The McGraw-Hill Companies

Business Research Methodology

ABOUT THE AUTHORS

Dr T N Srivastava has been a visiting faculty in NMIMS University for about 12 years, and has been on their board of studies for Decision Sciences. He is a retired Chief General Manager, Reserve Bank of India. He obtained his doctorate degree in Statistics under the guidance of Late Padmabhushan Dr V S Huzurbazar. He has authored/edited five books, and published 35 articles/papers in reputed American and Indian journals and *Economic Times*. He has about 35 years of teaching experience, and has taught a wide spectrum of officers from Banking and Finance, Defence, Corporate World, Indian Economic and Indian Statistical Services, as also students of management programmes. He has guided hundreds of participants of various programmes in carrying out studies involving use of statistical techniques.

Ms Shailaja Rego is a Faculty and Chairperson of Operations and Decision Sciences at NMIMS University, in the Department of Operations. She holds Master's degrees in Statistics as well as Business Administration. She has been teaching for the past 14 years, and has guided MBA students in several projects using statistical techniques discussed in the book. She has been invited by CIDA (Canadian International Communication Technology) to present a paper on "Forecasting of Information Communication Technology Growth in India" at an international conference in Canada. She has also been invited by Kingston University, London for Faculty Development. She has also attended a programme on GCPCL (Global Colloquium on Participant-Centered Learning) at Harvard University.

The McGraw-Hill Companies

Business Research Methodology

T N Srivastava
Visiting Faculty, NMIMS
Mumbai

Shailaja Rego
Chairperson,
Department of Operations and Design Sciences
Mumbai

Tata McGraw Hill Education Private Limited
NEW DELHI

McGraw-Hill Offices
New Delhi New York St Louis San Francisco Auckland Bogotá Caracas Kuala Lumpur Lisbon London Madrid Mexico City Milan Montreal San Juan Santiago Singapore Sydney Tokyo Toronto

The McGraw-Hill Companies

Tata McGraw-Hill

Published by the Tata McGraw Hill Education Private Limited,
7 West Patel Nagar, New Delhi 110 008.

Business Research Methodology

ISBN-13: 978-0-07-015910-5
ISBN-10: 0-07-015910-6

Vice President and Managing Director—McGraw-Hill Education: Asia/Pacific Region: *Ajay Shukla*
Head—Higher Education Publishing and Marketing: *Vibha Mahajan*
Publishing Manager—B&E/HSSL: *Tapas K Maji*
Associate Sponsoring Editor: *Piyali Ganguly*
Assistant Manager (Editorial Services): *Anubha Srivastava*
Senior Copy Editor: *Sneha Kumari*
Senior Production Manager: *Manohar Lal*
Production Executive: *Atul Gupta*
Deputy Marketing Manager: *Vijay S Jagannathan*
Senior Product Specialist: *Daisy Sachdeva*
General Manager—Production: *Rajender P Ghansela*
Assistant General Manager—Production: *B L Dogra*

Typeset at The Composers, 260, C.A. Apt., Paschim Vihar, New Delhi 110 063 and printed at
Lalit Offset Printer, 219, F.I.E., Patpar Ganj, Industrial Area, Delhi 110 092

Cover Design: Aishwarya Padhye

Cover Printer: SDR Printers

RAXCRRBZDLQL

The McGraw-Hill Companies

Dedicated to

My parents,
wife Nita, our jewels Pankaj, Meeta, Vijay and Poonam
and our grand jewels Simran, Sajay, Saluni and Sumil
–T N Srivastava

My parents and my husband Leslie Rego
–Shailaja Rego

SVKM's NMIMS
Transcending Horizons

Dr. Rajan Saxena *Ph.D.(Delhi)*
Vice Chancellor &
Distinguished Professor of Marketing

FOREWORD

The subject of business research methodology is of special significance to the MBA and research students as it provides them a blue print for carrying out research / project assignments that are an integral part of their curriculum. Although there are good books available in the market but there is a felt need for a book that is written in a simple language, presents the subject in a lucid manner without compromising the basics of the subject, and is replete with illustrations from Indian business environment. I commend the efforts of Dr T N Srivastava and Mrs Shailaja Rego in fulfilling this need for the MBA and research students in India. The book has been written with the laudable objective of developing the requisite competence and confidence among the students in identifying managerial issues that could be resolved by organising an appropriate research project and subsequent implementation. In fact, the contents of the book are designed to enable a student in playing the pivotal role as well as an advisory role in future. Towards this objective, the book explains the various concepts associated with research, in general, and highlights the importance of research in a business environment.

A unique feature of the book is that every topic is introduced with a case study which is very simple and facilitates easy understanding the concept.

An exclusive feature of the book is elaborate integration of the text with EXCEL and SPSS packages for facilitating the students in completing the research assignments right from conceptual stage to collection of requisite data and to the ultimate stage of drawing conclusions based on the collected data.

It is my pleasure to recommend this book for students of MBA, M. Phil., Ph.D. and other professional courses in educational, behavioural and social sciences

(RAJAN SAXENA)

SVKM's
Narsee Monjee Institute of Management Studies
(Declared as Deemed-to-be University under Section 3 of the UGC Act, 1956)
V. L. Mehta Road, Vile Parle (W), Mumbai-400 056 INDIA. T. 91-22-26134577 / 42355555 D. 66716279
F. 91-22-26114512 E. vc@nmims.edu / rajan.saxena@nmims.edu W. www.nmims.edu

Preface

The **motivation for writing this book** is the need that has been felt by a number of MBA students as well as faculty members for a text that

- has simple language and lucid presentation
- provides comprehensive coverage of both theory and practice oriented towards business book environment
- provides tools for manual as well as modern computing.

Thus, the book should be complete in all respects to undertake a research study, collect and analyse the relevant data, and prepare and present a report.

Accordingly, this book is designed to contain a judicious blend of the theory and practice of business research and understanding and applications of statistical methodology. It has reader-friendly illustrations, especially used in statistical packages for statistical analysis. The book should be self-sufficient for MBA students to understand and apply research methods for carrying out complete research projects from concepts to conclusions and finally, report writing.

Objectives

We have tried to meet several objectives during creation of this text. They are

1. To create an interest and motivation for studying the subject of research
2. To explain the concepts associated with research, and highlight its importance in a business environment
3. To provide the skills necessary for conducting research projects, which are an integral part of the curriculum
4. To provide expertise in the use of requisite statistical techniques—manually as well as computer oriented—which are used in research projects.
5. To develop competence and confidence among students in identifying managerial issues that could be resolved by organising an appropriate research project and subsequent implementation; in fact, the contents of the book are designed to enable a student in playing a pivotal role as well as an advisory role in future.

For Whom

This book is intended to serve as a textbook for MBA students who pursue the subject of Business Research Methodology.

This book is beneficial also for students in the field of education and behavioural sciences and other professional courses like M Phil.

The book might be useful to faculty members for conducting the course on Business Research Methodology. They would benefit a lot from exclusive teaching aids like powerpoint presentations given on the website of the book.

Contents

1. Chapter 1 encompasses the relevance and importance of research, in general, and business research, in particular. It also presents an overview of the entire process of conducting research—called *research process*—in a simple language to provide a general understanding of various aspects of business research. In fact, it lays the foundation for easy learning of various concepts and thorough understanding of topics in subsequent chapters.
2. Chapter 2 clarifies the concepts of various topics viz. constructs and concepts, variables, and deductive and inductive logic, that are an integral part of a research project. It also describes quantitative and qualitative research and case study method of research.
   Two innovative features of this chapter are the discussion on two very important topics viz. Creativity and PERT/CPM that are so relevant in conducting a research study which are generally not covered in the books on BRM. While creativity is integrated with any research or management activity or for that matter with any activity in life or work environment, PERT/CPM is essential to manage any activity or project including conducting a research study. Contrary to the wrong notion that it is a complicated technique useful only for big projects, it is a simple technique useful for any work right from preparing a cup of tea to conducting a research study to building a stadium or even organising Olympic games.
3. Chapter 3 describes various topics that are relevant for the actual conduct of research in an organisation. It provides an exhaustive and comprehensive view of the various steps of a research process, from problem identification to hypothesis development, i.e. developing the statement to be tested for acceptance or rejection. This facilitates conduct of further study involving collection of data, carrying out relevant analysis, etc. and ultimately resolving the problem that would have necessitated the conduct of research study.
   It also provides guidelines to students for carrying out research projects that are an integral part of their curriculum.
4. Chapter 4 provides a comprehensive knowledge about the various experimental and research designs, and their applications in business environment, in general, and conducting a research study, in particular.
5. Chapter 5 provides a comprehensive understanding of the four types of measurement scales viz. Nominal, Ordinal, Interval and Ratio. It is intended to equip with a basic toolkit of various comparative and non-comparative scales for research.
6. Chapter 6 explains primary and secondary types of data with respective advantages and limitations. It describes the various sources of primary and secondary data as also the various methods of collecting such data, including guidance for web-based searches.
7. Chapter 7 explains the process of collecting primary and secondary data. It provides requisite knowledge about various aspects associated with designing a questionnaire for collection of primary data. It also provides guidance in collecting/recording data from secondary sources and also outlines the steps for preparation of data.
8. Chapters 8 to 13 relate to statistical topics that are covered, to a varying degree, in various institutions, prior to the discussion on the subject of Business Research Methodology. These

topics have been discussed in the book with an orientation towards conducting business research studies in various areas/fields.

