree of correlation between two series.

The Pearsonian correlation coefficient between two variables x and y is the ratio of the covariance between x and y (written as COV (X, Y)) to the product of standard deviations of x and y and is expressed as

$$r \text{ (or) } r_{xy} = \frac{\text{COV}(X,Y)}{\sigma_x \sigma_y}$$

$$= \frac{\frac{\Sigma xy}{N}}{\sigma_x \sigma_y}$$

$$= \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

where:

$$x = X - \bar{X}$$

$$y = Y - \bar{Y}$$

$\sum xy$ = Sum of the product of deviations of x and y series calculated with reference to their arithmetic means

σ_x = Standard deviation of series X

$$= \sqrt{\frac{\Sigma X^2}{N}}$$

σ_y = Standard deviation of series Y

$$= \sqrt{\frac{\Sigma Y^2}{N}}$$

This method is to be applied only where the deviations of items are taken from actual means and not from assumed means.

Interpretation of Karl Pearson's method:

The value of the coefficient of correlation shall always lie between ±1, that is, $-1 < r < +1$.

1. When the value of $r = +1$, the correlation between the variables is perfect and positive.
2. When $r = -1$, the correlation is said to be perfect and negative.
3. When the value of $r = 0$, there exists no correlation between the variables.

Procedure for computing Karl Pearson's coefficient of correlation:

1. Calculate the means of the two series X and Y.
2. Calculate the deviations x, y in the two series from their respective means.
3. Square each deviation of x and y; then obtain the sum of the squared deviations, that is, $\sum x^2$ and $\sum y^2$. These are the sum of the squared deviations of X and Y series.
4. Multiply each deviation under x with each deviation under y and obtain the product xy. Then obtain the sum of the products of x and y, that is, $\sum xy$.
5. Substitute the values in the formula

$$r = \frac{\Sigma xy}{N\sigma x\sigma y}$$

Hence, $x = (X - \bar{X}) \quad y = (Y - \bar{Y})$.

This can be applied only where the deviations of items are taken from actual means and not from assumed means. The coefficient of correlation describes not only the magnitude of correlation but also its direction. The formula in Point 5 can be transformed to

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times y^2}}$$

where $x = (X - \bar{X}) \quad y = (Y - \bar{Y})$.

Exercise

Calculate coefficient of correlation from the data in Table 12.3.

TABLE 12.3

Data for Variables X and Y

X	100	200	300	400	500	600	700
Y	30	50	60	80	100	110	130

TABLE 12.4

Calculation of Coefficient of Correlation

		$\left(\frac{X-\bar{X}}{100}\right)$				$\left(\frac{Y-\bar{Y}}{10}\right)$		
X	$\mathbf{(X-\bar{X})}$	**x**	$\mathbf{x^2}$	**Y**	$\mathbf{(Y-\bar{Y})}$	**y**	$\mathbf{y^2}$	**xy**
100	−300	−3	9	30	−50	−5	25	15
200	−200	−2	4	50	−30	−3	9	6
300	−100	−1	1	60	−20	−2	4	2
400	0	0	0	80	0	0	0	0
500	100	1	1	100	20	2	4	2
600	200	2	4	110	30	3	9	6
700	300	3	9	130	50	5	25	15
$\mathbf{\Sigma X = 2800}$		$\mathbf{\Sigma x = 0}$	$\mathbf{\Sigma x^2 = 28}$	$\mathbf{\Sigma Y = 560}$		$\mathbf{\Sigma y = 0}$	$\mathbf{\Sigma y^2 = 76}$	$\mathbf{\Sigma xy = 46}$

Solution

See Table 12.4.

$$\bar{X} = \frac{\Sigma X}{N} = \frac{2800}{7} = 400$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{560}{7} = 80$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$\therefore r = \frac{46}{\sqrt{28 \times 76}}$$

$$\therefore r = 0.997$$

Exercise

Find the coefficient of correlation between sales and expenses of the 10 firms shown in Table 12.5. Interpret your result.

Solution

See Table 12.6.

$$\bar{X} = \frac{\Sigma X}{N} = \frac{580}{10} = 58$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{140}{10} = 14$$

TABLE 12.5

Sales and Expenses of 10 Firms

Firm	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sales	50	50	55	60	65	65	65	60	60	50
Expenses	11	13	14	16	16	15	15	14	13	13

TABLE 12.6

Calculation of Karl Pearson's Coefficient of Correlation

X	$\mathbf{X = X - \bar{X} = X - 58}$	$\mathbf{x^2}$	**Y**	$\mathbf{Y = Y - \bar{Y} = Y - 14}$	$\mathbf{y^2}$	**xy**
50	−8	64	11	−3	9	24
50	−8	64	13	−1	1	8
55	−3	9	14	0	0	0
60	2	4	16	2	4	4
65	7	49	16	2	4	14
65	7	49	15	1	1	7
65	7	49	15	1	1	7
60	2	4	14	0	0	0
60	2	4	13	−1	1	−2
50	−8	64	13	−1	1	8
$\mathbf{\Sigma X = 580}$		$\mathbf{\Sigma x^2 = 360}$	$\mathbf{\Sigma Y = 140}$		$\mathbf{\Sigma y^2 = 22}$	$\mathbf{\Sigma xy = 70}$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{70}{\sqrt{360 \times 22}}$$

$$= 0.7866$$

Thus, sales of the firm and its expenditure are positively correlated. Hence, we conclude that when sales increase, the expenditure of the firm also increases.

Exercise

Calculate coefficient of correlation between X and Y (Table 12.7).

Solution

Since the means of X and Y are whole numbers, we apply the actual mean method (Table 12.8).

TABLE 12.7

Data for X and Y Series

X	23	27	28	28	29	30	31	33	35	36
Y	18	20	22	27	21	29	27	29	28	29

TABLE 12.8

Calculation of Karl Pearson's Correlation Coefficient

X	x = (X – 30)	x^2	Y	y = (Y – 25)	y^2	xy
23	–7	49	18	–7	49	49
27	–3	9	20	–5	25	15
28	–2	4	22	–3	9	6
28	–2	4	27	2	4	–4
29	–1	1	21	–4	16	4
30	0	0	29	4	16	0
31	1	1	27	2	4	2
33	3	9	29	4	16	12
35	5	25	28	3	9	15
36	6	36	29	4	16	24
ΣX = 300	Σx = 0	$\Sigma x^2 = 138$	ΣY = 250	Σy = 0	$\Sigma y^2 = 164$	Σxy = 123

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{123}{\sqrt{138 \times 164}} = \frac{123}{150.438} = 0.818$$

12.7.3 Karl Pearson's Correlation Coefficient (Actual Mean Method)

This is also called the *product moment correlation coefficient*. This gives the degree of correlation along with whether it is positive or negative. This is denoted by *r*. It is defined as

$$r = \frac{\text{co variance between x and y}}{\sigma_x \times \sigma_y}$$

where covariance between x and y is

$$\text{Cov}\ (x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

$$\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}, \sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}$$

To simplify the calculation of the correlation coefficient, we use the formula

$$r = \frac{\Sigma dx\ dy - \left(\frac{\Sigma dx\ \Sigma dy}{n}\right)}{\sqrt{\Sigma(dx)^2 - \frac{\left[\Sigma(dx)\right]^2}{n}}\sqrt{\Sigma(dy)^2 - \frac{\left[\Sigma(dy)\right]^2}{n}}}$$

where:

$dx = x - A$

$dy = y - B$

B = assumed mean for y

The same formula can be represented by

$$r = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{\left(n\Sigma x^2 - (\Sigma x)^2\right)\left(n\Sigma y^2 - (\Sigma y)^2\right)}}$$

Note that

1. The value of the correlation coefficient always lies between –1 and +1.
2. If the value is positive, we say that there is a positive correlation between variables.
3. If the value is negative, we say that there is a negative correlation.
4. If the value is near 0, we say that no correlation exists.

Exercise

Calculate the correlation coefficient between x and y from the data in Table 12.9.

Solution

Calculation of correlation coefficient (by actual mean method): see Table 12.10.

$$\bar{X} = \frac{\Sigma X}{N} = \frac{198}{5} = 39.6 \qquad \therefore \text{Let } A = 40$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{260}{5} = 52 \qquad \therefore \text{Let } B = 50$$

TABLE 12.9

Data of Variables x and y

x	38	41	30	42	47
y	55	60	45	50	50

TABLE 12.10

Calculation of Correlation Coefficient

x_i	y_i	dx = x – A (A = 40)	dy = y – B (B = 50)	dx dy	$(dx)^2$	$(dy)^2$
38	55	–2	5	–10	4	25
41	60	1	10	10	1	100
30	45	–10	–5	50	100	25
42	50	2	0	0	4	0
47	50	7	0	0	49	0
$\Sigma X_i = 198$	$\Sigma X_i = 260$	$\Sigma dx = -2$	$\Sigma dy = 10$	$\Sigma dx\,dy = 50$	$\Sigma(dx)^2 = 158$	$\Sigma(dy)^2 = 150$

$$r = \frac{\Sigma dx\, dy - \left(\frac{\Sigma dx\, \Sigma dy}{n}\right)}{\sqrt{\Sigma(dx)^2 - \frac{\left[\Sigma(dx)\right]^2}{n}}\sqrt{\Sigma(dy)^2 - \frac{\left[\Sigma(dy)\right]^2}{n}}}$$

$$r = \frac{50 - \left(\frac{-2 \times 10}{5}\right)}{\sqrt{158 - \frac{(-2)^2}{5}}\sqrt{150 - \frac{(10)^2}{5}}} = \frac{50 - (-4)}{\sqrt{158 - 0.8}\sqrt{150 - 20}}$$

$$= \frac{54}{\sqrt{157.2}\sqrt{130}} = 0.3777$$

Note:
We have taken assumed means for x and y nearest to the actual means. We may take any different values also.

Exercise

Table 12.11 gives the distribution of the total population and those in the population who are wholly or partially blind. Find whether there is any relation between age and blindness.

Solution

Calculate the number of blind people out of a common base figure.

∴ No comparison can be made, as it is impossible to correlate the total number of persons with the number of blind persons (Table 12.12).

$$r = \frac{\Sigma dx\, dy - \frac{\Sigma dx\, \Sigma dy}{n}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{n}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{n}}}$$

TABLE 12.11

Data of Total Population and of Wholly or Partially Blind Persons

Age (Years)	**0–10**	**10–20**	**20–30**	**30–40**	**40–50**	**50–60**	**60–70**	**70–80**
Number of persons (thousands)	100	60	40	36	24	11	6	3
Number of blind persons	45	40	40	40	36	22	18	15

TABLE 12.12

Calculation of Correlation Coefficient

Age (Years)	Mid Value (x)	$dx = x - 35/10$	dx^2	No. Blind per 100,000 (y)	$dy = y - \bar{y} = y - 125$	dy^2	$dx\ dy$
0–10	5	−3	9	45/100,000 ×100,000 = 45	−140	19,600	−420
10–20	15	−2	4	40/60,000 ×100,000 = 67	−118	13,924	−236
20–30	25	−1	1	40/40,000 ×100,000 = 100	−85	7,225	−85
30–40	35	0	0	40/36,000 ×100,000 = 111	−74	5,476	0
40–50	45	1	1	36/24,000 ×100,000 = 150	−35	1,225	−35
50–60	55	2	4	22/11,000 ×100,000 = 200	15	225	30
60–70	65	3	9	18/6,000 ×100,000 = 300	115	13,225	345
70–80	75	4	16	15/3,000 ×100,000 = 500	315	99,225	1260
Total n = 8		**$\sum dx = 4$**	**$\sum dx^2 = 44$**		**$\sum dy = -7$**	**$\sum dy^2 = 160125$**	**$\sum dx\ dy = 2341$**

$$r = \frac{2341 - \frac{(4)(-7)}{8}}{\sqrt{44 - \frac{(4)^2}{8}}\sqrt{160125 - \frac{(-7)^2}{8}}}$$

$$r = 0.9$$

Advantages of the Pearsonian coefficient:

1. It is the most popular method.
2. It summarises, in one value, the degree of correlation and also the direction (positive or negative) of the correlation.

Limitations of the Pearsonian coefficient:

1. The correlation coefficient always assumes a linear relationship, regardless of whether or not that assumption is true.
2. Interpreting the value of *r* is very difficult, and thus, great care should be taken when interpreting the value of the correlation coefficient.
3. The value of the correlation coefficient is affected by extreme values.
4. The method is time consuming.

12.7.4 Correlation Coefficient when Deviations Are Taken from an Assumed Mean

Sometimes, the actual means of x and y series obtained are in fractions. In such cases, the calculation of coefficient of correlation by the previous method becomes complicated, as it involves too many calculations and is time consuming. To avoid this, the analyst can make use of an assumed mean method, where the deviations are taken from an assumed mean.

Formula for the assumed mean method:

$$r = \frac{N\,\Sigma dx\,dy - \Sigma dx\,\Sigma dy}{\sqrt{N\,\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\,\Sigma dy^2 - (\Sigma dy)^2}}$$

where:

dx = deviation of x series from the assumed mean
dy = deviation of y series from the assumed mean

12.8 Correlation of Bi-Variate Grouped Data

If the given data is grouped data, the formula for finding the correlation coefficient is given by

$$r = \frac{N \Sigma f\, dx\, dy - \Sigma\, f\, dx\, \Sigma\, f\, dy}{\sqrt{N \Sigma f\, dx^2 - (\Sigma f\, dx)^2}\sqrt{N \Sigma f\, dy^2 - (\Sigma f\, dy)^2}}$$

Note:
This formula is the same as the formula for calculating r using an assumed mean. The only difference is that the deviations are multiplied by the respective frequencies.

12.9 Caveat

We must be careful when interpreting the meaning of association. Although two variables may be associated, this association does not imply that variation in the independent variable is the cause of variation in the dependent variable.

For example, age and income are usually related. However, a further increase in age does not cause an increase in income. We can determine the presence or absence of association through statistics. However, there is no basis for concluding that a cause–effect relation exists between these variables.

12.10 Coefficient of Determination

This is given by r^2, that is, the square of the correlation coefficient. It explains to what extent the variation of the dependent variable is expressed by independent variables. A high value of r^2 shows a good relationship between two variables. If $r = 1$ and $r^2 = 1$, there is a perfect relationship between variables. The value of the coefficient of correlation between two variables is interpreted using the square of coefficient of determination.

r^2 is defined as the ratio of the explained variance to the total variance.

$$r^2 = \frac{\text{expected variance}}{\text{total variance}}$$

r^2 is to be preferred in presenting the results of correlation analysis. The coefficient of determination is a much more useful measure of the linear covariations of two variables. If the value of $r = 0.9$, r^2 is 0.81, meaning that 81% of the variation in the dependent variable has been explained by the independent variable.

The maximum value of r^2 is unity, because it is possible to explain all the variation in y, but it is not possible to explain more than all of it.

Exercise

$r_{12} = 0.6$, $r_{13} = -0.56$, $r_{23} = 0.80$. Find $r_{13.2}$.

Solution

$$r_{13.2} = \frac{r_{13} - r_{12} \times r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

$$r_{13.2} = \frac{-0.5 - 0.76 \times 0.8}{\sqrt{1 - (0.76)^2}\sqrt{1 - (0.80)^2}} = -2.99$$

12.11 Spearman's Rank Correlation Coefficient

Charles Edward Spearman (1904), a British psychologist, developed a method to find the coefficient of correlation by assigning a rank to each value of a variable or group.

This method is useful when a quantitative measure of certain factors involving quantity or quality cannot be fixed.

For example:

1. Evaluation of leadership skills.
2. Ability of an individual of the group.
3. Judgement of female beauty.

Here, the rank correlation coefficient is applied to a set of ordinal rank numbers, with 1 for the individual ranked first in quantity or quality, and so on, to N for the individual ranked last in a group of N individuals or N pairs of individuals.

1. Spearman's rank correlation coefficient is given by

$$R = 1 - \frac{6 \sum D^2}{N\left(N^2 - 1\right)}$$

where:
R = rank coefficient of correlation
D = difference of ranks between paired items in two series
N = total number of observations

2. Interpretation of Spearman's rank correlation coefficient
   i. The rank correlation coefficient lies between –1 and +1, that is, $-1 \leq R \leq +1$.
   ii. When R = +1, it can be inferred that there is complete agreement in the order of the ranks, and the ranks are in the same direction.
   iii. When R = –1, it can be inferred that there is complete agreement in the order of the ranks, and the ranks are in the opposite direction.
3. Procedure for solving the problem where actual ranks are given
   i. Calculate the differences (D) of the two ranks, that is, $(R_1 - R_2)$
   ii. Square the difference and calculate the sum of differences, that is, $\sum D^2$
   iii. Substitute the values in the formula.

4. Problems where ranks are not given
   i. When the actual data does not contain the actual ranks in it, then it is necessary to assign the ranks to the given data.
   ii. This can be done by assigning Rank 1 to the highest value or the lowest value of the series.
5. Equal ranks or tie in ranks

   In such cases, an average rank should be assigned to each individual or each entry.

   For example:

   If two individuals or entries are at the same place or same level and, say, a rank 4 is to be assigned, then the average of Rank 4 and its next rank, that is, 5, is taken and is assigned to such individuals.
6. Measure of rank correlation

   Professor Charles Spearman developed the measure of rank correlation. This correlation is calculated by finding the difference between ranks. It is denoted by P.
7. Step 1: Bracket rank method

   Here, when the same rank is given to two items, the next rank will be cancelled. For example,

Marks	44	54	91	74	84	71	84	95	96	39
$\mathbf{X_i}$	9	8	3	4	6	7	4	2	1	10

8. Correlation for tied ranks (average rank method)

   Here, the next rank will be cancelled, but is distributed to the rank holders who obtain the same rank. For example,

Marks $\mathbf{X_i}$	44	54	91	74	84	71	84	95	96	39
$\mathbf{Y_j}$	9	8	3	4.5	6	7	4.5	2	1	10

Step 2

The difference between $X_i - Y_j = d$ is calculated by deducting ranks of Y_j from X_i.

Step 3

Square the difference (d^2) and find its total $\sum d^2$.

Step 4

$$P = \frac{6\left[\sum d^2 + 1/12(m^3 - m) + \ldots\right]}{N(N^2 - N)}$$

where:

P = rank correlation coefficient
M = number of items having tied values
N = number of items

Note:
If two or more items are of equal value in the series arrangement, ranks are assigned to such tied values. An adjustment is to be made for such groups. The correlation factor $(m^3-m)/12$ is to be added to the formula for each repeated value.

Exercise

Compute the ranks in the subjects shown in Table 12.13 and the coefficient of correlation of ranks.

Solution

Calculation of coefficient of correlation (Table 12.14)

$$r_k = 1 - \frac{6\Sigma D^2}{N\left(N^2-1\right)} = 1 - \frac{6\times 12}{9\left(9^2-1\right)} = 0.014$$

∴ The coefficient of rank correlation is 0.014.

Exercise

Ten competitors in a beauty contest are ranked by three judges in the order shown in Table 12.15. Use the Spearman's rank correlation coefficient to determine which pair of judges has the nearest approach to common taste in beauty.

Solution

Calculation of rank correlation (Table 12.16)

TABLE 12.13

Data on the Marks in Two Subjects of Nine Students

Marks in mathematics (x)	35	23	47	17	10	43	09	06	28
Marks in statistics (y)	30	33	45	23	08	49	12	04	31

TABLE 12.14

Calculation of Coefficient of Correlation

Maths (x)	**R_1**	**Statistics (y)**	**R_2**	**$D = R_1 - R_2$**	**D^2**
35	7	30	5	2	4
23	5	33	7	–2	4
47	9	45	8	1	1
17	4	23	4	0	0
10	3	08	2	1	1
43	8	49	9	–1	1
09	2	12	3	–1	1
06	1	04	1	0	0
28	6	31	6	0	0
N = 9		**N = 9**			**$\Sigma D^2 = 12$**

TABLE 12.15

Data on Ten Competitors Ranked by Three Judges

First Judge	1	6	5	10	3	2	4	9	7	8
Second Judge	3	5	8	4	7	10	2	1	6	9
Third Judge	6	4	9	8	1	2	3	10	5	7

TABLE 12.16

Calculation of Rank Correlation

R_1	R_2	R_3	$D^2_{12} = (R_1 - R_2)^2$	$D^2_{23} = (R_2 - R_3)^2$	$D^2_{13} = (R_1 - R_3)^2$
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	4
N = 10			**$\Sigma D^2_{12} = 200$**	**$\Sigma D^2_{23} = 214$**	**$\Sigma D^2_{13} = 60$**

$$\text{Rank Correlation}(R) = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

Rank correlation between Judges 1 and 2

$$= 1 - \frac{6\Sigma D^2_{12}}{N(N^2 - 1)} = 1 - \frac{6(200)}{10\left[(10)^2 - 1\right]} = -0.21$$

Rank correlation between Judges 2 and 3

$$= 1 - \frac{6\Sigma D^2_{23}}{N(N^2 - 1)} = 1 - \frac{6(214)}{10\left[(10)^2 - 1\right]} = -0.29$$

Rank correlation between Judges 1 and 3

$$= 1 - \frac{6\Sigma D^2_{13}}{N(N^2 - 1)} = 1 - \frac{6(60)}{10\left[(10)^2 - 1\right]} = 0.64$$

∵ Coefficient of rank correlation is positive, in the judgements of Judge 1 and Judge 3.
∴ They have the nearest approach to common taste in beauty.

12.12 Coefficient of Correlation and Probable Error

The probable error of the coefficient of correlation helps in interpreting its value. With the help of probable error, it is possible to determine the reliability of the value of the coefficient in so far as it depends on conditions of random sampling.

$$PE = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$$

where:
PE = probable error
r = coefficient of correlation
N = number of pairs of items

If the value of r is less than PE, there is no evidence of correlation, that is, the value of r is not significant. If the value of r is more than 6 times the probable error, the existence of correlation is practically certain, that is, the value of r is significant. By adding and subtracting the value of probable error from the coefficient of correlation, we get, respectively, the upper and lower limits within which coefficient of correlation in the population can be expected to lie. Symbolically,

$$\rho = r \pm P.E$$

where ρ = correlation in the population.

12.12.1 Conditions for the Use of Probable Error

Data must approximate to a normal frequency curve (bell-shaped curve). The sample is selected in an unbiased manner, and individual items must be independent. If 0.6745 is omitted from the PE formula, we get the standard error of the coefficient of correlation.

$$S.E. = \frac{1 - r^2}{\sqrt{N}}$$

Exercise

If coefficient of correlation = 0.92 and number of observations = 7, calculate the probable error.

Solution

Given
Coefficient of correlation (r) = 0.92
No. of observations N = 7
To find Probable Error (PE_r),

$$PE_r = 0.6745\frac{1-r^2}{\sqrt{N}}$$

$$= 0.6745\frac{1-(0.92)^2}{\sqrt{7}}$$

$$= 0.6745\frac{0.1536}{2.645}$$

$$= 0.0391$$

12.13 Summary

In this chapter, the concept of correlation or the association between two variables has been discussed. The correlation coefficient r may assume values between –1 and 1. The sign indicates whether the association is direct (positive) or inverse (negative). A numerical value of r equal to unity indicates perfect association, while a value of zero indicates no association. Spearman's rank correlation for data with ranks is outlined.

REVIEW QUESTIONS

1. Explain various types of correlation.
2. State the difference between positive and negative correlation.
3. What do you mean by lag and lead in correlation?
4. What are the applications of correlation?
5. What do you understand by R^2?
6. Define Karl Pearson's coefficient of correlation.
7. If coefficient of correlation = 0.92 and number of observation = 7, calculate probable error.
8. What is meant by rank correlation?
9. What is linear correlation? Mention its various types.
10. What is the coefficient of correlation? Name the different methods of computing coefficient of correlation.
11. Explain what is meant by the correlation between two variables and comment on its interpretation. Bring out the usefulness of the concept by suitable examples.
12. Define Karl Pearson's coefficient of correlation. What is it intended to measure? How would you interpret the sign and magnitude of a correlation coefficient?
13. Explain the meaning and significance of the concept of correlation.

    Discuss the role of the concept of correlation coefficient between two variables in any empirical analysis.

13. What is meant by correlation? Does correlation always signify a cause and effect relationship between the variables?

    Even a high degree of correlation does not mean that a relationship of cause and effect exists between the two correlated variables. Discuss.

14. How is correlation between two variables measured? What is meant by
    a. Positive correlation and
    b. Perfect correlation?
    c. What would you infer if ∑xy turns out to be zero?
15. Distinguish, giving suitable examples, between
    a. Positive and negative correlation
    b. Linear and non-linear correlation
    c. Simple, partial and multiple correlations.
16. Explain multiple and partial correlations.

SELF-PRACTICE PROBLEMS

1. Find the Karl Pearson's coefficient of correlation from the following index numbers and interpret it.

Wages	98	100	101	102	103	98	96	95	94	95
Cost of living	100	98	96	98	94	04	98	96	92	90

2. Calculate the Karl Pearson's coefficient of correlation between age and playing habits from the data given below. Comment on the value.

Age	**20**	**21**	**22**	**23**	**24**	**25**
Number of students	400	300	200	250	300	250
Regular players	300	200	270	76	80	34

3. Find the correlation between age and playing habits of the following students.

Age	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Number of students	240	180	140	150	120	90	80
Number of players	190	140	80	78	50	72	54

4. Find the rank correlation coefficient:

Marks in statistics	98	43	57	57	59	69	68	64	34	75	86
Marks in physics	78	89	65	48	86	98	56	76	85	94	98

5. From the following data, calculate: (i) Correlation, (ii) SD of y.
   $x = 0.854y$, $y = 0.89x$ and SD of $x = 3$.
6. The following marks were obtained by a group of students in two papers. Calculate rank coefficient of correlation.

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Marketing	32	35	49	60	43	37	43	49	10	20
Quantitative techniques	40	30	70	20	30	50	72	60	45	25

7. Calculate coefficient of correlation between age and playing habits:

Age (Years)	**15–20**	**20–25**	**25–30**	**30–35**	**35–40**	**40–45**	**45–50**	**50–55**	**55–60**
Population	1500	1200	4000	3000	2500	1000	800	500	200
Number of players	1200	1560	2280	1500	1000	300	200	50	6

8. The following table gives the population in thousands of two towns at the time of the past seven censuses. Calculate Karl Pearson's coefficient of correlation between the populations of X and Y.

Year	**1941**	**1951**	**1961**	**1971**	**1981**	**1991**	**2001**
Population X	160	169	188	154	164	205	285
Population Y	155	194	203	179	216	243	287

9. Find Karl Pearson's coefficient of correlation between sales and expenses of the following 10 firms. Interpret your result.

Firm	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sales	50	50	55	60	65	65	65	60	60	50
Expenses	11	13	14	16	16	15	15	14	13	13

10. Ten competitors in a beauty contest are ranked by three judges in the following order. Use Spearman's rank correlation coefficient to determine which pair of judges has the nearest approach to common taste in beauty.

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

11. Find Karl Pearson's coefficient of correlation between X and Y from the following data:

X series	17	18	19	19	20	20	21	21	22	23
Y series	12	16	14	11	15	19	22	16	15	20

[Ans: r = +0.62]

12. The following table gives the results of a matriculation examination held in 2015.

Age of Candidates (Years)	**Percentage of Failures**	**Age of Candidates (Years)**	**Percentage of Failures**
13–14	39.2	18–19	39.2
14–15	40.6	19–20	48.9
15–16	43.4	20–21	47.1
16–17	34.2	21–22	54.5
17–18	36.6		

Calculate Karl Pearson's coefficient of correlation and its probable error. From your result, can you definitely assert that the failure is correlated with age?

[Ans: r = +0.682, $P.E_r = 0.12$]

13. Compute the coefficient of correlation from the following data:

City	A	B	C	D	E	F	G	H
Population (in thousands)	10	20	30	40	50	60	70	80
Accident rate (per million)	32	20	24	36	40	28	48	44

[Ans: r = +0.714]

14. Calculate Karl Pearson's coefficient of correlation and its probable errors from the following data of imports and exports:

Years	Value of Imports	Value of Exports
2005–2006	903	620
2006–2007	1036	625
2007–2008	904	573
2008–2009	961	640
2009–2010	1122	642
2010–2011	1092	661
2011–2012	1131	685
2012–2013	1190	793
2014–2014	1314	816
2015–2015	1350	805

[Ans: r = +0.917; PE = 0.034]

15. Given $r_{12} = +0.80$, $r_{13} = -0.40$ and $r_{23} = -0.56$, find the values of $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$.
[Ans: $r_{12.3} = 0.759$, $r_{13.2} = 0.097$ and $r_{23.1} = -0.436$]
16. If $r_{12} = +0.86$, $r_{13} = -0.65$ and $r_{23} = 0.72$, show that $r_{12.3} = 0.743$.
17. Is it possible to get the following from a set of experimental data?
    1. $r_{23} = 0.8$, $r_{31} = 0.5$, $r_{12} = 0.6$
    2. $r_{23} = 0.7$, $r_{12} = -0.4$, $r_{12} = 0.6$

[No, since $r_{12.3} > 1$. Hint: Calculate $r_{12.3}$.]

13

Regression Analysis

13.1 Introduction

For the system under study, there may be many variables, and it is of interest to examine the effects that some variables exert (or appear to exert) on others. The exact functional relationship between variables may be too complex, but we wish to approximate to this functional relationship by some simple mathematical function, such as a straight line or a polynomial, which approximates to the true function over certain limited ranges of the variables involved.

In industry and business, large amounts of data are continuously being generated. This may be data pertaining to a company's annual production, annual sales, capacity use, turnover, profits, manpower levels, absenteeism or some other variable of direct interest to management. Accumulated data may be used to gain information about the system. For example,

1. The study of the yields of crops grown with different amounts of fertilisers
2. The length of life of certain animals exposed to different amounts of radiation
3. The hardness of plastics that are heat-treated for different periods of time

In these problems, the variation in one measurement is studied for particular levels of the other variable selected by the experimenter. Thus, the independent variables in regression analysis are not assumed to be random variables. In correlation analysis, all variables are assumed to be random variables.

13.2 Regression

13.2.1 History, Meaning and Application

Regression means 'stepping back towards the average'. It was first used by a British biometrician, Sir Francis Galton, in 1877. The technique of regression analysis is used to determine the statistical relationship between two (or more) variables and to make predictions of one variable on the basis of the other(s). After having established the fact that two variables are closely related, it would be possible for us to make use of this relationship between two variables by estimating the unknown or the dependent variable on the basis of the other variables (the known or the independent variables).

13.2.2 Regression Analysis

Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of original units of data.

The statistical tool with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable is called *regression*.

With the help of regression analysis, we are in a position to find out the average probable change in one variable given a certain amount of change in another. Regression analysis is thus designed to examine the relationship of variable y to a variable x.

13.2.3 Advantages of Regression Analysis

1. Regression analysis provides estimates of values of the regressed variables from the values of regressor variables. Estimates of the regressed variable (dependent variable) can be made using the regression line, which describes the relationship existing between the X and Y variables.
2. Regression analysis helps to obtain a measure of the error involved in using the regression line as a basis for estimations. An estimate can be good if the regression line fits the data closely. If the observations are scattered to a great extent and do not lie close to the regression line, the line will not produce accurate estimates of the dependent variable.
3. Regression analysis helps in obtaining a measure of the degree of association or correlation that exists between the two variables.

13.2.4 Features of Regression

1. The objective of regression analysis is to study the 'nature of the relationship' between the variables so that we may be able to predict the value of one on the basis of another.
2. The cause and effect relation is clearly indicated through regression analysis – one variable is taken as dependent and the other as independent.
3. The variable whose value is influenced is called the *dependent variable* and is denoted by y; the variable that exerts the influence is called the *independent variable* and is denoted by x.

13.2.5 Assumptions in Regression Analysis

1. There exists an actual relationship between the dependent and independent variables.
2. The regression analysis is used to estimate the values within the range for which it is valid and not for values outside its range.
3. The relationship that exists between the dependent and independent variables remains the same till the regression equation is calculated.
4. The dependent variable may take any random value, but the values of the independent variables are fixed quantities without error.
5. In regression, we have only one dependent variable in our estimating equation. However, we can use more than one independent variable.

13.2.6 Application of Regression

1. In economics, it is the basic technique for measuring or estimating the relationship among the economic variables that constitute the essence of economic theory and economic life.
2. Regression analysis provides estimates of values of the dependent variable from values of the independent variable.
3. With the help of the regression coefficient, we can calculate the correlation coefficient. The square of the correlation coefficient (r), called the *coefficient of determination*, measures the degree of association of correlation that exists between the two variables.

13.2.7 Limitations of Regression Analysis

1. In making estimates from a regression equation, it is important to remember that the assumption is being made that the relationship has not changed since the regression equation was computed.
2. The relationship shown by the scatter diagram may not be the same if the equation is extended beyond the values used in computing the equation. For example, there may be a close linear relationship between the yield of a crop and the amount of fertiliser applied, with the yield increasing as the amount of fertiliser is increased. It is quite likely that if the amount of fertiliser were increased indefinitely, the yield would eventually decline as too much fertiliser was applied.

13.2.8 Regression Coefficient

The regression coefficient of x and y is denoted by the symbol b_{xy} or b_1. It measures the change in x corresponding to a unit change in y. When deviations are taken from the means of x and y, the regression coefficient of x and y is obtained as follows:

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$= \frac{\Sigma xy}{N\sigma_x \times \sigma_y} = \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2}$$

The regression coefficient of y and x is denoted by b_{yx} or b_2. It measures the change in y corresponding to a unit change in x. When the deviations are taken from actual means, the regression coefficient of y and x can be obtained as

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$= \frac{\Sigma xy}{N\sigma_x \times \sigma_y} = \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2}$$

The regression equation of Y on X is given as

$$Y = a + bX$$

where b = the slope of the line of regression of Y on X, called the *slope coefficient* or *regression coefficient*.

Similarly, the regression equation of X on Y is given as

$$X = c + dY$$

where d = the slope of the line of regression of X on Y, called the slope coefficient or regression coefficient.

The regression coefficient of Y on X is given by

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

The regression coefficient of X on Y is given by

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

where:

r = the coefficient of correlation between X and Y
σ_x = the population standard deviation of X
σ_y = the population standard deviation of Y

Theorem 13.1

The square root of the product of two regression coefficients gives us the value of the correlation coefficient; that is, $r = \sqrt{b_{xy} \times b_{yx}}$.

Proof:

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2$$

$$\therefore r = \sqrt{b_{xy} \times b_{yx}}$$

13.2.9 Properties of the Regression Coefficients

1. The coefficient of correlation is the geometric mean of the two regression coefficients. Symbolically,

$$r = \sqrt{b_{yx} \times b_{xy}}$$

Proof: The regression equation Y on X is given by

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \tag{13.1}$$

The regression equation X on Y is given by

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \tag{13.2}$$

Multiplying Equations 13.1 and 13.2, we have

$$\therefore b_{yx} \times b_{xy} = r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y} = r^2$$

$$\Rightarrow r = \pm\sqrt{b_{yx} \times b_{xy}}$$

2. The arithmetic mean of b_{yx} and bxy is equal to or greater than the coefficient of correlation. Symbolically,

$$\frac{b_{yx} + b_{xy}}{2} \geq r$$

3. If b_{yx} is positive, b_{xy} should also be positive. Thus, both regression coefficients must have the same sign. If b_{xy} and b_{yx} are both positive, r will also be positive, and if b_{xy} and b_{yx} are both negative, r will be negative.
4. If one of the regression coefficients is greater than one, the other must be less than one, since the value of the coefficient of correlation cannot be greater than one.

   For example, if $b_{yx} = 1.5$ and $b_{xy} = 1.2$,

$$r = \pm\sqrt{1.5 \times 1.2} = \pm\sqrt{1.80} = 1.342$$

   This is not possible. So, if $b_{yx} > 1$, then $b_{xy} < 1$.
5. The arithmetic mean of b_{yx} and b_{xy} is equal to or greater than the coefficient of correlation. Symbolically,

$$\frac{b_{yx} + b_{xy}}{2} \geq r$$

For example, if $b_{yx} = 0.8$ and $b_{xy} = 0.6$, the average of the two values would be $(0.8 + 0.6/2) = 0.7$,

$$r = \sqrt{0.8 \times 0.6} = 0.693$$

Thus, the value of r is less than average of b_{yx} and b_{xy}.

6. Regression coefficients are independent of origin but not on scale.
   Let $U = X - A$ and $V = Y - B$
   Where A and B are arbitrary means

$$\begin{aligned} X &= A + U & Y &= B + V \\ \bar{X} &= A + \bar{U} & \bar{Y} &= B + \bar{V} \\ (X - \bar{X}) &= (U - \bar{U}) & (Y - \bar{Y}) &= (V - \bar{V}) \\ \sigma_x^2 &= \sigma_u^2 & \sigma_y^2 &= \sigma_v^2 \end{aligned}$$

$$\sum (x - \bar{x})(y - \bar{y}) = \sum (u - \bar{u})(v - \bar{v})$$

Hence change of origin has no effect on the regression coefficient.

13.2.10 Features of Regression Coefficients

1. Both the regression coefficients will have the same sign; they will be either positive or negative. It is not possible for one of the regression coefficients to be negative and another positive.
2. The regression coefficients cannot be greater than one.
3. Since $b_{xy} = r\frac{\sigma_x}{\sigma_y}$, we can find out any of the four values given the other three.

13.2.11 Regression Line

According to J R Stockton, 'the tool used for estimating the value of one variable from the value of the other consists of a line through the points, drawn in such a manner as to represent the average relationship between the two variables. Such a line is called the line of regression'.

The regression line gives the best estimate of the value of one variable for any specific value of the other variable.

If we take the case of two variables x and y, we will have two regression lines: the regression of x on y and the regression of y on x. The regression line of y on x gives the most probable values of y for given values of x, and the regression line of x on y gives the most probable values of x for given values of y. Thus, we have two regression lines.

The equation for a straight line where the dependent variable Y is determined by the independent variable X is

$$Y = a + bX \tag{13.3}$$

where:
a = Y-intercept (intercept on Y-axis)
b = slope of the line
X = independent variable
Y = dependent variable

Similarly, if X is the dependent variable and Y is the independent, the equation of the straight line is given by

$$X = c + dY \tag{13.4}$$

where:
c = X-intercept (intercept on X-axis)
d = slope of the line
X = dependent variable
Y = independent variable

13.2.12 Interpretation of Regression line

1. As these two lines of regression are obtained from different sources and are based on different assumptions, these two regression equations are not reversible or interchangeable.
2. In cases where there is either perfect positive correlation or either perfect negative correlation between the two variables, the two regression lines will overlap (coincide with each other), and hence only one line appears.
3. If the two regression lines are far from each other, the degree of correlation is low.
4. If the two lines of regression are close to each other, the degree of correlation is high.
5. If the variables are not at all related, that is, if r = 0, the lines of regression are at right angles to each other, that is, one parallel to the X-axis and the other parallel to the Y-axis.
6. The two regression lines intersect at the point of average of X and Y.
7. If a perpendicular line is drawn from the point of intersection of the two regression lines (regression line Y on X and regression line X on Y) to the X-axis, the point where the perpendicular meets the X-axis is the mean value of X.

13.2.13 Role of Regression Analysis in Business Decision Making

In economics, it is the basic technique for measuring or estimating the relationship among the economic variables.

13.2.14 Correlation Analysis versus Regression Analysis

If we know that two variables, price (x) and demand (y), are closely related, we can find the most probable value of x for a given value of y or the most probable value of y for a given value of x.