9. Chapter 14 contains the **Multivariate Analysis** techniques which are included in the syllabus of most of the management institutes. However, different institutes prescribe different subsets of these techniques, out of the entire set covered in the book. **Incidentally, all these techniques are highly useful in designing and marketing of products and services**. Many of the research studies remain incomplete without the use of these techniques, mainly due to unawareness or expertise in use of **SPSS** software package for arriving at ultimate conclusions. The chapter tries to fulfil this need.
10. Chapter 15 provides guidelines for preparing a report of the research study.
11. Chapter 16 deals with the ethical issues associated with various levels of hierarchy involved with a research project, and with various stages of a research process.
12. Appendix I describes the Indicative Topics for Business Research.
13. Appendix 2 describes the role of **EXCEL** in statistical calculations. Templates have been provided that make the drudgery of calculations very simple as one is just required only to input the data in the corresponding template and the computer automatically calculates and displays the result on the screen. One is not even required to remember the statistical formulas as these have been in-built within the templates.
14. Appendix 3 describes the role of **SPSS** in statistical calculation packages in business research.

Exclusive Features

The contents of the proposed book are perceived to be ideal blend of theory and practice of Business Research and Research Methodology not only in size but also in depth. The exclusive features are indicated below:

- Business Research in an Organisation
- Relevance of Research for MBA Students
- Guide to Conducting Research Projects by Students
- Research in Management Institutions—Some Thoughts
- Dissemination of Research
- Indicative Topics for Business Research
- Time Scheduling—PERT and CPM
- Creativity and Research in an Organisation
- Research at Corporate and Sectoral Levels
- Guide for Conducting Good Business Research
- A Consultant's Approach to Problem-solving
- Cross-sectional Studies
- Longitudinal Studies
- Simulation
- Use of Graphs as Management Tools
- EXCEL
  - Relevant Details of the Package
  - Templates for Various Statistical Formulas
  - Using Templates for Solving Exercises

- SPSS
  - Relevant Details of the Package
  - Choice of Technique and Inputting Data
  - Interpretation of Output Generated by the Package
- CD
  - Examples
  - Data Sets
  - Excel Templates (to facilitate numerical calculations)
  - Cases
- Faculty Resource
  - Power Point Presentations
  - Additional examples to explain the concepts/topics
  - Solutions to questions/ problems

Acknowledgements

We are grateful to Dr Rajan Saxena, Vice Chancellor, NMIMS Deemed University, for the permission and encouragement to Shailaja Rego for writing the book, as also for writing the 'Foreword' to the book. We are also thankful to Prof Kavita Laghate of Bajaj Institute of Management Studies for encouraging and supporting the project by critical evaluation of the manuscript as also providing some illustrations and cases. We also express our thanks to Mr Leslie Rego and Mr Pankaj Srivastava for going through some parts of the manuscript and offering their valuable suggestions.

The reputed magazines like *Business Today*, *Businessworld*, *Business India* and *India Today*, and newspapers like *Economic Times*, *Hindustan Times* and *Times of India* publish live data about individuals and companies in well-researched articles. While some data is analysed by them, the other data is published just for dissemination among the readers. We have used their data to indicate the use of statistics in analysing live data that could arise in any organisation. However, while doing so, we have restricted ourselves only to analysing data without making any comments on the companies and the individuals. We would like to add that while we have taken due care to avoid any errors or omissions in recording the names of companies and individuals and the data relating to them, if any errors or omissions have occurred, these are totally inadvertent, and we extend our unsolicited apology for the same. The data and analysis provided by us could even stimulate thinking about collecting similar data for facilitating decision-making in the work environment.

We express our sincere appreciation of the efforts and contribution of Rohit Kumar Singh, Sarbani Choudhuri, Rohit Jain, students of MBA at NMIMS Deemed University, Mumbai, towards critically going through the various parts of the manuscript and their valuable advise. We are thankful to the students Akshay Cotha, Udit Sharma, Afreen Firdaus, Sajan John, Annu Asthana, Kinshuk Awasthi, Saurav Kumar, Gautam M, Anushital, Suman, Ramuni, Manu Priya, Mittal Modi, Devang Shah, Nirbhay Singhal, Sumit, Ahlawat, Kamal Kant Kaushik, Surya Sridhar A, Rishav Garg, Namit Saigal, and Sachchida Anand Sudhansu for helping in the case studies and sharing their project data. We are also thankful to R Vishwanathan and Prateek Gala, MBA students, for helping in the conduct of surveys.

We are also thankful to Mrs Carol Lobo for the valuable support provided by her.

We conclude the acknowledgements by recording our grateful thanks to the publishers and their dedicated team of officials viz. Ms Vibha Mahajan, for approval as well as generous support for

the project and Mr Tapas K Maji, for inspiration right from the conceptual stage. We are especially thankful to Ms Piyali Ganguly, for her guidance, in general, and editorial support, in particular. Her encouraging support, guidance, patience and prompt response to our queries have been largely responsible for this book being published in the stipulated time. Mr Manohar Lal, Ms Anubha Srivastava, and Ms Sneha Kumari have contributed significantly by their untiring efforts in bringing out this book.

We would also like to acknowledge the following reviewers for their invaluable feedback:

1. Sanjiwani Kumar, K.J.Somaiya College of Management, Maharashtra
2. Durga Surekha, SIES College of Management, Maharashtra
3. Vikas Nath, Jaipuria Institute of Management, Uttar Pradesh
4. S.Sivaiah, Malla Reddy PG College, Andhra Pradesh
5. K.Bharathi, St.Peters Engg College, Andhra Pradesh
6. Sourabh Bishnoi, Birla Institute of Management, Uttar Pradesh
7. Nisha Agarwal, Institute of Foreign Trade & Management, Uttar Pradesh
8. Sunita Tanwar, Ansal Institute of Technology, Haryana
9. Sanjeev Sharma, Apeejay School of Management, New Delhi
10. Susan Das, Asian School of Business Management, Orissa
11. RN Subudhi, KIIT School of Management, Orissa
12. Vijaya Bandyopadhyaya, KIIT School of Management, Orissa
13. P.Tony Joseph, Hindustan University, Tamilnadu
14. Tanmoy De, Institute of Management & Information Science, Orissa

Request to the Readers

We request the readers to kindly send their valuable suggestions for any modification or addition by way of text, illustration or case. These would be gratefully acknowledged in the next edition. The communication to authors could be sent to drtnsri@gmail.com or shailarego@gmail.com with subject marked as "BRM Suggestions".

T N Srivastava
Shailaja Rego

Contents

Introduction—Scope and Applications of Research

Research is Not an Option— It is Essential for Survival and Growth

1

Contents

LEARNING OBJECTIVES

The main purpose of this chapter is to provide basic understanding of the objective and motivation for conducting research. The concepts of research as also various types of researches are described in this chapter.

The criteria, characteristics and challenges of conducting any research are indicated. All the steps for conducting a research study constitute what is called a **'research process'.** This chapter provides an overview of the complete process of a research study, in a simple language. This is to lay the foundation for easy learning of various concepts and thorough understanding of topics in subsequent chapters.

A complete guide to students, especially those students pursuing MBA is provided to instill in them the confidence and to develop competence for understanding a research study.

Relevance

Mr. Raj Kumar retired gracefully, earning a great deal of respect from his juniors as well as seniors, as evidenced by the grand farewell function which he just attended.

While returning home with a cheque of Rs. 50,00,000 which he received as payment of PF, gratuity etc., he was thinking about his wife who was not mentally prepared like him for this unavoidable day. He reached home and took his wife out for dinner. On the way to the restaurant, at his wife's favourite jewellery shop, he purchased a diamond earring to boost her morale. While the day thus ended peacefully, the next morning Mr Kumar himself started pondering about the ways to manage his future through financial planning. Even though, retirement was a reality of life, he had never given a serious thought to it.

He started thinking of various investments options to ensure monthly inflow of income that would be sufficient to live a reasonable quality of life. He started thinking about fixed deposits of banks, post office and corporate bodies, monthly income schemes, mutual funds, etc. One simple investment in MIS of a bank came to his mind; if he invested all Rs. 50 lakh there, he would get a monthly income of about Rs. 33,500 (assuming 8% interest). While this amount was sufficient at present, he prudently visualised that at annual inflation of about 7%, this amount of equivalent to Rs. 33,500 will be Rs. 19546.93 after 10 years and Rs 9054.01 after 20 years. He figured out that this was not the ideal investment, and had to collect detailed information about all types of investments, enumerating their advantages and disadvantages. Since he could not invest on trial and error basis, he had to make a thorough study of all the options. While some investments assured interest/amount on maturity or on yearly basis, other investments like mutual funds indicated higher expected returns.

He also had to be financially prepared for contingencies like sickness, recession, etc.

He had to finalise a judicious mix of various types of deposits and investments in mutual funds, etc. so as to ensure a regular inflow of income to assure a financially secure life for him and his wife.

How, the subject of Business Research Mythology (BRM) can help in providing solution to such problem is indicated later, in Section 1.7.

1.1 INTRODUCTION—GLIMPSES OF PAST RESEARCH

At the outset, we would like to mention that even though the focus of this book is on Business Research, the basic concepts and features described in this chapter are relevant for research in any field including business.

Research in simple words means **'Something new—physical or conceptual (Knowledge)'**. The word 'new' is the underlying concept in any formal definition or meaning of research. The formal definitions and concepts of 'Research' are given in Section 1.2 titled 'Definitions of Research'.