TABLE 13.1

Correlation Analysis vs. Regression Analysis

Correlation Analysis	Regression Analysis
It is a measure of the degree of co-variability between x and y.	Studies the nature of the relationship between the variables so that we may be able to predict the value of one on the basis of another.
It is merely a tool of ascertaining the degree of relationship between two variables, and therefore, we cannot say that one variable is the cause and the other the effect.	The cause and effect relation is clearly indicated.
There may be nonsense correlation between two variables, which is purely due to chance and has no practical relevance, such as increase in income and increase in weight of a group of people.	There is nothing like nonsense regression.
Correlation coefficient is independent of change of scale and origin.	Regression coefficients are independent of change of origin but not of scale.

Similarly, if we know that the amount of tax and the rise in price of a commodity are closely related, we can find out the expected price for a certain amount of tax levy (Table 13.1).

13.3 The Least-Squares Method

There will be a discrepancy between most of the actual scores and the predicted score. The least-squares method uses the criterion of attempting to make the lowest total error in prediction of Y from X. More technically, the procedure used in the least-squares method generates a straight line that minimises the sum of squared deviations of the actual values from this predicted regression line.

The regression line should be drawn through the plotted points in such a way that the sum of the squares of the vertical deviations of the actual and estimated Y values is at a minimum. Thus, the line of regression becomes the line of 'best fit'.

According to the principle of least squares, the normal equations for estimating a and b, that is, the Y-intercept and slope of the best-fitting regression line, are

$$\Sigma Y = na + b\Sigma X \tag{13.5}$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \tag{13.6}$$

where:

n = total number of observed pairs of values

$\sum X$, $\sum Y$, $\sum XY$ and $\sum X^2$ totals computed from the observed pairs of values of the variables X and Y for which the line of least-squares estimate is to be fitted.

$$b = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

$$a = \bar{Y} - b\bar{X}$$

where:
$\bar{X}$ is the mean value of X
$\bar{Y}$ is the mean value of Y

Similarly, the normal equations for the regression equation of X on Y are

$$\Sigma Y = nc + d\Sigma Y \tag{13.7}$$

$$\Sigma XY = c\Sigma Y + d\Sigma Y^2 \tag{13.8}$$

where:

$$d = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma Y^2 - n\bar{Y}^2}$$

$$c = \bar{X} - b\bar{Y}$$

13.3.1 Application of Least-Squares Method

1. The method is used to generate a regression model by assigning data to a single line.
2. In this method, past demand data is used to form a linear model by regressing the data points to a single line.
3. Once the linear equation is formed, future demand (Y) can be predicted by substituting the value of X.
4. To calculate the value of constants b and a in the regression model, the following equations are used:

$$b = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$$

$$a = \frac{\Sigma Y}{n} - b\,\frac{\Sigma X}{n} \quad \text{or}$$

$$a = \bar{Y} - b\bar{X}$$

where n = sample size.

13.4 Standard Error of Estimate (SE)

1. This is the measure of variation of the observations around the computed regression line or the variability from the regression line.
2. The SE is analogous to the standard deviation (SD).

3. The SD measures the variation of the observations in a frequency distribution around the mean of that data.
4. The SE of Y measures the variability of the observed values of Y around the regression line. Here, the deviations are not from the arithmetic means; they are perpendicular distances of every point from the line of average relationships.

The SE indicates how precise the prediction of y is, based on x, or, conversely, how inaccurate the prediction might be. It is symbolised by $S_{y.x}$.

The formula for calculating the S.E of estimate is

$$S_{y.x} = \sqrt{\frac{\Sigma(y - y_e)^2}{N - 2}}$$

where $S_{y.x}$ = the SE of regression of y values from y_e.

A more convenient formula is

$$S_{y.x} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{N - 2}}$$

Similarly, the SE of regression of x values from x_e is

$$S_{xy} = \frac{\Sigma x^2 - a\Sigma x - b\Sigma xy}{N - 2}$$

The SE of estimate measures the accuracy of the estimated figures.

The smaller the value of SE of estimate, the closer the dots will be to the regression line, and the better the estimates based on the equation for this line. If SE of estimate is zero, there is no variation about the line, and the correlation will be perfect.

13.4.1 Standard Error of Estimate of Y on X

$$\text{SE of Y on X } (SE_{xy}) = \sqrt{\frac{\Sigma(Y - Ye)^2}{n - 2}}$$

where:

Y = observed value of y
Y_e = estimated values from the estimated equation that correspond to each y value
e = the error term ($Y–Y_e$)
n = number of observations in sample

13.4.2 Interpretation of SE of Estimates

1. The smaller the value of the SE of estimate, the closer are the observations (plotted on the graph) to the regression line, and the better are the estimates obtained from the equation of such a line.

2. The larger the value of the SE of estimate, the farther are the observations (plotted on the graph) from the regression line; that is, the observations are more scattered and dispersed widely, and estimates obtained from such equations are inaccurate.
3. If the value of the SE of estimate is zero, there is no variation from the regression line, and the correlation is perfect.

13.5 Multiple Regressions

1. In a simple regression, the dependent variable Y was assumed to be linearly related to a single variable X.
2. In real life, we often find that a dependent variable may depend on more than one independent variable. For example, the profit of a company would depend upon revenues and a number of costs, such as fixed costs and variable costs. Hence, the simple regression equation appears to be inadequate in representing such a situation.

13.5.1 Multiple Regression Equation

$$\hat{Y} = A + B_1X_1 + B_2X_2 + \ldots + B_nX_n$$

where:

$\hat{Y}$ = estimated value corresponding to the dependent variable
A = Y-intercept
$X_1, X_2, X_3, \ldots X_n$ = values of independent variables
$B_1, B_2, \ldots B_n$ = slopes associated with X_1, X_2

We can also write the equation as

$$\hat{Y} = A + B_1X_1 + B_2X_2 + \ldots + B_nX_n + \in$$

where $\in$ = random variable with mean 0.

13.5.2 Multicollinearity

1. As the degree of correlation between the independent variables increases, the regression coefficients become less reliable; that is, the independent variables may together explain the dependent variable, but because of multicollinearity, the coefficients of explanatory variables may be rejected.
2. It can happen that the model may be accepted (through the F-test), but the individual coefficients may be rejected (through the t-test).
3. If two variables are identical, the influence of each one in the model will be reduced.
4. Multicollinearity does not reduce the accuracy of the model (the predictive powers).

Exercise

Calculate the two regression equations and the correlation coefficient from the data given in Table 13.2.

Solution

Let marks in statistics be denoted by X and marks in mathematics by Y (Table 13.3).

$$\text{Regression equation of X on Y: } X - \bar{X} = r\frac{\sigma_x}{\sigma_y}\left(Y - \bar{Y}\right)$$

$$\bar{X} = \frac{\sum X}{N} = \frac{185}{5} = 37;\ \bar{Y} = \frac{\sum Y}{N} = \frac{170}{5} = 34$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2} = \frac{22}{82} = 0.268$$

$$X - 37 = 0.268\left(Y - 34\right)$$

$$X - 37 = 0.268Y - 9.112 \text{ or } X = 27.888 + 0.268Y \tag{13.9}$$

$$\text{Regression equation of Y on X: } Y - \bar{Y} = r\frac{\sigma_x}{\sigma_y}\left(X - \bar{X}\right)$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2} = \frac{22}{88} = 0.25$$

$$Y - 34 = 0.25\left(X - 37\right)$$

TABLE 13.2

Marks in Statistics and Mathematics

Marks in statistics (out of 50)	40	38	35	42	30
Marks in mathematics (out of 50)	30	35	40	36	29

TABLE 13.3

Calculation of Regression Equations and Correlation Coefficient

X	$x = X - \bar{X}$ (X−37)	x^2	Y	$y = (Y - \bar{Y})$ (Y−34)	y^2	xy
40	3	9	30	−4	16	−12
38	1	1	35	1	1	1
35	−2	4	40	6	36	−12
42	5	25	36	2	4	10
30	−7	49	29	−5	25	35
$\sum X = 185$	$\sum x = 0$	$\sum x^2 = 88$	$\sum Y = 170$	$\sum y = 0$	$\sum y^2 = 82$	$\sum xy = 22$

$$Y - 34 = 0.25X - 9.25 \text{ or } Y = 24.75 + 0.25Y \tag{13.10}$$

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{.268 \times .25} = 0.259.$$

Exercise

The following data gives the experience of machine operators and their performance rating as given by number of good parts out of 100 pieces.

Operator	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Experience (Years)	16	12	18	4	3	10	5	12
Performance	87	88	89	68	78	80	75	83

Calculate the regression line of performance ratings on experience and estimate the probable performance of an operator having 7 years' experience

Solution

Let the performance rating be denoted by y and experience by x.
We have to calculate the regression line of y on x (Table 13.4).
Regression equation of y on x:

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$r = \frac{\sigma y}{\sigma x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{247}{218} = 1.133$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{648}{8} = 81, \bar{x} = \frac{\Sigma x}{N} = \frac{80}{8} = 10$$

$$y - 81 = 1.133(x - 10)$$

TABLE 13.4

Calculation of Regression Line of y and x

Experience			**Performance**			
x	$\mathbf{x = x - \bar{x}}$ $\mathbf{\bar{x} = 10}$	$\mathbf{x^2}$	**y**	$\mathbf{y = y - \bar{y}}$ $\mathbf{\bar{y} = 81}$	$\mathbf{y^2}$	**xy**
16	+6	36	87	+6	36	+36
12	+2	4	88	+7	49	+14
18	+8	64	89	+8	64	+64
4	−6	36	68	−13	169	+78
3	−7	49	78	−3	9	+21
10	0	0	80	−1	1	0
5	−5	25	75	−6	36	+30
12	+2	4	83	+2	4	+4
$\sum x = 80$	$\sum x = 0$	$\sum x^2 = 218$	$\sum y = 648$	$\sum y = 0$	$\sum y^2 = 368$	$\sum xy = 247$

$$y - 81 = 1.133x - 11.33$$

$$= 1.133x + 69.67$$

When x = 7, y will be

$$y = 1.133(7) + 69.67$$

$$y = 7.931 + 69.67$$

$$y = 77.60$$

$$y = 78$$

Thus, the probable performance of an operator who has 7 years' experience = 78.

Exercise

The following table gives age and number of morning walkers. Determine regression equation and estimate number of walkers at the age of 55.

Age (years)	20	30	40	50	60	70	80
No. of walkers	4	6	8	10	9	4	1

Solution

Calculation of regression equations (Table 13.5)
We have

$$\bar{x} = \frac{\sum x}{N} = \frac{350}{7} = 50, \bar{y} = \frac{\sum y}{N} = \frac{42}{7} = 6$$

Regression coefficient of x on y,

$$bxy = \frac{\sum xy}{\sum y^2} = \frac{-120}{62} = -1.93$$

TABLE 13.5

Calculation of Regression Equations

Age (Years) x	**No. of Walkers y**	$x = x - \bar{x}$ $\bar{x} = 50$	$y = y - \bar{y}$ $\bar{y} = 6$	x^2	y^2	xy
20	4	−30	−2	900	4	60
30	6	−20	0	400	0	0
40	8	−10	2	100	4	−20
50	10	0	4	0	16	0
60	9	10	3	100	9	30
70	4	20	−2	400	4	−40
80	1	30	−5	900	25	−150
$\sum x = 350$	$\sum y = 42$	$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 2800$	$\sum y^2 = 62$	$\sum xy = -120$

Regression coefficient of y on x,

$$byx = \frac{\sum xy}{\sum x^2} = \frac{-120}{2800} = -0.042$$

Regression equation of x on y:

$$x = \bar{x} + bxy(y - \bar{y})$$

$$x = 50 + (-1.93)(y - 6)$$

$$x = 50 - 1.93y + 11.58$$

$$x = 61.58 - 1.93y$$

Regression equation of y on x:

$$y = \bar{y} + byx(x - \bar{x})$$

$$y = 6 + (-0.042)(x - 50)$$

$$y = 6 + (-0.042)(x - 50)$$

$$y = 8.1 - 0.042x$$

To estimate number of walkers at the age of 55, that is, when x = 55,

$$y_e = 8.1 - 0.042(55) = 5.79 \approx 6.$$

Exercise

Use the following data to obtain two regression equations.

Sales	91	97	108	121	67	124	51	73	111	57
Purchase	71	75	69	97	70	91	39	61	80	47

Solution

Obtaining regression equations (Table 13.6)
Regression equation of x and y:
Let x = a + by be the line on x on y
To find a and b, we have the following normal equations:

$$\sum x = Na + b\sum y$$

$$\sum xy = a\sum y + b\sum y^2$$

Substituting the values, we get

$$900 = 10a + 700b \quad (13.11)$$

$$66{,}900 = 700a + 51{,}868b \quad (13.12)$$

TABLE 13.6

Calculation of Regression Equations

x	y	xy	x^2	y^2
91	71	6,461	8,281	5,041
71	75	7,275	9,409	5,625
108	69	7,452	11,664	4,761
121	97	11,737	14,641	9,409
67	70	4,690	4,489	4,900
124	91	11,284	15,376	8,281
51	39	1,989	2,601	1,521
73	61	4,453	5,329	3,721
111	80	8,880	12,321	6,400
57	47	2,679	3,249	2,209
$\sum x = 900$	$\sum y = 700$	$\sum xy = 66{,}900$	$\sum x^2 = 87{,}360$	$\sum y^2 = 51{,}868$

Multiplying Equation 13.11 by 70, we have

$$63{,}000 = 700a + 49{,}000b \tag{13.13}$$

Taking Equations 13.12 and 13.13,

$$\begin{array}{rl} 66{,}900 = & 700a + 51{,}868b \\ 63{,}000 = & 700a + 49{,}000b \\ \hline 3900 = & 2868b \end{array}$$

$b = 1.36$
Substituting $b = 1.36$ in … Equation 13.11,

$$900 = 10a + 700(1.36)$$

$$900 = 10a + 952$$

$$a = -5.2$$

$\therefore$ Required equation of x and y:

$$x = -5.2 + 1.36y$$

Regression equation of x and y:
Let $x = a + bx$ be the line on x on y
To find a and b, we have the following normal equations:

$$\sum y = Na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Substituting the values, we get

$$700 = 10a + 900b \tag{13.14}$$

$$66{,}900 = 900a + 87{,}360b \tag{13.15}$$

Multiplying Equation 13.14 by 90, we have

$$63{,}000 = 900a + 81{,}000b \tag{13.16}$$

Taking Equations 13.15 and 13.16,

$$66{,}900 = 900a + 87{,}360b$$

$$\underline{63{,}000 = 900a + 81{,}000b}$$

$$3900 = \quad 6360b$$

$$b = 0.61$$

Substituting b = 0.61 in … 4, we get

$$700 = 10a + 900(0.61)$$

$$900 = 10a + 952$$

$$a = 15.1$$

∴ Required equation of y and x:

$$y = 15.1 + 0.61x$$

Exercise

Using regression analysis, find the advertising budget for achieving sales of 30 units. Sales of previous months and their corresponding advertising budgets are shown in Table 13.7.

Solution

By using the least-squares method, we can calculate the values of a and b. When the values of a and b are substituted in the linear regression equation with required levels of sales, the advertising budget needed to achieve the required level of sales can be forecast (Table 13.8).
Here, n = 6,

$$\bar{X} = \frac{\Sigma X}{n} = \frac{130}{6} = 21.67, \bar{Y} = \frac{\Sigma Y}{n} = \frac{49}{6} = 8.17$$

$$b = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$$

The equation can also be written as

$$\frac{(\Sigma XY) - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

TABLE 13.7

Sales and Advertising Budgets

Sales	21	11	37	22	24	15
Advertising budget	7	5	14	8	9	6

TABLE 13.8

Calculation of Regression Equations

X (Sales)	Y (Ad Budget)	XY	X^2	Y^2
21	7	147	441	49
11	5	55	121	25
37	14	518	1369	196
22	8	176	484	64
24	9	216	576	81
15	6	90	225	36
$\sum X = 130$	$\sum Y = 49$	$\sum XY = 1202$	$\sum X^2 = 3216$	$\sum Y^2 = 451$

Substituting the values, we get b = 0.35
Now,

$$a = \bar{Y} - b\bar{X} = 8.17 - (0.35 \times 21.67) = 0.58$$

Now, we can find the value of Y (dependent variable) for which the value of the independent variable is 30.

$$Y = a + bX \Rightarrow Y = 0.58 + 0.35 \times 30 = 11.08$$

So, to achieve a sales level of 30 units, the organisation needs to spend 11.08 units of capital on advertising.

Exercise

From the following information, calculate the regression equations:

$$\sum X = 30,\ \sum Y = 40,\ \sum XY = 214,\ \sum X^2 = 220,\ \sum Y^2 = 340 \text{ and } N = 5.$$

Also find the coefficient of correlation.

Solution

Regression equation of X on Y

$$X - \bar{X} = b_{xy}\left(Y - \bar{Y}\right)$$

where:

$$\bar{X} = \frac{\sum X}{N} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$$

$$b_{xy} = \frac{\Sigma XY - N\bar{X}\bar{Y}}{\Sigma Y^2 - N\left(\bar{Y}\right)^2} \quad \text{(when figures are given in original values)}$$

$$b_{xy} = \frac{214 - (5 \times 6 \times 8)}{340 - 5(8)^2}$$

$$b_{xy} = \frac{214 - 240}{340 - 320} = -\frac{26}{20} = -1.3$$

$$X - 6 = -1.3(Y - 8)$$

$$X - 6 = -1.3Y + 10.4$$

$$X = -1.3Y + 16.4$$

$$X = 16.4 - 1.3Y \tag{13.17}$$

Regression equation of Y on X:

$$\bar{Y} = 8, \bar{X} = 6$$

$$b_{yx} = \frac{\Sigma XY - N\bar{X}\bar{Y}}{\Sigma X^2 - N\left(\bar{X}\right)^2}$$

$$b_{yx} = \frac{214 - (5 \times 6 \times 8)}{220 - 5(6)^2}$$

$$b_{yx} = \frac{214 - 240}{220 - 180} = -\frac{26}{40} = -0.65$$

$$Y - \bar{Y} = b_{yx}\left(X - \bar{X}\right)$$

$$Y - 8 = -0.65(X - 6)$$

$$Y - 8 = -0.65X + 3.90$$

$$Y = -0.65X + 11.9$$

$$Y = 11.9 - 0.65X \tag{13.18}$$

$$\therefore r = \sqrt{b_{xy} \times b_{yx}}$$

$$\therefore r = \sqrt{(-1.3) \times (-0.65)}$$

$$\therefore r = \sqrt{0.845} = -0.92$$

TABLE 13.9

Data of Mean and SD

	x	y
Mean	40	45
SD	10	9

Exercise

Compute the two regression equations on the basis of the information in Table 13.9.

Karl Pearson coefficient of correlation between X and Y = 0.50. Also estimate the value of Y for X = 48 using the appropriate regression equation. (Mean and SD are known.)

Solution

Regression equation of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}\left(Y - \bar{Y}\right)$$

Here,

$$\bar{X} = 40,\ \bar{Y} = 45,\ \sigma_x = 10, \sigma_y = 9 \text{ and } r = 0.5$$

$$X - 40 = 0.5 \times \frac{10}{9}\left(Y - 45\right)$$

$$X = 0.556Y + 15$$

Regression equation of Y on X

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}\left(X - \bar{X}\right)$$

$$Y = \bar{Y} + r\frac{\sigma_y}{\sigma_x}\left(X - \bar{X}\right)$$

Substituting the values, we get

$$Y = 45 + 0.5 \times \frac{9}{10}\left(X - 40\right)$$

$$Y = 0.45X + 27$$

The appropriate regression equation for estimating the value of Y, for a given value of X, is

$$\hat{Y} = 0.45X + 27$$

Hence, if X = 48,

$$\hat{Y} = 0.45(48) + 27 = 48.6$$

Exercise

Table 13.10 shows the respective weights, X and Y, of a sample of 12 fathers and their eldest sons.

1. Find the least-squares regression line of Y on X.
2. Find the least-squares regression line of X on Y.

Solution

Calculation of regression lines (Table 13.11):

1. The regression line of Y on X is given by Y = a + bX, where a and b are obtained by normal equations:

$$\sum Y = aN + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

TABLE 13.10

Weights of X and Y (kg)

Weights of fathers in kg (X)	65	63	67	64	68	62	70	66	68	67	69	71
Weights of sons in kg (Y)	68	66	68	65	69	66	78	65	71	67	68	70

TABLE 13.11

Calculation of Regression Lines

X	Y	X^2	XY	Y^2
65	68	4,225	4,420	4,624
63	66	3,969	4,159	4,356
67	68	4,489	4,556	4,624
64	65	4,096	4,160	4,225
68	69	4,624	4,692	4,761
62	66	3,844	4,092	4,356
70	68	4,900	4,760	4,624
66	65	4,356	4,290	4,225
68	71	4,624	4,828	5,041
67	67	4,489	4,489	4,489
69	68	4,761	4,692	4,624
71	72	4,541	4,970	4,900
$\sum X = 800$	$\sum Y = 811$	$\sum X^2 = 53{,}418$	$\sum XY = 54{,}107$	$\sum Y^2 = 54{,}849$

$$a = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{N\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{(811)(53,418) - (800)(54,107)}{12(53,418) - (800)^2} = 35.82$$

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{12(54,107) - (800)(811)}{12(53,418) - (800)^2} = 0.476$$

Thus, Y = 35.82 + 0.476X.

2. The regression line of X on Y is given by X = a + bY, where a and b are obtained by normal equations:

$$\Sigma X = aN + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

$$a = \frac{(\Sigma X)(\Sigma Y^2) - (\Sigma X)(\Sigma XY)}{N\Sigma Y^2 - (\Sigma Y)^2}$$

$$= \frac{(800)(54,849) - (811)(54,107)}{12(54,849) - (811)^2} = -3.38$$

$$b = \frac{N\Sigma XY - (\Sigma Y)(\Sigma X)}{N\Sigma Y^2 - (\Sigma Y)^2}$$

$$= \frac{12(54,107) - (811)(800)}{12(54,849) - (811)^2} = 1.036$$

Thus, X = −3.38 + 1.036Y.

Exercise

Find the

1. Regression equation
2. Correlation coefficient
3. Value of Y when X = 30 (method of deviation from the actual means) (Table 13.12).

Solution

Calculation of regression equation by method of deviation from the means (Table 13.13)

TABLE 13.12

Data of X and Y

X	25	28	35	32	31	36	29	38	34	32
Y	43	46	49	41	36	32	31	30	33	39

TABLE 13.13

Calculation of Regression Equation

X	**Y**	$\mathbf{x = X - \bar{X}}$ **x = X – 32**	$\mathbf{y = Y - \bar{Y}}$ **y =Y – 38**	$\mathbf{x^2}$	$\mathbf{y^2}$	**xy**
25	43	–7	5	49	25	–35
28	46	–4	8	16	64	–32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	–1	–2	1	4	2
36	32	4	–6	16	36	–24
29	31	–3	–7	9	49	21
38	30	6	–8	36	64	–48
34	33	2	–5	4	25	–10
32	39	0	1	0	1	0
$\sum X = 320$	$\sum Y = 380$	$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 140$	$\sum x^2 = 398$	$\sum xy = -93$

$$\bar{X} = \frac{\sum X}{N} = \frac{320}{10} = 32, \bar{Y} = \frac{\sum Y}{N} = \frac{380}{10} = 38$$

$$\sigma_x = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{140}{10}} = 3.74, \sigma_y = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{398}{10}} = 6.31$$

$$r = \frac{\sum xy}{N\sigma_x\sigma_y} = \frac{-93}{(10)(3.74)(6.31)} = -0.394$$

∴ Regression equation of X on Y is

$$X = \bar{X} + r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$X = 32 + (-0.394)\left(\frac{3.74}{6.31}\right)(Y - 38)$$

$$X = 40.8806 - 0.2337Y$$

Now, regression equation of X on Y is

$$Y = \bar{Y} + r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$Y = 38 + (-0.394)\left(\frac{6.31}{3.74}\right)(X - 32)$$

$$X = 59.2576 - 0.6643X$$

Correlation coefficient:

$$r_{xy} = \sqrt{b_{xy} \times b_{yx}}$$

$$= \sqrt{\frac{\Sigma xy}{\Sigma y^2} \times \frac{\Sigma xy}{\Sigma x^2}}$$

$$= \sqrt{(-0.2337)(-0.6643)} = 0.394$$

Probable value of Y when X = 30:

$$Y = 59.2576 - 0.6643(30)$$

$$Y = 39.3286$$

Exercise

From the data in Table 13.14, find

1. Two regression coefficients
2. The correlation coefficient
3. Two regression equations
4. SD of X on Y (method of deviation from the actual mean).

Also, find the sales when the expenditure is Rs 16,000 (Table 13.15).

Solution

We have

$$\bar{X} = \frac{\Sigma X}{N} = \frac{56}{7} = 8, \bar{Y} = \frac{\Sigma Y}{N} = \frac{56}{7} = 8$$

Regression coefficient of X on Y:

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{21}{28} = 0.75$$

TABLE 13.14

Expenditure and Volume of Sales

Expenditure on advertisements (Rs. in thousands) X	11	7	9	5	8	6	10
Volume of sales (in lakhs) Y	10	8	6	5	9	7	11

TABLE 13.15

Calculation of Regression Lines

X	**Y**	$\mathbf{x = X - \bar{X}}$ $\mathbf{x = X - 8}$	$\mathbf{y = Y - \bar{Y}}$ $\mathbf{y = Y - 8}$	$\mathbf{x^2}$	$\mathbf{y^2}$	**xy**
11	10	3	2	9	4	6
7	8	–1	0	1	0	0
9	6	1	–2	1	4	–2
5	5	–3	–3	9	9	9
8	9	0	1	0	1	0
6	7	–2	–1	4	1	2
10	11	2	3	4	9	6
$\sum X = 56$	$\sum Y = 56$	$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 28$	$\sum y^2 = 28$	$\sum xy = 21$

Regression coefficient of Y on X:

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{21}{28} = 0.75$$

Correlation coefficient:

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$= \sqrt{(0.75)(0.75)} = \pm 0.75$$

Regression equation of X on Y:

$$X = \bar{X} + b_{xy}\left(Y - \bar{Y}\right)$$

$$= 8 + 0.75(Y - 8)$$

$$= 2 + 0.75Y$$

Regression equation of Y on X:

$$Y = \bar{Y} + b_{xy}\left(X - \bar{X}\right)$$

$$= 8 + 0.75(X - 8)$$

$$= 2 + 0.75X$$

SD of X:

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{28}{7}} = 2$$

SD of Y:

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{28}{7}} = 2$$

Calculation of sales (Y) when advertisement expenditure (X) is Rs 16,000:

$$Y = 2 + 0.75(16)$$

$$Y = 14$$

$$\therefore \text{When } X = 16,000$$

$$Y = 14 \times 1000 = 14,000 .$$

Exercise

From the data given below, compute the two regression coefficients and formulate the two regression equations.

$\sum X = 510$, $\sum Y = 7140$, $\sum XY = 54{,}900$, $\sum X^2 = 4150$, $\sum Y^2 = 740{,}200$ and $N = 102$.

Also determine the value of Y when X = 8 (value-based method).

Solution

By the value-based method,

Regression coefficient of X on Y:

$$b_{xy} = \frac{N\sum XY - (\sum Y)(\sum X)}{N\sum Y^2 - (\sum Y)^2}$$

$$= \frac{102(54,900) - (510)(7140)}{102(740,200) - (7140)^2} = 0.08$$

Regression coefficient of Y on X:

$$b_{yx} = \frac{N\sum XY - (\sum Y)(\sum X)}{N\sum X^2 - (\sum X)^2}$$

$$= \frac{102(54,900) - (510)(7140)}{102(4150) - (510)^2} = 12$$

$$\bar{X} = \frac{\sum X}{N} = \frac{510}{102} = 5$$

Regression equation of X on Y:

$$X = \bar{X} + b_{yx}\left(Y - \bar{Y}\right)$$

$$X = 5 + 0.08(Y - 70)$$

$$X = -0.6 + 0.08Y$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{7140}{102} = 70$$

Regression equation of Y on X:

$$Y = \bar{Y} + b_{xy}\left(X - \bar{X}\right)$$

$$Y = 70 + 12(X - 5)$$

$$Y = 10 + 12X$$

Value of Y when X = 8:

$$Y = 10 + 12(8) = 106$$

Coefficient of correlation:

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$= \sqrt{(0.08) \times (12)} = \pm 0.97$$

Exercise

From Table 13.16, obtain the equations of the two lines of regression. Also determine the value of the correlation coefficient between x and y (method of deviation from Assumed Mean).

Solution

Calculation of regression lines (Table 13.17):

$$b_{xy} = \frac{N\Sigma dx\, dx - (\Sigma dx)(\Sigma dy)}{N\Sigma dy^2 - (\Sigma dy)^2}$$

$$= \frac{12(-283) - (-17)(-18)}{12(728) - (-18)^2} = -0.44$$

$$b_{yx} = \frac{N\Sigma dxdy - (\Sigma dy)(\Sigma dx)}{N\Sigma dx^2 - (\Sigma dx)^2}$$

TABLE 13.16

Data of X and Y

X	43	44	46	40	44	42	45	42	38	40	42	57
Y	29	31	19	18	19	27	27	29	41	30	26	10

TABLE 13.17

Calculation of Regression Lines

X	dx = X − A = X − 45	dx²	Y	dy = Y − 27	dy²	dx dy
43	−2	4	29	2	4	−4
44	−1	1	31	4	16	−4
46	1	1	19	−8	64	−8
40	−5	25	18	−9	81	45
44	−1	1	19	−8	64	8
42	−3	9	27	0	0	0
45	0	0	27	0	0	0
42	−3	9	29	2	4	−6
38	−7	49	41	14	196	−98
40	−5	25	30	3	9	−15
42	−3	9	26	−1	1	3
57	12	144	10	−17	−289	−204
$\sum X = 523$	$\sum dx = -17$	$\sum dx^2 = 277$	$\sum Y = 306$	$\sum dy = -18$	$\sum dy^2 = 728$	$\sum dx\,dy = -283$

$$= \frac{12(-283) - (-17)(-18)}{12(277) - (-17)^2} = -1.22$$

$$\bar{X} = A + \frac{\sum dx}{N} = 45 + \frac{(-17)}{12} = 43.58$$

Regression equation of X on Y:

$$X = \bar{X} + b_{xy}\left(Y - \bar{Y}\right)$$

$$X = 43.58 + (-0.44)(Y - 25.5)$$

$$X = 54.80 - 0.44Y$$

$$\bar{Y} = A + \frac{\sum dy}{N} = 27 + \frac{(-18)}{12} = 25.5$$

Regression equation of Y on X:

$$Y = \bar{Y} + b_{xy}\left(X - \bar{X}\right)$$

$$Y = 25.5 + (-1.22)(X - 43.58)$$

$$Y = 78.67 - 1.22X$$

Correlation coefficient between X and Y:

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$= \sqrt{(-0.44)(-1.22)} = \pm 0.7326$$

Exercise

From the data in Table 13.18, find regression equations and SE of estimates (value-based method).

Also, determine the value of r on the basis of SE.

Solution

Calculation of the regression equation and the SE (Table 13.19)

By the value-based method, we have

$$b_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$$

$$= \frac{5(214) - (30)(40)}{5(340) - (40)^2} = -1.3$$

$$b_{yx} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{5(214) - (30)(40)}{5(220) - (30)^2} = 0.65$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6, \bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$$

Regression equation of X on Y:

$$X = \bar{X} + b_{xy}(Y - \bar{Y})$$

$$X = 6 + (-1.3)(Y - 8)$$

$$X = 16.4 - 1.3Y$$

TABLE 13.18

Data of X and Y

X	6	2	10	4	8
Y	9	11	5	8	7

TABLE 13.19

Calculation of the Regression Equation and the SE

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
$\sum X = 30$	$\sum Y = 40$	$\sum X^2 = 220$	$\sum Y^2 = 340$	$\sum XY = 214$

Regression equation of Y on X:

$$Y = \bar{Y} + b_{yx}\left(X - \bar{X}\right)$$

$$Y = 8 + (-0.65)(X - 6)$$

$$Y = 11.9 - 0.65X$$

Standard Error (SE)

1. Fundamental method
   Computation of estimated value of X on Y **(**Table 13.20)
   Computation of estimated value of Y on X **(**Table 13.21)
   Working table (Table 13.22)
   SE of estimated Y on X:

$$SE_{Y \text{ on } X} = \sqrt{\frac{\Sigma\left(Y - Y_e\right)^2}{N}} = \sqrt{\frac{3.1}{5}} = 0.787$$

TABLE 13.20

Computation of Estimates Value of X on Y

When Y = 9	**X_e = 16.4 – 1.3(9) = 4.7**
11	2.1
5	9.9
8	6
7	7.3

TABLE 13.21

Computation of Estimated Value of Y on X

When X = 6	**X_e = 11.9 – 0.65 (6) = 8**
2	10.6
10	5.4
4	9.3
8	6.7

TABLE 13.22

Calculation of SE of Estimates Y on X and X on Y

X	**Y**	**X_e**	**Y_e**	**$(X - X_e)^2$**	**$(Y - Y_e)^2$**
6	9	4.7	8	1.69	1
2	11	2.1	10.6	0.01	0.16
10	5	9.9	5.4	0.01	0.16
4	8	6.0	9.3	4	1.69
8	7	7.3	6.7	0.79	0.09
				$\sum(X - X_e)^2 = 6.20$	$\sum(Y - Y_e)^2 = 3.10$

SE of estimated X on Y:

$$SE_{X \text{ on } Y} = \sqrt{\frac{\Sigma (X - X_e)^2}{N}} = \sqrt{\frac{6.20}{5}} = 1.114$$

$$\sigma_y = \sqrt{\frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2} = \sqrt{\frac{340}{5} - \left(\frac{40}{5}\right)^2} = 2$$

Calculation of r

$$r = \sqrt{1 - \frac{(S.E._{Y \text{ on } X})^2}{(\sigma_y)^2}} = \sqrt{1 - \frac{(0.787)^2}{(2)^2}} = \pm 0.92$$

13.6 Summary

In this unit, fundamentals of linear regression have been highlighted. Broadly speaking, the fitting of any chosen mathematical function to given data is termed *regression analysis*. The estimation of the parameters of this model is accomplished by the least-squares criterion, which tries to minimise the sum of squares of the errors for all the data points.

How the parameters of a fitted straight-line model are estimated has been illustrated through an example.

After the model is fitted to data, the logical next question is to find out how good the quality of fit is. This question can best be answered by conducting statistical tests and determining the SE of estimate. This information permits us to make quantitative statements regarding confidence limits for estimates of the parameters as well as the forecast values.

Finally, it has been emphasised that the method of least squares used in linear regression is applicable to a wide class of models. In each case, the model parameters are obtained by the solution of the so-called 'normal equations'. These are simultaneous linear equations equal in number to the number of parameters to be estimated, obtained by partially differentiating the sum of squares of errors with respect to the individual parameters.

REVIEW QUESTIONS

1.
   a. What is regression analysis?
   b. Define the term *regression*. State its application.
   c. What are regression coefficients? List their features.
   d. What do you mean by regression lines?
   e. Point out the role of regression analysis in business decision making.
   f. Why are there two regression lines? When do we use one in preference to the other?

g. What are the limitations of regression analysis?

h. Prove that $r = \sqrt{bxy \times byx}$.

i. Write a note on standard error of estimate.

2.

a. What do you understand by linear regression?

b. What is regression? Why are there, in general, two regression lines? When do they coincide? Explain the use of regression equations in an economic inquiry.

3. Compare and contrast the roles of correlation and regression in studying the interdependence of two variables.

4. Define regression. Why are there two regression lines? Under what conditions can there be only one regression line?

5. Explain the concept of regression and comment on its utility.

6. Define the terms *regression*, *linear regression* and *curvilinear regression*. Describe the method of least squares and show how it can be used to fit a linear regression. How is linearity of regression tested?

7. Distinguish clearly between correlation and regression as concepts used in statistical analysis.

SELF-PRACTICE PROBLEMS

1. Two regression lines are $2x + 3y - 28 = 0$ and $x + 6y - 32 = 0$. Find the mean value and the correlation coefficient. If the variance of x is 27, find the SD of y.

2. From the following data, obtain the two regression lines:

x	7	3	11	6	10	12	14
y	10	12	7	10	8	10	9

3. The following table gives the ages (x) and blood pressure (y) of 12 women:

x	57	40	37	48	50	43	58	70	62	54	56	60
y	146	124	116	126	144	142	154	158	148	148	146	143

a. Obtain regression line y on x.

b. Estimate the blood pressure of a woman whose age is 65 years.

4. The following data gives the experience of machine operators and their performance rating as given by number of good parts out of 100 pieces:

Operator	1	2	3	4	5	6	7	8
Experience (Years)	14	10	16	5	6	12	4	16
Performance	67	86	88	64	74	86	74	87

Calculate the regression line of performance ratings on experience and estimate the probable performance of an operator having 7 years' experience.

5. The following table gives age and number of morning walkers. Determine regression equation and estimate number of walkers at the age of 45:

Age (years)	30	40	50	60	70	80	90
No. of walkers	5	3	6	13	6	5	3

6. Obtain the lines of regression for the following data:

X	1	2	3	4	5	6	7	8	9
Y	9	9	10	12	11	13	14	16	15

Obtain the estimate of Y which should correspond on the average to X = 6.2.
[**Ans: r** = 0.95; X = 0.95 Y = −6.4; Y = 0.95 X + 7.25; When X = 6.2, Y = 13.14]

7. Estimate the values of X corresponding to Y = 200 from the following data:

X	250	284	297	338	463	393
Y	137	147	184	196	276	260

[**Ans:** $\mathbf{X_{200}}$ = 331.51]

8. Given the following data, calculate the expected value of Y when X = 12.

	X	**Y**
Average	7.6	14.8
SD	3.6	2.5
		r = 0.99

[**Ans:** 17.83]

14

Time Series Analysis

14.1 Introduction

One of the most important tasks facing economists and business people these days is to make estimates for the future. The first step in making estimates for the future consists of gathering information from the past. In this connection, one usually deals with statistical data that is collected, observed or recorded at successive intervals of time. Such data is referred to as *time series*.

The term *time series* can be applied to all phenomena that are related to time, such as the number of accidents occurring in a day, the variation in the body temperature of a patient during a certain period, the number of marriages taking place during a certain period, and so on. For example,

1. A businessman is interested in finding out his likely sales in the year 2017 so that he can adjust his production accordingly and avoid the possibility of either unsold stocks or inadequate production to meet the demand.
2. An economist is interested in estimating the likely population in the coming year so that proper planning can be carried out with regard to food supply, jobs for the people and so on.

14.2 Time Series

14.2.1 Definition

When we observe numerical data at different points of time, the set of observations is known as a time series. If we observe production, population sales, or imports and exports at different points of time, say over the last 5–10 years, the set of observations formed will constitute a time series. In the analysis of time series, time is the most important factor, because the variable is related to time, which may be years, months, weeks, days, hours, or even minutes or seconds.

Times series can apply to phenomena related to time, such as

1. The number of accidents occurring in 24 hours
2. The variation in the temperature of a patient during a certain period
3. Study of profits over 15 years to estimate for the coming years
4. Estimation of food supply for the population in the next 10 years

14.2.2 Features of Time Series

1. Variables are determined with respect to time.
2. Time series consist of a homogeneous set of variables.
3. Time series deal with data over a long period.

For example, to study agricultural production, one may take 10–15 year data, while to study the pattern of rainfall, daily recording is essential.