Researchers have contributed in several ways to our knowledge enhancement and understanding. They have

- Demystified the mysteries of nature, universe and world around us
- Discovered the anatomy and functioning of our body
- Discovered medicines and surgical procedures to reduce our sufferings and also increase the life span
- Provided us the means of comfortable living in all respects, improved quality of life, mapped the world to a global village
- Provided tools for conducting business and trade in an efficient manner, and many other innumerable ways.

Here is a list of some of those great researchers, most of whom need no introduction.

Table 1.1 List of Famous Researchers and Important Researches:

Name	*Research*
Abraham Harold Maslow	Hierarchy of Needs (Physiological Safety, Social, Esteem, and Self-Actualisation)
Albert Einstein	Relativity Theory
Albert Humprey	SWOT Analysis (Strength, Weakness, Opportunities, Threats)
Alex Osborn	Brainstorming
Alfred Nobel	Dynamite. Received Nobel Prize in the areas of chemistry, physics, literature, international peace and medicine
Archimedes	The Archimedes Principle
Arya Bhatta	Digit '0'
Bill Smith at Motorola Juran at GE	Six Sigma Model
Chandrasekhara Venkata Raman	Raman Effect (Scattering of Light)
Charles Darwin	Darwin's Theory of Evolution
CK Prahalad	Fortune at the Bottom of the Pyramid Strategy
CK Prahalad	Bottom of the Pyramid Strategy
EJ McCarthy	Four 'Ps' (Product, Price, Place, Promotion) classification Model in Marketing
Galileo Galilei	Used telescope to prove that the earth revolves around the Sun
CK Prahalad and Gary Hamel	Core Competence Model
Hamel and Prahalad	Core Competence Model
Hammer and Champy	Business Process Reengineering (BPR)
Herman Ebbinghaus	Learning Curve
IBM	Smart Planet Project

(*Contd*)

(*Contd*)

James Maxwell	Electromagnetic Theory
Japanese Business Companies	Kaizen Philosophy of Continuous Incremental Improvements
John Logie Baird	Television
John Nash	Game Theory
Taiichi Ohno, Toyota Corporation	Just-In-Time Business
Marie Curie	Theory of radioactivity (the first person honoured with two Nobel Prizes)
Michael Porter	Competition and Company strategy (Five Competitive Forces Framework, Value Chain), Competition and Economic Development (Clusters, Diamond Model), and Competition and Societal Issues
Myron Samuel Scholes	Black–Scholes Equation (Valuing Derivatives/Options)
Norman Borlaug	Green Revolution in India (Wheat)
Peter Drucker	Management by Objectives, Knowledge Work Productivity
Raymond Vernon	Product Life Cycle
Sir Isaac Newton	Newton's Laws of Motion
Stephen R. Covey	Seven Habits of Highly Effective People
Taiichi Ohno, Toyota Corporation	Lean Manufacturing (Just-In-Time Supply Chain)
Thomas Alva Edison	Motion Picture Camera and Electric Bulb
Wilhelm Conrad Rontgen	X-rays
Wright Brothers, Orville and Wilbur	Airplane

1.2 DEFINITIONS OF RESEARCH

There are several formal definitions of research such as:

- Thorough systematic investigation
- Careful or diligent search, studious inquiry
- Endeavour to discover or collate old facts by the scientific study of a subject

Yet another formal definition of research is:

"It is organised systematic data-based scientific inquiry, or investigation into a specific problem undertaken with the purpose of finding answers or solutions to it."

In the context of the business environment, it is defined as:

"The research provides the need that guides managers to make informed decisions to deal with problems, successfully collect facts or observation and to present them in a systematic and logical manner."

Research is basically a human activity engaged in intellectual pursuit of discovering something new. It could be a product, a method, a service, a system, etc. It could also be in abstract forms like idea, thinking process, strategy, etc. However, the intellectual activity is not confined to highly educated persons or professionals as it pervades all sections of the society. In fact, for many, research is an intimidating term, even among those in academic field, and it is usually perceived to be in the realm of upper strata of knowledge.

We believe, in the broadest sense, research is simply the process of finding solutions to a problem after a thorough study and analysis of the situation.

If we take the above concepts to a logical conclusion, all of those who are engaged in finding out new items or/and better ways of doing things could be said to be doing research.

1.3 RESEARCH PERCEPTIONS

It is interesting to note that how different people have different perceptions about research. Several highly placed professionals were asked for a spontaneous response to the question "What comes to your mind when you read or hear the word "Research?". Some of their responses are reproduced below:

Research Perceptions

- "Deep dive into a subject; thirst for knowledge"
- "Comprehensively analysing data to look for trends, themes, revelations"
- "Curiosity and how it fuels the passion to discover or uncover truths, myths, trends, themes, etc."
- "Piles and Piles of data–can get too mired in details"
- "Needs someone with a strong ability to synthesise all the information and data to make it meaningful, cogent and relevant (relevance can often gets overlooked)"
- "Scientific research, labs, focus groups, numbers, statistics"
- "Ways of working out market research, production planning, budgeting, manpower forecasting"
- "Looking for data, and/or information to provide meaningful insights towards a business problem"
- "Ph.D."
- "Ability to deliver insight"
- "Doing analysis and background checks, examining options and alternatives in order to arrive at best choice for a business challenge/requirement"
- "A person in a white coat working in a lab"
- "Something for the betterment of civilisation"

1.4 OBJECTIVES OF RESEARCH

The objectives of research are different for different individuals and organisations. The following objectives which encompass most of the researches are listed. However, these are not comprehensive, and some may even be overlapping.

(i) To seek insight into an observed phenomena and explain its logic and reasoning of happening. For example, declining profitability.

(ii) To help the mankind in solving the problems faced from time to time, and make life more comfortable and entertaining. For example, telecommunication and e-ticketing.

(iii) To explore the possibility and methodology of doing things which have not been done so far but are useful for the mankind, in general, and an entity, in particular.

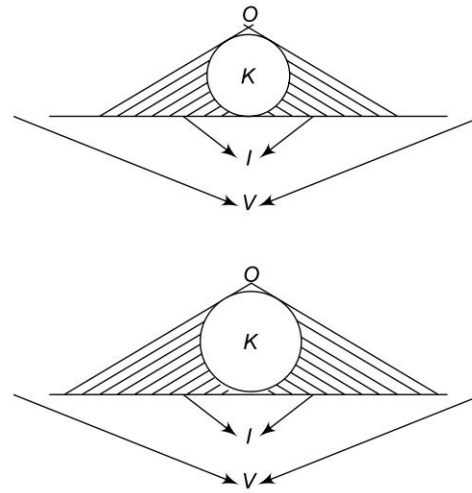
(iv) To continuously improve the effectiveness of present systems and procedures in any field. For example, compensation, recruitment and retention policies.

(v) Test or challenge existing beliefs, notions, etc. which have not been empirically proved so far, with flux of time and therefore need to be tested again for relevance in the changed context/ environment. For example, relationship between intelligence and creativity.

(vi) Explore into new areas that might have become relevant or even might become relevant in the near future. For example, alternative sources of energy to reduce carbon emissions.

(vii) Anlayse the past data for discovering trends, patterns and relationships. For example, business performance, prices of stock, oil, etc.

(viii) To expand the sphere of **knowledge (K)**, and increase the horizon of **vision (V)**. However, incidentally this simultaneously increases the realisation of **ignorance (I)**. This phenomenon is explained through the following diagram:



While standing at the top of the knowledge sphere, we cannot see the area below the surface of the sphere, but we can see farther towards the horizon.

From the above diagram, we note that as the sphere of knowledge grows, the vision increases but simultaneously, realisation of ignorance increases. It is perhaps because of this reason that the great philosophers and intellectuals are very humble as they realise that they do not know a lot!

1.5 MOTIVATION FOR RESEARCH

It is interesting to go behind the reasons for conducting research by an individual or an organisation. Behavioural scientists could term this as the factors that cause motivation, and are responsible for initiating research.

1.5.1 Motivation for Research by Individuals

The motivational factors for research for an individual and for an organisation are discussed separately, as per details described below.

If we go into the past and study innovations by **individuals**, we may list some of the factors that could have motivated them in pursuit of research.

- Inquisitiveness
- To demystify the mystery
- Desire to undertake a challenge in solving unsolved problems
- Reveal the secrets of nature and explaining natural phenomena
- Desire to help the mankind by discovering medicines, surgical procedures, improving productivity and quality of cereals, vegetables fruits, etc., improving modes of communication and entertainment, etc.

In the present day context, the following factors may be added:

- Desire for self-advancement
- Completion of mandatory assignment in professional courses and career
- Desire to get Ph.D. degree
- Desire to get recognition and distinguishing oneself from others by being the first to do something new
- Intellectual satisfaction
- Spirit of accepting challenge
- To enjoy the thrill of solving some unsolved problems or making something 'impossible' as 'possible'
- To realise one's dream or fulfilling parents' dream.
- To survive and grow in a competitive environment. Incidentally, this is the Darwinian Theory of Evolution of Species and is equally relevant in the competitive and fast changing world.

1.5.2 Motivation for Research in Organisations

In the context of **organisations,** it may be mentioned that all the research that is conducted in the present day world is not just due to the motivational factors. There are several other factors that induce a corporate body to conduct research studies. While it is difficult to enumerate all the factors, some of them are described below.

(i) Self-Motivation

The need for research is motivated by the desire to:

- Improve sales, profit, market share
- Improve rank among competitors
- Reduce cost of products/services/operations
- Increase ROI (Return on Investment)
- Diversification of products and services
- Entering new market
- Acquisition and Merger
- Improve quality of products/services
- Improve quality and productivity of systems

(ii) Regulatory

One may have to undertake research due to imposed regulatory conditions. For example, car manufacturers had to develop engines to meet the emission and fuel efficiency standards like Euro II. In the case of banks, they had to reshuffle their loan portfolio to meet the capital adequacy norms in the form of capital adequacy ratio (CAR) imposed by the Reserve Bank of India.

It is not only the imposition of regulatory requirements that cause the need for research, sometimes even regulatory relaxations also create the need for research to derive maximum advantage of the regulations which could be in the form of liberalisation, opening of new avenues, or even tax incentives. Following are some of the situations of this type:

- Liberalisation
  In general, liberalisation leads to increased competition which, in turn, leads to the need for improvement in quality of products and services, and lowering of costs. This can be achieved only through research at various levels in production, marketing, distribution systems and organisational structure.

Telecom and insurance sectors in India are classical illustrations.

- Tax Incentives
  Sometimes, the Government of a country might introduce certain tax incentives to encourage development in the desired direction. For example, when Government of India wanted to economise the use of petrol by cars, it introduced tax incentives for cars with engines of capacity less than 1000 cc. This measure induced the automobile industry to reorient their efforts towards designing of small cars. Incentive for energy efficient devices is another example.

(iii) Competition

It is perhaps the single most important factor that is responsible for continued research in any organisation. In fact, as mentioned earlier, these days research is not an option but is essential for survival and growth. It is equally true for an individual (especially professionals) as well as an organisation. The competition could be caused by:

- Regulatory measures to allow new entrants in a particular sector
- Entry of new ventures
- Entry into new market
- Shrinking of existing market

Competition usually leads to narrowing of margins which necessitates research to bring down the costs and prices and also to increase the volume of sales.

(iv) Customer Driven

One may have to engage in continued research to meet even growing expectations or new needs of customers.

(v) Failure

Failure of a product or a service or an advertisement campaign leads to research for finding out the reasons behind failure and accordingly drawing up a strategy.

(vi) Technological Innovations

One may have to undertake research for technological innovations. For example,

- In medicine, technological innovations are required to facilitate development of new medicines and surgical treatments to reduce suffering of human beings and to increase their lifespans.
- In agriculture, technological innovations are required to facilitate development of new seeds, cultivation and harvesting to improve quality and productivity of crops.
- Facilitate communication and entertainment options through innovations in telecom.

(vii) Environmental Considerations

Environment is also an important factor that is responsible for carrying out research. For example, the conventional bulbs and tubelights emit carbon. This led to the development of CFL (**Compact Fluorescent Light**) bulbs which are environment-friendly and also economical due to lesser consumption of electricity. The same logic has been responsible for the development of LED (**Light-Emitting Diode**) bulbs which are environment-friendly and consume lesser electricity than CFL. The extra cost of these bulbs is more than compensated by the lesser consumption coupled with long life of these bulbs.

(viii) Social

Social factors are also responsible for carrying out research. For example, high cost of gold led to designing lightweight jewellery. High cost of petrol led to designing of small cars. Increasing concern for financial security of families led to designing of several insurance products. Increasing concern for health led to developing drinks, snacks, oil, etc. with low fat contents/sugar, etc.

(ix) Economic

Economic factors also play an important role in conducting a research study. For example, high prices of petrol led to the development of cars with diesel engines. Higher value of dollar led to product substitution for imported items. Volatility in stocks and foreign exchange markets led to research in management of risk in financial investments.

(x) Infrastructure

Infrastructure is also important for conducting a research study. For example, increasing prices of residential property and higher home loan rates led to designing and development of low-cost housing projects.

(xi) Operations/Process Driven

With discerning customers opting for higher quality products and services, companies started implementing six sigma system in operations. Research is definitely needed to achieve the desired objectives in this direction. As regards the process, one illustration is from mechanical harvesting of tomato crops in USA. When it was observed that many tomatoes got spoiled due to their thin skin, the research was conducted to increase the thickness of tomato skin, thus resulting in considerably lesser percentage of spoiled tomatoes.

(xii) Coping with Changes

It is said that

"The world is moving so fast that if you just want to remain where you are, you have to run".

"The only thing that is constant in the world is the CHANGE".

We would like to add that:

The only thing that is certain in the world is that Rate of Change will keep on increasing.

The changes could occur in any one or all the factors described above.

The above factors are depicted in Fig. 1.1 on next page.

1.6 TYPES OF RESEARCH

There are several types of research depending on the purpose, methodology or field of application. Some of these are as follows:

- Basic, Pure, Conceptual and Fundamental
- Applied
- Empirical
- Historical
- Scientific
- Social and Behavioural
- **Business**
- Exploratory
- Descriptive
- Causal
- Normative

In addition, researches are named according to the subject in which it is applied. Examples are: Organisational Development, Economics, Operations Research, Agriculture, Medicine, etc.

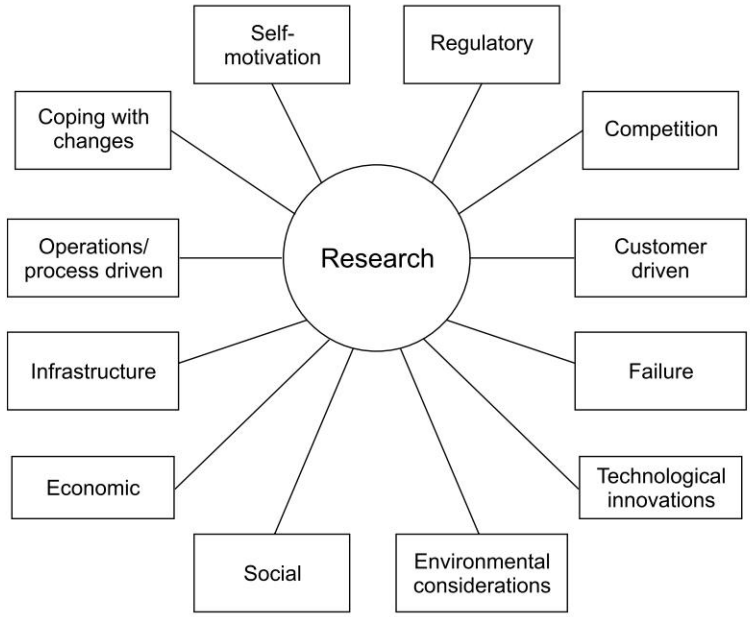


Fig. 1.1 *Motivational Factors for Research in Organisations*

A research is also classified as **Primary** and **Secondary** research depending on the type of data used. If the data is primary, i.e. collected for the study, it is called primary research; but if the data is copied/recorded from the published sources/Internet, etc., the research based on such data is called secondary research. Incidentally, market research involves both primary and secondary research, as indicated below:

Primary Research (based on)	*Secondary Research (based on)*
Data collected by government agencies like data on prices, production, bank deposits, etc.	Reports/Publications analysing and evaluating data collected by others
Data collected through consumer surveys, opinion polls, interviews, etc.	Tabular, pictorial, graphical presentations based on primary data
Financial data published in annual reports of companies	Compilation of data from a number of sources like done by Centre for Monitoring Indian Economy, and then publishing it
Original reports published by the collectors of data like Annual Survey of Industries by the Government, Annual Report by Reserve Bank of India	Articles interviewing those who conducted primary research

The details of collecting primary and secondary data are explained in Chapters 6 and 7.

Further, a research could be termed as **Quantitative** or **Qualitative** depending on the topic/aspect of research and type of data collected for analysis. If the data involves quantitative aspects like measurement and counting, the research is termed as quantitative. However, if the research

involves study of behaviour, attitude, etc., it is termed as qualitative. These are discussed, in detail, in Chapter 2.

1.6.1 Basic Research

Basic research is also known as **pure** or **fundamental** research. Its main objective is to extend the existing domain of knowledge about certain subject or topic either in physical form like sales, or in abstract form like human behviour. The knowledge itself could be in the form of trend, pattern or relationship. Individuals conduct basic research primarily out of their curiosity, inquisitiveness, conviction etc. Some examples from the management perspective are:

- Relationship between intelligence and creativity
- Relationship of analytical ability and verbal ability with the scores obtained in various subjects
- Relative impact of factors like salary, work environment, reward system affecting motivation of employees
- Relative impact of advertisements on various media such as newspaper, magazine and television
- The **Black–Scholes model** (mathematical model of the market for an equity)

Many of the pursuers of basic research are not concerned about the immediate applications of their research. However, most of the basic researches find their application sooner or later. In fact the first author, had visualised, that the failure rate* of electronic items could change periodically during day and night, summer and winter, etc. While his idea could not be used in India, the idea was found useful in USA in designing satellites which while revolving around the earth are subjected to periodic stress at different positions in different orbits.

*Interestingly, the human failure rate follows a bath-tub pattern as shown below:

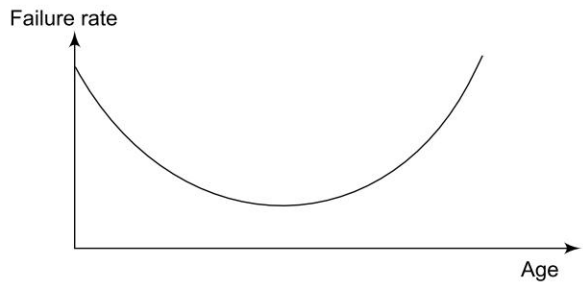


Fig. 1.2 *Human Failure Rate: Bath-tub Pattern*

The failure rate is highest in the beginning, and decreases with time. Thereafter, it is constant for quite some time, and then it starts rising.