14.2.3 Uses of Analysis of Time Series

Analysis of time series is of great significance not only to the economist and business person but also to the scientist, astronomer, geologist, sociologist, biologist, research worker and so on, for the following reasons:

1. Economic problems

   Changes in demand for a product over a period of time can be analysed using time series.
2. Study of past behaviour

   It helps in understanding past behaviour. By observing data over a period of time, one can easily understand changes that have taken place in the past. Such analysis will be helpful in predicting future behaviour.
3. Evaluation of current accomplishments

   The actual performance can be compared with the expected performance and the cause of variation analysed.

   For example, if expected sales for 2016 were 12,000 refrigerators and the actual sales were only 10,000, one can investigate the cause for the shortfall in achievement.

   Time Series analysis will enable us to apply the scientific procedure of `holding other things constant' as we examine one variable at a time. For example, if we know how much the effect of seasonality on business is, we may devise ways and means of ironing out the seasonal influence or decreasing it by producing commodities with complementary seasons.

   Economic and business activities depend on time series to review and find deviations in progress by studying the present situation.
4. Comparison

   Time series provide data in chronological order and present the information in a systematic way. This facilitates comparison between past and present.
5. Forecasting

   It helps in planning future operations. If the regularity of occurrence of any feature over a sufficient long period can be clearly established, then, within limits, prediction of probable future variations will become possible. It helps in making long-range forecasts and provides a base for long-term operational planning.

14.3 Components of Time Series

The forces affecting the values of phenomena in a time series are called *components* of the time series. For example, demand for product is influenced by substitutes, seasonal supply, competitors' prices, purchasing power, influence of business cycles and so on.

The effects of the various forces are:

1. Changes occurring as a result of the tendency of the data to increase or decrease are known as *secular movements*.
2. Changes taking place during a period of 12 months as a result of change in climate, weather conditions, festivals and so on are called *seasonal variations*.
3. Changes taking place as a result of booms and depressions are classified under the heading *cyclical variations*.
4. Changes taking place as a result of forces that could not be predicted, such as floods, earthquakes or famines, are classified under the heading *irregular or erratic variations*.

These four types of patterns, movements or, as they are often called, components or elements of a time series are

1. Secular trend
2. Seasonal variation
3. Cyclical variation
4. Irregular variation

14.3.1 Secular Trend or Long-Term Trend

The term *trend* is very commonly used in day-to-day conversation. For example, we often talk of a rising trend of population, prices and so on. Trend is the basic tendency of production, sales, income, employment and so on to grow or decline over a period of time.

The presence of more people means that more food, clothing and housing are necessary. Technological changes, discovery and exhaustion of natural resources, mass production methods, improvements in business organisation, and government intervention in the economy are the major causes for the growth or decline of economic time series.

In some cases, growth in one series involves decline in another. Examples include

1. Silk is displaced by rayon, terylene, etc.
2. Bullock carts, camel carts and tonga are displaced by other modes of transport such as trucks, auto-rickshaws, taxis and tempo.
3. Better medical facilities, improved sanitation, diet and so on reduce the death rate and contribute to a rise in birth rate.

Trends are divided in two headings:

1. Linear or straight-line trends
2. Non-linear trends

14.3.1.1 Meaning of Long Term

A particular period can be regarded as long or short in the study of secular trends depending on the nature of the data.

For example, if we are studying the figures of sales of a firm for 2012 and 2013, and we find that in 2013 the sales have gone up, this increase cannot be called a secular trend, because this is too short a period of time to conclude that the sales are showing an increasing tendency.

On the other hand, if we put a strong germicide into a bacterial culture and count the numbers of organisms still alive after each 3 seconds for 3 minutes, these 60 observations showing a general pattern would be called a secular movement.

It is clear from this example that in one case 2 years could not be regarded as a long period, whereas in another case even 3 minutes would constitute a long period. Hence, the nature of the data would dictate whether a particular period would be called long or not.

As a minimum safeguard, it may be said that to calculate trend, the period must cover at least two or three complete cycles. It is not necessary that the rise or fall must continue in the same direction throughout the period. We have to observe the general tendency of the data. As long as we can say that the period as a whole was characterised by an upward movement or by a downward movement, we say that a secular trend was present.

For example, if we observe the trend of prices over a period of 10 years and find that, except for a year or two, the prices have been continuously rising, we would call this a secular rise in prices.

14.3.1.2 Measurement of Secular Trends

1. In studying trend in and of itself, we ascertain the growth factor. For example, we can compare the growth in one firm of the textile industry with the growth in the industry as a whole.
2. The growth factor helps us in predicting the future behaviour of the data. If a trend can be determined, the rate of change can be ascertained and tentative estimates concerning future can be made.
3. The elimination of trend leaves us with seasonal, cyclical and irregular factors.

14.3.1.3 Features of Secular Trends

1. Long period

   In certain cases, the period extends to centuries and decades; for example, the study of rocks, minerals, the earth's crust and changes in temperature.

   For some phenomena, it takes 10–20 years; for example, production of steel by a steel industry, sales of cotton textiles.

   For some phenomena, hours, minutes and seconds constitutes a long period; for example, growth of bacteria or viruses, heart beat records, pulse rate records.

2. Tendency of trend curve or movement

   It may be smooth or with ups and downs, but shows an average tendency of growth (Figures 14.1 and 14.2).

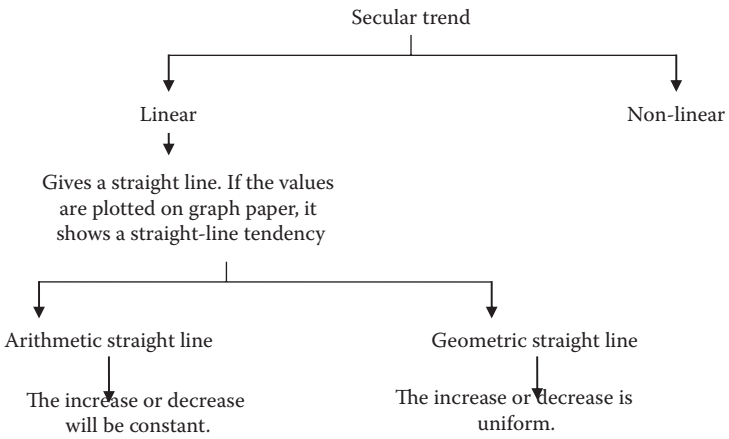


FIGURE 14.1
Tree diagram secular trend.

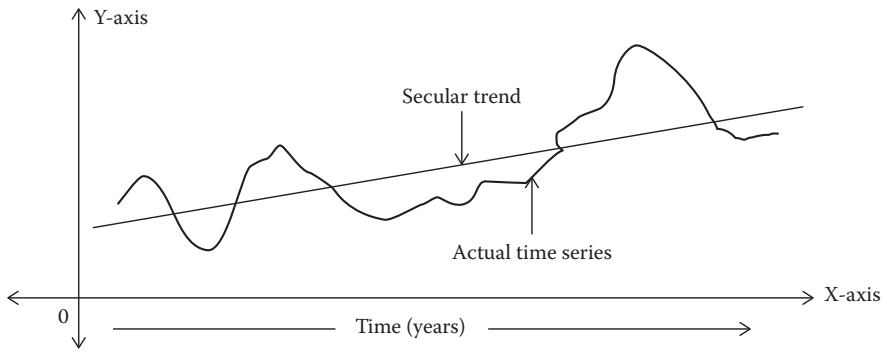


FIGURE 14.2
Secular trend.

14.3.1.4 Uses of Secular Trends

1. The study of the data over a long period of time gives a general idea about the behaviour of the data.
2. Measures of trend describe the features of the data in the past and also the pattern of growth or decline in trend.
3. By using secular trends, one can forecast future behaviour on the basis of the past data.

14.4 Seasonal Variations

Seasonal variations are those periodic movements in business activity that occur regularly every year and have their origin in the nature of the year itself. Since these variations repeat during a period of 12 months, they can be predicted fairly accurately. Seasonal

variation refers to rhythmic changes that occur in periodic movements. It occurs due to natural factors, seasons or man-made traditions such as festivals, social celebrations, marriages and so on. For example,

1. During summer in many cities, most textile showrooms sell a wide variety of cotton textiles, and the same shops sell synthetic and woollen products during winter.
2. During cropping seasons, vegetables or fruits are sold at a cheaper price, but during other seasons they are expensive.

Seasonal variation is evident when the data are recorded at weekly, monthly or quarterly intervals. Although the amplitude of seasonal variations may vary, their period is fixed at 1 year. As a result, seasonal variations do not appear in series of annual figures.

14.4.1 Factors that Cause Seasonal Variations

1. Climate and weather conditions

   The most important factor causing seasonal variation is the climate. Changes in the climate and weather conditions, such as rainfall, humidity and heat, affect different products and industries differently. For example,

   a. During winter, there is greater demand for woollen clothes, hot drinks and so on, whereas in summer, cotton clothes, cold drinks and ice cream have greater sales.
   b. Agriculture is influenced very much by the climate. The effect of the climate is that there are generally two seasons in agriculture – the growing season and the harvesting season – which directly affect the income of the farmer, which in turn, affects the entire business activity.

2. Customs, traditions and habits

   Though nature is primarily responsible for seasonal variations in time series, customs, traditions and habits also have their impact. For example,

   a. On certain occasions, such as Deepawali, Pongal, Dussehra and Christmas, there is a big demand for sweets, and also there is a high demand for cash before the festivals, because people want money for shopping and gifts.
   b. On the first of every month, there are heavy withdrawals, and the bankers have to keep plenty of cash to meet the possible demands on the basis of last month's experience.
   c. Most students buy books in the first few months of the opening of schools and colleges, and thus the sale of books, stationery and so on shows seasonal swings.

14.4.2 Application of Seasonal Variation

1. It analyses seasonal patterns of the variables in a short period and studies the variations in the behaviour of the data.
2. It helps in short-term forecasting and planning.
3. It helps in appraisal of business activities.

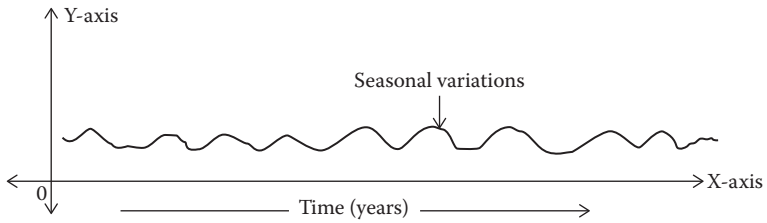


FIGURE 14.3
Seasonal variations.

4. Seasonal indexes are helpful in scheduling purchases, inventory control, personnel requirements, seasonal financing, and selling and advertising programmes.

For example, a housewife may buy fruits for canning or preserving at the peak of the season, when the prices are low and quality high. Seasonal fluctuations may also be ironed out in order that the intra-year fluctuations may be less pronounced. Thus, attempts were made in the United States to build up winter demand for ice cream by advertising: 'Ice-Cream is one of your best foods. Eat one plate a day'.

14.4.3 Features of Seasonal Variations

Seasonal variation in a time series is the repositive, recurrent pattern of change which occurs with a year or shorter time period. The term `seasonal' is meant to include any kind of variation which is periodic in nature and whose repeating cycles are of relatively short duration (Figure 14.3).

14.5 Cyclical Variations

Cyclical fluctuations or business cycle movements are recurrent up and down movements around secular trend levels which have a duration anywhere from 2 to 15 years.

Most of the time series relating to economics and business show some kind of cyclical variations. Cyclical fluctuations are long-term movements that represent consistently recurring rises and declines in activity.

14.5.1 Business Cycle

A business cycle consists of the recurrence of the up and down movements of business activity from some sort of statistical trend or 'normal'. By 'normal' we mean some kind of statistical average; we do not mean that there is anything very permanent or special.

14.5.1.1 Periods or Phases in the Business Cycle

1. Prosperity
2. Decline

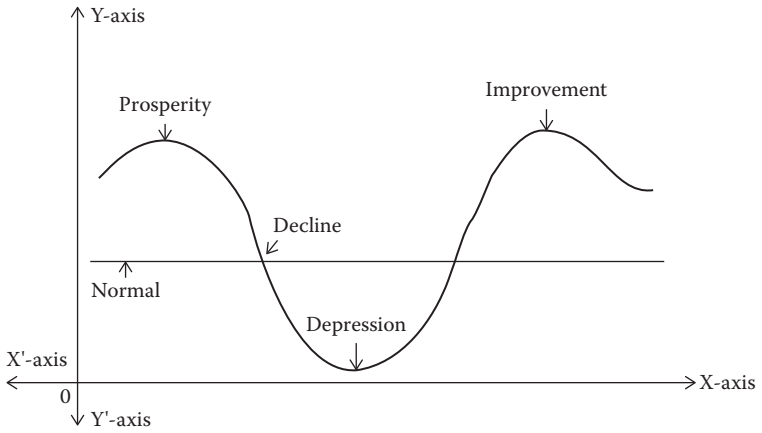


FIGURE 14.4
Phases of business cycle.

3. Depression
4. Improvement

Each phase changes gradually into the phase that follows it in the order given. Figure 14.4 illustrates a business cycle.

1. Prosperity phase

   In this phase, the public is optimistic – business is booming, prices are high and profits are easily made. There is a considerable expansion of business activity, which leads to an overdevelopment. It is then difficult to secure deliveries, and there is a shortage of transportation facilities, which has a tendency to cause large inventories to be accumulated during the time of highest prices.

2. Decline phase

   Wages increase and labour efficiency decreases. The strong demand for money causes interest rates to rise to a high level, while doubt enters the banker's mind as to the advisability of granting further loans. This situation causes businesses to make price discounts in order to secure the necessary cash. It then follows the expectation of further reductions, and the situation becomes critical instead of better. Buyers wait for lower prices, and this leads to a decline in business.

3. Depression phase

   Then follows a period of pessimism in trade and industry; factories close, businesses fail and there is widespread unemployment, while wages and prices are low. These conditions characterise the period of depression.

4. Improvement phase

   After a period of rigid economic liquidation and reorganisation, money accumulates and seeks a use. Then follows a period of increasing business activity with rising prices, a period of improvement or recovery. The improvement period generally develops into the prosperity period, and a business cycle is completed. These movements are constantly repeated in the order given as the cycle completes its swing (Figure 14.5).

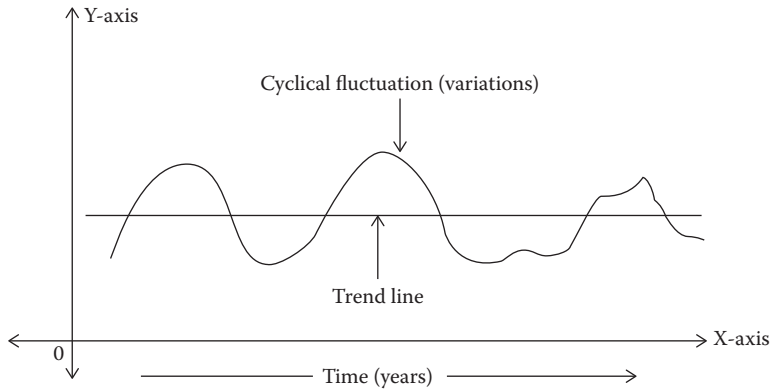


FIGURE 14.5
Cyclical variations.

14.5.2 Importance of Measuring Cyclical Variation

1. It is extremely useful in framing suitable policies for stabilising the level of business activity, that is, for avoiding periods of booms and depressions, as both are bad for an economy – particularly depression, which brings about a complete disaster and shatters the economy.
2. Study of past fluctuations is useful to determine the features of past behaviour and helps in the study of fluctuations of business.
3. Measures of cycles help to forecast the future and prepare plans with the help of projection of past cycles. In this way, future changes can be estimated.

14.5.3 Limitations of Measuring Cyclical Variation

1. Business cycles do not show regular periodicity – they differ widely in timing, amplitude and pattern, which makes their study very tough and tedious.
2. The cyclical variations are mixed with erratic, random or irregular forces, which makes it impracticable to isolate separately the effects of cyclical and irregular forces.

See Table 14.1 for a comparison of the business cycle versus seasonal variations.

14.6 Irregular Variations, Random Movements, Unpredictable Movements, Erratic Variations or Accidental Variations

Irregular variations refer to variations in business activities that do not repeat in a definite pattern.

Irregular variations are caused by such isolated special occurrences as floods, earthquakes, strikes and wars. Sudden changes in demand or very rapid technological progress may also be included in this category. By their very nature, these movements are

TABLE 14.1

Business Cycle vs. Seasonal Variations

Points	Cyclical Variations (Business Cycle)	Seasonal Variations
Duration	Longer than a year. May be of any duration, but normally the period is 2–10 years.	Occur during a year
Periodicity	Do not ordinarily exhibit regular periodicity, as successive cycles vary widely in timing, amplitude and pattern.	Exhibit regular periodicity
Fluctuations	Result from a different set of causes	No fluctuations
Effect of factors	The periods of prosperity, decline, depression and improvement, viewed as four phases, are generated by factors other than weather, social customs and those that create seasonal patterns.	Affected by factors such as weather, social customs and those that create seasonal patterns

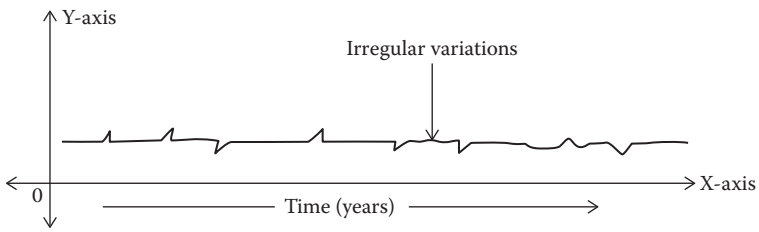


FIGURE 14.6
Irregular variations.

very irregular and unpredictable. Quantitatively, it is almost impossible to separate out the irregular movements from the cyclical movements (Figure 14.6).

14.6.1 Reasons for Recognising Irregular Movements

1. To suggest that on occasions it may be possible to explain certain movements in the data as due to specific causes and to simplify further analysis.
2. To emphasise the fact that predictions of economic conditions are always subject to a degree of error due to the unpredictable erratic influences that may occur.

14.7 Measurement of Trend

1. To find out trend characteristics in and of themselves.

   In studying trend in and of itself, we ascertain the growth factor.

For example, we can compare the growth in the steel industry with growth in the economy as a whole or with the growth in other industries, or we can compare the growth in one firm of the steel industry with the growth in the industry as a whole.

2. To enable us to eliminate trend in order to study other elements.

   The elimination of trend leaves us with seasonal, cyclical and irregular factors.

Methods for determining trend

1. Freehand or graphic method
2. Semi-average method
3. Moving average method
4. Method of least squares

14.7.1 Freehand or Graphic Method of Measuring Trend

Using this method, the given data are plotted on graph paper, and a trend line is fitted to the data just by inspecting the graph of the series.

When a trend line is fitted by the freehand method, an attempt should be made to make it conform to the following conditions For example,

1. It should be smooth – either a straight line or a combination of long gradual curves.
2. The sum of the vertical deviations from the trend of the annual observations above the trend should equal the sum of the vertical deviations from the trend of the observations below the trend.
3. The sum of the squares of the vertical deviations of the observations from the trend should be as small as possible.
4. The trend should bisect the cycles so that the area above the trend equals that below the trend, not only for the entire series but for each full cycle. This condition cannot always be met fully, but an attempt should be made to observe it as closely as possible.

Exercise

Fit a trend line to the data in Table 14.2 by the freehand method (Figure 14.7).

Solution

The trend line drawn by the freehand method can be extended to predict future values. However, since the freehand curve fitting is too subjective, this method should not be used as a basis for predictions.

TABLE 14.2

Production of Zinc for 9 Years

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012
Production of zinc (tonnes)	22	24	26	23	25	27	25	28	27

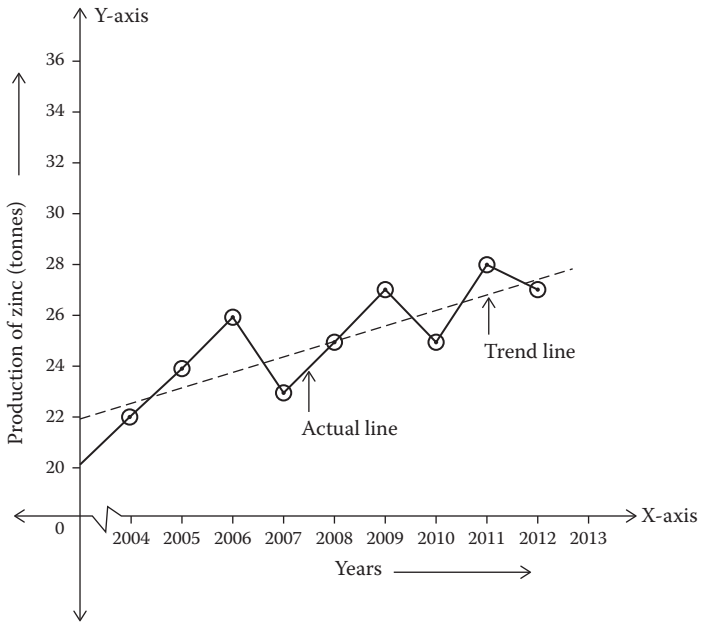


FIGURE 14.7
Trend by the freehand method.

Merits

1. This is the simplest method of measuring trend.
2. This method is very flexible in that it can be used regardless of whether the trend is a straight line or curve.
3. A trend line drawn by a statistician experienced in computing trends and having knowledge of the economic history of the concern or the industry under analysis may be a better expression of the secular movement than a trend fitted by the use of a rigid mathematical formula, which, while providing a good fit to the points, may have no other logical justification.

Limitations

1. This method is highly subjective, because the trend line depends on the personal judgement of the investigator, and therefore, different persons may draw different trend lines from the same set of data.
2. Since freehand curve fitting is subjective, it cannot have much value if it is used as a basis for predictions.
3. Although this method appears simple and direct, in actuality, as experienced statisticians can verify, it is very time consuming to construct a freehand trend if a careful and conscientious job is done.

14.7.2 Semi-Average Method

When this method is used, the given data is divided into two parts, preferably with the same number of years. For example, if we are given data from 2000 to 2017, that is, over a period of 18 years, the two equal parts will be 9 years, that is, from 2000 to 2008 and from 2008 to 2017. In the case of an odd number of years such as 7, 9 or 11, two equal parts can

be made simply by omitting the middle year. If data is given for 9 years from 2000 to 2008, the two equal parts will be from 2000 to 2003 and from 2005 to 2008; the middle year, 2004, will be omitted.

After the data has been divided into two parts, an average (arithmetic mean) of each part is obtained. We thus get two points. Each point is plotted at the midpoint of the class interval covered by the respective part, and then the two points are joined by a straight line, which gives us the required trend line. The line can be extended downwards or upwards to get intermediate values or to predict future values.

Exercise

Fit a trend line to the data in Table 14.3 by the method of semi-averages.

Solution

Since 7 years are given, the middle year will be left out, and an average of the first 3 years and the last 3 years will be obtained.

The average of the first 3 years is

$$\frac{104+107+116}{3}=\frac{327}{3}=109$$

And the average of the last 3 years is

$$\frac{110+118+114}{3}=\frac{342}{3}=114$$

Thus, we get two points, 109 and 114, which will be plotted corresponding to their respective middle years, that is, 2008 and 2012. By joining these two points, we obtain the required trend line. The line can be extended and can be used either for prediction or for determining intermediate values.

The actual data and the trend line are shown in Figure 14.8.

When there are even numbers of years, such as 6, 8, 10 or 12, two equal parts can easily be formed and an average of each part obtained. However, when the average is to be centred, there will be a problem. For example, if the data relates to 2002, 2003, 2004 and 2005, the average will be centred corresponding to 1 July 2003, that is, in the middle of 2003 and 2004.

Exercise

The sale of a commodity in tonnes varied from January 2015 to December 2015 as shown in Table 14.4.

Fit a trend line by the method of semi-averages.

TABLE 14.3

Sales of Firm P for 7 Years

Years	2007	2008	2009	2010	2011	2012	2013
Sales of Firm P (thousand units)	104	107	116	112	110	118	114

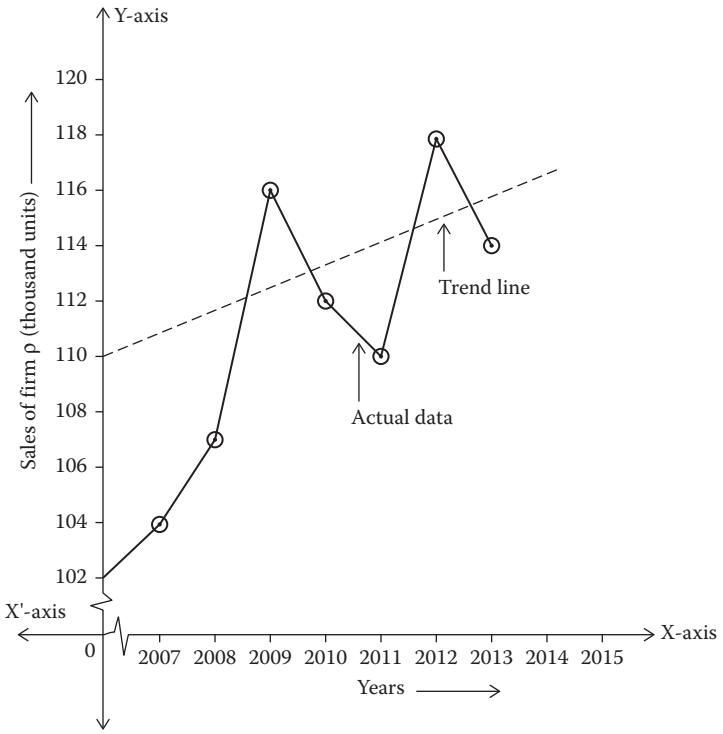


FIGURE 14.8
Trend by the method of semi-averages.

TABLE 14.4

Sales of Commodity for 12 Months

290	310	290	290	280	250
240	240	230	210	220	210

TABLE 14.5

Calculation of Trend Value by the Method of Semi-Averages

Month	Sales (Tonnes)		Month	Sales (Tonnes)	
January	290	1710 (Total) of first 6 months	July	240	1350 (Total) of last 6 months
February	310		August	240	
March	290		September	230	
April	290		October	210	
May	280		November	220	
June	250		December	210	

Solution

Calculation of trend value by the method of semi-averages (Table 14.5)

$$\text{Average of the first half} = \frac{1710}{6} = 285 \text{ tonnes}$$

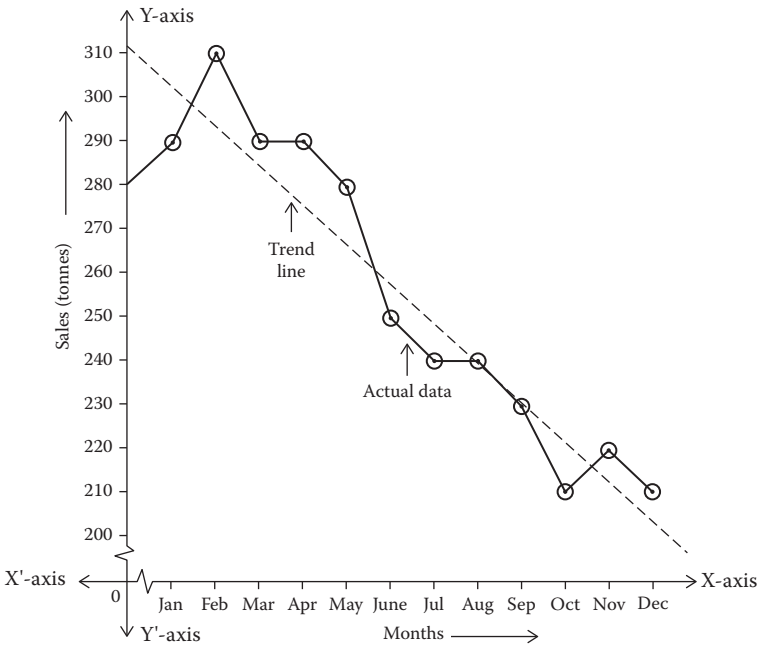


FIGURE 14.9
Trend by the method of semi-averages.

$$\text{Average of the second half} = \frac{1350}{6} = 225 \text{ tonnes}$$

These two figures, namely, 285 and 225, will be plotted at the middle of their respective periods, that is, at the middle of March–April and of September–October 2015.

By joining these two points, we get a trend line that describes the given data (Figure 14.9).

Merits

1. This method is simple to understand compared with the moving average method and the method of least squares.
2. This is an objective method of measuring trend, as everyone who applies the method is bound to get the same result (leaving aside arithmetical mistakes).

Limitations

1. This method assumes a straight-line relationship between the plotted points regardless of whether that relationship exists or not.
2. If there are extremes in either half or both halves of the series, then the trend line is not a true picture of the growth factor.

14.7.3 Moving Average Method

When a trend is to be determined by the method of moving averages, the average value for a number of years (or months or weeks) is secured, and this average is taken as the normal or trend value for the unit of time falling at the middle of the period covered in the

calculation of the average. The effect of the averaging is to give a smoother curve, lessening the influence of the fluctuations that pull the annual figures away from the general trend.

Selection of period

While applying this method, it is necessary to select a period for the moving average, such as 3-yearly, 5-yearly or 8-yearly moving average.

The period of moving average is to be decided in the light of the length of the cycles. Since the moving average method is most commonly applied to data that is characterised by cyclical movements, it is necessary to select a period for moving average that coincides with the length of the cycle; otherwise, the cycle will not be entirely removed. This danger is more severe the shorter the time period represented by the average.

When the period of moving average and the period of the cycle do not coincide, the moving average will display a cycle that has the same period as the cycle in the data.

Often, we find that the cycles in the data are not of uniform length. In such a case, we should take a moving average period equal to or somewhat greater than the average period of the cycle in the data. Ordinarily, the necessary period will range between 3 and 10 years for general business series, but even longer periods are required for certain types of data.

The formula for a 3-yearly moving average will be

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3}, \frac{d+e+f}{3} \cdots$$

and for a 5-yearly moving average

$$\frac{a+b+c+d+e}{5}, \frac{b+c+d+e+f}{5}, \frac{c+d+e+f+g}{5} \cdots$$

Exercise

Find the trend of bank clearances by the method of moving average (assume a 5-yearly cycle) (see Table 14.6).

Solution

Calculation of trend values by the method of moving averages (Table 14.7).

Exercise

Calculate 5-yearly and 7-yearly averages for the data of number of commercial industrial failures in a country during 2000–2015 (Table 14.8).
Also plot the actual and trend values on a graph.

TABLE 14.6

Bank Clearances for 13 Years

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Bank clearances (Crores of Rs.)	53	79	76	66	69	94	105	87	79	104	97	92	101

TABLE 14.7

Calculation of Trend Values by the Method of Moving Averages

Year	Bank Clearances (Rs. Crores)	5-Yearly Moving Totals	5-Yearly Moving Averages
2000	53	—	—
2001	79	—	—
2002	76	343	68.6
2003	66	384	76.8
2004	69	410	82.0
2005	94	421	84.2
2006	105	434	86.8
2007	87	469	93.8
2008	79	472	94.4
2009	104	459	91.8
2010	97	473	94.6
2011	92	—	—
2012	101	—	—

TABLE 14.8

Number of Commercial Industrial Failures in a Country during 2000–2015

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
No. of failures	23	26	28	32	20	12	12	10	9	13	11	14	12	9	3	1

Solution

Calculation of 5-yearly and 7-yearly moving average (Table 14.9)

Even Period and Moving Average

If the moving average is an even-period moving average, say, 4-yearly or 6-yearly, the moving total and moving average, which are placed at the centre of the time span from which they are computed, fall between two time periods.

This placement is inconvenient, since the moving average so placed would not coincide with an original time period. We, therefore, synchronise moving averages and original data. This process is called *centring* and always consists of taking a two-period moving average of the moving averages.

Exercise

Work out the centred 4-yearly moving average for the data in Table 14.10.

Solution

Calculation of the centred 4-yearly moving average (Table 14.11).

Exercise

Assume a 4-yearly cycle, calculating the trend by the method of moving averages from the data in Table 14.12 relating to the production of tea in India.

TABLE 14.9

Calculation of 5-Yearly and 7-Yearly Moving Average

Year	No. of Failures	5-Yearly Moving Total	5-Yearly Moving Average	7-Yearly Moving Total	7-Yearly Moving Average
2000	23	—	—	—	—
2001	26	—	—	—	—
2002	28	129	25.8 or 26	—	—
2003	32	118	23.6 = 24	153	21.9 or 22
2004	20	104	20.8 = 21	140	20.0 = 20
2005	12	86	17.2 = 17	123	17.6 = 18
2006	12	64	12.6 = 13	108	15.4 = 15
2007	10	56	11.2 = 11	87	12.4 = 12
2008	9	55	11.0 = 11	81	11.6 = 12
2009	13	57	11.4 = 11	81	11.6 = 12
2010	11	59	11.8 = 12	78	11.1 = 11
2011	14	59	11.8 = 12	71	10.1 = 10
2012	12	42	9.8 = 10	63	9.0 = 9
2013	9	39	7.9 = 8	—	—
2014	3	—	—	—	—
2015	1	—	—	—	—

TABLE 14.10

Tonnage of Cargo Cleared for 12 Years

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Tonnage of cargo cleared	1102	1250	1180	1340	1212	1317	1452	1549	1586	1476	1624	1586

TABLE 14.11

Calculation of Centred 4-Yearly Moving Average

Year	Tonnage of Cargo Cleared	4-Yearly Moving Total	4-Yearly Moving Average	4-Yearly Centred Moving Average
2002	1102	—	—	—
2003	1250	← 4872	1218.00	—
2004	1180	← 4982	1245.50	← 1213.75
2005	1340	← 5049	1262.25	← 1253.87
2006	1212	← 5321	1330.25	← 1296.25
2007	1317	← 5530	1382.50	← 1356.37
2008	1452	← 5904	1476.00	← 1429.25
2009	1549	← 6063	1515.75	← 1495.87
2010	1586	← 6235	1558.75	← 1537.25
2011	1476	← 6272	1568.00	← 1563.37
2012	1624	—	—	—
2013	1586	—	—	—

TABLE 14.12

Production of Tea in India for 10 Years

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Production of tea (tonnes)	464	515	518	467	502	540	557	571	586	612

TABLE 14.13

Calculation of Trend by the Moving Average Method

Year	Production (Tonnes)	4-Yearly Moving Total	4-Yearly Average	4-Yearly Moving Average Centred
2004	464	—	—	—
2005	515	←1964	491.0	
2006	518	←2002	500.50	←495.7
2007	467	←2027	506.75	←503.6
2008	502	←2066	516.50	←511.6
2009	540	←2170	542.50	←529.5
2010	557	←2254	563.50	←553.0
2011	571	←2326	581.50	←572.5
2012	586	—	—	—
2013	612	—	—	—

Solution

Calculation of trend by the moving average method (Table 14.13).

Merits

1. This method is simple as compared with the method of least squares.
2. It is a flexible method of measuring trend for the reason that if a few more figures are added to the data, the entire calculations are not changed – we only get some more trend values.
3. If the period of moving average happens to coincide with the period of cyclical fluctuations in the data, such fluctuations are automatically eliminated.
4. The moving average follows the general movements of the data, and its shape is determined by the data rather than the statistician's choice of a mathematical function.
5. It is particularly effective if the trend or a series is very irregular.

Limitations

1. Trend values cannot be computed for all the years. The longer the period of moving average, the greater the number of years for which trend values cannot be obtained.
2. Great care has to be exercised in selecting the period of moving average. No hard and fast rules are available for the choice of the period, and one has to use one's own judgement.
3. Since the moving average is not represented by a mathematical function, this method cannot be used in forecasting, which is one of the main objectives of trend analysis.

4. Theoretically, we say that if the period of moving average happens to coincide with the period of the cycle, the cyclical fluctuations are completely eliminated, but in practice, since the cycles are by no means perfectly periodic, the lengths of the various cycles in any given series will usually vary considerably, and, therefore, no moving average can completely remove the cycle.
5. When the trend situation is not linear (a straight line), the moving average lies either above or below the true sweep of the data. Consequently, the moving average is appropriate for trend computations only when
   a. The purpose of investigation does not call for current analysis or forecasting
   b. The trend is linear
   c. The cynical variations are regular in both period and amplitude.

Unfortunately, these conditions are encountered very infrequently.

14.7.4 The Method of Least Squares

This method is most widely used in practice. It is a mathematical method, and with its help a trend line is fitted to the data in such a manner that the following two conditions are satisfied.

1. $\Sigma(Y - Y_e) = 0$

   That is, the sum of deviations of the actual values of Y and the computed values of Y is zero.

2. $\Sigma(Y - Y_e)^2$ is least

   That is, the sum of the squares of the deviations of the actual and computed values is least from this line. That is why this method is called the method of least squares. The line obtained by this method is known as the *line of best fit*.

   This method of least squares may be used to fit either a straight-line trend or a parabolic trend.

The straight-line trend is represented by the equation

$$Y_e = a + bX$$

where:

Y_e = Used to designate the trend values to distinguish them from actual Y values.
a = Y-intercept or the value of the Y variable when $X = 0$.
b = The slope of the line, or the amount of change in the Y variable that is associated with a change of one unit in the X variable.
X = Variable in time series analysis representing time.

Important points for fitting of straight-line trend by the least-squares method

1. Select year as origin.
2. Decide unit of time represented by X, such as half year, 1 year or 5 years.
3. Decide type of units measured by Y, such as production (in tonnes quintals etc.), sales (in rupees, dollars, pounds, etc.), prices (in rupees, dollars, pounds, etc.) and employment of workers (in thousands, lakhs, crores etc.).

Calculation of constants a and b

The values of the constants a and b are determined by solving the following two normal equations:

$$\Sigma Y = Na + b\Sigma X \tag{14.1}$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \tag{14.2}$$

where:

N = Number of years (months or any other period) for which data is given
Equation 14.1 = Nearly the summation of the given function
Equation 14.2 = The summation of X multiplied into the given function
a = Arithmetic mean of Y
b = Rate of change

We can measure the variable X from any point of time as the origin, such as the first year. But the calculations are very much simplified when the midpoint in time is taken as the origin, because in that case, the negative values in the first half of the series balance out the positive values in the second half, so that $\Sigma X = 0$.

In other words, the time variable is measured as a deviation from its mean.

Since $\Sigma X = 0$, Equations 14.1 and 14.2 would take the form

$$\Sigma Y = Na$$

$$\Sigma XY = b\Sigma X^2 \tag{14.3}$$

Now, the values of a and b can be determined easily.

Since $\sum Y = Na$

$$\therefore a = \frac{\Sigma Y}{N} = \bar{Y}$$

Since $\Sigma XY = b\Sigma X^2$

$$\therefore b = \frac{\Sigma XY}{\Sigma X^2}$$

The constant a gives the arithmetic means of Y, and the constant b indicates the rate of change.

Note:

1. In the case of an odd number of years, when the deviations are taken from the middle year, ΣX will always be zero, provided there is no gap in the data given.

2. In the case of an even number of years, ΣX will be zero if the X origin is placed midway between the two middle years.

For example, if the years are 2011, 2012, 2013, 2014, 2015 and 2016, we can take deviations from the middle year, 2013.5. Thus,

1. The deviations would be –2.5, –1.5, –0.5, +0.5, +1.5 and +2.5 for the various years and
2. The total ΣX would be zero.

Hence, in both odd as well as even numbers of years, we can use the simple procedure of determining the values of the constants a and b.

Exercise

Table 14.14 shows the figures of production (in thousands of maunds) of a sugar factory.

1. Fit a straight-line trend to these figures.
2. Plot these figures on a graph and show the trend line.

Solution

1. Fitting the straight-line trend (Table 14.15)/
   The equation of the straight-line trend is

$$Y_e = a + bX$$

TABLE 14.14

Production of a Sugar Factory

Year	2010	2011	2012	2013	2014	2015	2016
Production (thousand maunds)	80	90	92	83	94	99	92

TABLE 14.15

Fitting the Straight-Line Trend

Year	Production (Thousand Maunds) (Y)	X	XY	X^2	Trend values (Ye)
2010	80	–3	–240	9	84
2011	90	–2	–180	4	86
2012	92	–1	–92	1	88
2013	83	0	0	0	90
2014	94	1	94	1	92
2015	99	2	198	4	94
2016	92	3	276	9	96
N = 7	**ΣY = 630**	**ΣX = 0**	**ΣXY = 56**	**$\Sigma X^2 = 28$**	**$Y_e = 630$**

Since $\Sigma X = 0$,

$$a = \frac{\Sigma Y}{N}, b = \frac{\Sigma XY}{\Sigma X^2}$$

Here, $\Sigma Y = 630$, $N = 7$, $\Sigma XY = 56$, $\Sigma X^2 = 28$

$$\therefore \quad a = \frac{\Sigma Y}{N} = \frac{630}{7} = 90$$

And

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{56}{28} = 2$$

Hence, the equation of the straight-line trend is

$$Y_e = 90 + 2X$$

Origin, 2013
X units, one year
Y units, production in thousands of maunds

$$\text{For } X = -3,\ Y_0 = 90 + 2\ (-3) = 84$$

$$\text{For } X = -2,\ Y_0 = 90 + 2\ (-2) = 86$$

$$\text{For } X = -1,\ Y_0 = 90 + 2\ (-1) = 88$$

Similarly, by putting $X = 0, 1, 2, 3$, we can obtain other trend values. However, since the value of b is constant, only the first trend value need be obtained, and then, if the value is positive, we may continue adding the value of b to every preceding value.

2. The graph of this data is shown in Figure 14.11.

For example, in the above case, for 2013, it will be 86+2=88, and so on. If b is negative, then instead of adding, we will deduct.

Exercise

Apply the method of least squares to obtain the trend values from the following data and show that $\Sigma(Y - Y_e) = 0$ (Table 14.16).

Also, predict the sales for the year 2014.

Solution

Calculation of trend values by method of least squares (Table 14.17).

The equation of the straight-line trend is

$$Y_e = a + bX$$

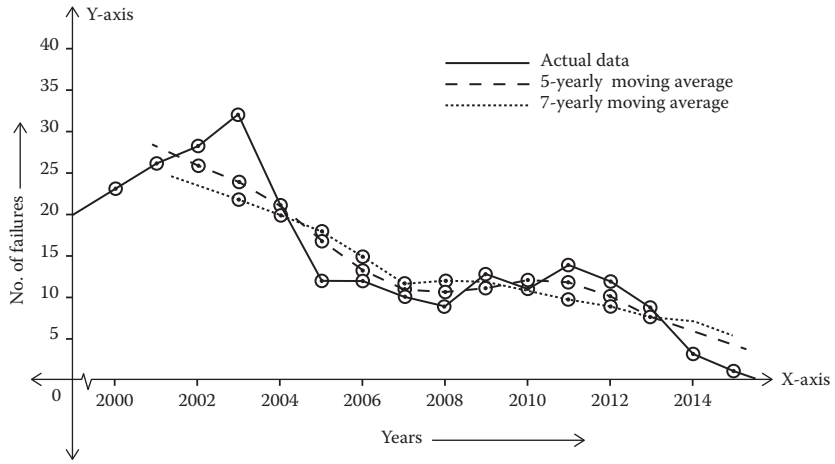


FIGURE 14.10
Trend by the method of moving averages.

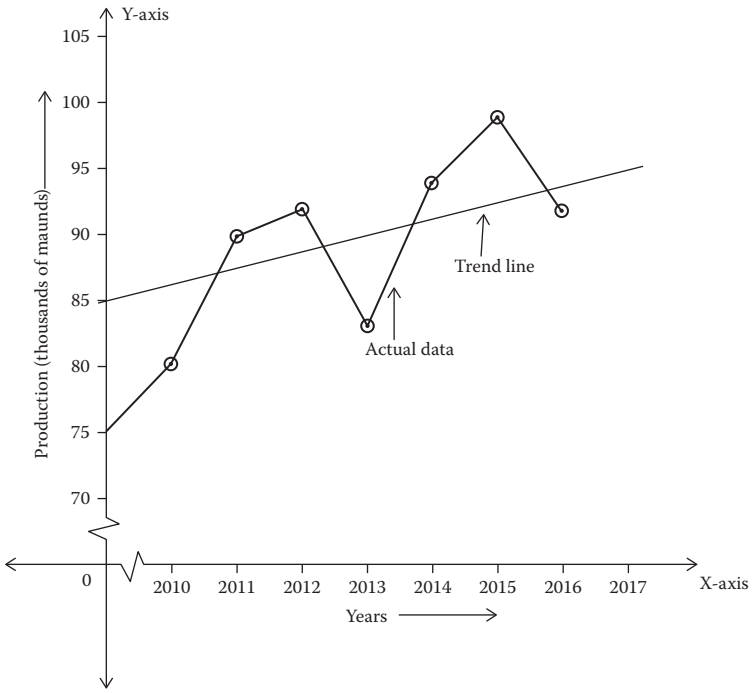


FIGURE 14.11
Linear trend by the method of least squares.

TABLE 14.16

Data of Sales for 5 Years

Year	**2004**	**2005**	**2006**	**2007**	**2008**
Sales (tonnes)	100	120	110	140	80

TABLE 14.17

Calculation of Trend Values by Method of Least Squares

Year	Sales (Y)	Deviations from Middle Year (X)	XY	X^2	Y_e	$(Y - Y_e)$
2004	100	−2	−200	4	114	−14
2005	120	−1	−120	1	112	+18
2006	110	0	0	0	110	0
2007	140	1	140	1	108	32
2008	80	2	160	4	106	−26
N = 5	**ΣY = 550**	**ΣX = 0**	**ΣXY = −20**	**$\Sigma X^2 = 10$**	**$\Sigma Y_e = 0$**	**Σ(Y − Ye) = 0**

Since $\Sigma X = 0$

$$a = \frac{\Sigma Y}{N}, b = \frac{\Sigma XY}{\Sigma X^2}$$

$$\Sigma Y = 550,\ N = 5,\ \Sigma XY = -20,\ \Sigma X^2 = 10$$

Substituting the values

$$a = \frac{550}{5} = 110 \quad b = -\frac{20}{10} = -2$$

The required equation is

$$Y_e = 110 - 2X$$

For $X = -2,\ Y_e = 110 - 2(-2) = 114$

Now, the other trend values will be obtained by deducting the value of b from the preceding value. Thus, for 2005, the trend value will be 114 − 2 = 112 (since the value of b is negative). For 2014, likely sales = 94 tonnes (since X would be +8 for 2014).

Exercise

Fit a straight-line trend by the method of least squares to the data in Table 14.18. Assuming that the same rate of change continues, what would be the predicted earnings for the year 2012?

TABLE 14.18

Earnings (Rs. in Thousands) for 8 Years

Year	2003	2004	2005	2006	2007	2008	2009	2010
Earnings (Rs. in thousands)	38	40	65	72	69	60	87	95

TABLE 14.19

Fitting of Straight-Line Trend by the Method of Least Squares

Year	Earnings (in Thousands) (Y)	Deviations from 2006.5	Deviations Multiplied by 2 (X)	XY	X^2
2003	38	−3.5	−7	−266	49
2004	40	−2.5	−5	−200	25
2005	65	−1.5	−3	−195	9
2006	72	−0.5	−1	−72	1
2007	69	0.5	1	69	1
2008	60	1.5	3	180	9
2009	87	2.5	5	435	25
2010	95	3.5	7	665	49
N = 8	**ΣY = 526**		**ΣX = 0**	**ΣXY = 616**	**$\Sigma X^2 = 168$**

Solution

Fitting of straight-line trend by the method of least squares (Table 14.19)

$$Y_c = a + bX$$

$$a = \frac{\Sigma Y}{N} = \frac{526}{8} = 65.75$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{616}{168} = 3.66$$

$$Y = 65.75 + 3.66X$$

For 2012, X will be +11.
When X is +11, Y will be

$$Y = 65.75 + 3.66(11)$$

$$= 65.75 + 40.37 = 106.$$

Thus, the estimated earnings for the year 2012 are Rs. 106.12 thousands.

Note:

The same result will be obtained if we do not multiply the deviations by 2. But in that case, our computations would be more difficult, as seen here (Table 14.20):

$$a = \frac{\Sigma Y}{N} = \frac{526}{8} = 65.75$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{308}{48} = 7.33$$

TABLE 14.20

Fitting of Straight-Line Trend by the Method of Least Squares

Year	Sales (Thousands of Rupees) (Y)	Deviations from 2006.5 (X)	XY	X^2
2003	38	−3.5	−133.00	12.25
2004	40	−2.5	−100.00	6.25
2005	65	−1.5	−97.50	2.25
2006	72	−0.5	−36.00	0.25
2007	69	0.5	34.50	0.25
2008	60	1.5	90.00	2.25
2009	87	2.5	217.50	6.25
2010	95	3.5	332.50	12.25
N = 8	$\Sigma Y = 526$	$\Sigma X = 0$	$\Sigma XY = 308$	$\Sigma X^2 = 42.00$

TABLE 14.21

Production of a Sugar Factory

Year	2004	2005	2006	2007	2008	2009	2010
Production (thousands of maunds)	77	88	94	85	91	98	90

The advantage of this method is that the value of b gives an annual increment of change rather than six monthly increments, as in the first method. Hence, we will not have to double the value of b to obtain a yearly increment. It is clear from this problem that in the first case, the value of b is half of what we obtain from the second method (b was 3.66 in the first case and 7.33 in the second case).

Exercise

The figures of production (in thousands of maunds) of a sugar factory are given in Table 14.21.

1. Fit a straight line by the least-squares method, and tabulate the trend values.
2. Eliminate the trend. What components of the time series are thus left over?
3. What is the monthly increase in the production of sugar?

Solution

1. The equation of the straight-line trend is

$$Y_e = a + bX$$

Since ΣX is not zero, the values of a and b will be obtained directly by solving the following two normal equations (Table 14.22):

$$\Sigma Y = Na + b\Sigma X \tag{14.4}$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \tag{14.5}$$

TABLE 14.22

Fitting of Straight-Line Trend by the Method of Least Squares

Year	Production (Y)	Taking 2007 as Origin (X)	XY	X^2	Trend Values (Y_e)	$Y - Y_e$
2004	77	–4	–308	16	83.283	–6.283
2005	88	–2	–176	4	86.043	1.957
2006	94	–1	–94	1	87.423	6.577
2007	85	0	0	0	88.803	–3.803
2008	91	1	91	1	93.183	0.817
2009	98	2	196	4	91.563	6.437
2010	90	5	450	25	95.703	–5.703
N = 7	$\Sigma Y = 623$	$\Sigma X = 1$	$\Sigma XY = +159$	$\Sigma X^2 = 51$	$\Sigma Y_e = 623$	$\Sigma(Y - Y_e) = 0$

$$623 = 7a + b \tag{14.6}$$

$$159 = a + 51b \tag{14.7}$$

Multiplying Equation 14.7 by 7, we get

$$1113 = 7a + 357b \tag{14.8}$$

Deducting Equation 14.8 from Equation 14.6,

$$-490 = -356b$$

$$b = \frac{490}{356} = 1.38$$

Substituting the value of b in Equation 14.6,

$$623 = 7a + 1.38$$

$$7a = 623 - 1.38 = 621.62$$

$$A = 88.803$$

So, the equation of the straight-line trend is

$$Y = 88.803 + 1.38X$$

When

$$X = -4, Y = 88.803 + 1.38(-4)$$

$$= 88.803 - 5.52 = 83.283$$

When

$$X = -2, Y = 88.803 + 1.38(-2)$$

$$= 88.803 - 2.76 = 86.043$$

When $X = -1, Y = 88.803 - 1.38 = 87.428$
When $X = 1, Y = 88.803 + 1.38 = 90.183$
When $X = 2, Y = 88.803 + (1.38 \times 2) = 91.563$
When $X = 5, Y = 88.803 + (1.38 \times 5) = 95.703$

2. After eliminating the trend, we are left with cyclical and irregular variations.
3. The monthly increase in the production of sugar is b/12, that is, 1.38/12 = 0.115 thousands of maunds.

Merits

1. This is a mathematical method of measuring trend, and as such, there is no possibility of subjectiveness.
2. The line obtained by this method is called the *line of best fit* because it is this line from where the sum of the positive and negative deviations is zero and the sum of the square of the deviations is least; that is, $\Sigma(\mathbf{Y} - \mathbf{Y}_e) = 0$ and $\Sigma(Y - \mathbf{Y}_e)^2$ is least.

Limitations

Mathematical curves are useful to describe the general movement of a time series, but it is doubtful whether any analytical significance should be attached to them, except in special cases. It is seldom possible to justify on theoretical grounds any real dependence of a variable on the passage of time. Variables do change in a more or less systematic manner over time, but this can usually be attributed to the operation of other explanatory variables.

Mathematical methods of fitting trend are not foolproof; in fact, they can be the source of some of the most serious errors that are made in statistical work. They should never be used unless rigidly controlled by a separate logical analysis.

14.8 Second-Degree Parabola

The simplest example of the non-linear trend is the second-degree parabola, the equation of which is written in the form

$$Y_c = a + bX + cX^2$$

When numerical values for a, b and c have been derived, the trend value for any year may be computed by substituting in the equation the value of X for that year. The values of a, b and c can be determined by solving the following three normal equations simultaneously:

$$(a) \quad \Sigma Y = Na + b\Sigma X + c\Sigma X^2$$

$$(b) \quad \Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

$$(c) \quad \Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

Note:

1. The first equation is merely the summation of the given function.
2. The second is the summation of X multiplied into the given function.
3. The third is the summation of X^2 multiplied into the given function.

When time origin is taken between two middle years, ΣX will be zero. In that case, the above equations are reduced to

$$(a) \quad \Sigma Y = Na + c\Sigma X^2$$

$$(b) \quad \Sigma XY = b\Sigma X^2$$

$$(c) \quad \Sigma X^2 Y = a\Sigma X^2 + c\Sigma X^4$$

The value of b can now directly be obtained from Equation (b) and that of a and c by solving (a) and (b) simultaneously. Thus,

$$a = \frac{\Sigma(Y) - c\Sigma\left(X^2\right)}{N}$$

$$b = \frac{\Sigma XY}{\Sigma X^2}$$

$$c = \frac{N\left(\Sigma X^2 Y\right) - \Sigma\left(X^2\right)\Sigma(Y)}{N\Sigma\left(X^4\right) - \left(\Sigma X^2\right)^2}$$

14.9 Measurement of Seasonal Variations

Most of the phenomena in economics and business show seasonal patterns. When data is expressed annually, there is no seasonal variation. However, monthly or quarterly data frequently exhibits strong seasonal movements, and considerable interest attaches to devising a pattern of average seasonal variation. For example, it is necessary to decide whether weekly, quarterly or monthly indexes are required. This will be decided in the light of the nature of the problem and the type of data available.

14.9.1 Seasonal Index

To obtain a statistical description of a pattern of seasonal variation, it will be desirable to first free the data from the effects of trend, cycles and irregular variations. Once these other components have been eliminated, we can calculate, in index form, a measure of seasonal variations, which is usually referred to as a *seasonal index*.

14.9.2 Criteria for Computing an Index of Seasonal Variation

1. It should measure only the seasonal forces in the data. It should not be influenced by the forces of trend or cycle that may be present.
2. It should modify the erratic fluctuations in the data with an acceptable system of averaging.
3. It should recognise slowly changing seasonal patterns that may be present and modify the index to keep up with these changes.

14.9.3 Methods Used for Measuring Seasonal Variations

1. Method of simple averages (weekly, monthly or quarterly).
2. Ratio-to-trend method or percentage-to-trend method
3. Ratio-to-moving average method or percentages of moving average method
4. Link relative method

14.9.3.1 Method of Simple Averages (Weekly, Monthly or Quarterly)

This is the simplest method of obtaining a seasonal index.

1. Steps for calculating the index:
   a. Arrange the unadjusted data by years and months (or quarters if quarterly data is given).
   b. Find the totals of January, February, March and so on.

      Divide each total by the number of years for which data is given. For example, if we are given monthly data for 4 years, we first obtain the total for each month for 5 years and divide each total by four to obtain an average.
   c. Obtain an average of monthly averages by dividing the total of monthly averages by 12.
   d. Taking the average of monthly averages as 100, compute the percentages of various monthly averages as follows:

Seasonal index for January

$$= \frac{\text{Monthly average for January}}{\text{Average of monthly averages}} \times 100$$

Note:
If, instead of the average of each month, the totals of each month are obtained, we will get the same result.

Exercise

Consumption of monthly electric power in millions of kilowatt hours for street lighting in the United States during 2011–2015 is given in Table 14.23.

Find the seasonal variations by the method of monthly averages.

TABLE 14.23

Monthly Consumption of Electric Power

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
2011	348	281	278	250	231	216	223	245	269	302	325	347
2012	342	309	299	268	249	236	242	262	288	321	342	364
2013	367	328	320	287	269	251	259	284	309	245	367	394
2014	392	349	342	311	290	273	282	305	328	364	379	417
2015	420	378	370	334	314	296	305	330	356	396	422	452

TABLE 14.24

Calculation of Seasonal Indices by the Method of Monthly Averages

	Monthly Consumption of Electric Power					Monthly Total for 5 Years	5-Yearly Average	Percentage
Month	2011	2012	2013	2014	2015			
1	2	3	4	5	6	7	8	9
Jan.	318	342	367	392	420	1,839	367.8	116.1
Feb.	281	309	328	349	378	1,645	329.0	103.9
Mar.	278	299	320	342	370	1,609	321.8	101.6
Apr.	250	268	287	311	334	1,450	290.0	91.4
May	231	249	269	290	314	1,353	270.6	85.5
June	216	236	251	273	296	1,272	254.4	80.3
July	223	242	259	282	305	1,311	262.2	82.8
Aug.	245	262	284	305	330	1,426	285.2	90.0
Sept.	269	288	309	328	356	1,550	310.0	98.0
Oct.	302	321	245	364	396	1,728	345.6	109.1
Nov.	325	342	367	379	422	1,845	369.0	116.6
Dec.	347	364	394	417	452	1,974	394.8	124.7
Total						19,002	3,800.4	1,200
Average						1,583.5	316.7	100

Solution

Calculation of seasonal indices by the method of monthly averages (Table 14.24). The calculations in the table are explained as follows:

1. Column 7 gives the total for each month for the 5 years.
2. In Column 8, each total of Column 7 has been divided by 5 to obtain an average for each month.
3. The average of monthly averages is obtained by dividing the total of monthly averages by 12.
4. In Column 9, each monthly average has been expressed as a percentage of the average of monthly averages. Thus, the percentage for January

$$= \frac{367.8}{316.7} \times 100 = 116.1$$

Percentage for February

$$= \frac{329.0}{316.7} \times 100 = 103.9$$

Percentage for March

$$= \frac{321.8}{316.7} \times 100 = 101.6$$

And so on.

If, instead of monthly data, we are given weekly or quarterly data, we will compute weekly or quarterly averages by following the same procedure.

Merits

This method is the simplest of all methods of measuring seasonality.

Limitations

1. It is not a very good method. It assumes that there is no trend component in the series, that is, computation of seasonal indices (CSI) = 0. But this is not a justified assumption.
2. Most economic series have trends, and, therefore, the seasonal index computed by this method is actually an index of trends and seasonality.
3. The effects of cycles on the original values may or may not be eliminated by the averaging process. This depends on the duration of the cycle and the term of the average, that is, on the number of months included in the average. Thus, this method is seldom of any value. In its simplest form, the method only serves a purpose where no definite trend exists.

14.9.3.2 Ratio-to-Trend Method or Percentage-to-Trend Method

This method of calculating a seasonal index is relatively simple and yet an improvement over the method of simple averages. It is based on the multiplicative model of time series analysis. It assumes that seasonal variation for a given month or quarter is a constant fraction of trend value. Random variations are assumed to disappear when ratios are changed.

Seasonal variation for a given month is a constant fraction of trend. The ratio-to-trend presumably isolates the seasonal factor in the following manner. Trend is eliminated when the ratios are computed. In effect,

$$\frac{T \times S \times C \times I}{T} = S \times C \times I$$

where:

T = trend
S = seasonal component
C = cyclical component
I = irregular component

Random elements are supposed to disappear when the ratios are averaged. A careful selection of the period of years used in the computation is expected to cause the influences

of prosperity or depression to offset each other and thus remove the cycle. For series that are not subject to pronounced cyclical or random influences and for which trend can be computed accurately, this method may suffice.

Steps in the calculation of a seasonal index

1. Trend values are obtained by applying the method of least squares.
2. The next step is to divide the original data month by month by the corresponding trend values and to multiply these ratios by 100.
3. The values so obtained are now free from trend and the problem that remains is to free them also of irregular and cyclical movements.
4. In order to free the values from irregular and cyclical movements, the figures given for the various years for January, February, March and so on are averaged with any one of the usual measures of central value, for instance, the median or the mean. If the data are examined month by month, it is sometimes possible to ascribe a definite cause to usually high or low values.
5. The seasonal index for each month is expressed as a percentage of the average month. The sum of 12 values must equal 1200 or 100%. If it does not, an adjustment is made by multiplying each index by a suitable factor (1200/the sum of the 12 values). This gives the final seasonal index.

Exercise

Compute seasonal variations by the ratio-to-trend method from the information given in Table 14.25.

Solution

For determining seasonal variation by the ratio-to-trend method, first we will determine the trend for yearly data and then convert it to quarterly data (Table 14.26).

$$a = \frac{\Sigma y}{N} = \frac{170}{5} = 34$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{100}{10} = 10$$

Calculation of quarterly trend values

$$\text{Quarterly increment} = \frac{\text{Yearly increment}}{4} = \frac{10}{4} = 2.5$$

TABLE 14.25

Quarterly Data for 5 Years

Year	Quarter I	Quarter II	Quarter III	Quarter IV
2010	10	12	18	20
2011	20	23	27	30
2012	24	26	32	38
2013	38	42	48	52
2014	46	54	56	64

TABLE 14.26

Calculation of Trend by Method of Least Squares

Year	Yearly Total	Yearly Average (y)	x	xy	X^2	$Y = a\,bx$	Trend Value
2010	60	15	−2	−30	4	34 + 10(−2)	14
2011	100	25	−1	−25	1	34 + 10(−1)	24
2012	120	30	0	0	0	34 + 10(0)	34
2013	180	45	1	45	1	34 + 10(1)	44
2014	220	55	2	110	4	34 + 10(2)	54
Total		$\sum y = 170$		$\sum Xy = 100$	$\sum X^2 = 10$		170

Consider 2010. The trend value for the middle quarter that is half of 2nd and half of 3rd is 14. So, the trend value of the 2nd quarter will be

$$14 - \frac{2.5}{2} = 14 - 1.25 = 12.75$$

For the third quarter, it will be

$$14 + \frac{2.5}{2} = 14 + 1.25 = 15.25$$

For the first quarter, it will be

$$12.75 - 2.5 = 10.25$$

For the fourth quarter, it will be

$$15.25 + 2.5 = 17.75$$

Quarterly trend values (Table 14.27)

The given values are to be expressed as the percentages of the corresponding trend values $= \frac{O}{T} \times 100$ (Table 14.28)

The average of quarterly average of trend figures:

$$\frac{92.32 + 95.05 + 104.64 + 108.23}{4} = \frac{400.24}{4} = 100.06$$

Quarterly seasonal index

For $\text{Quarter I} = \frac{92.32}{100.06} \times 100 = 92.26$

For $\text{Quarter II} = \frac{95.05}{100.06} \times 100 = 94.99$

TABLE 14.27

Calculation of Quarterly Trend Values

Year	Quarter I	Quarter II	Quarter III	Quarter IV	Total
2010	10.25	12.75	15.25	17.75	56
2011	20.25	22.75	25.25	27.75	96
2012	30.25	32.75	35.25	37.75	136
2013	40.25	42.75	45.25	47.75	176
2014	50.25	52.75	55.25	57.75	216
Total					680

TABLE 14.28

Given Quarterly Values as Percentage of Trend Values

Year	Quarter I	Quarter II	Quarter III	Quarter IV
2010	$\frac{10}{10.25}\times 100 = 97.56$	94.12	118.03	112.67
2011	98.76	101.10	106.93	108.11
2012	79.33	79.39	90.78	100.66
2013	94.41	98.25	106.08	108.90
2014	91.54	102.37	101.36	110.80
Total	461.60	475.23	523.18	541.14
Average	92.32	95.05	104.64	108.23

For $\text{Quarter III} = \frac{104.64}{100.06}\times 100 = 104.58$

For $\text{Quarter IV} = \frac{108.23}{100.06}\times 100 = 108.17$

Merits

1. The method is more logical and useful.
2. The method considers all the values.
3. When the period of study is short, this method is more useful to obtain seasonal indices.
4. It has an advantage over the moving average procedure too, for it has a ratio-to-trend value for each month for which data are available. Thus, there is no loss of data, as occurs in the case of moving averages.
5. It is simple to compute and easy to understand.

Limitations

If there are pronounced cyclical swings in the series, the trend, whether a straight line or a curve, can never follow the actual data as closely as a 12-month moving average does. In consequence, a seasonal index computed by the ratio-to-moving-average method may be less biased than one calculated by the ratio-to-trend method.

14.9.3.3 Ratio-to-Moving Average Method or Percentages of Moving Average Method

This is the most widely used method of measuring seasonal variations.

Steps

1. Eliminate seasonality from the data by ironing it out of the original data. Since seasonal variations recur every year, that is, since the fluctuations have a time span of 12 months, a centred 12-month moving average tends to eliminate these fluctuations.
2. Express the original data for each month as a percentage of the centred 12-month moving average corresponding to it.
3. Divide each monthly item of the original data by the corresponding 12-month moving average, and list the quotients as 'percent of moving average'.
4. By using the median as an average, we can obtain the typical seasonal relative for each month, which will not be affected by irregular factors.

   Sometimes, a so-called *modified mean* is used as an average for each month. Here, extreme values are omitted before the arithmetic mean is taken. In any array of seasonal relatives for each month, a value, or several values, on one end or both ends that may be relatives is taken. A separate table is prepared, in which the calculations involved in this step are shown. These means are preliminary seasonal indexes. They should average 100% or total 1200 for 12 months by definition.
5. If the total is not equal to 1200 or 100%, an adjustment is made to eliminate the discrepancy. The adjustment consists of multiplying the average of each month obtained in Step 4 by

$$\frac{1200}{\text{the total of the modified mean for 12 months}}$$

This adjustment is made not only to achieve accuracy, but also because when we come to eliminate seasonality from the original data, we do not wish to raise or lower the level of the data unduly. Thus, if a seasonal index aggregates more than 1200 (or averages more than 100), the original data adjusted in terms of the index will total less than the unadjusted original data. If it totals less than 1200, the opposite will be true.

The logical reasoning behind this method follows from the fact that the 12-month moving average can be considered to represent the influence of cycle and trend (C × T). If the actual value for any month is divided by the 12-month moving average centred to that month, presumably cycle and trend are removed. This may be represented by the expression

$$\frac{T \times S \times C \times I}{T \times C} = S \times I$$

Thus, the ratio to the moving average, from which this method gets its name, represents irregular and seasonal influences. If the ratios for each worked over a period of years are then averaged, most random influences will usually be eliminated. Hence, in effect,

$$\frac{S \times I}{1} = S$$

Merits

1. This method is considered to be most satisfactory and, as such, is more widely used in practice than other methods.
2. The index obtained by the ratio-to-moving average method ordinarily does not fluctuate as much as the index based on straight-line trends.
3. Mathematical methods of avoiding the effects of the business cycle are not usually needed, for the 12-month moving average follows the cyclical course of the actual data quite closely. Therefore, the index ratios are often more representative of the data from which they are obtained than is the case in the ratio-to-trend method.
4. Also, the ratio-to-moving average method allows greater flexibility.

Limitations

Seasonal indices cannot be obtained for each month for which data is available. When a 12-month moving average is taken, 6 months at the beginning and 6 months at the end are left out, for which we cannot calculate seasonal indices.

14.9.3.4 Link Relative Method

Among all the methods of measuring seasonal variations, the link relative method is the most difficult.

Steps

1. Calculate the link relatives of the seasonal figures. Link relatives are calculated by dividing the figure of each season by the figure of the immediately preceding season and multiplying it by 100.

$$\frac{\text{Current season's figure}}{\text{Previous season's figure}} \times 100$$

   There percentages are called *link relatives* since they link each month (or quarter or other time period) to the preceding one.
2. Calculate the average of the link relatives for each season. While calculating the average, we might take the arithmetic average, but the median is probably better. The arithmetic average would give undue weight to extreme cases which were not due primarily to seasonal influences.
3. Convert these averages into chain relatives on the basis of the first season.
4. Calculate the chain relatives of the first season on the basis of the last season. There will be some difference between the chain relative of the first season and the chain relative calculated by the previous method (arithmetic average). This difference will be due to the effect of long-term changes. It is therefore, necessary to correct these chain relatives.
5. For correction, the chain relative of the first season calculated by the first method is deducted from the chain relative (of the first season) calculated by the second

method (chain relatives). The difference is divided by the number of seasons. The resulting figure multiplied by 1, 2, 3 (and so on) is deducted from the chain relatives of the second, third, fourth (and so on) seasons, respectively. These are corrected chain relatives.

6. Express the corrected chain relatives as percentages of their averages. These provide the required seasonal indices by the method of link relatives.

Exercise

Apply the method of link relatives to the data in Table 14.29 and calculate seasonal indices.

Solution

Calculation of seasonal indices by method of link relatives (Table 14.30).

In Table 14.30, the figure for correlation has been calculated as follows:

Chain relative of the first quarter (on the basis of the first quarter) = 100
Chain relative of the first quarter (on the basis of the last quarter)

TABLE 14.29

Quarterly Data for 5 Years

Quarter	2001	2002	2003	2004	2005
I	6.0	5.4	6.8	7.2	6.6
II	6.5	7.9	6.5	5.8	7.3
III	7.8	8.4	9.3	7.5	8.0
IV	8.7	7.3	6.4	8.5	7.1

TABLE 14.30

Calculation of Seasonal Indices by Method of Link Relative

Year	Quarter I	Quarter II	Quarter III	Quarter IV
2001	—	108.3	120.0	111.5
2002	62.1	146.3	106.3	86.9
2003	93.2	95.6	143.1	68.8
2004	112.5	80.6	129.3	113.3
2005	77.6	110.6	109.6	88.8
Arithmetic average	$\frac{345.4}{4} = 86.35$	$\frac{541.4}{5} = 108.28$	$\frac{608.3}{5} = 121.66$	$\frac{469.3}{5} = 93.86$
Chain relatives	100	$\frac{100 \times 108.28}{100} = 108.28$	$\frac{121.66 \times 108.28}{100} = 131.7$	$\frac{93.86 \times 131.7}{100} = 123.6$
Corrected chain relatives	100	$108.28 - 1.675 = 106.605$	$131.70 - 3.35 = 128.35$	$123.6 - 5.025 = 118.575$
Seasonal indices	100	$\frac{106.28}{113.385} \times 100 = 93.73$	$\frac{128.35}{113.385} \times 100 = 113.2$	$\frac{118.575}{113.385} \times 100 = 104.6$

$$= \frac{86.35 \times 123.6}{100} = 106.7$$

Difference between these chain relatives

$$= 106.7 - 100 = 6.7$$

Difference per quarter

$$= \frac{6.7}{4} = 1.675$$

Adjusted chain relatives are obtained by subtracting 1×1.675, 2×1.675 and 3×1.675 from the chain relatives of the second, third and fourth quarters, respectively.

Calculation of seasonal variations indices
Average of corrected chain relatives

$$= \frac{100 + 106.605 + 128.35 + 118.575}{4} = \frac{435.54}{4} = 113.385$$

Seasonal variation index

$$= \frac{\text{Correct chain relatives} \times 100}{113.85}$$

14.10 Summary

Some procedures for time series analysis have been described in this chapter with a view to making more accurate and reliable forecasts of the future. Quite often, the question that puzzles a person is how to select an appropriate forecasting method. Many times, the problem context or time horizon involved will decide the method or limit the choice of methods.

The decomposition method has been discussed. Here, the time series is broken up into seasonal, trend, cycle and random components from the given data and reconstructed for forecasting purposes. A detailed example to illustrate the procedure is also given.

REVIEW QUESTIONS

1. Give the meaning of time series.
2. What are the components of time series?
3. What do you mean by trend?

4. What do you understand by seasonal indices?
5. What is irregular variation?
6. What are the methods used for measuring trend?
7. Define the method of least squares to measure trend.
8. What are the methods used for measuring seasonal variations?
9. Explain clearly the meaning of time series analysis. Indicate the importance of such analysis in business.
10. What is a time series? Mention its important components. Explain these components with examples. Discuss briefly the methods of smoothing a time series.
11. a. What are the different components of fluctuation in a time series? Elucidate the methods available for measuring the trend components and their relative merits and demerits.
    b. What are the limitations and advantages of the moving averages method of trend fitting?
12. What is trend? How would you find out the trend values by the method of least squares? Illustrate by a numerical example.
13. Explain how a growth factor, a decline factor, a seasonal factor and a cyclical factor affect a variable over a period of time.
14. What do you understand by the term *moving average*? Indicate its uses.
15. What is a secular trend? Discuss any two methods of isolating trend values in a time series.
16. What is meant by seasonal fluctuation? State the procedure for obtaining a seasonal index by the method of monthly averages.
17. Explain the different methods of measuring seasonal components in time-series data. How will you eliminate the seasonal components?

SELF-PRACTICE PROBLEMS

1. Find the trend line by using the method of least squares and estimate the trend value for the year 2015.

Year	2008	2009	2010	2011	2012	2013
Production	30	60	40	70	50	65

2. Fit a trend line by the method of least squares for the following data and estimate the trend value for the year 2016.

Year	2009	2010	2011	2012	2013	2014	2015
Sales	50	80	95	65	30	40	45

3. Fit a trend line by the method of least squares for the following data:

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015
Income	630	714	883	837	817	749	729	886	869

4. Fit a straight-line trend by the method of least squares to the following data and obtain the trend value for the year 2017:

Year	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
Production (Lakhs of tonnes)	3.6	3.8	4.4	4.7	5.6	7.3
Year	2010	2011	2012	2013	2014	2015
Production (Lakhs of tonnes)	7.1	7.6	7.7	9.0	9.0	10.1

[Ans: Y = 6.66 + 0298x; Trend value for 1972 = 10.534]

5. Compute the trend value by the method of least squares from the data given below:

Year	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Number of goats (lakhs)	56	55	51	47	42	38	35	32

[Ans: Y = 445–1.86x]

6. Find trend values (mixed with cyclical movements, if any) from the following data of output, by the method of moving averages:

	Year			
Quarter	**2012**	**2013**	**2014**	**2015**
I	29	40	47	45
II	37	42	51	49
III	43	55	63	60
IV	34	43	53	48

15

Research Methodology

15.1 Introduction

In the context of planning and development, the quality of the supporting research is of utmost importance. It is therefore necessary to design and adhere to an appropriate research methodology. The research methodology may differ from problem to problem, but the basic approach remains the same.

In the present fast-track business environment marked by cutthroat competition, many organisations rely on business research to gain a competitive advantage and greater market share. There is a need to generate confidence that the information/data obtained from a business unit will not be misused.

Good research is systematically structured with specified steps, to be taken in a specified sequence in accordance with a well-defined set of rules. Care should be taken that the systematic characteristics of research do not rule out creative thinking, but it certainly does reject the use of guessing and intuition in arriving at conclusions.

Good research is empirical, as it is related basically to one or more aspects of a real situation and deals with concrete data that provides a basis for external validity to research results.

Good research is replicable, as it allows research results to be verified by replicating the study and thereby building a sound basis for decisions.

Business research is the pursuit of truth with the help of study, observation, comparison and experiment. It is the search for knowledge through objective and systematic methods of finding solutions to a problem; to discover answers to questions through the application of scientific procedures; to find truth that is hidden and has not been discovered as yet; to gain familiarity with a phenomenon; and to portray accurately the characteristics of a particular individual, situation or group.

Research comprises the following:

1. The definition and redefinition of problems
2. The fountain of knowledge for the sake of knowledge, an important source providing guidelines for solving different business, governmental and social problems
3. An art of formal training that enables one to understand the new developments in one's field in a better way
4. Search for knowledge; scientific and systematic search for pertinent information on a specific topic; the art of scientific investigation
5. Careful investigation or inquiry, especially through the search for new facts in any branch of knowledge
6. The systematised effort to gain new knowledge

7. A movement from the known to the unknown
8. An original contribution to the existing stock of knowledge, making for its advancement

The business research problem is as follows:

1. Scientific training in the methodology of research is lacking.
2. Most business units do not have confidence that the material supplied by them to researchers will not be misused. They are often reluctant to supply the needed information to researchers. The concept of secrecy seems to be sacrosanct to business organisations.
3. This proves an impermeable barrier to researchers.
4. Research studies overlapping one another are quite often undertaken for want of adequate information – this duplication of results fritters away resources.
5. Research, to many researchers, is mostly a scissor-and-paste job, without any insight into the collated materials. Due to this, the research results quite often do not reflect the reality or realities.
6. There is insufficient interaction between research departments and business establishments.
7. There is no code of conduct for researchers; inter-university and interdepartmental rivalries are quite common.
8. It is difficult to obtain adequate and timely secretarial assistance, which causes unnecessary delays in the completion of research studies.
9. The management and functioning of libraries are not satisfactory in many places, and hence, much of the time and energy of researchers is spent in sourcing books, journals, reports and so on, rather than in extracting relevant material from them.
10. There is difficulty with timely availability of published data, that is, secondary data, as various government and other agencies are doing this job in our country.

Objective of Business Research: To determine the frequency with which something occurs or with which it is associated with something else and to test a hypothesis of a causal relationship between variables. To understand the market trends, find the optimal marketing mix, devise effective human resources policies, or find the best investment options.

15.2 Types of Research

15.2.1 Application of Descriptive Research

This is used for descriptive studies, in which the researcher seeks to measure items such as frequency of shopping, preferences of people and so on. The methods of research used are survey methods of all kinds, comparative methods and co-relational methods.

15.2.2 Analytical Research

The researcher has to use information that is already available, and analyse facts to make a critical evaluation of the material.

15.2.3 Applied Research

Its main purpose is to find a solution for an immediate problem facing a society or an industrial/business organisation and to discover a solution for some pressing practical problems. Examples include research studies concerning human behaviour carried out with a view to making generalisations about human behaviour, and research aimed at certain conclusions facing a concrete social or business problem.

15.2.4 Fundamental Research

This is concerned with generalisation and with the formulation of a theory: for example, research concerning natural phenomena or relating to pure mathematics.

15.2.5 Quantitative Research

This is based on the measurement of quantity or amount. It is applicable to phenomena that can be expressed in terms of quantity.

15.2.6 Attitude or Opinion Research

This is qualitative research designed to find out how people feel or what they think about a particular subject or institution.

15.2.7 Qualitative Research

This has importance in the behavioural sciences. Its aim is to discover the underlying motives of human behaviour. It analyses the various factors that motivate people to behave in a particular manner or make people like or dislike a particular thing.

15.2.8 Motivation Research

This is a type of qualitative research aimed at discovering underlying motives and desires, using in-depth interviews. It is concerned with the determination of motivations underlying consumer (market) behaviour. Techniques of research are word association tests, sentence completion tests and story completion tests.