Many applied fields like electronics, medicine, pharmaceuticals and biotechnology, etc. owe their growth due to basic research in basic sciences such as physics, chemistry, etc.

*Incidentally, the failure rate of physical items and living ones increases with time or age.

1.6.2 Applied Research

It is the research relating to a specific product, service or system or campaign. For example, a branch manager may like to streamline the functioning of its customer counters to reduce waiting time for the customers. A company may like to evaluate impact of its advertising campaign. It may also be used to promote a product or class of products through favourable results obtained through research. Applied research aims at solving any problem or resolving any issue in a scientific and systematic manner. The problems or issues could vary from organisation to organisation. For example, in a business enterprise, it could relate to designing and marketing of products and services. The terminology 'Applied Research' gets its name depending on the area or field where it is applied. Some of these examples are: Social, Behavioural Science, Human Resource Development, Organisational Development, Economic, Marketing , Operations, Business. **The last type viz. Business Research is at the central stage of the contents of this book.**

1.6.3 Empirical Research

Empirical research is a research which is based on observed data without the support of any theory or model. Such research seeks to resolve an issue or reach any conclusion by using empirical or data-based evidence.

For example, all the research about the heavenly bodies is based on data collected about their movement. Empirical research is usually resorted to when the event or phenomenon cannot be explained by logic or scientific reasoning.

A classical example of empirical research is the study that was conducted by Francis Galton in 1838. The study related to the relationship between heights of fathers and sons.

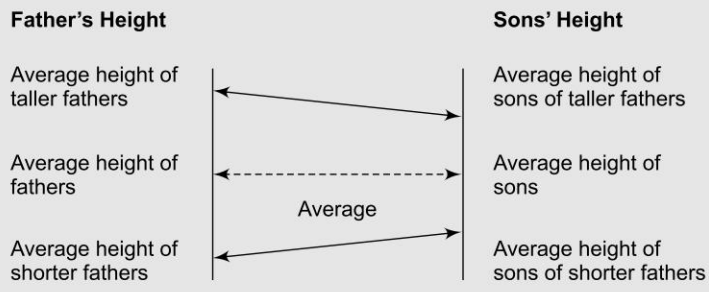
Relationship between Heights of Fathers and Sons

In day-to-day family conversations, it happens quite often that when a child is taller than the average child, people remark that it is going on the father/mother. Francis Galton was, perhaps, the first person who ventured to study this feature based on actual data. In fact, this is amongst the first reported studies on relationships between any two variables.

He collected data about the heights of 1000 fathers and their 1000 sons. The data revealed that the average height of sons was the same as that of their fathers – implying that there is no change over a generation. However, he discovered an interesting feature!

He classified the fathers in two categories viz. tall and short fathers. The taller fathers were those whose heights were more than the average height of all the fathers. Similarly, shorter fathers were those whose heights were less than the average height of all the fathers. He calculated the average heights of taller as well as shorter fathers. Galton, then calculated average heights of sons of taller and sons of shorter fathers, as shown below:

It was observed that the average height of sons of taller fathers was less than those of their fathers, and the average height of sons of shorter fathers was more than that of their fathers. Based on this, Galton concluded that it is not necessary that taller fathers would continue to have taller sons or shorter fathers would continue to have shorter sons, and remarked that height regresses i.e. the height tends to go back to the average. Hence, even today, the phrase 'Regression Analysis' is used for the study of quantifying relationships described in detail in Chapter 10.



Yet another example is the **frequency** of occurrence of different **alphabets** in the **English** text. All the alphabets viz. a, b, c ……, y, z are not used equally frequently in the English language. The frequency of their occurrence has been found to be as follows:

Frequency of Occurrence of Alphabets in English Text

In the English language the alphabets viz. a, b, c, ……., y and z are not used equally frequently. The frequency of their usage has been observed to be is as follows:

Alphabet	*Frequency of Usage (%)*	*Alphabet*	*Frequency of Usage (%)*
a	8.2	n	6.7
b	1.5	o	7.5
c	2.8	p	1.9
d	4.3	q	0.1
e	12.7	r	6.0
f	2.2	s	6.3
g	2.0	t	9.1
h	6.1	u	2.8
i	7.0	v	1.0
j	0.1	w	2.4
k	0.7	x	0.1
l	4.0	y	2.0
m	2.4	z	0.1

It may be noted that the alphabet '**e**' is used most frequently to the extent of about **13%** and the alphabet 'z' is used least with a frequency of only about **0.1%**.

F.B. Morse used this empirical analysis in designing codes for various alphabets for transmitting of messages. These codes comprise two symbols viz. dots ('.') and dashes ('–'). The assignment of codes to different alphabets was based on the criteria that more frequently an

alphabet is used, the lesser should be the time for its transmission. That is how, he assigned the code '.' to the letter 'e', the most frequently used alphabet so that it would take least time for transmission. The present-day binary codes comprising of two digits viz. zeros (**0**'s) and ones (**1**'s), ideal for transmission through computer media, are also designed on the above criteria.

Anyone engaged in printing or manufacturing or using various alphabets for any other purpose, also uses the above information, as each of the alphabets need not be produced or used in equal quantity.

1.6.4 Scientific Research

Scientific research has two connotations. One is the research conducted in science subjects such as physics, chemistry, biology, etc., and the other is the scientific process of conducting a research study.

The first connotation i.e. scientific research relates to explain the natural phenomena, functioning of body elements of living ones, preventing and curing diseases and body disorders, and inventing means for providing more comforts, long and healthy life, variety of entertainment, etc.

As regards the **second connotation**, we refer to the dictionary where the word 'scientific' connotes logical, systematic and unbiased. This is how we refer to the desirability of doing anything in a scientific manner. Even though, the word scientific implies 'systematic', quite often we use the phrase 'in a scientific and systematic manner'.

Characteristics of Scientific Research Whenever we mention that a study has to be conducted in a scientific and systematic manner, we imply the following characteristics:

- Clarity of purpose
- Objectivity without any bias and prejudice
- Using proven and established methodology explained with the help of diagrams and charts, etc. without any trace of ambiguity
- Collecting evidence in acceptable form
- The assumptions made for prescribed analysis are justified in the actual analysis
- Conclusions/Recommendations to flow naturally from the analysis without implanting own subjectivity or views. These have to be totally new or modified form or extension of earlier research. Even contradiction of earlier research is also research.
- Should be replicable by others to verify or modify or extend or even contradict
- Scope and limitations of the entire study to be brought out clearly

1.6.5 Social/Behavioural Research

It refers to research conducted by social and behavioural researchers in sociology, political science, behavioural science, education, etc. Social researchers use several methods to explore, understand and describe social life.

Behavioural researchers attempt to analyse reasoning for behavioural issues leading to criteria for selection, preference, etc. In fact these researches led to the development of **"Qualitative Research",** described in detail in Chapter 2. These are now highly useful in studying the consumer behaviour and preferences for various products and services.

1.6.6 Historical Research

Historical research is defined as:

"The process of systematically examining past events to give an account; may involve interpretation to recapture the nuances, personalities, and ideas that influenced these events; to communicate an understanding of past events."

Such research comprises methodology used by historians for discovering evidence about certain reported events, developments, etc., and using the same for writing history. For other academicians, like economists, in business environment, it facilitates discussion of the past events—linking them to the present events, and allows one to reflect and provide possible answers to the current issues. Like managerial planning exercise, it attempts to discuss the following questions:

- What is the present issue problem?
- What happened earlier in the similar situation?
- Relevance of past conditions to the present situation.
- What will happen now?

Learning form Historical Research – Emulate success and Avoid failure!

1.6.7 Business Research

It is the research relating to problems or issues relating to business entities or business environment or a group of business entities. It has been formally defined as follows:

"Provides information to guide managerial decisions about conduct of business".

"It is the process of planning, acquiring, analysing and disseminating relevant data, information and insights to decision-makers in an organisation to take appropriate actions that lead to maximising business performance".

"The basic objective of BRM is to get the most useful information for decision-making in the most cost-effective yet realistic manner!"

Each manager or businessman is continuously engaged in research—trying to solve a problem—small or big, consciously or unconsciously. While sometimes, the problem may be due to extraneous reasons, many a times, the problem is due to internal phenomenon or self-generated by the inner desire to improve or excel. Thus, since they are constantly engaged in studying and analysing issues in their own ingenious ways, it may not be an exaggeration to say that managers and businessmen are continuously engaged in researching all the time. Some of the topics related to business research are given in the following table:

Table 1.2 Topics of Business Research

Area	*Topic for Research*
Marketing	Branding
	Pricing
	Effectiveness of Launching a product
	Effectiveness of Advertisement campaign

(*Contd*)

(*Contd*)

	Assessment of Demand for the Product or a Service
	Customer Profiling
	Customer Relationship Management (CRM)
Finance and Banking	Risk Management
	Credit Rating
	Evaluation of Investment
	Quality of Assets
	Brand Evaluation
	Equity Research
	Derivatives/Futures and Options
	Mergers and Acquisitions
Operations	Operations Research
	Six Sigma—Controlling and Improving Production Process and Quality
	Technology Absorption
	Supply Chain Management
HRD	Recruitment and Retention Policies
	Performance Appraisal and Reward System
	Training
IT	Website Management
	Network Management
	Decision Support System
	Business Intelligence and Data Mining
Retail Management	Identifying Customer Buying Behaviour: Preferences and Patterns
Insurance	Designing Policies of Various Types
	Impact of Different Factors on Health and Life
Telecom	Criteria for Selection of phone and service provider among different age/income/professional groups

A detailed list of indicative topics for business research is given at the end of the book.