15.2.9 Conceptual Research

This is related to abstract ideas or theory. It is mainly used by philosophers and thinkers to develop new concepts or to reinterpret existing ones.

15.2.10 Empirical Research

This relies on experience or observation alone. It is data-based research, coming up with conclusions that are capable of being verified by observation or experiment. It is necessary

to get at facts first-hand, at their source, and actively to go about doing certain things to stimulate the production of the desired information. Researchers must first provide themselves with a working hypothesis or guess as to the probable results. They then work to get enough facts (data) to prove or disprove this hypothesis. They then set up experimental designs which they think will manipulate the persons or the materials concerned so as to bring forth the desired information.

15.2.11 Descriptive Research or Ex Post Facto Research

This is an attempt by researchers to discover causes even when they cannot control the variables. Its main purpose is description of the state of affairs as it exists at present. Researchers have no control over the variables. They can only report what has happened or what is happening.

15.2.12 Categorical Research

Research is confined to a single time period.

15.2.13 Longitudinal Research

This is carried on over several time periods.

15.2.14 Field-Setting or Laboratory or Simulation Research

Laboratory research is conducted in a room or building equipped for scientific experimentation. It attempts to investigate naturally occurring behaviours under controlled conditions with manipulated variables.

15.2.15 Clinical or Diagnostic Research

This follows case-study methods or in-depth approaches to reach basic causal relations. Studies go deeply into the causes of things or events that interest us, using very small samples and very deeply probing data-gathering devices.

15.2.16 Exploratory Research

Its main objective is the development of hypotheses rather than their testing.

15.2.17 Formalised Research

Studies are those with substantial structure and with specific hypotheses to be tested.

15.2.18 Historical Research

This uses historical sources such as documents, remains and so on. It studies events or ideas of the past, including the philosophy of persons and groups at any remote point in time.

15.2.19 Target-Oriented Research

Researchers are free to pick up a problem and redesign the inquiry as they proceed and prepare to conceptualise accordingly.

15.2.20 Decision-Oriented Research

This is for the needs of a decision maker. Researchers are not free to embark on research according to their own inclinations. For example, see Operation Research.

15.2.21 Operation Research

This is a scientific method of providing executive departments with a quantitative basis for decisions regarding operations under their control. It refers to the application of mathematical, logical and analytical techniques to the solution of business cost-minimisation, profit-maximisation or optimisation problems.

15.2.22 Market Research

This is the investigation of the structure and development of a market for the purpose of formulating efficient policies for purchasing, production and sales.

15.3 Types of Research Approach

15.3.1 Quantitative Approach

This involves the generation of data in quantitative form, which can be subjected to rigorous quantitative analysis in a formal and rigid fashion.

15.3.2 Inferential Approach

This is carried out to form a data base from which to infer characteristics or relationships in a population. It means survey research, whereby a sample of the population is studied (questioned or observed) to determine its characteristics. Finally, it is inferred that the population has the same characteristics.

15.3.3 Experimental Approach

There is greater control over the research environment, and some variables are manipulated to observe their effect on other variables.

15.3.4 Simulation Approach

Simulation is the imitation of the operation of a real-world process or system over time. Simulations aim to provide an experience as close to the 'real thing' as possible. It involves the construction of an artificial environment within which relevant information and data can be generated. It permits an observation of the dynamic behaviour of a system (or its subsystem) under controlled conditions. It is useful in building models for understanding future conditions.

15.3.5 Qualitative Approach

This is concerned with subjective assessment of attitudes, opinions and behaviour. Techniques used are focus group interviews, projective techniques and depth interviews.

15.4 Benefits of Research

15.4.1 Benefits in Business and Industry

1. Research provides the basis for nearly all government policies in our economic system: the plight of cultivators, the problems of distribution, the size and nature of defence services and so on.
2. It collects information on the economic and social structure of the nation, which indicates what is happening in the economy and what changes are taking place.
3. Research has special significance in solving various operational and planning problems of business and industry.
4. It helps people in business and industry who are responsible for taking business decisions.
5. Given knowledge of future demand, it is generally not difficult for a firm or an industry to adjust its supply schedule within the limits of its projected capacity.
6. Research with regard to demand and market factors has great utility in business.
7. A properly structured budget points out what aspects of the business produce money and which ones use it, which should drop some parts of the business, or expand others.
8. Market analysis has become an integral tool of business policy these days.
9. Once sales forecasting is done, efficient production and investment programmes can be set up, around which are grouped the purchasing and financing plans.

15.4.2 Benefits to Society

1. It is important for social scientists in studying social relationships and in seeking answers to various social problems.
2. It includes benefits to the activity, attitude, awareness, behaviour, capacity, opportunity, performance, policy, practice, process or understanding.
3. A well-cultivated critical thinker gathers and assesses relevant information, using abstract ideas to interpret it effectively, and comes to well-reasoned conclusions and solutions, testing them against relevant criteria and standards.

15.4.3 Benefits for Professions, Philosophers and Thinkers

1. The responsibility of research as a science is to develop a body of principles that make possible the understanding and prediction of the whole range of human interactions.
2. Because of its social orientation, it is increasingly being looked to for practical guidance in solving immediate problems of human relations.

3. To those students who are to write a master's or PhD thesis, research may mean careerism or a way to attain a high position in the social structure.
4. To professionals in research methodology, research may mean a source of livelihood.
5. To philosophers and thinkers, research may mean the development of new styles and creative work.
6. To analysts and intellectuals, research may mean the development of new styles and creative work.

15.5 Contents of Research Plan

An example format for the report is shown in Figure 15.1.

Other formats for reports may be acceptable, but the method illustrated is a method of presentation that may be generally useful. Also, for some uses, parts of the report format that is given may not be applicable, or may appear in a different order, or combined with another part of the report.

15.5.1 Layout of the Report

Anybody who is reading the research report must necessarily be conveyed enough information about the study so that he or she can place it in its general scientific context, judge the adequacy of its methods, and thus form an opinion of how seriously the findings are to be taken. Following are some suggestions, though styles may differ.

15.5.2 Preliminary Pages

In its preliminary pages, the report should carry a title and date, with acknowledgements at the end in the form of a preface or foreword. There should be a table of contents, list of tables and illustrations so that anybody can easily locate the required information in the report.

15.5.3 Main Text

This provides a complete and detailed outline of the research report. The title of the research study is repeated at the top of the first page of the main text and then follows

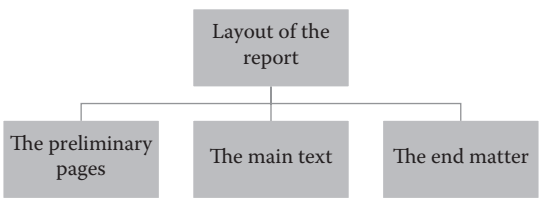


FIGURE 15.1
Layout of the report.

the other details on pages numbered consecutively, beginning with the second page. The main text has the following sections:

1. Introduction
2. Statement of findings and recommendations
3. Results
4. Implications drawn from the results
5. Summary

The main body of the report should be presented in logical sequence and broken down into readily identifiable sections. The key findings are concisely presented in the executive summary, running into 100–200 words or a maximum of two pages. The major thrust should be on highlighting the objectives, salient features and analysis of the results, including the recommendations.

15.5.3.1 Introduction

This includes

1. A clear statement of the objectives of research
2. A brief summary of other relevant research so that the present study can be seen in that context.
3. The hypotheses of the study
4. Definitions of concepts
5. The methodology adopted
6. The statistical analysis adopted
7. The scope of the study
8. Limitations

15.5.3.2 Statement of Findings and Recommendations

This must be in non-technical language so that it can be easily understood by all concerned. If the findings happen to be extensive, at this point they should be put into a summarised form.

15.5.3.3 Results

This includes

1. The main body of the report
2. Statistical summaries

All relevant results must find a place in the report and must be presented in logical sequence and split into readily identifiable sections.

15.5.3.4 Implications of the Results

This includes

1. A statement of the inferences drawn from the present study that may be expected to apply in similar circumstances
2. The conditions of the present study that may limit the extent of legitimate generalisations of the inferences drawn from the study
3. The relevant questions that still remain unanswered, or new questions raised by the study, along with suggestions for the kind of research that would provide answers for them

15.5.3.5 Summary

It has become customary to conclude the research report with a brief summary, describing in brief the research problem, the methodology, the findings and conclusions drawn from the research results.

15.5.4 End Matter

At the end, appendices should be included in respect of all technical data such as questionnaires, sample information, mathematical derivations and so on. A bibliography of sources consulted should be given. An index should invariably be given at the end of the report. The value of the index lies in the fact that it works as a guide for the reader to the contents of the report.

15.6 Criteria of Good Research

1. The purpose **s**hould be clearly defined, and common concepts should be used. Statements should be short and direct.
2. Research method should be defined in a clear manner with sufficient detail. This will allow the repetition of the study in future for further advancement, while maintaining the continuity of what has been done in the past.
3. Pictures and graphs should accompany tables.
4. Graphics and animations should accompany the presentation of the report.
5. The procedure should be described in sufficient detail to permit another researcher to repeat the research for further advancement, keeping the continuity of what has already been attained.

15.7 Features of a Research Report

The major thrust should be as outlined in the following subsections.

15.7.1 Problem Definition

A problem is any situation that requires further investigation. Decisions made on judgements may not always turn out to be correct, but the problem may not be important enough to justify spending substantial time, money and effort on solving it.

It is correctly said that a problem well defined is half solved. A clear, precise, to-the-point statement of the problem itself provides clues to its solution. On the other hand, a vague, general or inaccurate statement of the problem only confuses the researcher and can lead to the wrong problem being researched and useless results generated.

15.7.2 Research Objectives

1. Form the heart of the study
2. Address the purpose of the project

   The basic purpose of research is to facilitate the decision-making process. A manager has a number of alternative solutions to choose from in response to every problem and situation. In the absence of information, he or she may make the choice on the basis of a hunch. By doing so, the manager is taking a risk, because he or she has no concrete evidence to evaluate this alternative in comparison with others or to assess its possible outcome. But with the help of information provided by research, the manager can reduce the number of alternative choices to one, two or three, and the possible outcome of each choice is also known. Thus, the decision-making process becomes a little easier.

   Research helps to reduce the risk associated with the process of decision making. The risk arises because of two types of uncertainty. Uncertainties about the expected outcome of the decisions will always remain, no matter how much information you may have collected to base your decision on hard facts. Unforeseen factors have the uncanny ability of upsetting even the most stable apple cart.

   Despite the best research efforts, the outcome can still be unpredictable. The risk also arises because of the uncertainty of what will happen in the future, the way the customer or distributor will behave, the manner in which the competition will react and so on. To the extent that research provides information about the future, it anticipates the future, thus providing the manager with a solid basis for decision making. However, it cannot provide perfectly exact or accurate information. But since the techniques of research are based on scientific methods of collecting, analysing and interoperating data, its findings and projections, at least, provide a definite trend of scenarios for future decision making.

   The third purpose of research is that it helps firms to discover opportunities that can be profitably exploited. These opportunities may exist in the form of untapped customer needs or wants not catered to by existing firms. A food company has introduced into the market a dairy whitener (as a substitute for milk) called 'Every day', to be used for making tea and coffee. The product has proved to be a success because it is most convenient for use in offices, where tea and coffee are consumed in large quantities but milk is not easy to procure. 'Every day' fills a gap in the market for powdered milk that was not being catered to by the existing milk powders.
3. Follow a set of well-planned objectives.

4. The purpose of the quantitative objectives is to optimize certain performance measures of the research system. For example, research to predict the product demand with high precision, and research on advertising budgets in relation to other expenditures such that the incremental sales revenue is maximised.

15.7.3 Background Material

1. Include a review of the previous research or descriptions of the conditions that caused the project to be authorised.
2. This may include preliminary results from an experience survey or secondary data from various sources.
3. The references from secondary data, definitions and assumptions are included.

15.7.4 Methodology

1. Sampling design
2. Research design
3. Data collection
4. Data analysis
5. Limitations
6. Findings
7. Conclusions
8. Recommendations
9. Appendices
10. Bibliography

15.7.4.1 Sampling Design

Here, the researcher defines the target population and the sampling methods put to use. The section contains other necessary information, such as

1. Types of sampling (probability or non-probability) used
2. Types of probability sampling (simple random or complex random) or non-probability sampling (quota sampling or snowball sampling) used
3. The factors influencing the determination of sample size and selection of the sampling elements
4. The levels of confidence and the margin of acceptable error

15.7.4.2 Research Design

This should be carefully planned to yield results that are as objective as possible. It should contain information on

1. Nature of research design
2. Design of questionnaires

3. Questionnaire development and pre-testing
4. Data that was gathered
5. Definition of interview and type of interviewers
6. Sources (both primary and secondary) from which data was collected
7. Scales and instruments used
8. Designs of sampling, coding and method of data input
9. Strengths and weaknesses
10. Copies of materials used and the technical details could be placed in the appendix

15.7.4.3 Data Collection

The collection of primary data for business research is of paramount importance to assist management in making decisions. Generally, information regarding a large number of characteristics is necessary to analyse any problem pertaining to management. The collection of primary data requires a great deal of deliberation and expertise.

15.7.4.4 Data Analysis

This provides information on the different methods used to analyse the data. It should justify the choice of methods based on assumptions. It should be adequate to reveal significance, and the methods of analysis used should be appropriate. The validity and reliability of the data should be checked carefully.

It provides details on

1. Data handling
2. Groundwork analysis
3. Rational statistical testing and analysis

15.7.4.5 Limitations

Certain researchers tend to avoid this section, but this is not a sign of professionalism. There should be a tactful combination of reference to and explanation of the various methodologies and their limitations or implementation problems. The limitations need not be explained in detail. Details of limitations do not detract from the research. They help the reader to acknowledge its honesty and validity.

15.7.4.6 Findings

1. It is better to report one finding per page and support it with quantitative data.
2. This section presents all the relevant data but makes no attempt to draw any inferences.
3. It attempts to bring to the fore any pattern in the industry.
4. Charts, graphs and tables are generally used to present quantitative data.

15.7.4.7 Conclusions

Conclusions are inferences drawn from the findings. They should be directly related to the research objectives or hypotheses. The researcher should always present the conclusions, as he or she has first-hand knowledge of the research study. It is wrong to leave the inference of the conclusions to the reader. Conclusions should be confined to those justified by the data of the research and limited to those for which the data provides an adequate basis. Towards the end of the main text, the researcher should again state the results of his or her research clearly and precisely. In fact, it is the final summing up.

15.7.4.8 Recommendations

Recommendations are a few corrective actions presented by the researcher. The actions the report calls for as per the researcher should be highlighted; they should be in line with the results of the report; they should be explicit; they may even contain plans of how future research into the same can proceed. Recommendations should be given if the client wants them; otherwise, they should be avoided, because some decision makers do not want their thought processes to be limited to the recommendations given. In such a case, the report should not include any recommendations.

15.7.4.9 Appendices

Appendices are optional. They are used to present details that were part of the research but were not necessary to the presentation of the findings or conclusions. They include

1. Raw data.
2. Calculations.
3. Graphs.
4. Copies of forms and questionnaires.
5. Complex tables.
6. Instructions to field workers.
7. Quantitative material that would look inappropriate in the main text. The reader can refer to them if required.

Care should be taken that they do not exist in isolation and reference to each appendix is given in the text.

15.7.4.10 Bibliography

This is a list of citations or references to books or periodical articles on a particular topic. Any journals consulted should also be listed.

15.7.4.11 Index

An index is an alphabetical listing of keywords and concepts in the text. It contains `pointers' to those words and concepts, which are usually page, section or paragraph numbers. It generally appears at the end of a work. It is essential to the readability and usability of technical articles and books.

15.8 Summary

Research will require the collection of new data through surveys and/or the use of existing data, as is often the case in the application of econometrics. These approaches are often associated with survey statisticians and economists, respectively.

The report should be written in a concise and objective style in simple language, avoiding vague expressions such as 'it seems', 'there may be' and the like.

Charts and illustrations in the main report should be used only if they present the information more clearly and forcibly.

Efforts should be made to provide short-duration intensive courses for meeting these requirements, and to develop satisfactory liaison among all concerned for better and more realistic research.

There is a need for developing some mechanisms of a university–industry interaction programme, so that academics can get ideas from practitioners on what needs to be researched and practitioners can apply the research done by the academics.

There should be proper compilation and revision, at regular intervals, of a list of subjects on which research is ongoing and the places where this is happening.

Research problems in various disciplines of applied science that are of immediate concern to industry should be identified.

There is a need to develop a code of conduct for researchers, which, if adhered to sincerely, can overcome these problems.

Efficient secretarial assistance should be made available to researchers in time.

Efforts should be made to provide all governmental publications to our libraries. Our libraries must get copies of old and new acts/rules, reports and other government publications in good time.

The sampling methods used should be explained, and calculations could be placed in the appendix rather than in the body of the report.

Confidence in research is warranted if the researcher is experienced, has a good reputation in research and is a person of integrity.

REVIEW QUESTIONS

1. List the guidelines for effective report writing.
2. What are the objectives of research?
3. What are the features of good research design?
4. Write any four essential characteristics of presentations.
5. What are the different types of questions in questionnaires?
6. Describe briefly the features of a good questionnaire. How will you prepare a questionnaire for studying awareness of Five-Year Plans in your area?
7. Explain what precautions must be taken while drafting a questionnaire in order that it may be really useful. Illustrate your points.
8. Describe briefly the questionnaire method of collecting primary data. State the essentials of a good questionnaire.
9. Define the term *research*. What should be the objectives of good research? Discuss in the context of management science.

10. Explain the importance of formulating the research hypothesis. What is the procedure for hypothesis testing?
11. Define the term *research design*. What are the various types of research design?
12. Explain the different steps involved in the research process.

16

Case Studies for Highlighting Quantitative Techniques

16.1 Application of Hypothesis Testing in Industry

16.1.1 Introduction

This study is based on the applications of hypothesis testing in the situation given. Hindustan Lever Limited Company (HLL) wants to find out the mean spending on bath soaps by an average Indian family per month. HLL management assumed that the average expenditure on bath soaps per month is Rs. 120 and population standard deviation is 8.

By the application of hypothesis testing, we have to check whether the assumption made by HLL management is correct or not.

16.1.2 Company Profile

Hindustan Unilever Limited

Hindustan Lever Limited

Type	Public
Founded	1933
Headquarters	Mumbai, India
Key people	Mr Harish Manwani, chairman; Douglas Baillie, CEO
Industry	Fast moving consumer goods (FMCG)
Products	Tea, soap, detergents
Employees	41,000
Parent	Unilever
Website	www.hll.com

HLL, formerly Hindustan Lever Limited, headquartered in Mumbai, is India's largest consumer products company, formed in 1933 as Lever Brothers India Limited. Its 41,000 employees are headed by Mr Harish Manwani, the non-executive chairman of the board. HLL is the market leader in Indian products such as tea, soaps, detergents, and its products have become daily household names in India. The Anglo-Dutch company Unilever owns a majority stake in HLL.

Recently, in February 2007, the company was renamed Hindustan Unilever Limited to provide the optimum balance between maintaining the heritage of the company and the

future benefits and synergies of global alignment with the corporate name of Unilever. This decision will be put to the shareholders for approval at the next annual general meeting.

16.1.3 Brands

Some of its brands include Kwality Walls ice cream, Lifebuoy, Lux, Breeze, Liril, Rexona, Hamam, Moti soaps, Pureit Water Purifier, Lipton tea, Brooke Bond tea, Bru Coffee, Pepsodent and Close Up toothpaste and brushes, Surf, Rin and Wheel laundry detergents, Kissan squashes and jams, Annapurna salt and atta, Pond's talcs and creams, Vaseline lotions, Fair & Lovely creams, Lakmé beauty products, Clinic Plus, Clinic All Clear, Sunsilk and Lux shampoos, Vim dishwash, Ala bleach and Domex disinfectant.

16.1.4 Marketing

In February 2003, HLL launched a new division called Hindustan Lever Network. This division markets a wide range of FMCG through network marketing. Network marketing was pioneered in the United States in the 1940s by companies such as Amway Corporation and operates by recruiting individuals as consultants. These consultants are paid a commission on the purchases made by them and on the purchases made by those recruited by them.

Objective

1. To know the preference of customers towards various bath soaps
2. To study the average expenditure on bath soaps per month by the average family
3. To study whether the assumption made by HLL management is correct or not

16.1.5 Area of Study

The place chosen for the study was Nagpur District of Maharashtra State of India, where people come from all over India. The reason behind collecting the data here is that the sample is not confined to a particular region. It represents the entire population.

16.1.6 Data Source

Primary data has been used for the purpose of the project. The data is taken in the form of questionnaires collected from a population of 45 teachers. The data collected is expressed in the form of data tables and charts.

Questionnaire

1. Name of the teacher
2. Name of college/school

3. Which bath soap do they prefer?
4. What number of bath soaps are consumed in a month?

16.1.7 Data Analysis

For data analysis, statistical tools such as mean and standard deviation, tables and graphs are used.

Sample size: 45

Tabulation of Data

Sr No.	Faculty Name	City	Kind of Bath Soap Consumed	No. of Bath Soaps Consumed in a Month
1	Amit Welekar	Mathura	Cinthol	3
2	Jafar Sheikh	Banaswadi	Mysore Sandal Soap	4
3	Kanchan Dhote	Belgaum	Lux	3
4	Snehal Khadge	Silguri	Pears	4
5	Nikhil Bangde	Baroda	No Marks Scrub Soap	2
6	Amit Tajne	Kota	No Marks Scrub Soap	1
7	Rajesh Babu	Trichur	Lifebuoy	4
8	Prashant Kshirsagar	Gondia	Dove	2
9	Meenal Wankhde	Sarangpur	Abiance	1
10	Anant Kumar	Burhampur	Lux	4
11	Raghuveer Pawde	Warangal	Santoor	4
12	Rachit Singh	Banglore	Lifebuoy	4
13	Amol Wankhede	Gaziabad	Lux	12
14	Ravi Fate	Panipat	Pears	5
15	Anuj Dubey	Medhi Patnam	Mysore Sandal Soap	4
16	Vedant Awari	Batinda	Dove	4
17	Vishal Tiwari	Nagpur	Dove	5
18	Shrikant Moje	Bijapur	Lifebuoy	2
19	Ashish Dhamke	Nagpur	Medimix	4
20	Akash Kumare	Raipur	Lux	12
21	Pooja Badole	Bhopal	Lux	5
22	Nitesh Kolhe	Vijayawada	Lux	2
23	Amol Kolhe	Warangal	Medimix	4
24	Namdeo Masram	Kakinada	Dove	4
25	Pravin Signe	Guwahati	Dettol	4
26	Sandeep Palsodkar	Hubli	Mysore Sandal Soap	3
27	N. Pradeep	Kota	Lux	4
28	T. Raju	Hubli	Cinthol	6
29	Sairaj	Sri Ganganagar	Lux	4
30	Archana Malewar	Rohtak	Lifebuoy	3
31	Durgashankar Malewar	Noida	Lux	6
32	Pinky Dubey	Warangal	Cinthol	4
33	Parul	Trivendrum	Santoor	4
34	Nayan	Mathura	Growning Glory	5
35	Tanvi	Aizwal	Lifebuoy	4
36	Uganti	Belgaum	Santoor	6

(*Continued*)

Sr No.	Faculty Name	City	Kind of Bath Soap Consumed	No. of Bath Soaps Consumed in a Month
37	Anup Dubey	Meerut	Pears	6
38	Ajay	Eluru	Rexona	5
39	Vijay	Burdwan	Lux	3
40	Jay	Ooty	Park Avenue	2
41	Kanchan	Burdwanan	Lux	4
42	Monika	Bhopal	Pears	4
43	Rekha	Patiala	Lux	2
44	Ankit	Sikar	Lux	10
45	Neeta	Medhi Patnam	Lifebuoy	4

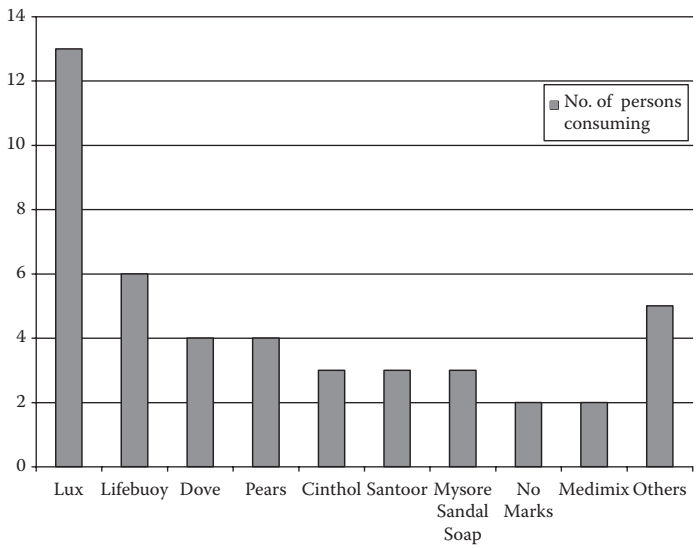
16.1.8 Findings

Brand Preference

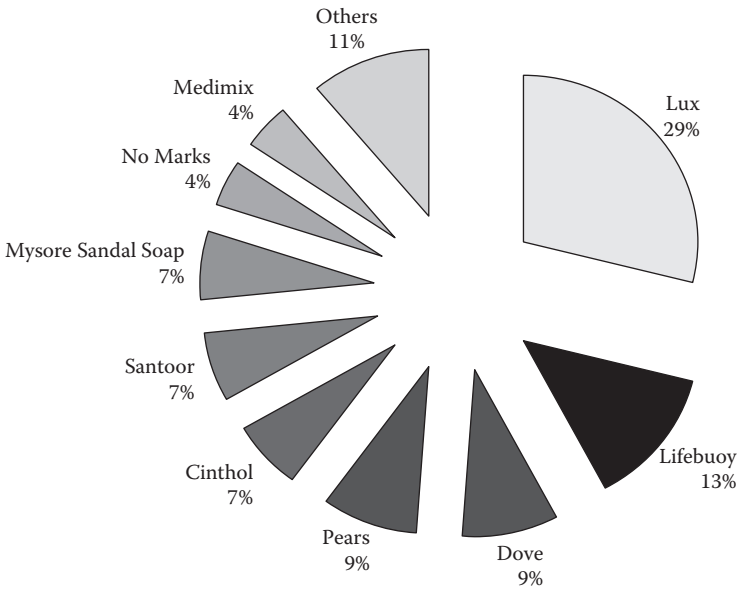
Bath Soap	No. of Persons Consuming
Lux	13
Lifebuoy	6
Dove	4
Pears	4
Cinthol	3
Santoor	3
Mysore Sandal Soap	3
No Marks	2
Medimix	2
Others	5
Total	45

The brand preference of the customers may be expressed with the help of the following charts.

Bar Graph



Pie Graph



From the bar chart and the pie chart, we find that major brand preference is towards Lux Soap.

Computing Total Cost of Bath Soaps Brand-Wise

Bath Soap	No. of Persons Consuming	No. of Units	Price per Soap	Total Cost
Lux	13	71	13	923
Lifebuoy	6	21	12	252
Dove	4	15	27	405
Pears	4	19	23	437
Cinthol	3	13	13	169
Santoor	3	14	13	182
Mysore Sandal Soap	3	11	20	220
No Marks	2	3	20	60
Medimix	2	8	15	120
Abiance	1	1	28	28
Dettol	1	4	18	72
Growning Glory	1	5	18	90
Park Avenue	1	2	28	56
Rexona	1	5	13	65
Total	45			3079

16.1.9 Thrust Area of the Project

As already discussed, HLL wants to find out the mean spending on bath soaps by an average Indian family per month. HLL management assumed that the average expenditure on bath soaps per month is Rs. 120 and the population standard deviation is 8.

To solve this problem, we have taken a sample of 45 faculty members, keeping in mind that people come from all over India and have a salary between Rs. 10,000 and Rs. 12,000. The very reason behind collecting the data here is that the sample is not confined to a particular region. It represents the entire population.

Here, HLL management made an assumption relating to the characteristics of the population. They assumed that average expenditure on bath soaps per month is Rs. 120. Now, the question is: will this assumption made by HLL management be correct or incorrect?

For the management, it is essential to know whether their assumption is correct or not. In general, all the marketing strategies are based on the conclusions drawn from quantitative methods. This is one of the instances when quantitative methods are known to be a powerful decision-making tool.

The HLL management assumption can be evaluated using the statistical technique called *hypothesis testing procedures*. Hypothesis testing enables the researcher to determine the validity of his or her hypothesis concerned with a particular issue. Hypothesis testing enables the researcher to decide whether data from a sample will provide support to a particular hypothesis, based on which it can be generalised to the overall population.

There are certain steps to solve the problem. So, let me follow them one by one.

Step 1: Formulation of the hypothesis.

The hypothesis testing starts with the formulation of the hypothesis. Therefore, H_0 and H_1 are

H_0: Average expenditure on bath soaps per month is Rs. 120 ($\mu = 120$)

H_1: Average expenditure on bath soaps per month differs from Rs. 120 ($\mu \neq 120$).

Step 2: Selection of the statistical test to be used.

Here, the size of the sample is large (i.e. $n = 45$), and the population standard deviation is known. Therefore, we can use a Z-test to solve the problem.

Step 3: Selection of the significance level.

The level of significance is $\alpha = 5\%$.

Step 4: Calculation of standard error of the sample statistic and standardisation of the sample statistic.

The test statistic is given by

$$Z = \frac{X - \mu}{SE}$$

where:

$$SE = \sigma/\sqrt{n}$$

$$= 8/\sqrt{45}$$

$$SE = 1.1926$$

$$X = \sum X_i / n$$

$$= 3079/45$$

$$X = 68.4222$$

Substituting above values in the test statistic, we get

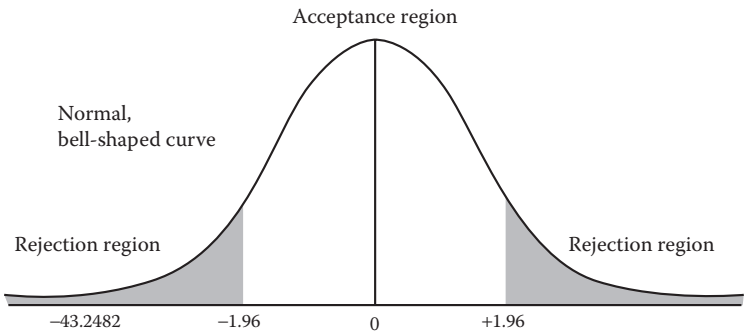
$$Z = (68.4222 - 120)/1.1926$$

$$Z = -43.2482$$

Step 5: Determination of the critical values.

Since the test is two-tailed and the level of significance specified is 5%, the critical values are –1.96 and +1.96.

Step 6: Comparison of the value of the sample statistic with the critical value and determination of whether the value falls within the acceptance or the rejection region.



The calculated Z value lies in the rejection region. Therefore we reject H_0.

The calculated Z value is less than –1.96. Therefore we reject H_0.

Step 7: Deduction of the business research conclusion.

Since H_0 is rejected, we can conclude that average expenditure on bath soaps per month differs from Rs. 120; that is, the assumption made by HLL management is incorrect.

16.2 Universal Home Care Products

16.2.1 Introduction

Name of company	**Universal Home Care Products Ltd.**
Producer	Detergents and cleaning agents
Sales (crores)	Rs. 1775
Net income (crores)	Rs. 112.3

16.2.2 Recent Development

The company had developed an all-purpose household cleaner, 'Sparkle'. It could clean a variety of surfaces, such as wood, glass, metal, plastic and ceramic. Universal projected market share:

1. 6% for the first year
2. 10% by the second year
3. 14% after 2 years

Competitor:

1. Day-chem Ltd. might already be planning
2. To launch a similar multipurpose household cleaner
3. With similar positioning
4. Targeting the same segments

Estimation of profit for new product:

1. Daychem's ability to bring out a competing product was estimated at 60%.
2. Universal estimated the profits for its new product for three different prices.
3. If Universal set a low price, the estimated profits were Rs. 60,000.
4. At a medium price, the estimated profits were Rs. 75,000.
5. At a high price, the estimated profits were Rs. 90,000.

Data Estimated Profits (Thousands of Rs)

If Universal's Price Is	If Competitor's Price Is			
	Low	Medium	High	Total
Low	20	30	30	80
Medium	80	70	50	200
High	60	40	20	120
Total	160	140	100	400

16.2.3 The Problem

Is there any relationship between the firm's pricing structure for Sparkle and the competitor's pricing structure for the competing product?

16.2.4 Procedure

Step 1: Formulating the null and alternate hypothesis

- H_0→ There is no relationship between the firms' pricing structure for Sparkle and the competitor's pricing structure for the competing product.
- H_1→ There is a relationship between the firms' pricing structure for Sparkle and the competitor's pricing structure for the competing product.

Step 2: Calculation of expected frequency (calculated expected frequency in brackets)

$$E_{ij} = \frac{n_i \times n_j}{n}$$

where:

E_{ij} = Expected frequency of a cell
n_i = Row total
n_j = Column total
n = Total sample size

If Universal's Price Is	If Competitor's Price Is			
	Low	**Medium**	**High**	**Total**
Low	(26.1890)	(45.7479)	(49.0630)	120.9999
Medium	(28.7863)	(50.2849)	(53.9288)	133
High	(24.0247)	(41.9671)	(45.0082)	111
Total	79	137.9999	148	364.9999

Step 3: Calculation of χ^2

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\left[\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}\right]$$

where:

O_{ij} = The observed frequency in *i*th row and *j*th column
E_{ij} = The expected frequency in *i*th row and *j*th column

O_i	E_i	$O_i - E_i/E_i$	$(O_i - E_i)^2/E_i$
32	26.1890	5.8110	1.2894
40	45.7479	–5.7479	0.7222
49	49.0630	–0.0630	0.0001
35	28.7863	6.2137	1.3413
48	50.2849	–2.2849	0.1038
50	53.9288	–3.9288	0.2862
12	24.0247	–12.0247	6.0185
50	41.9671	8.0329	1.5376
49	45.0082	3.9918	0.3540
Total: 365	364.9999		11.6531

Step 4: Decision on the level of significance and degrees of freedom

Level of significance is 1%, that is, 01.

Degrees of freedom (ν) is given by

$$\nu = (n - 1)\,(r - 1)$$

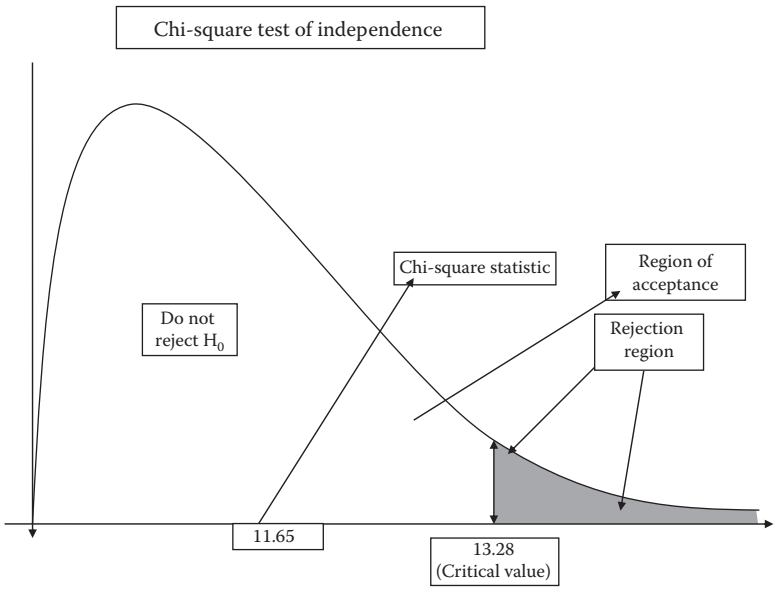
where:

ν = degrees of freedom
n = number of rows
r = number of columns

Here, n = 3 and k = 4

∴ Degrees of freedom = (3 − 1) (3 − 1) = 4

Step 5: Determination of the critical value and comparing it with the calculated χ^2 value. From the χ^2 distribution table, the χ^2 tabulated value for 4 degrees of freedom at the 1% level of significance was found to be 13.28.



Step 6: Deducing the business conclusion

Since the calculated χ^2 (11.65) < the tabulated χ^2 (13.28), there is no evidence to reject the H_0. Therefore, the null hypothesis is accepted. Therefore, we found that there is no relationship between the firm's pricing structure for Sparkle and the competitor's pricing structure for the competing product.

16.3 Model Pertaining to Heart Attack

16.3.1 Introduction

According to the statistics about heart disease published in 2006 by the Centers for Disease Control (CDC), 24.7 million adults have been diagnosed with heart disease.

As per CDC (the agency that publishes statistics about heart disease), 'Diseases of the heart' is listed as the No.1 cause of death.

According to statistics about heart disease, approximately 500,000 new cases are diagnosed each year.

The statistics for heart disease may be depressing, but even more depressing is the fact that 66.30% of the population over the age of 20 are overweight, another risk factor for developing heart disease.

16.3.2 The Problem

The National Vital Statistics Report, which is responsible for publishing these statistics on heart disease deaths, does not include subcategories. It is impossible to tell how many of such deaths were caused by coronary artery disease, heart failure or another disease, such as diabetes, affecting the heart.

To evaluate whether there is any association between the variables under study, like heart attack, which is segregated according to age group, we need to consider three major factors, such as coronary artery disease, heart failure and another disease affecting the heart, so that we are in a position to answer questions like 'How many persons from a particular population, in a particular age group, get an attack due to a particular responsible factor?' and 'What is the proportion of persons in the population prone to heart attack?'

16.3.3 My Idea

To investigate this, I visited about 400 households where someone had lost their beloved due to death from heart disease. My data source was two cardiologists from my city, from whom I collected the data, from the Nanded District of Maharashtra state, and I made a personal visit to each household of the population (sample size 400).

16.3.4 Details

The sample unit is age group of persons/individuals: 'Age below 30', 'Age between 30 and 40' and 'Age above 45'.

16.3.5 Objectives

To find whether there is significance/a significant association between the age group and the factor responsible.

16.3.6 Methodology

To evaluate the statistical significance of association among the variables involved in the cross-tabulation, I used the χ^2 test of independence because

- The χ^2 test can be performed on the actual numbers
- The expected frequency of cell (f_e) must be more than five
- The sample size is large enough … here it is 400
- Observations drawn need to be random and independent

Table 16.1 shows the data collected.

TABLE 16.1

Factors Responsible for Heart Disease Death and Distribution according to Age Group

Factors Age Group	Coronary Artery Disease	Heart Failure	Cholesterol Level	Total
Below 30	20	30	30	80
Between 30 and 40	80	70	50	200
45 years and above	60	40	20	120
Total	160	140	100	400

16.3.7 Related Works

Step 1: Formulation of null and alternate hypotheses

H_0: There is no association between the age group and factors responsible for heart disease death.

H_1: There is an association between the age group and factors responsible for heart disease death.