1.6.8 Exploratory Research

Exploratory research, as the main word 'explore' suggests, is conducted to explore a problem, at its preliminary stage, to get some basic idea about the solution at preliminary stage of a research study. It is usually conducted when there is no earlier theory or model to guide us or when we wish to have some preliminary ideas to understand the problem to be studied, as also the approach towards arriving at the solution. It might help in modifying the original objective of the study or might even lead to a new perspective rather than the earlier perceived problem.

Exploratory research is used to finalise the questionnaire or format/schedule to ensure coverage of all possibilities that could arise while filling up a questionnaire or schedule during conduct of full-fledged research. Exploratory analysis is recommended for designing questionnaires especially for deciding the number of items/categories on which the data is to be collected.

For example, a shirt manufacturer sponsored a survey to find the percentage of executives purchasing different sizes of a shirt. The interviewer (researcher) was asked to record the sizes 36, 38, 40, 42, 44 as indicated by executives. The exploratory survey indicated that quite a good percentage of executives indicated the size as 39 and 41 (which were either imported or tailor-made). This information led to change the questionnaire to include these options.

In a telecom survey, respondents were required to indicate the criteria for selecting a service provider. A number of categories of users were listed, however, while conducting an exploratory survey it was observed that there were about 5% respondents (selected by criteria of contacting every 20^{th} person entering the mall) who were retired persons. Thereafter, the category of user was modified to include retired category.

Similarly, exploratory analysis is considered mandatory while designing **Management Information System (MIS)** for an organisation.

One interesting feature of exploratory research is the **flexibility** in selection of units for recording information. For example, if the information/data is to be collected though interviewing persons who could be either experts in the area or who could be subjects/participants for whom the research is being conducted. In such cases, the selection of personnel is somewhat flexible and not rigid. It depends on the researcher's perception as to who would be able to provide the relevant and requisite information.

In the context of exploratory research involving use of quantitative data, the following features may be studied:

- Range of the variables being studied.
- Variability in the data – helps in deciding sample size. More variability requires more sample size.
- Proportion of units having a particular characteristic. Lower or higher proportion than 0.5 indicates need for bigger sample size. As indicated in Chapter 10, the sample size required for estimating proportion is 0.5. Thus, if in the exploratory sample the proportion of units having particular characteristic is 0.5, the sample size required will be maximum. It goes on decreasing as the observed sample proportion moves away from 0.5 on either side.
- Proportion of units in various categories/groups (significantly higher or lower proportion may lead to redefining categories/groups.

  For example: In the credit card survey of a bank the variable for the average amount they would be ready to spend (per month) if the cash-back of 5% is offered to them, was categorical, variables ranging from: less than 5000, between 5000 and 10000, between 10000 and 20000, between 20000 and 50000 and >50000. For conducting exploratory research, a preliminary sample of cardholders was selected. It was found during the exploratory analysis that the category less than 5000 had 40% of the responses, and >50000 had no responses. This prompted the researcher to redesign the categories as <2500, 2500 to 5000, 5000 to 10000 and >10000.
- Trend Analysis/Pattern – For example, plotting of sales of selected items on various days of a week may indicate the trend and pattern of sales.
- Association – Study at a garment store may be used to discover association between colour of shirts and age group of customers.

- Distribution – Return of equity of a group of companies, stock indices like Sensex, Nifty – Symmetrical or Skewed
- Study the impact of proposed changes
- Study the factors contributing to substantial decrease or increase in business

We may associate three issues with an exploratory research viz.

- Why it is conducted?
- When it is conducted?
- How it is conducted?

These are explained in the following table.

Table 1.3 Issues Associated with an Exploratory Research

Why	• Formulate the precise problem for more detailed or deeper study • To arrive at some theory or hypotheses to be tested in detailed study • Discovering – Ideas – Insights – Trends – Patterns • Designing Questionnaire/Schedule • Designing MIS
When	• There is no prior research done in the similar field or the field is ever changing • There are limitations of resources like Time and Money
How	Methodology of Conducting • Qualitative • Quantitative • Primary • Secondary

1.6.9 Descriptive Research

In statistical methods, we study measures of location and dispersion like mean, median, mode, standard deviation, coefficient of variation, measure of skewness, etc. All these measures are used to **describe** the characteristics of data. Thus, any research which aims to describe the above characteristics of data is labelled as Descriptive Research or, sometimes even statistical research. We hasten to add that the scope of statistical research is much wider and comprehensive as it includes topics such as correlation/regression analysis, statistical inference, forecasting, designing of products and services, etc. Descriptive research is only one of the components of statistical research. In fact, descriptive research is usually the first step towards any full-fledged statistical analysis.

Descriptive research is said to answer the questions ***who (or which)***, ***what***, ***where***, ***when*** **and** ***how.*** It may be noted that it does not answer the question 'Why'?

This is illustrated in the following table.

Table 1.4 Questions Answered by Descriptive Research

Adjective Typifying the Research	*Illustrative Question*
Who	Who has been most consistent batsman among Sachin, Dravid and Ganguly, in the test matches?
Which	Which is the cricket ground where maximum number of centuries have been scored?
	Which are the companies that have declared more than 50% dividend for the year 2009-10?
What	What is the average salary offered to MBA students with marketing specialisation?
	What proportion/percentage of engineering graduates opt for specialisation in Finance?
Where	Where the response to a particular advertisement was most favourable, among all the major cities where the test marketing was carried out?
When	When did the manufacturing process go out of control?
	When does the peak festive sales start?
How (Much, Many)	How much productivity increased in an organisation after implementing new financial package?
	How many Mutual Funds have paid more than 10% dividend for their Tax Saver Scheme?

In short, **descriptive research** deals with everything that can be measured or counted.

It is the most commonly used research in business. Many a times, this research is carried out to investigate the reasoning or logic of occurrence of a certain phenomenon/event.

- Relationship between two or more factors
  Example: Sale of ice cream and temperature
- Comparison of two or more factors
  Example: Out of marketing ability and financial management, which is more important for the success of a company?
- Identifying the most important or common factor
  Example: Which income group is availing the revolving credit facility of a credit card provider, to the maximum extent?
  Which category of credit card users is defaulting maximum (a) in number and (b) in amount?

Descriptive research is mainly done when a researcher wants to have a quantitative idea(s) of the variable(s) under study. However, while this research is highly accurate and useful, it does not provide the causes for the finding behind a situation as such.

For example, while analysing marks obtained by the batches of 2008 and 2009 MBA students, it was observed that

'Average salary offered for the batch of 2009 students had decreased but the standard deviation had increased as compared to 2008'.

While the above finding is factual, it does not indicate any reasons for the observed change. On further probing the reasons (descriptive research), it was noted that most students of the 2008 batch

had a work experience of 2.5 years, at the time of joining MBA as compared to 0.5 years for 2009 batch. Further, the employment scenario in 2009 was not as good as in 2008.

1.6.10 Causal Research

The causal research is concerned with finding the root cause of a symptom. For example, if the sale of a product is declining, or if customers prefer a product over other similar product(s), one may like to know the cause(s) for the same. Thus, this type of study encompasses situations where we study the impact or influence of one factor (cause) on some other factor (effect). The influencing factors could be one or more than one.

Some of the examples of such research are

- The factors influencing buying behaviour of customers
- The factors influencing the motivation of an employee
- Identifying factors affecting NPAs (Non-Performing Assets) of a financial institution

1.6.11 Normative Research

Normative Research is usually conducted while developing a new product, service or system to assure whether desired objective/standard has been achieved. Some examples are:

- Productivity of staff or a production system to be increased to a specified level (say, from 2 to 2.5 units)
- Waiting time for customers at counters to be reduced to the desired level (say, from 15 to 10 minutes)
- Processing time of application for dispatching an item or providing a service to be reduced by specified amount (say, from 2 to 1 day)
- Processing time for redeeming units in mutual funds (say, from 15 to 10 days)
- Time taken for declaring result from last day of examination at an institute (say, from 4 weeks to 3 weeks)

Normative Exploratory Research This is the normative research of exploratory nature.

As mentioned above, normative studies aim at achieving some desired goals or targets which are usually improvement over earlier level. Exploratory studies do assume certain theory or model which explains the present level, and build up the normative exploratory approach from that theory or model.

"For example, if a bank is required to conduct a normative study to increase the productivity of staff from existing level of business from 1.5 cr per employee to 2.0 cr per employee within one year, the need for research arises from the fact that the productivity of a bank employee depends on skills and attitudes of the staff, marketing efforts by the concerned executives, services and schemes evolved for the customers, level of computerisation, etc. The management may like to have an idea about the impact of these factors on the productivity in the bank. This would enable them to bring about requisite changes in each of the factors to attain the new **'norm'** in respect of productivity.

A similar study could be conducted for improving profitability or any other parameter of a company or operations or a client, etc.

1.7 RESEARCH PROCESS: THE PROCESS OF CONDUCTING BUSINESS RESEARCH–AN OVERVIEW

The research process is the methodology of conducting a research assignment/project/study in a scientific and systematic manner. It takes into account all the relevant factors that are important in ensuring that the objectives of the research study are achieved with optimum utilisation of resources. It also ensures that the approach is quite comprehensive with the involvement of all those who are

- Experts in the area
- Associated with the management of the project, and
- Associated with the execution or implementation of the project, based on the results or findings of the research.

Generally, formal research follows a well-structured process and comprises well-defined steps that could be listed as follows:

(i) Specifying the area like 'credit cards' in a bank, and the objective of the study (e.g. improving profitability of credit card business)

(ii) Defining and Refining Problem/Opportunity includes two steps:
- Defining Problem: Translating the objective into a specific problem or issue that becomes the specific topic for the study (e.g. customers' acceptability of the proposal to charge certain fees for certain specified additional facilities on credit cards.)
- Refining/Redefining Problem: This can be done by three methods:
  - Literature Review: Searching relevant literature about the features and schemes of other existing and potential competitors information) on the internet or through other published sources
  - Interviewing relevant people to better understand the problem
  - Group discussion with relevant people.

This would lead to understanding the problem well and refining the defined problem to fulfil the objectives.

(iii) Hypotheses Development: It involves making some intelligent assumptions based on the objective and review of literature. This gives focus/direction to the study leading to insights/information about fulfilling the objectives of the research. Generally, the objectives are converted into hypotheses that can be tested after refinement.

(iv) Preparing Research Design i.e. the blueprint of study (like exploratory)/descriptive/normative/causal, designing questionnaire and the sampling plan (like on-line survey/telephonic survey /personal interview survey), etc.

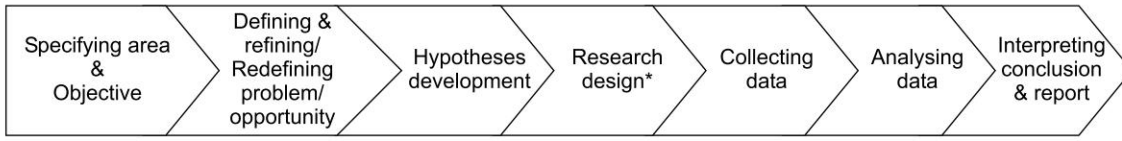
(v) Collecting data e.g. primary data in the form of responses to the questions included in the questionnaire

(vi) Analysing the data (using analytical tools)

(vii) Interpreting the results and drawing conclusions based on data.

These steps involved in a research process, are presented in the following diagram:

We hasten to add that all the technical terms in the depicted process are explained, in detail, in Chapter 3 titled "Research Process". The objective at this stage is only to provide a general understanding of the various steps and components of a research process.



* The blueprint of appropriate study and methodology for collection of data, measurement of data elements/ units, and analysis of data.

Illustration 1.1 Research Process – Explanation of Steps

We may recall the relevance given in the beginning of this chapter. The guide for Mr. Raj Kumar to conduct a study for the above objective, describing all the steps of business research methodology is as follows:

1. **Genesis of the problem (Objective):** Need for regular monthly income to lead a financially secured life for him and his wife.
2. **Managerial Problem:** To find out avenues of investments taking into account the monthly income flow, maturity, liquidity, maturity and risk associated with various investments to take care of monthly income for day-to-day expenses, medical expenses, other family obligations and contingency/emergency situations that might arise.
3. **Defining and Refining the problem**: Translating objective into specific problem **(Research Problem).**

   To design a portfolio of investments that would ensure regular monthly income, medical insurance schemes and allocation of funds for meeting contingency requirements by taking inflation into account. The investment avenues to include bank and post office saving schemes, mutual funds, equity instruments, real estate, etc.

   Later refining the problem by literature review, study of investments schemes, medical insurance schemes, real estate market, equity market, mutual fund schemes, etc. Noting down the associated risks. Further, refining the problem by discussing with people who had similar problem and understanding from their experiences.
4. **Developing Hypotheses:** Making some assumptions which are to be tested, after collection of data which would lead to insights/information about fulfilling the objectives of the research. In this study the hypotheses could be:
   - Property price increases more than the inflation rate
   - Return on selected mutual funds is more than return on fixed deposits
   - Mutual funds have steady growth in the long run, etc.
5. **Research Design:** Developing a detailed plan of what type of data is to be collected, how and when it will be collected, how to strike a balance between the available resources [man, money and minutes (time)] and deciding upon the depth of the study. If any sampling is involved in the study, then the sampling plan, etc. also has to be evolved. In this case, research design will include the decision on the type of data that is required to be collected (primary or secondary) to test the hypotheses which are developed in the previous stage, from when (i.e. time span for collection of data) the data is collected, etc.
6. **Data Collection:** Rates of return, liquidity, maturity, risk, past records of various instruments interviews with individuals.

7. **Data Analysis:** Calculation of fixed income and estimation of income associated with investments in equity, mutual fund and real estate, etc.
8. **Interpreting and Conclusion:** In the form of various options, their liquidity and returns with risk assessment.
9. Final Decision by Raj Kumar

1.8 CRITERIA, CHARACTERISTICS AND CHALLENGES FOR GOOD/IDEAL RESEARCH

There is no quantitative measure for judging the goodness of a research project. However, there are certain qualitative features that are desirable for a research to be labelled as 'ideal'. There are no specific criteria to be satisfied for a research to be labelled as 'ideal'. All one can attempt is to eliminate subjective approaches.

Criteria for Ideal Research In general, following criteria are desirable for a relevant research to qualify for being labelled as 'ideal':

(i) Objective should be clearly defined—both its conceptual framework as well as its practical aspects.
(ii) The objective should be translated into clearly defined problem amenable to be resolved through a research study.
(iii) The research process should be detailed to ensure comprehensiveness.
(iv) The research design should be selected based on sound logic and realistic assumptions, and should be planned to collect reliable data without any prejudice. It should be explained in a manner so that the same or similar study could be conducted by another researcher. The sources of secondary data should be indicated to establish credibility.
(v) Analysis should be appropriate leading to results that would be relevant and adequate for decision-making. However, the conclusions should be derived and based on the data only without reflecting researcher's ideology and philosophy. Statistical terminology used in the design and analysis should be explained so as to appreciate the level of confidence and reliability by the decision-maker.
(vi) All the assumptions made in the research design and analysis should be stated. These have impact on defining scope and limitations of the study.
(vii) The findings and conclusions should be consistent with the results obtained in the analysis without adding extraneous views.
(viii) Scope should be clearly stated without any exaggeration.
(ix) Limitations should be stated without any inhibition or suppression/omission of facts.
(x) Adequate ethical standards, as outlined in Chapter 16, adopted for planning and execution of the study at all levels, should be stated.

Characteristics of Ideal Research

(i) It is conducted in a scientific and systematic manner with sound logic.
(ii) The methodology and steps followed are as per the stated plan.
    (a) The methodology is transparent and 'visible'.

(b) This enables another researcher to repeat or replicate the research without any ambiguity.

(iii) Establishes credibility of research and its conclusions.

(iv) All the assumptions are stated and the sources of data indicated.

It enables readers to get further insights by referring to the sources and assess the applicability of the conclusions.

(v) Scope and limitations are clearly brought out. It enables to draw the conclusions in a realistic manner.

Challenges of Research Research faces challenges at every stage as illustrated below:

- The first challenge is to clearly define the objective that can be translated into hypotheses. Many times, it is quite difficult to find the cause of the problem from a symptom. As a doctor tries to find the root cause of the symptom, a researcher should target to identify the root cause and formulate the right hypotheses.
- The literature review has following challenges:
  - To identify which literature is relevant for the study;
  - How much to review (when to stop and proceed for next stage).
- The process of developing hypotheses faces following challenges:
  - Formulating the right hypothesis that suits the objective.
- At data collection stage, the major challenges are:
  - Availability of adequate, reliable and relevant data;
  - Maintaining the accuracy with limited resources;
  - Not changing the methodology to suit some vested interests or using a new methodology because it is 'comfortable' though not applicable;
  - To have the courage to explain or accept deviations from plan, and own the responsibility.
- Choosing appropriate sampling design.
- At analysis stage, the challenges are:
  - Using right tools and techniques;
  - Considering and testing if assumptions of the tools/techniques are satisfied. For example, if one uses regression analysis, one has to test the assumptions of linearity and normality of the independent and dependent variables (refer to Chapter 10).
- At conclusions/interpretation stage, the challenges are:
  - To avoid temptation to suppress or exaggerate the conclusions for making the study 'sensational';
  - Not to manipulate the results to serve one's purpose;
  - Withstand pressure to change conclusions to suit some vested interests.

Even though the above points are guiding principles for any research, in practice, a research study carried out for or within an organisation is as good as it is understood by decision-makers and implemented to get desired results. The researchers may like to keep this in mind without compromising on ethical standards described in Chapter 16.

1.9 QUALITATIVE REQUIREMENTS FOR RESEARCHERS

It is well appreciated that the quality of any research study depends on the quality of researchers. It is imperative, therefore, that the quality of researchers should be commensurate with the importance or criticality of the research. In general, following traits are desired in researchers:

(i) **Inquisitiveness** – Having a questioning and challenging mind to think beyond what is obvious

(ii) **Identity** – Having the inner urge to establish one's identity

(iii) **Perseverance** – Determination to overcome obstacles which incidentally are an integral part of any pursuit

(iv) **Tolerance for Failures** – Failures, at various steps are an integral part, occurring now and then in most of the projects. One should realise that each failure is a source of learning and a failure should be taken as a stepping stone towards the venture and should be pursued with renewed vigour and determination.

1.10 BUSINESS RESEARCH IN ORGANISATIONS

Some of the perennial issues that are resolved in the organisations, on continuous basis, are:

- What affects the productivity of employees and systems?
- How to design organisational structure and work systems that promote innovation?
- How to measure organisational effectiveness?
- How to channelise organisational conflicts in a positive direction?
- How to exploit the full potential of staff?
- How to attract and retain the professional and skilled talents?
- How to improve decision-making under uncertain environment and volatile changes?