Step 2: Calculation of expected frequency values (Table 16.2)

$$E_{ij} = \frac{n_i \times n_j}{n}$$

where:

E_{ij} = expected frequency of a cell corresponding to a particular age group and a factor
n_i = row total
n_j = column total
n = Total sample size

Step 3: Calculation of χ^2 using the following formula (Table 16.3):

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

where:

O_{ij} = the observed frequency in ith row and jth column
E_{ij} = the expected frequency in ith row and jth column

Step 4: Level of significance (α) is 1%

TABLE 16.2

Expected Frequency (in Brackets) and Observed Frequency (without Brackets)

Factors → Age	Coronary Artery Disease	Heart Failure	Another Disease Affecting the Heart	Total
<30	(32)	(28)	(20)	80
Between 30 and 40	(80)	(70)	(50)	200
≥45	(48)	(40)	(30)	120
Total	160	140	100	400

TABLE 16.3

Calculation of χ^2

O_i	E_i	$O_i - E_i$	$(O_i - E_i)2/E_i$
20	32	−12	4.5
30	28	2	0.1428571
30	20	10	5
80	80	0	0
70	70	0	0
50	50	0	0
60	48	12	3
40	42	−2	0.095238
20	30	−10	3.333333
Total: 400	400	0	16.071428 ≈16.08

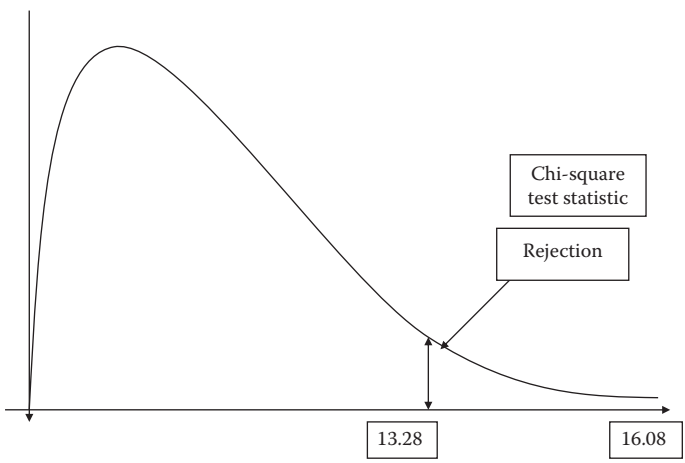
$$\text{Degrees of freedom} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$
$$= (3-1)(3-1)$$
$$= 4$$

Step 5: Determination of the critical value and its comparison with the calculated value of χ^2.

From χ^2 distribution table, tabulated value of χ^2 for 4 degrees of freedom, at 1% level of significance, was found to be 13.28.

The graph shows the χ^2 square test, for 4 degrees of freedom, at 1% level of significance, showing the region of rejection.

Figure: Chi-square test –test of independence



Step 6: Research conclusion

Since $\chi^2_{\text{calculated value, 1\%}} > \chi^2_{\text{tabulated}}$, there is evidence to reject the null hypothesis (H_0).

Therefore, the null hypothesis, that there is no significant association between the age group and the factors responsible for heart disease death, is rejected.

16.3.8 Conclusion and Further Work

We found that there is an association between the variables, that is, age group and factors such as coronary artery disease, heart failure and another disease affecting the heart.

This may contribute to the National Vital Statistics Report by the inclusion of subcategories in the report, making it possible to tell how many heart disease deaths were caused by coronary artery disease, heart failure and another disease such as diabetes affecting the heart, so that we can suggest that you should

1. Take care of your heart
2. Prevent cholesterol from affecting the quality of your life
3. Avoid air pollution, which is dangerous for your heart, and maintain a clean environment

16.4 A Study of Mall Intercept Survey by the Application of Purchase Intercepts Technique

16.4.1 Introduction

The study is based on the applications of personal interviews in the situation given and applies a coding process to them. Ganesh, a theatre proprietor, wants to find out whether or not the customers are satisfied in relation to which service he has provided to them.

16.4.2 Problem Statement

The data were collected by the survey method, which involves in-theatre observation and interviewing, whereby consumers are intercepted and interviewed when they have come to see a movie.

Objective:

1. To intercept customers and interview them when they come to see a movie
2. To find out whether or not the customers are satisfied regarding the services which they get, such as
   a. Seating arrangement
   b. Booking facility
   c. Parking facility
   d. Price of the ticket
3. To find out whether or not the customers come to the theatre frequently

16.4.3 Company Profile

• Proprietor name	Ranaji Sarada
• Establishment	1977
• Area of theatre	12,000 sq. feet
• No. of seats available	800

Family circle	175
Balcony	325
Upper circle	100
Lower circle	200
• Size of screen	45×20 sq. feet
• Distance from screen to first row	15 feet
• Distance between two rows	18 inches
• No. of employees	10
• Ticket rates:	
Family circle	Rs. 45
Balcony	Rs. 40
Upper circle	Rs. 30
Lower circle	Rs. 20
• Industry	Service

16.4.4 Methodology

16.4.4.1 Sampling Design

I used simple random sampling, because people come to see a movie from all over Nanded, and they belong to various age groups.

16.4.4.2 Research Design

- I used primary data collection and the personal interview method.
- In personal interviews, I had face-to-face interaction with interviewees with the help of a questionnaire, which helped me to collect quality data.
- This helps to reduce non-response error.
- I used a coding process for the responses given by the interviewees to analyse customer behaviour.

Coding the responses

Particular	Coding
Yes	@
No	*
Weekly	!
Monthly	#
Frequently	$
Discount facility	+
Parking facility	×
Booking facility	■
Fresh and quality food	▫
Comedy	®
Inspiring	€
Horror	©
Family	∞

16.4.5 Data Collection

Data was collected from 50 people during a period of 2 weeks.

16.4.6 Data Analysis

Analysis was done by a coding process and graph.

16.4.7 Primary Data

Sr. No.	Name and Address	Q. 2	Q. 3	Q. 4	Q. 5	Q. 6	Q. 7	Q. 8
1	Shailesh, Saptgiri Nagar	!	@	×	@	®	@	@
2	Sachin, Anand Nagar	#	@	■	*	€	@	@
3	Sumit, Vazirabad	$	@	■	*	€	@	@
4	Dilshad, Itawara	#	@	□	@	®	@	@
5	Amit, Baba Nagar	#	@	×	*	€	@	@
6	Prasad, Bhagya Nagar	!	@	×	*	€	@	@
7	Sumit, Vidya Nagar	!	@	×	@	®	*	@
8	Hemlata, Mantri Nagar	!	@	■	*	∞	*	@
9	Pallawi, Patbandhare Nagar	$	@	□	*	∞	@	@
10	Ratna, Sadar Mount Road	#	@	□	*	€	@	@
11	Jaishri, Magan Pura	$	@	×	*	®	@	@
12	Trupti, Mantri Nagar	!	@	×	*	®	@	@
13	Manisha, Saptagiri Colony	!	@	■	@	∞	*	@
14	Shirish, Vithal Nagar	!	@	■	*	∞	@	@
15	Kumar, Chowk	#	@	□	*	®	*	@
16	Imran, Sarafa	$	@	□	*	®	*	@
17	Sarfaraz, Sarafa	#	@	■	@	∞	*	@
18	Jamal, Killa	!	@	×	*	©	*	@
19	Lilaben, Vazirabad	#	@	■	*	∞	*	@
20	Prashant, Sri Nagar	!	@	□	*	®	@	@
21	Prafful, Work shop	$	@	□	*	©	@	@
22	Ganesh, Itwara	!	@	×	@	®	@	@
23	Kanchan Singh, Gurudwara	#	@	×	@	©	*	@
24	Inder Pal Singh, Gurudwara	$	@	×	*	©	*	@
25	Savita, Sharda Nagar	!	@	■	*	®	*	@
26	Pranita, Sri Nagar	$	@	■	@	∞	*	@
27	Sarita, Ram Nagar	$	@	■	*	€	@	@
28	Sony, Taroda Naka	#	@	□	*	®	@	@
29	Sarika, Ganesh Nagar	!	@	×	@	€	*	@
30	Arti, Gurudwara	!	@	□	@	€	*	@
31	Sarang, Bhagya Nagar	!	@	×	*	®	@	@
32	Ravi, Datta Nagar	!	@	■	*	®	@	@
33	Purushottam, Samrat Nagar	!	@	■	*	®	*	@
34	Aniket, Vasant Nagar	#	@	×	*	€	@	@
35	Manoj, Anand Nagar	$	@	□	*	∞	*	@
36	Mohan, Taroda Naka	#	@	□	*	©	@	@
37	Vijay, Anand Nagar	!	@	□	*	®	*	@
38	Amol, Sarpanch Nagar	!	@	■	*	®	@	@

Sr. No.	Name and Address	Q. 2	Q. 3	Q. 4	Q. 5	Q. 6	Q. 7	Q. 8
39	Vishnu, Anand Nagar	#	@	×	*	∞	*	@
40	Nisha, Shrinagar	$	@	×	*	∞	*	@
41	Jaya, Kailas Nagar	#	@	×	*	€	@	@
42	Amol, Kailas Nagar	!	@	▫	*	∞	*	@
43	Rita, Vazirabad	!	@	×	*	∞	*	@
44	Shilpa, Sharada Nagar	#	@	×	*	®	*	@
45	Pranali, Anand Nagar	#	@	■	*	®	*	@
46	Prerana, Chikalwadi	!	@	■	*	∞	*	@
47	Mangesh, Ganraj Nagar	$	@	×	*	€	@	@
48	Ajit, Sarpanch Nagar	$	@	▫	*	∞	@	@
49	Varsha, Geeta Nagar	!	@	▫	*	®	*	@
50	Rinky, Saptgiri Nagar	!	@	■	*	∞	*	@

Data Analysis with Coding

Sr. No.	@	*	!	#	$	+	×	■	▫	®	€	©	∞
2.			23	15	12								
3.	50												
4.							19	16	15				
5.	10	40											
6.										19	11	5	15
7.	25	25											
8.	50												
Total	135	65	23	15	12	0	19	16	15	19	11	5	15

Limitations

1. Customers chose more than one option for the same question.
2. Hence, it was very difficult to decide which to write.
3. This may cause sampling error.

16.4.8 Findings

From this analysis, it was found that

1. Customers visit weekly to see a movie
2. Customers are satisfied with the available sitting arrangement
3. Customers are mostly satisfied with the parking facility
4. Customers prefer to see a comedy movie
5. Customers are also satisfied with the price of the ticket.

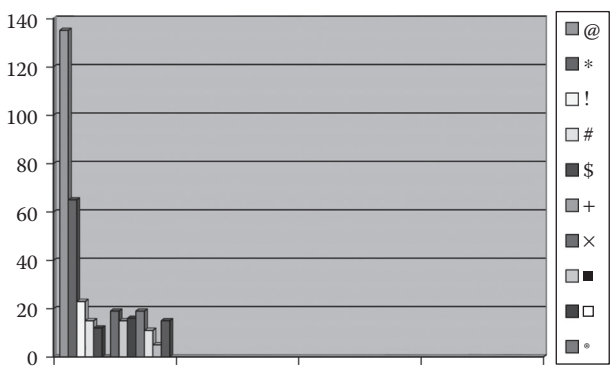
16.4.9 Conclusion

1. Ganesh talkies is good.
2. The services provided to the customers by Ganesh talkies are satisfactory.

16.4.10 Appendix: Questionnaires

1. May I know your name and address, please?
2. Are you coming here frequently?
   (Weekly/Monthly/Frequently)
3. Are you satisfied with the present sitting arrangement?
   (Yes/No)
4. Are they providing any extra facility?
   (Discount/Parking/Fresh and quality food/Booking)
5. Do you think that price of ticket is high?
   (Yes/No)
6. Which type of movie do you like?
   (Comedy/Family/Inspiring/Horror)
7. Do find any difficulty to reach the destination?
   (Yes/No)
8. Does the advertisement in newspaper or through the poster helps you to know about film whether it is doing good business or not.
   (Yes/No)

Sr. No	Particular	Coding
1	Yes	@
2	No	*
3	Weekly	!
4	Monthly	#
5	Frequently	$
6	Discount facility	+
7	Parking facility	×
8	Booking facility	■
9	Fresh and quality food	□
10	Comedy	®
11	Inspiring	€
12	Horror	©
13	Family	∞



17

Multiple Choice Questions with Answers and Necessary Explanation

Questions **Answers**

1. If the Bernoulli experiment is repeated with replacement, we get — d
   a. t-distribution
   b. z-distribution
   c. F-distribution
   d. Hyper-geometric distribution
   e. Normal distribution
2. Suppose there were 200 students in your high school graduating class. Current statistics indicate that 0.5%, or 0.005, of the population will become millionaires. What is the probability that at least one student from your classroom will become a millionaire? — e
   a. 0.5321
   b. 0.4321
   c. 0.3321
   d. 0.2321
   e. 0.6321

   Explanation:

   $\mu = np$

   $= 200(0.005)$

   $= 1.00.P(X \geq 1)$

   $= 1 - P(X = 0)$

   $= 1 - 0.3679$

   $= 0.6321$
3. A college basketball coach has 12 players on his roster. Eight of the players are receiving basketball scholarships, and four are not. Recently, the team have been losing most of their games. The coach decided to draw the names of five of their games. The coach decided to draw the names of five players out of a hat and designate them as the starting line-up. What is the probability that four of the five players selected are on a scholarship? — c
   a. 0.154
   b. 0.254
   c. 0.354

Questions **Answers**

d. 0.454

e. 0.554

Explanation:

$P(4) = ({}^{8}C_4)\,({}^{4}C_1)/{}^{12}C_5$

$= 280/792$

$= 0.354$

4. For a normal distribution where $\mu = 10$, $\sigma = 2$ and $X = 13$, the corresponding Z-value for X is — b

a. 3.00

b. 1.50

c. 4.50

d. 6.50

e. 2.00

Explanation:

$Z = (13 - 10)/2 = 1.50$

5. For a binomial random variable, $p = 0.70$ and $n = 50$. The mean and variance are, respectively, — a

a. 35, 10.5

b. 25, 9.5

c. 15, 18.5

d. 45, 11.5

e. 55, 12.5

6. The binomial distribution may be approximated by the normal distribution when np and n(1 – p) are both greater than — c

a. 3

b. 4

c. 5

d. 6

e. 7

Explanation: np and n(1 – p) should be greater than 5.

7. For a discrete probability distribution, the outcome must be — d

a. Exhaustive

b. Dependent

c. Independent

d. Mutually exclusive

e. Finite

8. In which of the following distributions is the probability of a success not the same for each trial? — d

a. Binomial

Questions	**Answers**
b. Poisson	
c. Normal	
d. Hyper-geometric	
e. Log-normal	
9. What does 6! 2!/4! 3! equal?	b
a. 9	
b. 10	
c. 12	
d. 14	
e. 15	
10. The means of all the sample means and population means are	b
a. Different	
b. Equal	
c. Smaller	
d. Larger	
e. Negligible	
11. The central limit theorem states that if the population is not normal, the sampling distribution of the mean approaches normal as the size of the sample	b
a. Decreases	
b. Increases	
c. Is constant	
d. Approaches ∞	
e. Approaches $-\infty$	
12. In which of the following sampling methods is the population divided into primary units and then samples drawn from the primary units?	b
a. Stratified	
b. Cluster	
c. Systematic	
d. Simple random	
e. Area	
13. Sampling distributions of the means shows all possible sample means and their probabilities of occurrence?	c
a. Frequency distribution	
b. Relative frequency distribution	
c. Probability distribution	
d. Cumulative frequency distribution	
e. Normal distribution	

Questions	Answers
14. The difference between the population parameter and the sample statistics is called the a. Sampling distribution b. Frequency distribution c. Sampling error d. Standard error e. Estimation	c
15. The distance between a selected value and the population mean divided by the standard deviation is the a. F-value b. t-value c. Normal value d. Z-value e. χ^2-value	d
16. The Poisson distribution is often used to approximate binomial probabilities when a. n is large and p is large b. n is small and p is large c. n is large and p is small d. n is small and p is small e. n is zero and p is zero	c
17. The population mean is equal to the a. Variance of the sample means b. Standard deviation of the sample means c. Standard error of the sample means d. Mean of the sample means e. Standard error of the population	d
18. In which of the following sampling methods is the first element of the population randomly selected to begin the sampling? a. Simple random sampling b. Stratified random sampling c. Systematic sampling d. Cluster sampling e. Area sampling	c
19. Sampling is always done a. With replacement b. Without replacement c. With and without replacement	b

Questions	**Answers**
d. Depends on population	
e. Depends on sample size	
20. Standard error is used as a tool in	a
a. Tests of hypothesis	
b. Tests of significance	
c. Both tests of hypothesis and tests of significance	
d. Either tests of hypothesis or tests of significance	
e. The calculation of difference in means	
21. Usually, which size of sample is considered a 'large sample'?	c
a. Greater than and equal to 20	
b. Greater than and equal to 25	
c. Greater than and equal to 30	
d. Greater than and equal to 35	
e. Greater than and equal to 40	
22. If the population standard deviation is not known, which distribution do we use?	c
a. F	
b. Z	
c. t	
d. χ^2	
e. Normal	
23. What is the condition for the difference between the hypothesised population parameter and the actual value for automatic rejection of our hypothesis? The sample size should be	a
a. Large	
b. Small	
c. Either large or small	
d. Both large and small	
e. Between –1 and +1	
24. Each new employee is given an identification number. The personnel files are arranged sequentially starting with employee number 0001. To sample the employees, the number 0153 was first selected. Then, numbers 0253, 0353, 0453 and so on became members of the sample. This type of sampling is called	b
a. Simple random sampling	
b. Systematic sampling	
c. Stratified random sampling	
d. Cluster sampling	
e. Area sampling	

Questions	**Answers**
25. 'Sampling error' as used in statistical inference a. Indicates that a Type I error has been made b. Indicates that a Type II error has been made c. Is the difference between a large statistic and its corresponding population parameter d. Indicates that the Z-test must be used e. Indicates that the t-test must be used	c
26. A Type II error is committed if we a. Reject a true null hypothesis b. Accept the null hypothesis when it is actually false c. Reject the alternate hypothesis when it is actually false d. Accept the alternate hypothesis when it is actually false e. Accept both the null and alternate hypotheses at the same time	b
27. A machine is programmed to produce tennis balls so that the mean bounce is 36 inches when the ball is dropped from a platform. The supervisor suspects that the mean bounce has changed and is less than 36 inches. An experiment is to be conducted using 42 tennis balls. The 5% significance level is to be used to test the hypothesis. The sample mean is computed to be 35.5 inches and the standard deviation of the sample 0.9 inches. What is the value of Z? a. –1.60 b. –2.60 c. –3.60 d. –4.60 e. –5.60 Explanation: H_0: $\mu = 36$ H_1: $\mu < 36$ H_0 is rejected if the computed value of Z is less than –1.645 $Z = (35.5 - 36)/(0.9/\sqrt{42})$ $= -3.60$	c
28.1 to 28.3. For the following problem: A firm with plant in two metropolitan areas adjusts the hourly wages paid to its employees in one area if there is a significant difference between the two population mean hourly wages. Based on the following sample data, is there a difference between the two mean wages? To solve the problem, answer these questions. 28.1. What are the null and alternate hypotheses? a. H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$ b. H_0: $\mu_1 < \mu_2$ H_1: $\mu_1 > \mu_2$	a

Questions	Answers
c. $H_0: \mu_1 > \mu_2$ $H_1: \mu_1 < \mu_2$	
d. $H_0: \mu_1 \neq \mu_2$ $H_1: \mu_1 = \mu_2$	
e. $H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 \geq \mu_2$	
28.2. This test belongs to which category?	a
a. Two-tailed test	
b. One-tailed test	
c. Right-tailed test	
d. Left-tailed test	
e. Either right-tailed or left-tailed test	
Explanation: This is a two-tailed test, because no direction is specified.	
28.3. Using the 0.05 level, what is the critical value or values?	e
a. H_0 is rejected if $Z < -1.96$	
b. H_0 is rejected if $Z > -1.96$	
c. H_0 is rejected if $Z \leq -1.96$ and $Z \geq 1.96$	
d. H_0 is rejected if $Z \geq -1.96$ and $Z \leq 1.96$	
e. H_0 is rejected if $Z < -1.96$ or $Z > 1.96$	
Explanation:	
H_0 is rejected if $Z < -1.96$ or $Z > 1.96$	
29. To test a hypothesis involving proportions, both np and n(1 − p) should	b
a. Exceed 30	
b. Exceed 5	
c. Lie in the range from 0 to 1	
d. Be at least –2.58	
e. Be at least +2.58	
30. The 0.01 level is used in an experiment, and a one-tailed test is applied. Computed Z is –1.8. This indicates that	a
a. H_0 should not be rejected.	
b. We should reject H_0 and accept H_1.	
c. We should take a larger sample.	
d. We should take a smaller sample.	
e. We should have used the 0.05 level of significance.	
31. The hypotheses are $H_0: \mu = 240$ inches of pressure; $H_1: \mu \neq 240$ inches of pressure.	b
a. A one-tailed test is being applied.	
b. A two-tailed test is being applied.	
c. A three-tailed test is being applied.	
d. The right test is being applied.	
e. The wrong test is being applied.	

Questions	**Answers**
32. What level of measurement is required for the goodness of fit test? a. Ordinal level and nominal level b. Ordinal level or nominal level c. At least nominal level d. At least ordinal level e. Any other level	c
33.1 to 33.3. For the following problem: A firm manufacturing two types of electrical items, A and B, can make a profit of Rs. 20 per unit of A and Rs. 30 per unit of B. Both A and B make use of two essential components: a motor and a transformer. Each unit of A requires three motors and two transformers, and each unit of B requires two motors and four transformers. The total supply of components per month is restricted to 210 motors and 300 transformers. Type B also requires a voltage stabiliser, the supply of which is restricted to 65 units per month. Answer the following: 33.1. The objective function will be a. Minimise $Z = 20x + 30y$ b. Maximise $Z = 20x + 30y$ c. Minimise $Z = 30x + 20y$ d. Maximise $Z = 30x + 20y$ e. Maximise or minimise $Z = 20x \pm 30y$	b
33.2. On which of the following sets of degrees of freedom is the F-distribution based? a. 2 b. 3 c. 1 d. 4 e. 5	a
33.3. The source of variation in the data is a. Treatment b. Error c. Mean square treatment d. Mean square error e. Sum of squares due to error	a
34. For the ANOVA procedure, the population should be skewed a. Negatively b. Positively c. Normally	c

Questions **Answers**

d. Log-normally

e. Exponentially

35. Rejecting the null hypothesis in an ANOVA procedure indicates that — d

a. All pairs of means differ.

b. All pairs of means are equal.

c. Some pairs of means are equal.

d. At least one pair of means is different.

e. At least one pair of means is equal.

36. If the significance level is 0.05 and there are 3 degrees of freedom in the numerator and 12 in the denominator, the critical value of F is — c

a. 1.49

b. 2.49

c. 3.49

d. 4.49

e. 5.49

37.1 to 37.4. — b

The following is a two-way ANOVA table.

Source	Sum of Squares	Degrees of Freedom	Mean Square
Treatment	50	2	25
Blocks	24	3	8
Error	48	6	8
Total	122	11	

37.1. How many treatments are there?

a. 2

b. 3

c. 4

d. 5

e. 6

37.2. How many blocks are there? — c

a. 2

b. 3

c. 4

d. 5

e. 6

37.3. How many samples are there in the problem? — a

a. 12

b. 11

c. 10

Questions **Answers**

d. 9

e. 8

37.4. Conduct a test for treatments. Is there a significant difference among the treatment means? Use the 0.05 significance level. a

a. No difference in the treatment means

b. Difference in the treatment means

c. No difference in the block means

d. Difference in the block means

e. Difference in treatment and block means

Explanation:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: Not all treatment means are the same.

H_0 is rejected if $F > 5.14$ (using the 0.05 significance level). $F = 25/8 = 3.125$.

H_0 is rejected. There is no difference in the treatment means.

38. The square of the coefficient of correlation is called c

a. Coefficient of correlation

b. Coefficient of regression

c. Coefficient of determination

d. Coefficient of non-determination

e. Pearson's product moment correlation

39. The variable used to predict another variable is called the b

a. Dependent variable

b. Independent variable

c. Correlation variable

d. Student's t-variable

e. Random variable

40. The method used to arrive at the 'best-fitting' straight line in regression analysis is d

a. Freehand method

b. Non-determination method

c. Determination method

d. Least-squares method

e. Correlation method

41. In the regression equation for the straight line, the value of b would be about

a. –1.00

b. +1.00

c. 0

Questions	Answers
d. −0.50	c
e. +0.50	
42. The seasonal variation is the pattern in a time series within a	b
a. Week	
b. Month	
c. Year	
d. 3 months	
e. Half year	
43. The behaviour of a variable is considered to be cyclic only if the movements recur after a period of more than	c
a. 1 month	
b. 6 months	
c. 1 year	
d. 18 months	
e. 2 years	
44. Seasonal factors are usually computed on a	c
a. Monthly basis	
b. Quarterly basis	
c. Monthly or quarterly basis	
d. Yearly basis	
e. Monthly and quarterly basis	
45. The choices available to the decision maker are called the	d
a. Options	
b. Either	
c. Or	
d. Alternatives	
e. Except	
46. Future events that cannot be controlled are called the	c
a. Uncontrolled nature	
b. Uncertain nature	
c. States of nature	
d. Condition of nature	
e. Uncertainty of nature	
47. The combination of a particular state of nature and a particular decision alternative is called the	b
a. Pay-off table	
b. Pay-off	
c. State of nature	
d. Opportunity loss	
e. Maximax	

Questions	**Answers**
48. All possible combinations of states of nature and decision alternatives is called a a. Pay-off table b. Pay-off c. State of nature d. Opportunity loss e. Maximax	a
49. Expected value, E(X) = a. $X\ \Sigma P(X)$ b. $X^2\ \Sigma P(X)$ c. $X\ \Sigma P(X)^2$ d. $\Sigma\ XP(X)$ e. $\Sigma\ XP(X)^2$	d
50. For the following sample of paired observations X: 2 4 6 8 10 Y: 10 8 6 4 2 Covariance between X and Y is a. −8 b. +8 c. −10 d. +10 e. ±8	c
51. For the following paired data of a population X: 4 4 5 6 7 10 Y: 6 7 5 6 4 1 f: 1 3 5 6 4 1 Covariance between X and Y is a. −10,125 b. −1.225 c. +1.225 d. −1.425 e. +1.425	d
52. In the case of a binomial distribution, the trials are statistically a. Mutually exclusive b. Not mutually exclusive c. Independent d. Dependent e. Either mutually exclusive or not mutually exclusive	c

Questions	**Answers**
53. The purpose of regression analysis is to arrive at the regression equation to predict the value of a. One variable b. Two variables c. n – 1 variables d. n variables e. n + 1 variables	a
54. For the binomial distribution, mean and variance are a. n and np b. np and np^2 c. np and nq^2 d. np and npq e. n and np	d
55. Because of the symmetry of the normal probability distribution, which of the following distributions are also at the centre? a. Mean and mode b. Mean and standard deviation c. Median and variance d. Mean and mode e. Median and mode	e
56. If σ is the standard deviation of the normal distribution, what percentage of the observation will be in the interval μ – 1.28 σ to μ + 1.28 σ? a. 98% b. 95% c. 99% d. 80% e. 97%	d
57. What is the probability that an observation from a standard normal distribution will lie in the interval –1.96 to 1.96? a. 98% b. 95% c. 99% d. 80% e. 97%	b
58. What is the probability that an observation from a standard normal distribution will lie in the interval –2.33 to 2.33? a. 98% b. 95% c. 99%	a

Questions	Answers
d. 80% e. 97%	
59. What is the probability that an observation from a standard normal distribution will lie in the interval –1.28 to 1.28? a. 98% b. 95% c. 99% d. 80% e. 97%	d
60. A normal variable X has a mean of 56 and a standard deviation of 12. The Z value corresponding to the X value of –5 is a. –1.08 b. –2.08 c. –3.08 d. –4.08 e. –5.08	e
61. A normal variable has a mean of 10 and a standard deviation of 5. What is the probability that the normal variable will take a value in the interval 0.2 to 19.8? a. 98% b. 95% c. 99% d. 80% e. 97%	b
62. In which sampling method does each element have an equal chance of being selected but each sample not have the same chance of being selected? a. Stratified b. Simple random c. Systematic d. Cluster e. Random sampling	c
63. A factory produces a certain type of output by three types of machine. The respective daily production figures are Machine I: 3000 Units; Machine II: 2500 Units; Machine III: 4500 Units. Past experience shows that 1% of the output produced by Machine I is defective. The corresponding fractions of defective produce for the other two machines are 1.2% and 2%, respectively. An item is drawn at random from the day's production run and is found to be defective. What is the probability that it comes from the output of random from the day's production run and is found to be defective. The probability that defective items comes from the output of Machine I, Machine II and Machine III, respectively?	a

Questions **Answers**

a. 1/5, 1/5, 3/5
b. 3/5, 1/5, 1/5
c. 3/5, 3/5, 1/5
d. 1/5, 3/5, 1/5
e. 3/5, 3/5, 3/5
f. 1/5, 1/5, 1/5

64. In the throw of a die, the six faces numbered 1, 2, 3, 4, 5 and 6 are — c
    a. Equally likely
    b. Independent
    c. Mutually exclusive
    d. Dependent
    e. Events or cases

    Explanation:

    Here, the happening of any one of them excludes the happening of all others in the same experiment.

65. A Poisson distribution may be obtained as a limiting case of binomial probability distribution under the following conditions. Which one of the following is not a condition? — c
    a. The number of trials is indefinitely large.
    b. The probability of success is small.
    c. The probability of success is large.
    d. np is finite.
    e. m = np.

66. In the summer, a truck driver gets on average one puncture in 1000 km. Find the probability that he will have no puncture. — c
    a. e^{-1}
    b. e^{-2}
    c. e^{-3}
    d. e^{-4}
    e. e^{-5}

67. Simple random sampling is not a popular method because — d
    a. It requires an up-to-date frame.
    b. It is time consuming and costly.
    c. It may not reflect the true characteristics of the population.
    d. It is not a probability sampling.
    e. It gives highly improbabilistic results.

68. If the two lines of regression are $4x - 5y + 30 = 0$ and $20x - 9y - 107 = 0$, which of these is the line of regression of x on y? — b
    a. First
    b. Second

Questions **Answers**

c. First and second both

d. Neither first nor second

e. Can't say

Explanation:

Suppose that the first equation is the equation of the line of regression of x and y, and the second is the equation of the line of regression of y on x. Bring in the form of $y = mx + c$ and find the slopes, denote them by b_{xy} and b_{yx}, and use the formula $r^2 = b_{yx}. b_{xy}$. If $r^2 < 1$, then our supposition is correct; otherwise, it is wrong.

69. In systematic sampling, which unit is selected at random? d

a. All

b. Some

c. Odd

d. First

e. Even

70. The equations of two lines of regression obtained in a correlation analysis are $2x = 8 - 3y$ and $2y = 5 - x$. Obtain the value of the correlation coefficient. b

a. $r = +0.866$

b. $r = -0.866$

c. $r = +0.086$

f. $r = -0.086$

g. $r = \pm 0.0086$

71. If two regression coefficients are 0.8 and 1.2, what would be the value of the coefficient of correlation? a

a. 0.9798

b. 0.9879

c. 0.9978

d. 0.9897

e. 0.9798

72. Primary data is collected by _____ methods. c

a. 2

b. 3

c. 4

d. 5

e. 6

73. Name the data that is originally collected for any investigation. b

a. Secondary

b. Primary

c. Primary and secondary

Questions **Answers**

d. Primary or secondary

e. Neither primary nor secondary

74. A factory is producing 50,000 pairs of shoes daily. From a sample of 500 pairs, 2% were found to be of substandard quality. Estimate the number of pairs that can be reasonably expected to be spoiled in daily production and assign limits at the 95% level of confidence. e

a. (0.0087, 0.0322)

b. (0.0887, 0.3322)

c. (0.0787, 0.2322)

d. (0.0787, 0.3322)

e. (0.0078, 0.0322)

75. Two dice are thrown, and the sum of the numbers on the faces up is obtained. The probability of this sum being 2 is b

a. 1/6

b. 1/36

c. 1/18

d. 1/9

e. 1/3

76. The chance of drawing a white ball in the first draw and again a white ball in the second draw without replacement of the ball in the first draw from a bag containing six white and four red balls is e

a. 2/10

b. 6/10

c. 5/10

d. 36/100

e. 1/3

77. For two independent events, A and B, for which P(A) = 1/2 and P(B) = 1/3, the probability that at least one of them occurs is a

a. 5/6

b. 2/3

c. 1/2

d. 4/3

e. 5/3

Explanation: $P(H_2|H_1) = P(H_2) = 1/2$ (since H_1 and H_2 are independent).

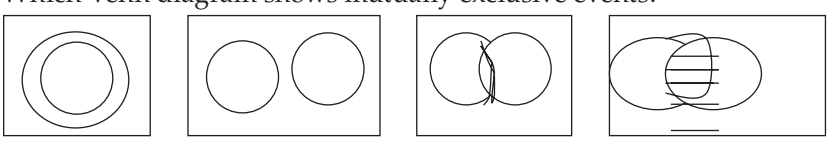
78. Which of the following is an example of a Bernoulli process? c

a. Tossing a coin a random number of times

b. Tossing a coin till a head is obtained

c. Tossing a coin for a fixed number of times

d. Tossing a coin till 50% success is obtained

Questions	**Answers**
79. The value of r^2 for a particular situation is 0.49. What is the coefficient of correlation? a. 0.49 b. 0.7 c. 0.07 d. Cannot be determined for the information given.	d
80. For solving a problem using quantitative methods, the following steps are involved: a. Defining the problem b. Observing the organisational environment c. Data collection d. Arriving at the solution e. Constructing a model f. Presenting the solution Which is the correct order of solving the problem in Q.M.? i. abcdef ii. acbedf iii. baecdf iv. fedcba v. abdcef	iii
81. What is the general form of relationship among A.M, G.M. and H.M.? a. A.M. $\leq$ G.M. $\geq$ H.M. b. A.M. $\geq$ G.M. $\geq$ H.M. c. A.M. $\leq$ G.M. $\leq$ H.M. d. A.M. = H.M. $\leq$ G.M.	b
82. Which Venn diagram shows mutually exclusive events?  (a) (b) (c) (d)	b
83. If two events, A and B, are not mutually exclusive, then the probability that neither A nor B occurs is equal to a. 1 – P(AB) b. P(AB) – 1 c. 1 – [P(A) = P(B) – P(AB)] d. P(A) = P(B) – P(AB) – 1	c

Questions	**Answers**
84. One card is drawn from a deck of 52 cards. What is the probability of the card being either red or a king? a. 7/13 b. 1/26 c. 1/13 d. 1/2	a
85. The variation due to events such as floods, strikes and war in time-series data can be classified as a. Secular trend b. Cyclical variation c. Seasonal variation d. Irregular variation	d

Bibliography

Adguna, W., and Labuschagne, M. (2002). Genotypeenvironment interactions and phenotypic stability analysis of linseed in Ethiopia. *Plant Breeding*, *121*, 66–71.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley.

Albert, J., and Bennett, J. (2003). *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game.* New York: Springer.

Allison, T., and Cicchetti, D. (1976). Sleep in mammals: Ecological and constitutional correlates. *Science*, *194*, 732–734.

Anderson, T. W. (1958). *An Introduction to Multivariate Analysis*. New York: Wiley.

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W., and Tukey, J. (1972). *Robust Estimates of Location*. Princeton, NJ: Princeton University Press.

Andrews, D., and Herzberg, A. M. (1985). *Data*. New York: Springer-Verlag.

Anscombe, F. J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika, 37,* 358–382.

Armitage, P. (1983). *Statistical Methods in Medical Research*. Boston, MA: Blackwell.

Bailey, C., Cox, E., and Springer, J. (1978). High pressure liquid chromatographic determination of the intermediate/side reaction products in FD&C Red No. 2 and FD&C Yellow No. 5; Statistical analysis of instrument response. *J. Assoc. Offic. Anal. Chem., 61,* 1404–1414.

Barlow, R. E., Toland, R. H., and Freeman, T. (1984). A Bayesian analysis of stress-rupture life of Kevlar/epoxy spherical pressure vessels. In *Proceedings of the Canadian Conference in Applied Statistics*. T. D. Dwivedi (ed.). New York: Marcel-Dekker.

Barnothy, J. M. (1964). Development of young mice. In *Biological Effects of Magnetic Fields*. M. Barnothy (ed.). New York: Plenum Press.

Beecher, H. K. (1959). *Measurment of Subjective Responses*. Oxford, England: Oxford University Press.

Beller, G., Smith, T., Abelmann, W., Haber, E., and Hood, W. (1971). Digitalis intoxication: A prospective clinical study with serum level correlations. *N. Eng. J. Med., 284,* 989–997.

Benjamin, J., and Cornell, C. (1970). *Probability, Statistics, and Decision for Civil Engineers*. New York: McGraw-Hill.

Bennett, C., and Franklin, N. (1954). *Statistical Analysis in Chemistry and the Chemical Industry.* New York: Wiley.

Bernstein, P. (1998). *Against the Gods: The Remarkable Story of Risk*. New York: Wiley.

Berkson, J. (1966). Examination of randomness of alpha particle emisssions. In *Research Papers in Statistics*. F. N. David (ed.). New York: Wiley.

Berry, W. L. et al. (1980). *Management Decision Sciences*. Homewood, IL: R.D. Irwin.

Bevan, S., Kullberg, R., and Rice, J. (1979). An analysis of cell membrane noise. *Ann. Stat., 7,* 237–257.

Bhattacharjee, C., Bradley, P., Smith, M., Scally, A., and Wilson, B. (2000). Do animals bite more during a fall moon? Retrospective observational analysis. *BMJ, 321,* 1559–1561.

Bickel, P., Chen, C., Kwon, J., Rice, J., van Zwet, E., and Varaiya, P. (2004). Measuring traffic. Berkeley Department of Statistics Technical Report 664.

Bickel, P., and Doksum, K. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Oakland, CA: Holden-Day.

Bickel, P., and Doksum, K. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*. Upper Saddle River, NJ: Prentice-Hall.

Bickel, P., and O'Connell, J. W. (1975). Is there a sex bias in graduate admissions? *Science*, *187*, 398–404.

Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press.

Bjerkdal, T. (1960). Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli. *Am. J. Hyg., 72,* 130–148.

Bliss, C., and Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics*, *9*, 174–200.

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Royal Stat. Soc., Series B, 26*, 211–246.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. New York: Wiley.

Box, G. E. P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Brownlee, K. A. (1960). *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley.

Brunk, H. D. (1975). *An Introduction to Mathematical Statistics*. Gardena, CA: Xerox.

Burr, I. (1974). *Applied Statistical Methods*. New York: Academic Press.

Carey, J. R., Liedo, P., Orozco, D., and Vaupel, J. W. (1992). Slowing of mortality rates at older ages in large medfly cohorts. *Science*, *258*, 457–461.

Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P. (1983). *Graphical Methods for Data Analysis*. Boston, MA: Duxbury.

Chance, W. (1969). *Statistical Methods for Decision Making*. Homewood, IL: R.D. Irwin.

Chang, A. E., et al. (1979). Delta-9-Tetrahydrocannabinol as an antiemetic in cancer patients receiving high dose methotrexate. *Ann. Intern. Med.*, *91*, 819–824.

Chang, S. L. (1945). Sedimentation in water and the specific gravity of cysts of *Entamoeba histolytica*. *Am. J. Hyg., 41*, 156–163.

Chapman, H., and Demeritt, D. (1936). *Elements of Forest Mensuration*. Nashville, TN: Williams Press.

Chernoff, H., and Lehman, E. (1954). The use of maximum likelihood estimates in tests for goodness of fit. *Ann. Math. Statist., 23*, 315–345.

Chou, Y. (1972). *Probability and Statistics for Decision Making*. New York: Holt, Rinehart and Winston.

Clancy, V. J. (1947). Empirical distributions in chemistry. *Nature*, *159*, 340.

Clark, T. C., and E. W. Jordan (1985). *Introduction to Business and Economic Statistics.* Cincinnati, OH: South-Western Publishing.

Cleveland, W., Greadel, T., Kleiner, B., and Warner, J. (1974). Sunday and workday variations in photochemical air pollutants in New Jersey and New York. *Science, 186*, 1037–1038.

Cobb, L., Thomas, G., Dillard, D., Merendino, J., and Bruce, R. (1959). An evaluation of internal mammary artery ligation by a double blind technique. *N. Eng. J. Med.*, *260*, 1115–1118.

Cochran, W. G. (1977). *Sampling Techniques*. New York: Wiley.

Cogswell, J. J. (1973). Forced oscillation technique for determination of resistance to breathing in children. *Arch. Dis. Child.*, *48*, 259–266.

Converse, P., and Traugott, M. (1986). Assessing the accuracy of polls and surveys. *Science, 234*, 1094–1098.

Cook, R. D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Crowe, W. R. (1965). *Index Numbers: Theory and Applications.* London: Macdonald & Evans.

Cummings, K. M., Giovino, G., Sciandra, R., Koenigsberg, M., and Emont, S. (1987). Physician advice to quit smoking: Who gets it and who doesn't? *Am. J. Prev. Med.*, *3*, 69–75.

Czitrom, V., and Reece, J. (1997). *Statistical Case Studies for Process Improvement*. Philadelphia, PA: SIAM-ASA, 87–103.

Dahiya, R., and Gurland, J. (1972). Pearson chi-squared test of fit with random intervals. *Biometrika*, *59*, 147–153.

Dahlquist, G., and Bjorck, A. (1974). *Numerical Methods*. Englewood Cliffs, NJ: Prentice-Hall.

David, H. (1981). *Order Statistics*. New York: Wiley.

Davies, O. (1960). *The Design and Analysis of Industrial Experiments*. London: Oliver and Boyd.

De Forina, M., Armanino, C., Lanteri, S., and Tiscornia, E. (1983). Classification of olive oils from their fatty acid composition. In *Food Research and Data Analysis*, 189–214. H. Martens and H. Russwurm Jr. (eds.). London: Applied Science Publishers.

DeHoff, R., and Rhines, F. (eds.). (1968). *Quantitative Microscopy*. New York: McGraw-Hill.

Deming, W. (1960). *Sample Design in Business Research*. New York: Wiley.

Diamond, G., and Forrester, J. (1979). Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. *N. Eng. J. Med., 300*, 1350–1358.
Doksum, K., and Sievers, G. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, *63*, 421–434.
Dongarra, J. (1979). *LINPACK Users' Guide*. Philadelphia, PA: SIAM.
Donoho, A., Donoho, D., and Casko, M. (1986). *Mac-Spin: Dynamic Data Display*. Belmont, CA: Wadsworth.
Dorfman, D. (1978). The Cyril Burt question: New findings. *Science, 201*, 1177–1186.
Dowdall, J. A. (1974). Women's attitudes towards employment and family roles. *Soc. Anal., 35*, 251–262.
Draper, N., and Smith, H. (1981). *Applied Regression Analysis*. New York: Wiley.
Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In *Judgment under Uncertainty: Heuristics and Biases*, 249–267. D. Kahneman, P. Slovic, and A. Tversky (eds.). Cambridge: Cambridge University Press.
Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psych. Rev., 70,* 193–242.
Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
Enns, P. G. (1985). *Business Statistics.* Homewood, IL: R.D. Irwin.
Evans, D. (1953). Experimental evidence concerning contagious distributions in ecology. *Biometrika, 40*, 186–211.
Fechter, J. V., and Porter, L. G. (1979). *Kitchen Range Energy Consumption*. Washington, DC: Prepared for Office of Conservation, U. S. Department of Energy, NBSIR 78–1556.
Feller, W. (1957). *An Introduction to Probability Theory and Its Application*. New York: Wiley.
Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.
Filliben, J. (1975). The probability plot correlation coefficient test for normality. *Technometrics, 17*, 111–117.
Finkner, A. (1950). Methods of sampling for estimating commercial peach production in North Carolina. *N. C. Agri. Exp. Stn. Tech. Bull., 91*, 29.
Fisher, R. A. (1936). Has Mendel's work been rediscovered? *Ann. Sci., 1*, 115–137.
Fisher, R. A. (1958). *Statistical Methods for Research Workers*. New York: Hafner.
Freedman, D., Pisani, R., and Purves, R. (1978). *Statistics*. New York: Norton.
Gardner, M. (1976). Mathematical games. *Sci. Am.*, 234, 119–123.
Gastwirth, J. (1987). The statistical precision of medical screening procedures. *Statistical Science, 3*, 213–222.
Geissler, A. (1889). Beitrage zur Frage des Geschlechtsverhaltnisses der Gerbornen *Z. K. Sachs. Stat. Bur., 35*, 1–24.
Gerlough, D., and Schuhl, A. (1955). *Use of Poisson Distribution in Highway Traffic*. Saugatuck, CT: Eno Foundation for Highway Traffic Control.
Glass, D., and Hall, J. (1954). A study of intergeneration changes in status. In *Social Mobility in Britain*. D. Glass (ed.). Glencoe, IL: Free Press.
Gosset, W. S. (1931). The Lanarkshire milk experiment. *Biometrika, 23*, 398.
Grace, N., Muench, H., and Chalmers, T. (1966). The present status of shunts for portal hypertension in cirrhosis. *Gastroenterology, 50*, 684–691.
Gupta, S. P., and M. P. Gupta. (1988). *Business Statistics.* New Delhi: Sultan Chand & Sons.
Haberman, S. (1978). *Analysis of Qualitative Data*. New York: Academic Press.
Hampson, R., and Walker, R. (1961). Vapor pressures of platinum, iridium, and rhodium. *J. Res. Nat. Bur. Stand., 65 A*, 289–295.
Hanley, J. A., and Shapiro, S. H. (1994). Sexual activity and the lifespan of male fruitflies: A dataset that gets attention. *J. Stat. Edu., 2*(1).
Harbaugh, J., Doveton, J., and Davis, J. (1977). *Probability Methods in Oil Exploration*. New York: Wiley.
Hartley, H. O., and Ross, A. (1954). Unbiased ratio estimates. *Nature, 174*, 270–271.
Heckman, M. (1960). Flame photometric determination of calcium in animal feeds. *J. Assoc. Offic. Anal. Chem., 43*, 337–340.

Hennekens, C., Drolette, M., Jesse, M., Davies, J., and Hutchison, G. (1976). Coffee drinking and death due to coronary heart disease. *N. Eng. J. Med.*, *294*, 633–636.

Herson, J. (1976). An investigation of the relative efficiency of least-squares prediction to conventional probability sampling plans. *J. Amer. Stat. Assoc., 71*, 700–703. From the National Center for Health Statistics Hospital Discharge Survey (January 1968).

Hill, R. A., and Barton, R. A. (2005). Red enhances human performance in contests. *Nature*, *435*, 293.

Hoaglin, D. (1980). A poissonness plot. *Amer. Stat., 34*, 146–149.

Hoaglin, D., Mosteller, F., and Tukey, J. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

Hockersmith, T., and Ku, H. (1969). Uncertainties associated with proving ring calibration. In *Precision Measurement and Calibration*. H. Ku (ed.). Washington, DC: U.S. National Bureau of Standards Special Publication 300, Vol. I.

Hollander, M., and Wolfe, D. (1973). *Nonparametric Statistical Methods*. New York: Wiley.

Hopper, J. H., and Seeman, E. (1994). The bone density of female twins discordant for tobacco use. *N. Eng. J. Med.*, *330*, 387–392.

Horvitz, D., Shah, B., and Simmons, W. (1967). The unrelated randomized response model. *Proc. Soc. Stat. Sect. Amer. Stat.*, 65–72.

Houck, J. C. (1970). Temperature coefficient of the bismuth I-II transition pressure. *J. Res. Nat. Bur. Stand., 74 A*, 51–54.

Huber, P. (1981). *Robust Statistics*. New York: Wiley.

Huie, R. E., and Herron, J. T. (1972). Rates of reaction of atomic oxygen III, spiropentane, cyclopentane, cyclohexane, and cycloheptane. *J. Res. Nat. Bur. Stand., 74 A*, 77–80.

Johnson, S., and Johnson, R. (1972). Tonsillectomy history in Hodgkin's disease. *N. Eng. J. Med.*, *287*, 1122–1125.

Joiner, B. (1981). Lurking variables: Some examples. *Amer. Stat.*, 35, 227–233.

Kacprzak, J., and Chvojka, R. (1976). Determination of methyl mercury in fish by flameless atomic absorption spectroscopy and comparison with an acid digestion method for total mercury. *J. Assoc. Offic. Anal Chem.*, *59*, 153–157.

Kamal, A. A., Eldamaty, S. E., and Faris, R. (1991). Blood lead level of Cairo traffic policeman. *Sci. Total Environ.*, *105*, 165–170.

Kiefer, M. J., Buchan, R. M., Keefe, T. J., and Blehm, K. D. (1987). A predictive model for determining asbestos concentrations for fibers less than five micrometers in length. *Environ. Res.*, *43*, 31–38.

Kirchoefer, R. (1979). Semiautomated method for the analysis of chlorpheniramine maleate tables: Collaborative study. *J. Assoc. Offic. Anal. Chem., 62*, 1197–1120.

Kiser, C. V., and Schaefer, N. L. (1949). Demographic characteristics of women in Who's Who. *Milbank Mem. Fund Q.*, 27, 422.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Knafl, G., Spiegelman, C., Sacks, J., and Ylvisaker, D. (1984). Nonparametric calibration. *Technometrics*, *26*, 233–241.

Ku, H. (1969). *Precision Measurement and Calibration*. Washington, DC: National Bureau of Standards Special Publication 300.

Ku, H. (1981). Personal communication.

Lagakos, S., and Mosteller, F. (1981). FD&C Red No. 40 experiments. *J. Natl. Cancer Inst., 66*, 197–213.

Lawson, C. L., and Hanson, R. J. (1974). *Solving Least Squares Problems*. Englewood Cliffs, NJ: Prentice-Hall.

Lazarsfeld, P., Berelson, B., and Gaudet, H. (1948). *The People's Choice: How the Voter Makes UP His Mind in a Presidential Election*. New York: Columbia University Press.

Le Cam, L., and Neyman, J. (eds.). (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume V: Weather Modification*. Berkeley, CA: University of California Press.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Oakland, CA: Holden-Day.

Lehmann, E. L. (1983). *Theory of Point Estimation*. New York: Wiley.

Lehmann, E. L., and Casella, G. (1998). *Theory of Point Estimation*. New York: Springer.
Levine, J. D., Gordon, N. C., and Fields, H. L. (1978). The mechanism of placebo analgesia. *Lancet*, *312*, 654–657.
Levine, P. H. (1973). An acute effect of cigarette smoking on platelet function. *Circulation, 48*, 619–623.
Levin, R. I. (1979). *Statistics for Management*. New Delhi: Prentice Hall of India.
Lin, S.-L., Sutton, V., and Quarashi, M. (1979). Equivalence of microbiological and hydroxylamine methods of analysis for ampicillin dosage forms. *J. Assoc. Offic. Anal. Chem.*, *62*, 989–997.
Lynd, R. S., and Lynd, H. M. (1956). *Middletown: A Study in Modern American Culture*. New York: Harcourt-Brace.
MacFarquhar, L. (2004). The pollster. *New Yorker*. October 18.
Malkiel, B. G. (2004). *A Random Walk Down Wall Street: Completely Revised and Updated Eighth Edition*. W. W. Norton.
Marshall, C. G., Ogden, D. C., and Colquhoun, D. (1990). The actions of suxamethonium (succinylcholine) as an agonist and channel blocker at the nicotinic receptor of frog muscle. *J. Physiol.*, *428*, 155–174.
Martin, H., Gudzinowicz, B., and Fanger, H. (1975). *Normal Values in Clinical Chemistry*. New York: Marcel-Dekker.
McCool, J. (1979). Analysis of single classification experiments based on censored sample from the two-parameter Weibull distribution. *J. Stat. Planning and Inference*, *3*, 39–68.
McNish, A. (1962). The speed of light. *IRE Trans. on Instrumentation*, *11*, 138–148.
Mecklenburg, R. S., Benson, E. A., Benson, J. W., Fredlung, P. N., Guinn, T., Metz, R. J., Nielson, R. L., and Sanner, C. A. (1984). Acute complications associated with insulin pump therapy: Report of experience with 161 patients. *J. Amer. Med. Assoc.*, *252*(23), 3265–3269.
Miller, R. (1981). *Simultaneous Statistical Inference*. New York: Springer-Verlag.
Mittleman, M. A., Maclure, M., Tofler, G. H., et al. (1993). Triggering of acute myocardial infarction by heavy exertion. *N. Eng. J. Med.*, *329*, 1677–1683.
Morton, A. Q. (1978). *Literary Detection*. New York: Scribner's.
Mudgett, B. D. (1951). *Index Numbers*. New York: Wiley.
Natrella, M. (1963). *Experimental Statistics*. Washington, DC: National Bureau of Standards Handbook 91.
Olsen, A., Simpson, J., and Eden, J. (1975). A Bayesian analysis of a multiplicative treatment effect in whether modification. *Technometrics, 17*, 161–166.
Orcutt, R. H. (1970). Generation of controlled low pressures of nitrogen by means of dissociation equilibria. *J. Res. Nat. Bur. Stand.*, *74 A*, 45–49.
Overfield, T., and Klauber, M. R. (1980). Prevalence of tuberculosis in Eskimos having blood group B gene. *Hum. Bio., 52*, 87–92.
Patridge, L., and Marion, M. (1981). Sexual activity and the lifespan of male friutflies. *Nature*, *294*, 580–581.
Pearson, E., D'Agostino, R., and Bowman, K. (1977). Tests for departure from normality: Comparison of powers. *Biometrika*, *64*, 231–246.
Pearson, E. S., and Wishart, J. (eds.). (1958). *Student's Collected Works*. Cambridge, England: Cambridge Press.
Pearson, K., and Hartley, H. (1966). *Biometrika Tables for Statisticians*. Cambridge, England: Cambridge University Press.
Peck, R., Casella, G., Cobb, G., Hoerl, R. Nolan, D., Starbuck, R., and Stern, H. (2005). *Statistics: A Guide to the Unknown*. Duxbury Press.
Perry, L., Van Dyke, R., and Theye, R. (1974). Sympathoadrenal and hemodynamic effects of isoflurane, halothane, and cyclopropane in dogs. *Anesthesiology, 40*, 465–470.
Phillips, D. P., and King, E. W. (1988). Death takes a holiday: Mortality surrounding major social occasions. *Lancet*, *2*, 728–732.
Phillips, D. P., and Smith, D. G. (1990). Postponement of death until symbolically meaningful occasions. *J. Amer. Med. Assoc., 263*, 1947–1961.

Plato, C., Rucknagel, D., and Gerschowitz, H. (1964). Studies of the distribution of glucose-6-phosphate dehydrogenase deficiency, thalassemia, and other genetic traits in the coastal and mountain villages of Cyprus. *Am. J. Hum. Genet*., 16, 267–283.

Preston-Thomas, H., Turnbull, G., Green, E., Dauphinee, T., and Kalra, S. (1960). An absolute measurement of the acceleration due to gravity at Ottawa. *Can. J. Phys., 38*, 824–852.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika*, *43*, 353–360.

Raftery, A., and Zeh, J. (1993). Estimation of bowhead whale, *Balaena mysticetus*, population size. In *Case Studies in Bayesian Statistics, Springer Lecture Notes in Statistics, 83,* 163–240. C. Gatsonis, J. Hodges, R. Kass, and N. Singpurwalla (eds.). New York: Springer.

Raiffa, H. (1970). *Decision Analysis*. Reading, MA: Addison-Wesley.

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. New York: Wiley.

Ratcliff, J. (1957). New surgery for ailing hearts. *Reader's Digest*, *71*, 70–73.

Redelmeier, D. A., and Tibshirani, R. J. (1997a). Association between cellular-telephone calls and motor vehicle collisions. *N. Eng. J. Med., 336*, 453–458.

Redelmeier, D. A., and Tibshirani, R. J. (1997b). Is using a car phone like driving drunk? *Chance Magazine*, *10*(2), 5–9.

Rice, J. R. (1983). *Numerical Methods, Software, and Analysis*. New York: McGraw-Hill.

Robson, G. (1929). *Monograph of the Recent Cephalopoda, Part I*. London: British Museum.

Rosner, B. (2006). *Fundamentals of Biostatistics*. Duxbury Press.

Rosen, B., and Jerdee, T. (1974). Influence of sex role stereotypes on personnel decisions. *J. Appel. Psych., 59*, 9–14.

Rudemo, H. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Stat., 9*, 65–78.

Ryan, T., and Joiner, B. L. (unpublished ms.). *Normal Probability Plots and Tests for Normality*. Pennsylvania State University.

Ryan, T., Joiner, B., and Ryan B. (1976). *Minitab Student Handbook*. Boston, MA: Duxbury Press.

Sachs, R. K., van den Engh, G., Trask, B., Yokota, H., and Hearst, J. E. (1995). A random-walk/giant-loop model for interphase chromosome. *Proc. Natl. Acad. Sci. USA*, *92*, 2710–2714.

Schachter, A. (1959). *The Psychology of Affiliation*. Stanford, CA: Stanford University Press.

Scheffe, H. (1959). *The Analysis of Variance*. New York: Wiley.

Schlairfer, R. (1959). *Probability and Statistics for Business Decision*. New York: McGraw-Hill.

Schlairfer, R. (1969). *Analysis of Decision under Uncertainty*. New York: McGraw-Hill.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory and Practice*. New York: Wiley.

Shoemaker, A. L. (1996). What's normal? Temperature, gender and heart rate. *J. Stat. Edu*., *3*(2).

Simiu, E., and Filliben, J. (1975). *Statistical Analysis of Extreme Winds*. Washington, DC: National Bureau of Standards. Tech. Note No. 868.

Simpson, J., Olsen, A., and Eden, J. (1975). A Bayesian analysis of a multiplicative treatment effect in weather modification. *Technometrics*, *17*, 161–166.

Smith, D. G., Prentice, R., Thompson, D. J., and Herrmann, W. L. (1975). Association of exogenous estrogen and endometrial carcinoma. *N. Eng. J. Med., 293*, 1164–1167.

Snyder, R. G. (1961). The sex ratio of offspring of pilots of high performance military aircraft. *Hum, Biol*., *3*, 1–10.

Stanley, W., and Walton, D. (1961). Trifluoperazine ("Stelazine"): A controlled clinical trial in chronic schizophrenia. *J. Mental Sci., 107*, 250–257.

Steel, E., Small, J., Leigh, S., and Filliben, J. (1980). Statistical considerations in the preparation of chrysotile filter standard reference materials. Washington, DC. NBS Technical Report.

Steering Committee of the Physicians' Health Study Research Group. (1989). Final report on the aspirin component of the ongoing physicians' health study. *N. Eng. J. Med., 32*(3), 129–135.

Stigler, S. M. (1977). Do robust estimates work with real data? *Ann. Stat*., *5*, 1055–1098.

Strang, G. (1980). *Linear Algebra and Its Applications*. New York: Academic Press.

Student. (1907). On the error of counting with a haemacytometer. *Biometrika*, *5*, 351.

Suboticanec, K., Folnegovic-Smalc, V., Turcin, R., Mestrovic, B., and Buzina, R. (1986). Plasma levels and urinary vitamin C excretion in schizophrenic patients. *Hum. Nut.: Clin. Nut., 40C*, 421–428.

Taguma, Y., Kitamoto, Y., Furaki, G., Ueda, H., Monma, H., Ishisaki, M., Takahashi, H., Sekino, H., and Sasaki, Y. (1985). Effects of catopril on heavy proteinuria in axotemic diabetes. *N. Eng. J. Med.*, *313*(26), 1617–1620.

Tanur, J., Mosteller, F., Kruskal, W., Link, R., Pieters, R., and Rising, G. (1972). *Statistics: A Guide to the Unknown*. Oakland, CA: Holden-Day.

Thomas, H. A. (1948). Frequency of minor floods. *J. Boston Soc. Civil Engineers*, *34*, 425–442.

Tukey, J. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Tversky, A., and Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases in judgements reveal some heuristics of thinking under uncertainty. *Science*, *185*, 1124–1131.

Udias, A., and Rice, J. (1975). Statistical analysis of microearthquake activity near San Andreas Geophysical Observation, Hollister, California. *Bull. Seismol. Soc. Am., 65*, 809–828.

Van Atta, C., and Chen, W. (1968). Correction measurements in grid turbulence using digital harmonic analysis. *J. Fluid Mech.*, *34*, 497–515.

Veitch, J., and Wilks, A. (1985). A characterization of Arctic undersea noise. *J. Acoust. Soc. Am., 77*, 989–999.

Velleman, P., and Hoaglin, D. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston, MA: Duxbury Press.

Vianna, N., Greenwald, P., and Davies, J. (1971). Tonsillectomy and Hodgkin's disease: The lymphoid tissue barrier. *Lancet, 1*, 431–432.

Warner, S. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.*, *60*, 63–69.

Weindling, S. (1977). Statistics report: Math 80B.

Weisburg, S. (1980). *Applied Linear Regression*. New York: Wiley.

Welch, W. J. (1987). Rerandomizing the median in matched-pairs designs. *Biometrika*, *74*, 609–614.

White, J., Riethof, M., and Kushnir, I. (1960). Estimation of microcrystalline wax in beeswax. *J. Assoc. Offic. Anal. Chem., 43*, 781–790.

Wilk, M., and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, *55*, 1–17.

Williams, E. J. (1959). *Regression Analysis*. New York: Wiley.

Williams, W. (1978). How bad can "good" data really be? *Amer. Stat., 32*, 61–67.

Wilson, E. B. (1952). *An Introduction to Scientific Research*. New York: McGraw-Hill.

Wood, B. A. (1972). Modulus of natural bubber crosslinked by dicumly perioxide I: Experimental observations. *J. Res. Nat. Bur. Stand., 76A*, 51–59.

Woodward, P. (1948). A statistical theory of cascade multiplication. *Proc. Cambridge Philos. Soc.*, *44*, 404–412.

Woolf, B. (1955). On estimating the relation between blood group disease. *Ann. Hum. Genet., 19*, 251–253.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. New York: Hafner.

Yip, P., Chao, A., and Chiu, C. (2000). Seasonal variation in suicides: Diminished or vanished. Experience from England and Wales, 1982–1996. *Br. J. Psychiatry, 177*, 366–369.

Yokota, H., van den Engh, G., Hearst, J. E., Sachs, R. K., and Trask, B. (1995). Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. *J. Cell Biol.*, *130*, 1239–1249.

Youden, J. (1972). Enduring values. *Technometrics, 14*, 1–11.

Youden, W. J. (1962). *Experimentation and Measurement*. Washington, DC: National Science Teachers Association.

Youden, W. J. (1974). *Risk, Choice and Prediction*. Boston, MA: Duxbury Press.

Glossary

administration error: is an error caused by improper administration or execution of the research task

analysis of variance for regression: the procedure for computing the F-ratio used to test the significance of the regression as a whole

arithmetic mean: is equal to sum of the values divided by the number of values

attitudes: learned predispositions that project a positive or negative behaviour consistently towards various objects of the world

audit: an in-depth analysis of the existing situation in a firm

auto-correlation: similar to correlation in that it describes the association between values of the same variable but at different time periods. Auto-correlation coefficients provide important information about the underlying patterns in the data

average absolute deviation: in a data set, the average distance of the observations from the mean

average deviation: the arithmetic mean of the absolute deviations from the mean or the median

Bayes' theorem: the formula for conditional probability under statistical dependence

binary questions: those questions that permit only two possible answers

binomial distribution: a probability distribution expressing the probability of one set of the dichotomous alternatives, that is, success or failure

business research design: it can be defined as the plan and structure of inquiry formulated to obtain answers to research questions on business aspects

business research process: the business research process involves a series of steps that systematically investigate a problem or an opportunity facing the organisation

census: the measurement or examination of every element in the population. It is the collection of each and every item in the given population or universe

central limit theorem: a rule ensuring that the sampling distribution of the mean approaches normal as the sample size increases, regardless of the shape of the population distribution from which the sample is selected

certainty: the decision environment in which only one state of nature exists

checklist questions: checklist questions allow the participants to choose one or more of the response options available

chi-square distribution: a family of probability distributions, differentiated by their degrees of freedom, used to test a number of different hypotheses about variance, proportions and distributional goodness of fit

class limits: class limits are the smallest and largest observations (data, events, etc.) in each class. Therefore, each class has two limits: a lower and upper

class mark: the midpoint of a class in a frequency distribution, that is, the average of the lower and upper limits

close-ended questions: close-ended questions refer to those questions that restrict the interviewee's answers to predefined response options

cluster: within a population, groups that are essentially similar to each other, although the groups themselves have wide internal variation

cluster analysis: cluster analysis can be defined as a set of techniques used to classify objects into relatively homogeneous groups called clusters

cluster sampling: a probability sampling method in which the population is first divided into clusters, and then one or more clusters are selected for sampling

cluster sampling: a method of random sampling in which the population is divided into groups, or clusters, of elements and then a random sample of these clusters is selected

coefficient of correlation: the square root of the coefficient of determination. Its sign indicates the direction of the relationship between two variables: direct or inverse. A number lying between –1 (perfect negative correlation) and +1 (perfect positive correlation) to quantify the association between two variables

coefficient of variation: the ratio of standard deviation to mean expressed as a percentage

coefficient of variation: a relative measure of dispersion, comparable across distributions, which expresses the standard deviation as a percentage of the mean

cohort panel: defined as those people within a geographically or otherwise delineated population who experienced the same significant life event within a given period of time

collective exhaustive events: the list of events that represents all the possible outcomes of an experiment

computed: to compute is to calculate, either literally or figuratively

conditional probability: the probability of one event occurring, given that another event has occurred

conditional profit: the profit that would result from a given combination of decision alternative and state of nature

confidence level: the probability that we associate with an interval estimate of a population parameter indicating how confident we are that the interval estimate will include the population parameter

consumer panels: continuous consumer panels are those panels that monitor shifts in individual or specific household behaviours and attitudes over a period of time

consumer price index: the US government prepares this index, which measures changes in the prices of a representative set of consumers or fractions

continuous data: data that may progress from one class to the next without a break and may be expressed by either whole numbers or fractions

continuous random variable: a random variable allowed to take on any value within a given range

continuous scales: in continuous scales, respondents are asked to rate items being studied by marketing at an appropriate place on a line drawn from one extreme of the scale to the other

control group: a control group is a group of test units that are not exposed to the change in the independent variable

convenience sampling: a type of non-probability sampling method in which selection of units from the population is based on their easy availability and accessibility to the researcher

correlation: degree of association between two variables

correlation analysis: a technique to determine the degree to which variables are linearly related

covariance: this is the joint variation between the variables X and Y, mathematically defined as

$$\frac{\Sigma\left(X_i - \bar{X}\right)\left(Y_j - \bar{Y}\right)}{n}$$

for n data points

cumulative frequency distribution: a tabular display of data showing how many observations lie above, or below, certain values

curvilinear relationship: an association between two variables that is described by a curved line

data: a collection of any number of related observations on one or more variables

data array: the arrangement of raw data by observations in either ascending or descending order

data point: a single observation from a data set

data-processing error: an error that occurs because of incorrect data entry, incorrect computer programming or any other error during data analysis

data set: a collection of data

decision point: a branching point that requires a decision

decision tree: a graphic display of the decision environment, indicating decision alternatives, states of nature, probabilities attached to those states of nature, and conditional benefits and losses

decomposition: identifying the trend, seasonality, cycle and randomness in a time series

Delphi: a method of collecting information from experts, useful for long-term forecasting. It is iterative in nature and maintains anonymity to reduce subjective bias

dependent variable: the variable we are trying to predict in regression analysis. The variable of interest or focus that is influenced by one or more independent variable(s)

depth interview: a depth interview is a type of qualitative research approach in which a trained moderator conducts interviews with individuals, rather than with groups, to obtain information about a product or brand

descriptive models: models that are used to describe the behaviour of a system based on data

descriptive statistics: concerned with the analysis and synthesis of data so that a better description of the situation can be made

descriptive studies: a type of formal research in which the objectives are clearly established. In descriptive studies, a researcher gathers details about all aspects of a problem situation

diary panel: a panel of households who continuously record their purchases of selected products in a diary

direct relationship: a relationship between two variables such that as the independent variable's value increases, so does the value of the dependent variable

discrete data: data that does not progress from one class to the next without a break; that is, where classes represent distinct categories or counts and may be represented by whole numbers

discrete probability distribution: a probability distribution in which the variable is allowed to take on a limited number of values

discrete random variable: a random variable that is allowed to take on only a limited number of values

dispersion: the scatter or variability in a set of data

dummy variable: in statistics, particularly in regression analysis, a dummy variable, also known as an indicator variable or qualitative variable, is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome

estimate: a specific observed value of an estimator. A value obtained from data for a certain parameter of the assumed model or a forecast value obtained from the model

estimating equation: a mathematical formula that relates the unknown variable to the known variables in regression analysis

event: one or more of the possible outcomes of doing something, or one of the possible outcomes of an experiment

expected frequencies: the frequencies we would expect to see in a contingency table or frequency distribution if the null hypothesis is true

expected marginal loss: the marginal loss multiplied by the probability of not selling that unit

expected marginal profit: the marginal profit multiplied by the probability of selling that unit

expected profit: the summation of the conditional profits for a given decision alternative, each weighted by the probability that it will happen

expected profit with perfect information: the expected value of the profit with perfect certainty about the occurrence of the states of nature

expected value: a weighted average of outcomes of an experiment

expected value criterion: a criterion requiring the decision maker to calculate the expected value for each decision alternative (the sum of the weighted pay-offs for that alternative in which the weights are the probability values assigned by the decision maker to the states of nature that can happen)

expected value of a random variable: the sum of the products of each value of the random variable with that value's probability of occurrence

expected value of perfect information: the difference between expected profits (under conditions of risk) and expected profit with perfect information

experiment: the activity that results in, or produces, an event

experiment: an experiment refers to the process of manipulating one or more variables and measuring their effect on the other variables, while controlling external variables

experimental group: a group of test units that are exposed to a change in the independent variable

explanatory models: models that are used to explain the behaviour of a system by establishing relationships between its various components

exploratory research: research conducted for a problem that has not been clearly defined. It often occurs before we know enough to make conceptual distinctions or posit an explanatory relationship. It helps to determine the best research design, data collection method and selection of subjects

exponential smoothing: a short-term forecasting method based on weighted averages of past data so that the weightage declines exponentially as the data recedes into the past, with the highest weightage being given to the most recent data

finite population: a population having a stated or limited size

finite population multiplier: a factor used to correct the standard error of the mean for studying a population of finite size that is small in relation to the size of the sample

focus group: a group of individuals selected and assembled by researchers to discuss and comment on, from personal experience, a topic that is the subject of the research

forecasting: it is the process of making predictions of the future based on past and present data and analysis of trends. Businesses utilize forecasting to determine how to allocate their budgets or plan for anticipated expenses for an upcoming period of

time. This is typically based on the projected demand for the goods and services they offer

fractile: in a frequency distribution, the location of a value at, or above, a given fraction of the data

frequency curve: a frequency polygon smoothed by adding classes and data points to a data set

frequency distribution: a line graph connecting the midpoints of each class in a data set, plotted at a height corresponding to the frequency of the class

generalisability: refers to the amount of flexibility in interpreting the data in different research designs

geometric mean: a measure of central tendency used to measure the average rate of change or growth for some quality, computed by taking the nth root of the product of n values representing change. The geometric mean of N observations is the Nth root of the product of the given value observations

goodness-of-fit test: a statistical test for determining whether there is a significant difference between an observed frequency distribution and a theoretical probability distribution hypothesised to describe the observed distribution

grand mean: the mean for the entire group of subjects from all the samples in the experiment

harmonic mean: of N observations is the reciprocal of the arithmetic mean of the reciprocals of the given values of N observations

histogram: a graph of a data set, composed of a series of rectangles, each proportionate in width to the range of values in a class and proportional in height to the number of items falling in the class, or the fraction of items in the class

hypothesis: a statement based on some presumptions about the existence of a relationship between two or more variables that can be tested through empirical data

independent variable: a variable over which the researcher is able to exert some control for studying its effect upon a dependent variable

independent variables: the known variable or variables in regression analysis. Variables that can either be set to a desirable value or takes values that can be observed but not controlled. Parameters of the model are estimated by minimising the sum of squares of error (discrepancy between fitted and actual value)

index number: a ratio that measures how much a variable changes over time

index of industrial production: prepared monthly by the Federal Reserve Board, the IIP measures the quantity of production in the areas of manufacturing, mining and utilities

inductive statistics: concerned with the developments of scientific criteria that can be used to derive information about a group of data by examining only a small portion (sample) of that group

interquartile range: the difference between the values of the first and the third quartiles, this difference indicates the range of the middle half of the data set

interval estimate: a range of values used to estimate an unknown population parameter

interval scale: a scale where there is equal distance between the points on the scale. Examples of interval scales are Fahrenheit and Celsius scales. They are similar to ordinal scales, but the intervals between the points on the scale are equal

interviewer error: interviewer error refers to an administrative error caused by mistakes committed by the interviewer while administering the questionnaire or recording the responses

inverse relationship: a relationship between two variables such that, as the independent variable increases, the dependent variable decreases

itemised category scales: scales in which respondents have to select an answer from a limited number of ordered categories

joint probability: the probability of two events occurring together or in succession

judgement sampling: a method of selecting a sample from a population in which personal knowledge or expertise is used to identify those items from the population that are to be included in the sample

judgement sampling: a non-probability sampling method in which the selection of a unit from the population is based on the judgement of an experienced researcher or an expert

Laspeyres method: to weight an aggregates index, this method uses as weights the quantities consumed during the base period

least-squares method: a technique for fitting a straight line through a set of points in such a way that the sum of the squared vertical distances from the n points to the line is minimised

Likert scales: consist of a series of statements to which the respondent provides answers in the form of degree of agreement or disagreement. This expresses attitude towards the concept under study. The respondent selects a numerical score for each statement to indicate the degree of agreement or otherwise. Each such score is finally added up to measure the respondent's attitude

linear regression: fitting of any chosen mathematical model, linear in unknown parameters, to given data

linear relationship: a particular type of association between two variables that can be described mathematically by a straight line

mail survey: a survey in which questionnaires are sent to qualified respondents by mail or e-mail

marginal loss: the loss incurred from stocking a unit that is not sold

marginal probability: the unconditional probability of one event occurring; the probability of a single event

marginal profit: the profit earned from selling one additional unit

mean: a central tendency measure representing the arithmetic average of a set of observations

measure of central tendency: a measure indicating the value to be expected of a typical or middle data point

measure of dispersion: a measure describing how scattered or spread out the observations in a data set are

measurement: the process of assigning numbers or labels to different objects under study to represent them quantitatively or qualitatively

median: the middle point of a data set, a measure of location that divides the data set into halves. It is that value of the variable which divides the distribution into equal parts

median class: the class in a frequency distribution that contains the median value for a data set

minimum probability: the probability of selling at least an additional unit that must exist to justify stocking that unit

mode: the value most often repeated in the data set. It is represented by the highest point in the distribution curve of a data set. It is that value of the variable which occurs the maximum number of times

model: a general mathematical relationship relating a dependent (or response) variable Y to independent variables $X_1, X_2 \ldots, X_k$ by a form $Y = f(X_1, X_2 \ldots, X_k)$

moving average: an average computed by considering the K most recent (for a K-period moving average) demand points, commonly used for short-term forecasting

multiple regression: the statistical process by which several variables are used to predict another variable

mutually exclusive events: events that cannot happen together

node: a point at which a chance event takes place on a decision tree

non-linear regression: fitting of any chosen mathematical model, non-linear in unknown parameters, to given data

non-parametric tests: statistical techniques that do not make restrictive assumptions about the shape of a population distribution when performing a hypothesis test

non-probability sample: a sample in which the selection of units is based on factors other than random chance, for example, convenience, prior experience or the judgement of the researcher

non-response errors: these errors arise when the survey does not include one or more pieces of information from a unit that has to be part of the study. They include failure to respond completely or even failure to respond to one or more questions of the surveyor

non-sampling errors or systematic errors: a systematic error refers to an error that occurs due to the nature of the research design and the precision of execution

normal distribution: a distribution of a continuous random variable with a single-peaked, bell-shaped curve. The mean lies at the centre of the distribution and the curve is symmetrical around a vertical line erected at the mean. The two tails extend indefinitely, never touching the horizontal axis

nominal scale: a nominal scale segregates data into categories that are mutually exclusive and collectively exhaustive. This scale assigns numbers to each of these categories, and these numbers do not stand for any quantitative value; hence, they cannot be added, subtracted or divided

obsolescence loss: the loss occasioned by stocking too many units and having to dispose of unsold units

ogive: a graph of a cumulative frequency distribution

omnibus surveys: surveys conducted using questionnaires that contain a pool of questions which are of interest to different clients

open-ended class: a class that allows either the upper or the lower end of a quantitative classification scheme to be timeless

open-ended question: a type of question that requires a participant to respond in his or her own words without being restricted to predefined response choices

operations research: a scientific method of providing executive departments with a quantitative basis for decision regarding the operations under control

opportunity loss: the profit that could have been earned if stock had been sufficient to supply a unit that was demanded

ordinal scale: in an ordinal scale, the measurement is done with order, indicating that the classifications are different, and also that the variables in the ordinal scale can be ranked

Paasche method: in weighing an aggregates index, the Paasche method uses as weights the quantities consumed during the current period

paired comparison scales: in paired comparison scales, respondents are asked to select one of two items in a pair based on preset criteria

panel: a group of individuals or organisations that have agreed to provide information to a researcher over a period of time

panel survey: a type of longitudinal survey that involves collecting data from the same sample of individuals or households across time

parameters: the constant terms of the chosen model that have to be estimated before the model is completely specified

parameters: numerical values that describe the characteristics of a whole population, commonly represented by Greek letters

parameter values: are values that tell you something about a population. This is the opposite of a statistic, which tells you something about a small part of the population

pay-off: the benefit that accrues from a given combination of decision alternative and state of nature

percentage relative: ratio of a current value to a base value with the result multiplied by 100

percentiles: fractiles that divide the data into 100 equal parts

periodic surveys: surveys that are conducted at regular intervals – weekly, monthly, quarterly or annually

personal interviews: those interviews that involve the physical presence of a person acting as a mediator on behalf of the researcher

pictorial scales: in pictorial scales, respondents are asked to rate a concept or statement based on their intensity of agreement or disagreement, on a scale. These scales are generally used for respondents who cannot analyse complex scales, such as young children or illiterates

pilot studies: these involve collecting data from the actual respondents in order to gain insight into the topic; this will help the researcher in conducting a larger study. Here, the data is collected informally in order to obtain the opinions of the respondents

point estimate: a single number that is used to estimate an unknown population parameter

Poisson distribution: a discrete distribution in which the probability of the occurrence of an event within a very small time period is a very small number, the probability that two or more such events will occur within the same small time interval is effectively 0, and the probability of the occurrence of the event within one time period is independent of where that time period is

population: a collection of all the elements we are studying and about which we are trying to draw conclusions. It is the collection of items on which information is required

population: the entire aggregation of items from which samples are drawn

population specification error: an error that results from an incorrect definition of the universe or population from which the sample is chosen

posterior probability: a probability that has been revised after additional information was obtained

precision: the degree of accuracy with which the sample means can estimate the population mean, as revealed by the standard error of the mean

prediction: a term denoting the estimate or guess of a future variable that may be arrived at by subjective hunches or intuition

predictive models: models that are used to predict the status of a system in the near future based on data

price index: compares changes in price from one period to another

primary data: the collection of data by the investigator himself or herself

primary research: the collection of the information by interview or questionnaire designed for a specific need. It involves collection of data for the first time

probability: the chance that something will happen

probability distribution: a list of the outcomes of an experiment with the probabilities mean we would expect to see associated with these outcomes

probability sampling: a type of sampling process in which every member of the population has a computable and non-zero probability of being included in the sample

probability tree: a graphical representation showing the possible outcomes of a series of experiments and their respective probabilities

profile analysis: a process in which two or more objects are rated by respondents on a scale

projective technique: an unstructured, indirect form of questioning that encourages respondents to project their underlying motivations, beliefs, attitudes, or a feeling regarding the issue of concern

Q-sort scales: in a Q-sort scale, respondents are asked to sort various characteristics or objects that are being compared into various groups so that the distribution of the number of objects or characteristics in each group follows a normal pattern

qualifying questions: those questions that evaluate the respondent and qualify him or her for further questioning

quantiles: those values which divide the distribution into a fixed number of equal parts; for example, quartiles divide the distribution into four equal parts

quantitative techniques: the name given to the group of statistical and operations research (or programming) techniques

quantity index: a measure of how much the number or quantity of a variable changes over time

quartiles: fractiles that divide the data into four equal parts

quartile deviation: half of the interquartile range; a measure of the average range of one-fourth of the data

quartile deviation: one-half the distance between the first and third quartiles

questionnaire: a device for getting answers to questions by using a form to which the respondent responds

questionnaire: a set of questions to be asked of respondents in an interview, with appropriate instructions indicating which questions are to be asked, and in what order

quota sampling: a non-probability sampling method that is constrained to include a predetermined number of units from each specified subgroup in the population, regardless of their actual probability of inclusion

random or probability sampling: a method of selecting a sample from a population in which all the items in the population have an equal chance of being chosen in the sample

random sampling error: the difference between the sample results and the results of a census conducted by identical procedures

random variable: a variable that takes on different values as a result of the outcomes of a random experiment

range: the distance between the highest and lowest values in a data set

rank correlation: a method for doing correlation analysis when the data are not available to use in numerical form, but when information is sufficient to rank the data

rank correlation coefficient: a measure of the degree of association between two variables that is based on the ranks of observations, not their numerical values

rank order scales: comparative scales, where the respondent is asked to rate an item in comparison with another item or a group of items against each other based on a common criterion

ranking questions: those questions which require the participant to rank the response options listed on a continuum basis in order of preference

ratio scales: ratio scales have a fixed zero point and also have equal intervals. Unlike the ordinal scale, the ratio scale allows the comparison of two variables measured on the scale. This is possible because the numbers or units on the scales are equal at all levels of the scale

raw data: information before it is arranged or analysed by statistical methods

reference database: a reference database provides a bibliography of documents, abstracts or locations of original information

regression: It is a statistical device with the help of which we are in a position to estimate or predict the unknown values of one variable from known values of another variable. The variable which is used to predict the variable of interest is called the independent variable, and the variable we are trying to predict is called the dependent variable

regression line: a line fitted to a set of data points to estimate the relationship between two variables

relative frequency distribution: the display of a data set that shows the fraction or percentage of the total data set that falls into each of a set of mutually exclusive classes

relative variation: used to compare two or more distributions by relating the variation of one distribution to the variation of the other

relevance: refers to the appropriateness of using a particular scale for measuring a variable. It can be represented as relevance = reliability × validity

reliability: it is considered that when the outcome of a measuring process is reproducible, the measuring instrument is reliable. Reliable measuring scales provide stable measures at different times under different conditions

representative sample: a sample that contains the relevant characteristics of the population in the same proportion as they are included in that population

research design: the actual framework of research that provides the specific details regarding the process that needs to be followed in conducting the research

response bias: a survey error that results from the inclination of people to answer a question falsely, through either deliberate misrepresentation or unconscious falsification

response variable: see **dependent variable**

sample: a collection of some, but not all, of the elements of the population under study, used to describe the population

sample: a portion of the elements in a population chosen for direct examination or measurement. A sample is any group of measurements selected from a population

sample selection error: a systematic error that occurs because of an inaccuracy in either the sample design stage or the execution of the sampling procedure, resulting in an unrepresentative sample

sample space: the set of all possible outcomes of an experiment

sampling: the act, process or technique of selecting a representative part of a population for the purpose of determining the characteristics of the whole population

sampling distribution: a statistic for a given population, a probability distribution of all the possible values a statistic may take on for a given sample size

sampling error: error or variation among sample statistics due to chance; that is, difference between each sample and the population and among several samples, which are due to solely to the elements we happened to choose for the sample

sampling frame error: an error in which the list of members of the sample does not correspond exactly with the target population

sampling unit: a basic unit that contains a single element or a group of elements of the population to be sampled

scale: a scale can be defined as a set of numbers or symbols developed in a manner so as to facilitate the assigning of these numbers or symbols to the units under research following certain rules

scaling: the process of measuring quantitative aspects of subjective or abstract concepts. It is a method to assign numbers or symbols to some attributes of an object

scatter diagram: an ungrouped plot of two variables on the X and Y axes.

A graph of points on a rectangular grid; the X and Y coordinates of each point correspond to the two measurements made on some particular sample element, and the pattern of points illustrates the relationship between the two variables

seasonal index: a number with a base of 1.00 that indicates the seasonality for a given period in relation to other periods

secondary data: the collection of data compiled by someone other than the user

secondary data: data that already exists, which has been collected by some other person or organisation for their use, but which is later made available to other researchers free or at a concessional rate

secondary data: data that has already been collected previously for some other research purpose. It can be obtained from magazines, journals, online articles, company literature and so on

secondary data study: a secondary data study is concerned with the analysis of already existing data that is related to the research topic in question

selection bias: an improper assignment of respondents to treatment conditions

semantic differential scales: scales that are used to describe a set of beliefs that underline a person's attitude towards an organisation, product or brand. This scale is based on the principle that individuals think dichotomously, or in terms of polar opposites, such as reliable–unreliable, modern–old fashioned, cold–warm and so on

sensitivity: refers to an instrument's ability to accurately measure variability in stimuli or responses

significance level: a value indicating the percentage of sample values that are outside certain limits, assuming the null hypothesis is correct; that is, the probability of rejecting the null hypothesis when it is true

simple random sampling: methods of selecting samples that allow each possible sample an equal probability of being picked and each item in the entire population an equal chance of being included in the sample

simple random sampling: a type of probability sampling method in which each element in the target population has an equal chance or probability of inclusion in the sample

single item scales: scales with only one item to measure. Itemised category scales are most commonly used under single item scales

situational errors: the errors arising due to various situational factors

skewness: lack of symmetry

slope: a constant for any given straight line, whose value represents how much each unit change of the independent variable changes the dependent variable

source databases: source databases usually publish numerical data, full text or a combination of both

standard deviation: the positive square root of the variance; a measure of dispersion in the same units as the original data, rather than in the squared units of the variance. It is the root mean square deviation of a given set of data

standard error: the standard deviation of the sampling distribution of a statistic

standard error of estimate: a measure of the reliability of the estimating equation, indicating the variability of the observed points around the regression line; that is, the extent to which observed values differ from their predicted values on the regression line

standard error of the mean: the standard deviation of the sampling distribution of the mean; a measure of the extent to which we expect the means from different samples to vary from the population mean due to the chance error in the sampling process

standard error of the regression coefficient: a measure of the variability of sample regression coefficients around the true population regression coefficient

standard normal probability distribution: a normal probability distribution, with mean $\mu = 0$ and standard deviation $\sigma = 1$

state of nature: future event not under the control of the decision maker

statistical data: numerical description of quantitative aspects of things. These descriptions may take the form of counts or measurements

statistical decision theory: it is concerned with the establishment of rules and procedures for choosing the course of action from alternative courses of action in a situation of uncertainty

statistical methods: these methods include all those devices of analysis and synthesis by means of which statistical data is systematically collected and used to explain or describe a given phenomenon

statistics: numerical measures describing the characteristics of a sample

strata: groups within a population formed in such a way that each group is relatively homogeneous, but wider variability exists among the separate groups

stratified random sampling: a type of probability sampling method in which the population is separated into groups (strata), usually based on some internal similarities, and then a random sample is selected within each stratum

stratified sampling: a method of random sampling in which the population is divided into homogeneous groups, or strata, and elements within each stratum are selected at random according to one of two rules: (1) a specified number of elements are drawn from each stratum corresponding to the proportion of that stratum in the population or (2) an equal number of elements are drawn from each stratum and the results are weighted according to the stratum's proportion of the total population

subjective probability: probability based on the personal beliefs of the person making the probability estimate

survey: a research technique that is used to gather information from a sample of respondents by employing a questionnaire

survey research: a method of collecting information that involves asking a set of preformulated questions in a predetermined sequence in a structured questionnaire to a sample of individuals drawn so as to be representative of a defined population

symmetrical: a characteristic of a distribution in which each half is the mirror image of the other half

syndicated data: data produced by a market research firm, which provides a body of similar data compiled from a large number of sources, organised into a common format for a fee to its subscribers

systematic sampling: a method of random sampling used in statistics in which elements to be sampled are selected from the population at a uniform interval that is measured in time, order or space

systematic sampling: a type of probability sampling method in which sample selection begins at a random starting point but subsequently selects additional sampling units at equal intervals along a stated gradient or numbered list

Thurstone scales: in Thurstone scales, researchers select a group of 80–100 items indicating the different degrees of favourable attitude towards a concept under study. Once items are selected, they are given to a group of judges, who are asked to categorise them according to how much they favour or disfavour them

time lag: the length between two time periods, generally used in time series, where one may test, for instance, how values of periods 1, 2, 3, 4 correlate with values of periods 4, 5, 6, 7 (time-lag 3 periods)

time series: information accumulated at regular intervals and the statistical methods used to determine patterns in such data. Sets of observations at equal time intervals, which may form the basis of future forecasting. Any data on demand, sales or consumption taken at regular intervals of time constitutes a time series.

time-series analysis: comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data

time-series model: a model that predicts the future by expressing it as a function of the past

transformations: mathematical manipulations for converting one variable into a different form, so we can fit curves as well as lines by regression

trend: a growth or decline in the mean value of a variable over the relevant time span

type I error: rejecting a null hypothesis when it is true

type II error: accepting a null hypothesis when it is false

unconscious misrepresentation: a situation in which a respondent gives wrong or estimated information due to ignorance and forgetfulness even though he or she has no intention of doing so

unweighted aggregates index: uses all the values considered, where each value is of equal importance

unweighted average of relatives methods: to construct an index number, this method first finds the ratio of the current price to the base price for each product, adds the resulting percentage relatives, and then divides by the number of products

utility: the value of a certain outcome or pay-off to someone; the pleasure or displeasure someone derives from an outcome

validity: the ability of a scale or a measuring instrument to measure what it is intended to measure can be termed the validity of the measurement

variance: a measure of the average squared distance between the mean and each item in the population. It is the square of the standard deviation, defined as the arithmetic mean of the squared deviations from the mean

Venn diagram: it is diagram that shows all possible logical relations between a finite collection of different sets. Typically overlapping shapes, usually circles, are used

warm-up questions: those questions that tend to make the respondent think and recollect past experiences necessary for the completion of a questionnaire

weighted aggregates index: using all the values considered, this index assigns weights to these values

weighted average of relatives method: to construct an index number, this method weights by importance and uses the value of each element in the composite

weighted mean: an average calculated to take into account the importance of each value to the overall total, that is, an average in which each observation value is weighted by some index of its importance

*Appendix I: Areas under the Normal Curve Corresponding to Given Value of z**

TABLE B Areas under the Normal Curve Corresponding to Given Values of z^a

Column 2 gives the proportion of the area under the entire curve which is between the mean ($z = 0$) and thepositive value of z. Areas for negative values of z are the same as for positive values, since the curve is symmetrical.

Column 3 gives the proportion of the area under the entire curve which falls beyond the stated positive value of z. Areas for negative values of z are the same, since the curve is symmetrical.

z	Area between mean and z	Area beyond z	z	Area between mean and z	Area beyond z
1	2	3	1	2	3
0.00	.0000	.5000	0.25	.0987	.4013
0.01	.0040	.4960	0.26	.1026	.3974
0.02	.0080	.4920	0.27	.1064	.3936
0.03	.0120	.4880	0.28	.1103	.3897
0.04	.0160	.4840	0.29	.1141	.3859
0.05	.0199	.4801	0.30	.1179	.3821
0.06	.0239	.4761	0.31	.1217	.3783
0.07	.0279	.4721	0.32	.1255	.3745
0.08	.0319	.4681	0.33	.1293	.3707
0.09	.0359	.4641	0.34	.1331	.3669
0.10	.0398	.4602	0.35	.1368	.3632
0.11	.0438	.4562	0.36	.1406	.3594
0.12	.0478	.4522	0.37	.1443	.3557
0.13	.0517	.4483	0.38	.1480	.3520
0.14	.0557	.4443	0.39	.1517	.3483
0.15	.0596	.4404	0.40	.1554	.3446
0.16	.0636	.4364	0.41	.1591	.3409
0.17	.0675	.4325	0.42	.1628	.3372
0.18	.0714	.4286	0.43	1664	.3336
0.19	.0753	.4247	0.44	.1700	.3300
0.20	.0793	.4207	0.45	.1736	.3264
0.21	.0832	.4168	0.46	.1772	.3228
0.22	.0871	.4129	0.47	.1808	.3192
0.23	.0910	.4090	0.48	.1844	.3156
0.24	.0948	.4052	0.49	.1879	.3121

[a] From Appendix 2 of : R. Clarke, A. Coladarci, and J. Caffrey, *Statistical Reasoning and Procedures.* Charles E. Merrill Books, Inc., Columbus, Ohio, 1965, with permission of the publisher.

TABLE B (*Continued*)

z	Area between mean and z	Area beyond z	z	Area between mean and z	Area beyond z
1	2	3	1	2	3
0.50	.1915	.3085	1.00	.3413	.1587
0.51	.1950	.3050	1.01	.3438	.1562
0.52	.1985	.3015	1.02	.3461	.1539
0.53	.2019	.2981	1.03	.3485	.1515
0.54	.2054	.2946	1.04	.3508	.1492
0.55	.2088	.2912	1.05	.3531	.1469
0.56	.2123	.2877	1.06	.3554	.1446
0.57	.2157	.2843	1.07	.3577	.1423
0.58	.2190	.2810	1.08	.3599	.1401
0.59	.2224	.2776	1.09	.3621	.1379
0.60	.2257	.2743	1.10	.3643	.1357
0.61	.2291	.2709	1.11	.3665	.1335
0.62	.2324	.2676	1.12	.3686	.1314
0.63	.2357	.2643	1.13	.3708	.1292
0.64	.2389	.2611	1.14	.3729	.1271
0.65	.2422	.2578	1.15	.3749	.1251
0.66	.2454	.2546	1.16	.3770	.1230
0.67	.2486	.2514	1.17	.3790	.1210
0.68	.2517	.2483	1.18	.3810	.1190
0.69	.2549	.2451	1.19	.3830	.1170
0.70	.2580	.2420	1.20	.3849	.1151
0.71	.2611	.2389	1.21	.3869	.1131
0.72	.2642	.2358	1.22	.3888	.1112
0.73	.2673	.2327	1.23	.3907	.1093
0.74	.2704	.2296	1.24	.3925	.1075
0.75	.2734	.2266	1.25	.3944	.1056
0.76	.2764	.2236	1.26	.3962	.1038
0.77	.2794	.2206	1.27	.3980	.1020
0.78	.2823	.2177	1.28	.3997	.1003
0.79	.2852	.2148	1.29	.4015	.0985
0.80	.2881	.2119	1.30	.4032	.0968
0.81	.2910	.2090	1.31	.4049	.0951
0.82	.2939	.2061	1.32	.4066	.0934
0.83	.2967	.2033	1.33	.4082	.0918
0.84	.2995	.2005	1.34	.4099	.0901
0.85	.3023	.1977	1.35	.4115	.0885
0.86	.3051	.1949	1.36	.4131	.0869
0.87	.3078	.1922	1.37	.4147	.0853
0.88	.3106	.1894	1.38	.4162	.0838
0.89	.3133	.1867	1.39	.4177	.0823
0.90	.3159	.1841	1.40	.4192	.0808
0.91	.3186	.1814	1.41	.4207	.0793
0.92	.3212	.1788	1.42	.4222	.0778
0.93	.3238	.1762	1.43	.4236	.0764
0.94	.3264	.1736	1.44	.4251	.0749
0.95	.3289	.1711	1.45	.4265	.0735
0.96	.3315	.1685	1.46	.4279	.0721
0.97	.3340	.1660	1.47	.4292	.0708
0.98	.3365	.1635	1.48	.4306	.0649
0.99	.3389	.1611	1.49	.4319	.0681

TABLE B (*Continued*)

z	Area between mean and z	Area beyond z	z	Area between mean and z	Area beyond z
1	2	3	1	2	3
1.50	.4332	.0668	2.00	.4772	.0228
1.51	.4345	.0655	2.01	.4778	.0222
1.52	.4357	.0643	2.02	.4783	.0217
1.53	.4370	.0630	2.03	.4788	.0212
1.54	.4382	.0618	2.04	.4793	.0207
1.55	.4394	.0606	2.05	.4798	.0202
1.56	.4406	.0594	2.06	.4803	.0197
1.57	.4418	.0582	2.07	.4808	.0192
1.58	.4429	.0571	2.08	.4812	.0188
1.59	.4441	.0559	2.09	.4817	.0183
1.60	.4452	.0548	2.10	.4821	.0179
1.61	.4463	.0537	2.11	.4826	.0174
1.62	.4474	.0526	2.12	.4830	.0170
1.63	.4484	.0516	2.13	.4834	.0166
1.64	.4495	.0505	2.14	.4838	.0162
1.65	.4505	.0495	2.15	.4842	.0158
1.66	.4515	.0485	2.16	.4846	.0154
1.67	.4525	.0475	2.17	.4850	.0150
1.68	.4535	.0465	2.18	.4854	.0146
1.69	.4545	.0455	2.19	.4857	.0143
1.70	.4554	.0446	2.20	.4861	.0139
1.71	.4564	.0436	2.21	.4864	.0136
1.72	.4573	.0427	2.22	.4868	.0132
1.73	.4582	.0418	2.23	.4871	.0129
1.74	.4591	.0409	2.24	.4875	.0125
1.75	.4599	.0401	2.25	.4878	.0122
1.76	.4608	.0392	2.26	.4881	.0119
1.77	.4616	.0384	2.27	.4884	.0116
1.78	.4625	.0375	2.28	.4887	.0113
1.79	.4633	.0367	2.29	.4890	.0110
1.80	.4641	.0359	2.30	.4893	.0107
1.81	.4649	.0351	2.31	.4896	.0104
1.82	.4656	.0344	2.32	.4898	.0102
1.83	.4664	.0336	2.33	.4901	.0099
1.84	.4671	.0329	2.34	.4904	.0096
1.85	.4678	.0322	2.35	.4906	.0094
1.86	.4686	.0314	2.36	.4909	.0190
1.87	.4693	.0307	2.37	.4911	.0089
1.88	.4699	.0301	2.38	.4913	.0087
1.89	.4706	.0294	2.39	.4916	.0084
1.90	.4713	.0287	2.40	.4918	.0082
1.91	.4719	.0281	2.41	.4920	.0080
1.92	.4726	.0274	2.42	.4922	.0078
1.93	.4732	.0268	2.43	.4925	.0075
1.94	.4738	.0262	2.44	.4927	.0073
1.95	.4744	.0256	2.45	.4929	.0071
1.96	.4750	.0250	2.46	.4931	.0069
1.97	.4756	.0244	2.47	.4932	.0068
1.98	.4761	.0239	2.48	.4934	.0066
1.99	.4767	.0233	2.49	.4936	.0064

TABLE B (*Continued*)

z	Area between mean and z	Area beyond z	z	Area between mean and z	Area beyond z
1	2	3	1	2	3
2.50	.4938	.0062	2.90	.4981	.0019
2.51	.4938	.0060	2.91	.4982	.0018
2.52	.4941	.0059	2.92	.4982	.0018
2.53	.4943	.0057	2.93	.4983	.0017
2.54	.4945	.0055	2.94	.4984	.0016
2.55	.4946	.0054	2.95	.4984	.0016
2.56	.4948	.0052	2.96	.4985	.0015
2.57	.4949	.0051	2.97	.4985	.0015
2.58	.4951	.0049	2.98	.4986	.0014
2.59	.4952	.0048	2.99	.4986	.0014
2.60	.4953	.0047	3.00	.4987	.0013
2.61	.4955	.0045	3.01	.4987	.0013
2.62	.4956	.0044	3.02	.4987	.0013
2.63	.4957	.0043	3.03	.4988	.0012
2.64	.4959	.0041	3.04	.4988	.0012
2.65	.4960	.0040	3.05	.4989	.0011
2.66	.4961	.0039	3.06	.4989	.0011
2.67	.4962	.0038	3.07	.4989	.0011
2.68	.4963	.0037	3.08	.4990	.0010
2.69	.4964	.0036	3.09	.4990	.0010
2.70	.4965	.0035	3.10	.4990	.0010
2.71	.4966	.0034	3.11	.4991	.0009
2.72	.4967	.0033	3.12	.4991	.0009
2.73	.4968	.0032	3.13	.4991	.0009
2.74	.4969	.0031	3.14	.4992	.0008
2.75	.4970	.0030	3.15	.4992	.0008
2.76	.4971	.0029	3.16	.4992	.0008
2.77	.4972	.0028	3.17	.4992	.0008
2.78	.4973	.0027	3.18	.4993	.0007
2.79	.4974	.0026	3.19	.4993	.0007
2.80	.4974	.0026	3.20	.4993	.0007
2.81	.4975	.0025	3.21	.4993	.0007
2.82	.4976	.0024	3.22	.4994	.0006
2.83	.4977	.0023	3.23	.4994	.0006
2.84	.4977	.0023	3.24	.4994	.0006
2.85	.4978	.0022	3.30	.4995	.0005
2.86	.4979	.0021	3.40	.4997	.0003
2.87	.4979	.0021	3.50	.4998	.0002
2.88	.4980	.0020	3.60	.4998	.0002
2.89	.4981	.0019	3.70	.4999	.0001

TABLE C Values of z Corresponding to Divisions of the Area under the Normal Curve into a Larger Proportion and a Smaller Proportion.[a]

The larger area 4	z 2	The smaller area 3	The larger area 4	z 2	The smaller area 3
.500	.0000	.500	.725	.5978	.275
.505	.0125	.495	.730	.6128	.270
.510	.0251	.490	.735	.6280	.265
.515	.0376	.485	.740	.6433	.260
.520	.0502	.480	.745	.6588	.255
.525	.0627	.475	.750	.6745	.250
.530	.0753	.470	.755	.6903	.245
.535	.0878	.465	.760	.7063	.240
.540	.1004	.460	.765	.7225	.235
.545	.1130	.455	.770	.7388	.230
.550	.1257	.450	.775	.7554	.225
.555	.1383	.445	.780	.7722	.220
.560	.1510	.440	.785	.7892	.215
.565	.1637	.435	.790	.8064	.210
.570	.1764	.430	.795	.8239	.205
.575	.1891	.425	.800	.8416	.200
.580	.2019	.420	.805	.8596	.195
.585	.2147	.415	.810	.8779	.190
.590	.2275	.410	.815	.8965	.185
.595	.2404	.405	.820	.9154	.180
.600	.2533	.400	.825	.9346	.175
.605	.2663	.395	.830	.9542	.170
.610	.2793	.390	.835	.9741	.165
.615	.2924	.385	.840	.9945	.160
.620	.3055	.380	.845	1.0152	.155
.625	.3186	.375	.850	1.0363	.150
.630	.3319	.370	.855	1.0581	.145
.635	.3451	.365	.860	1.0803	.140
.640	.3585	.360	.865	1.1031	.135
.645	.3719	.355	.870	1.1264	.130
.650	.3853	.350	.875	1.1503	.125
.655	.3989	.345	.880	1.1750	.120
.660	.4125	.340	.885	1.2004	.115
.665	.4261	.335	.890	1.2265	.110
.670	.4399	.330	.895	1.2536	.105
.675	.4538	.325	.900	1.2816	.100
.680	.4677	.320	.905	1.3106	.095
.685	.4817	.315	.910	1.3408	.090
.690	.4959	.310	.915	1.3722	.085
.695	.5101	.305	.920	1.4051	.080
.700	.5244	.300	.925	1.4395	.075
.705	.5388	.295	.930	1.4757	.070
.710	.5534	.290	.935	1.5141	.065
.715	.5681	.285	.940	1.5548	.060
.720	.5828	.280	.945	1.5982	.055

TABLE C (*Continued*)

The larger area 1	z 2	The smaller area 3	The larger area 1	z 2	The smaller area 3
.950	1.6449	.050	.990	2.3263	.010
.955	1.6954	.045	.995	2.5758	.005
.960	1.7507	.040			
.965	1.8119	.035	.996	2.6521	.004
.970	1.8808	.030	.997	2.7478	.003
			.998	2.8782	.002
.975	1.9600	.025	.999	3.0902	.001
.980	2.0537	.020	.9995	3.2905	.0005
.985	2.1701	.015			

Appendix II: Student's t-Distribution

TABLE D Student's t Distribution[a]

The first column identifies the specific t distribution according to its number of degrees of freedom. Other columns give the proportion of the area under the entire curve which falls beyond the tabled positive value of t. Areas for negative values of t are the same, since the curve is symmetrical.

	Area in one tail					
df	.25	.10	.05	.025	.01	.005
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.500
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.696	1.356	1.782	2.179	2.681	3.054
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.132	2.602	2.947
16	0.690	1.337	1.746	2.120	2.584	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.540	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.320	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.312	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
31	0.682	1.310	1.696	2.040	2.453	2.744
32	0.682	1.309	1.694	2.037	2.449	2.738
33	0.682	1.308	1.692	2.034	2.445	2.733
34	0.682	1.307	1.691	2.032	2.441	2.728
35	0.682	1.306	1.690	2.030	2.438	2.724
36	0.681	1.306	1.688	2.028	2.434	2.720
37	0.681	1.305	1.687	2.026	2.431	2.715
38	0.681	1.304	1.686	2.024	2.429	2.712
39	0.681	1.304	1.685	2.023	2.426	2.708
40	0.681	1.303	1.684	2.021	2.423	2.704
45	0.680	1.301	1.679	2.014	2.412	2.690
50	0.679	1.299	1.676	2.009	2.403	2.678
55	0.679	1.297	1.673	2.004	2.396	2.668

TABLE D (*Continued*)

	Area in one tail					
df	.25	.10	.05	.025	.01	.005
60	0.679	1.296	1.671	2.000	2.390	2.660
70	0.678	1.294	1.667	1.994	2.381	2.648
80	0.678	1.292	1.664	1.990	2.374	2.639
90	0.677	1.291	1.662	1.987	2.368	2.632
100	0.677	1.290	1.660	1.984	2.364	2.626
120	0.676	1.289	1.658	1.980	2.358	2.617
150	0.676	1.287	1.655	1.976	2.352	2.609
200	0.676	1.286	1.652	1.972	2.345	2.601
300	0.675	1.284	1.650	1.968	2.339	2.592
400	0.675	1.284	1.649	1.966	2.336	2.588
500	0.675	1.283	1.648	1.965	2.334	2.586
1000	0.675	1.282	1.646	1.962	2.330	2.581
∞	0.674	1.282	1.645	1.960	2.326	2.576

[a]Modified from Section 2.1, *Handbook of Statistical Tables* by Donald B. Owen. Copyright © 1962. Addison-Wesley, Reading, Mass. Courtesy of U.S. Atomic Energy Commission.

Appendix III: The χ^2 Distribution

TABLE G The x^2 Distribution*.

The first column identifies the specific x^2 distribution according to its number of degrees of freedom. Other columns give the proportion of the area under the entire curve which falls above the tabled value of x^2.

	Area in the upper tail									
df	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	.000039	.00016	.00098	.0039	.016	2.71	3.84	5.02	6.63	7.88
2	.010	.020	.051	.10	.21	4.61	5.99	7.38	9.21	10.60
3	.072	.11	.22	.35	.58	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	107.56	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

*Modified from Table 8: E. Person, and H. Hartley, *Biometrika Tables for Statisticians*, Vol. 1, 3rd ed., University Press, Cambridge, 1966, with permission of the Biometrika Trustees.

Note: When $df > 30$, the critical value of x^2 may be found by the following approximate formula: $x^2 = df\,[1 - (2/9df) + z\sqrt{2/9df}]^3$, where z is the normal deviate above which lies the same proportionate area in the normal curve. For example, to find the value of x^2 which divides the upper 1% of the distribution from the remainder when $df = 30$, we calculate: $x^2 = 30\,(1 - .00741 + 2.3263\sqrt{.0074074})^3 = 50.91$ which compares closely with the tabled value of 50.89.

Appendix IV: The F-Distribution

TABLE H The F Distribution. * Values of F Corresponding to 5% (Roman Type) and 1% (Boldface Type) of the Area in the Upper Tail

The specific F distribution must be identified by the number of degrees of freedom characterizing the numerator and the denominator of F.

Degrees of Freedom: Numerator

Degrees of Freedom: Denominator	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254
	4,052	**4,999**	**5,403**	**5,625**	**5,764**	**5,859**	**5,928**	**5,981**	**6,022**	**6,056**	**6,082**	**6,106**	**6,142**	**6,169**	**6,208**	**6,234**	**6,258**	**6,286**	**6,302**	**6,323**	**6,334**	**6,352**	**6,361**	**6,366**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.34**	**99.36**	**99.38**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**	**99.46**	**99.47**	**99.48**	**99.48**	**99.49**	**99.49**	**99.49**	**99.50**	**99.50**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**	**26.60**	**26.50**	**26.41**	**26.35**	**26.27**	**26.23**	**26.18**	**26.14**	**26.12**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	22.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**	**13.93**	**13.83**	**13.74**	**13.69**	**13.61**	**13.57**	**13.52**	**13.48**	**13.46**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.27**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**	**9.47**	**9.38**	**9.29**	**9.24**	**9.17**	**9.13**	**9.07**	**9.04**	**9.02**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**	**7.31**	**7.23**	**7.14**	**7.09**	**7.02**	**6.99**	**6.94**	**6.90**	**6.88**
7	5.59	4.47	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**	**6.07**	**5.98**	**5.90**	**5.85**	**5.78**	**5.75**	**5.70**	**5.67**	**5.65**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**	**5.28**	**5.20**	**5.11**	**5.06**	**5.00**	**4.96**	**4.91**	**4.88**	**4.86**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**	**4.73**	**4.64**	**4.56**	**4.51**	**4.45**	**4.41**	**4.36**	**4.33**	**4.31**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**	**4.33**	**4.25**	**4.17**	**4.12**	**4.05**	**4.01**	**3.96**	**3.93**	**3.91**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**	**4.02**	**3.94**	**3.86**	**3.80**	**3.74**	**3.70**	**3.66**	**3.62**	**3.60**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**	**3.78**	**3.70**	**3.61**	**3.56**	**3.49**	**3.46**	**3.41**	**3.38**	**3.36**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**	**3.59**	**3.51**	**3.42**	**3.37**	**3.30**	**3.27**	**3.21**	**3.18**	**3.16**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.37	2.24	2.21	2.19	2.16	2.14	2.13
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.92**	**2.89**	**2.87**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**	**3.45**	**3.37**	**3.25**	**3.18**	**3.10**	**3.01**	**2.96**	**2.89**	**2.86**	**2.80**	**2.77**	**2.75**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.95**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**	**3.08**	**3.00**	**2.92**	**2.86**	**2.79**	**2.76**	**2.70**	**2.67**	**2.65**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**	**3.00**	**2.91**	**2.83**	**2.78**	**2.71**	**2.68**	**2.62**	**2.59**	**2.57**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**	**2.92**	**2.84**	**2.76**	**2.70**	**2.63**	**2.60**	**2.54**	**2.51**	**2.49**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	190	1.87	1.85	1.84
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**	**2.86**	**2.77**	**2.69**	**2.63**	**2.56**	**2.53**	**2.47**	**2.44**	**2.42**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**	**2.80**	**2.72**	**2.63**	**2.58**	**2.51**	**2.47**	**2.42**	**2.38**	**2.36**

TABLE H (*Continued*)

Degrees of freedom: denominator	Degrees of Freedom: Numerator																					
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.23	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.46	2.42	2.37	2.33	2.31
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	3.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.97	2.89	2.78	2.70	2.62	2.53	2.48	2.41	2.37	2.32	2.28	2.26
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23	2.21
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	3.05	2.99	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.32	2.29	2.23	2.19	2.17
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69
	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96	2.86	2.77	2.66	2.58	2.50	2.41	2.36	2.28	2.25	2.19	2.15	2.13
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67
	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98	2.93	2.83	2.74	2.63	2.55	2.47	2.38	2.33	2.25	2.21	2.16	2.12	2.10
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95	2.90	2.80	2.71	2.60	2.52	2.44	2.35	2.30	2.22	2.18	2.13	2.09	2.06
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64
	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92	2.87	2.77	2.68	2.57	2.49	2.41	2.32	2.27	2.19	2.15	2.10	2.06	2.03
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84	2.74	2.66	2.55	2.47	2.38	2.29	2.24	2.16	2.13	2.07	2.03	2.01
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86	2.80	2.70	2.62	2.51	2.42	2.34	2.25	2.20	2.12	2.08	2.02	1.98	1.96
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57
	7.44	5.29	4.42	3.93	3.61	3.38	3.21	3.08	2.97	2.89	2.82	2.76	2.66	2.58	2.47	2.38	2.30	2.21	2.15	2.08	2.04	1.98	1.94	1.91
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
	7.39	5.25	4.38	3.89	3.58	3.35	3.18	3.04	2.94	2.86	2.78	2.72	2.62	2.54	2.43	2.35	2.26	2.17	2.12	2.04	2.00	1.94	1.90	1.87
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.91	2.82	2.75	2.69	2.59	2.51	2.40	2.32	2.22	2.14	2.08	2.00	1.97	1.90	1.86	1.84
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73	2.66	2.56	2.49	2.37	2.29	2.20	2.11	2.05	1.97	1.94	1.88	1.84	1.81
42	4.07	3.22	3.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	7.27	5.15	4.29	3.80	3.49	3.26	3.10	2.96	2.86	2.77	2.70	2.64	2.54	2.46	2.35	2.26	2.17	2.08	2.02	1.94	1.91	1.85	1.80	1.78
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	7.24	5.12	4.26	3.78	3.46	3.24	3.07	2.94	2.84	2.75	2.68	2.62	2.52	2.44	2.32	2.24	2.15	2.06	2.00	1.92	1.88	1.82	1.78	1.75
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	7.21	5.10	4.24	3.76	3.44	3.22	3.05	2.92	2.82	2.73	2.66	2.60	2.50	2.42	2.30	2.22	2.13	2.04	1.98	1.90	1.86	1.80	1.76	1.72
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	2.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	7.19	5.08	4.22	3.74	3.42	3.20	3.04	2.90	2.80	2.71	2.64	2.58	2.48	2.40	2.28	2.20	2.11	2.02	1.96	1.88	1.84	1.78	1.73	1.70
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	7.17	5.06	4.20	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62	2.56	2.46	2.39	2.26	2.18	2.10	2.00	1.94	1.86	1.82	1.76	1.71	1.68
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41
	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.59	2.53	2.43	2.35	2.23	2.15	2.06	1.96	1.90	1.82	1.78	1.71	1.66	1.64
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.40	2.32	2.20	2.12	2.03	1.93	1.87	1.79	1.74	1.68	1.63	1.60
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37
	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.79	2.70	2.61	2.54	2.47	2.37	2.30	2.18	2.09	2.00	1.90	1.84	1.76	1.71	1.64	1.60	1.56
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51	2.45	2.35	2.28	2.15	2.07	1.98	1.88	1.82	1.74	1.69	1.62	1.56	1.53
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55	2.48	2.41	2.32	2.24	2.11	2.03	1.94	1.84	1.78	1.70	1.65	1.57	1.52	1.49
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43	2.36	2.26	2.19	2.06	1.98	1.89	1.79	1.73	1.64	1.59	1.51	1.46	1.43
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
	6.84	4.78	3.94	3.47	3.17	2.95	3.79	2.65	2.56	2.47	2.40	2.33	2.23	2.15	2.03	1.94	1.85	1.75	1.68	1.59	1.54	1.46	1.40	1.37

TABLE H (*Continued*)

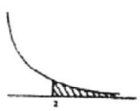
Degrees of Freedom: Denominator	Degrees of Freedom: Numerator																					
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22
	6.81	4.75	3.91	3.44	3.14	2.92	2.76	2.62	2.53	2.44	2.37	2.30	2.20	2.12	2.00	1.91	1.83	1.72	1.66	1.56	1.51	1.43	1.37	1.33
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
	6.76	4.71	3.88	3.41	3.11	2.90	2.73	2.60	2.50	2.41	2.34	2.28	2.17	2.09	1.97	1.88	1.79	1.69	1.62	1.53	1.48	1.39	1.33	1.28
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
	6.70	4.66	3.83	3.36	3.06	2.85	2.69	2.55	2.46	2.37	2.29	2.23	2.12	2.04	1.92	1.84	1.74	1.64	1.57	1.47	1.42	1.32	1.24	1.19
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
	6.66	4.62	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.26	2.20	2.09	2.01	1.89	1.81	1.71	1.61	1.54	1.44	1.38	1.28	1.19	1.11
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
	6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24	2.18	2.07	1.99	1.87	1.79	1.69	1.59	1.52	1.41	1.36	1.25	1.15	1.00

*Reproduced by permission from *Statistical Methods*, 5th ed., by George W. Snedecor, Copyright © 1956 by The Iowa State University Press.

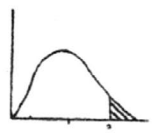
Appendix V: Proportions of Area for the χ^2 Distribution

n_2 / n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
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
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	12
	—	—	—	—	—	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

TABLE T Proportions of Area for the χ^2 Distribution



For df = 1.2



For df ≥ 30

	Proportion of area										
df	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
1	0.00004	0.00016	0.00098	0.00393	0.0158	0.455	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	1.386	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	2.366	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	3.357	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	4.251	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	5.35	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80

	Proportion of area										
df	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.83	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.34	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.43	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	51.17	60.39	64.28	79.33	98.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	89.33	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	99.33	118.5	124.3	129.6	135.8	140.2

TABLE U Table of probabilities associated with values as small as observed values of U in the Mann-Whitney test

Table $n_2 = 3$

n_1 / U	1	2	3
0	.250	.100	.050
1	.500	.200	.100
2	.750	.400	.200
3		.600	.350
4			.560
5			.650

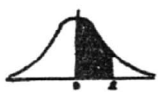
Table $n_2 = 4$

n_1 / U	2	3	4
0	.200	.067	.028	.014
1	.400	.133	.057	.029
2	.600	.267	.114	.057
3		.400	.200	.100
4		.600	.314	.171
5			.429	.243
6			.571	.343
7				.443
8				.557

Appendix VI: Area under Normal Curve

16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

TABLE X Area Under Normal Curve.
An entry in the table is the proportion under the entire curve which is between z = 0 and a positive value of z. Areas for negative values for z are obtained by symmetry



Areas of a standard normal distribution

z	.0	0.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2903	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.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	.4657	.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	.4993	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Index

S

T

U

V

W

Z