For resolving the above issues, very often the help of business research methodology is taken.

Following are some of the rationale for the top management, to encourage business research activities in their organisations:

(i) To demonstrate the use of business research studies in the decision-making process in the organisation

(ii) To create healthy environment for flourishing of research-oriented thinking at all levels and creating synergies

(iii) To create an environment wherein the ideas and suggestions emanating from lower levels are not suppressed

(iv) To provide the requisite intellectual support and guidance

(v) To recognise and reward at least those who have conducted studies of exceptional quality. (For example, in an organisation, a middle-level officer conducted a study for sharing of revenue between his and some other organisation. The study which was held valid by an independent agency resulted in a claim of several crores of rupees from the other organisation. The Chairman of the organisation, instead of receiving the cheque by courier, sent that officer to go and collect the cheque personally from the Chairman of the other organisation.)

1.11 RELEVANCE OF BUSINESS RESEARCH METHODOLOGY FOR MBA STUDENTS

During the course of studies, MBA students are required to carry out research assignments in various subjects. These are in addition to the research project which they are required to carry out usually towards the end of the course. The knowledge of BRM equips a student to carry out the assignment in a professional manner with better quality of research report.

The subject of BRM is highly useful even after joining service or undertaking any entrepreneurial venture. Even though, we have outlined the relevance and need for research in any organisation at a macro level, it is equally relevant at any level in the organisation at micro level.

The knowledge of BRM helps to approach or accept any problem with confidence and proceed to solve it in a scientific manner. This maximises the chances of success.

Relevance of MBA

While Late Mr. Dhirubhai Ambani was having chat with the MBA students, after delivering convocation address, he was asked as to why doing MBA was necessary when he could set up such a big business empire without MBA degree. Mr. Ambani is reported to have replied, in all humbleness, that he was blessed by the GOD with business acumen and visionary approach, etc., and therefore he could succeed in life. If they (students) were also so blessed, they need not go far MBA; but if they were not, then doing MBA will certainly help them in managing their pursuits better.

1.12 GUIDE TO CONDUCTING RESEARCH PROJECTS BY STUDENTS

Research projects are an integral part of MBA curriculum. In addition to a major project in the last phase of the course, students are required individually as well as in groups to submit research assignments in many papers. Here are some recommended steps to be taken for the completion of assignments (the steps will be fully comprehended after studying all the Chapters from 1 to 4):

(i) Define the problem or topic in a general form;
(ii) Search the literature for studies in this area. Avoid selecting a topic which has been rather over-researched unless the approach is radically different from the earlier studies;
(iii) Discuss with fellow students and faculty before finalising the topic or problem. The scope and limitations of the study to be decided depending on the resources including time and data available for the study;
(iv) Formulate the problem and develop hypothesis based on research question, investigation question and measurement question;
(v) Finalise the research design in consultation with the concerned faculty;
(vi) Before designing the questionnaire (if required) and collection of data, ensure that it is suitable for the statistical analysis to be carried out which itself is decided based on the objective of the study;
(vii) Collect relevant data;
(viii) Carry out appropriate analysis;
(ix) Draw conclusions;

(x) Prepare project report comprising following sections:
  (a) Summary
  (b) Index of contents
  (c) Preamble (State the problem and the reason for selecting the problem—highlight importance of the project)
  (d) Research design including sampling plan
  (e) Collection of data, presentation and analysis
  (f) Conclusions
  (g) Epilogue: Indicate scope and limitations of the study
  (h) Acknowledgements
  (i) References

1.13 RESEARCH IN MANAGEMENT INSTITUTIONS — SOME THOUGHTS

We would like, in all humbleness, to highlight the importance of nurturing academic environment in a management institute as pointed out by a management expert.

"An academic institution can run into decadence in the absence of an academic culture."

The following five areas have been identified for focusing towards attaining academic excellence:
(i) Creating an atmosphere of inclusiveness
(ii) Encourage critical thinking
(iii) Developing reflective practice
(iv) Developing support system
(v) Introducing system of recognition

Incidentally, for encouraging excellence in the research projects, an organisation could institute awards for the 'Best' project in each field viz. Marketing, Finance, Operations, Human Resources Development, etc.

The institute could organise a **'Creativity Festival'** towards the end of the final semester just like cultural festivals. The students should be encouraged to come out with their creative ideas relevant for business entities both global and Indian.

1.14 DISSEMINATION OF RESEARCH

In the present context, Internet is the best media for dissemination of research. The basic philosophy in dissemination is "Give to Get". If no one gives no one gets and if everyone gives everyone gets—that is the philosophy of placing information on the net.

At the level of a management institute, there could be a website of its own containing all the news and information that it would like to share with the world. In addition, the site should contain pages for storing the relevant parts of all the research projects carried out by the students. While full project reports could be made accessible to the Faculty Members, the students could be given access only to the methodology used for collecting and analysing of data and the conclusions.

In this connection, we would like to mention a highly successful experience by the authors.

Dissemination of Project Reports

The research reports by the students of a programme were presented by them in the presence of three outside experts (other than visiting faculty—the idea was that they should not have any bias towards any student). They ranked the first three reports. All the reports were bound together, circulated among all the students, and two copies were kept in the library for perusal by anyone as also the next batch of students. The following caption was printed on first page of the bound volume:

"This volume may be deemed as a garland whose beads were provided by the students. The institute only threaded the beads together".

The student, whose report was adjudged the best, was invited to present the report to the next batch. This gesture by the institute provided a lot of appreciation, not only among the students but also at the place where he was serving. In fact, the letter of invitation to present the report was sent directly to the Chairman of his company.

In addition to successful research carried out by the students, there are certain cases of unsuccessful research projects due to various factors such as the use of inappropriate methodology, non-availability of desired data, collection of data not amenable to relevant software package, etc. Compilation of even such unsuccessful experiences could provide a 'learning' experience for the students. This aspect is illustrated with a research project undertaken by a group of students at a management institute.

Dissemination of Ideas and Expertise

Each one of us generates ideas and develops expertise in our own fields of operation, but most of them are not recorded for dissemination. It is a mute point for us to ponder as to how many people have gone out of the world without recording their ideas and passing on their expertise. Perhaps, the world would have been a far better place to live in if all such ideas and expertise were recorded and made available for the betterment of the mankind.

SUMMARY

The genesis of any research lies in the urge to improve/innovate—either on one's own or necessitated by outside environment. Similar is the case with an organisation. Various motivating factors that cause an individual or an organisation to conduct a research are described, in detail. These factors are initiated on one's own initiative or are caused or sometimes even 'forced' by external environment that is beyond control.

Research is of various types like basic, applied, business, exploratory, descriptive, causal and normative. Various kinds of research are used in studying various aspects of the functioning of a business organisation or an entity.

For any research to be effective and useful, it has to fulfill certain criteria, have certain characteristics and meet certain challenges.

Any research is as good as the researcher(s). Therefore, one has to possess certain qualities that enable him/her to pursue the mission successfully.

DISCUSSION QUESTIONS

1. Discuss the various motivating factors for conducting a research study. Explain with an illustration.
2. Discuss the various types of researches with examples.
3. Describe the various criteria, characteristics and challenges of a good research.
4. Describe the qualitative requirement desired in a researcher.
5. Describe the different steps of research process and illustrate with a suitable example.

Concepts and Tools for Business Research

2

Contents

LEARNING OBJECTIVES

Business Research Methodology (BRM) is an improvement over the earlier methods of conducting research that relied merely on statistical methods and common sense. Over a period of time, the researchers, based on their experience of certain shortcomings and also with an urge to make the research more productive and useful, made optimum use of resources by extending the scope of statistical research. The BRM is also more comprehensive in in-depth situation analysis that is useful for decision-making. For example, a company, apart from analysing varying demands for its various products vis-a-vis that for its competitors, would also like to study behavioural aspects of its employees with respect to latter's performance in duties and perception of customers. It could also have suggestions from the employees about improving the systems and procedures and, expectations and preferences from customers and dealers. Further, the company could explore the ways and means of improving the quality and reducing costs in view of the competition through such studies.

In the process of improving the scope and utility of BRM, several tools and techniques have been developed to improve the relevance and scope of BRM. The purpose of this chapter is to describe some of these, in an exclusive way for their better understanding. While discussing them with the text as and where they are relevant, their application and potential get marginalised as they are useful, in addition to BRM, in other areas that are relevant for a researcher/executive. For example, the concepts and techniques of PERT and CPM are useful in any sphere of managing work, in addition to BRM.

Relevance

Mr. Anand had taken over as CEO of Modern Electronics, only a day back when he was informed that a senior partner of a multinational consultancy firm was to visit him for presenting the report prepared by the consultancy firm about the new MIS proposed by them. The meeting was fixed by the earlier CEO who had to leave immediately for taking up foreign assignment. During the presentation, Mr. Anand, to hide his ignorance, nodded his head, to several suggestions without fully understanding several terminologies used by the team of consultants. He was especially at sea when reference was made to the studies made by the consultants within the Modern Electronics. After the meeting was over, Mr. Anand called the Head of MIS Department to make a presentation for explaining all the terminologies, before he went through the report. In fact, he issued a circular that, in future, before any discussion in the management committee meeting, the concerned executive should come to him and explain the relevant concepts and terminologies, so that he could guide the discussions with full understanding.

Incidentally, it was reported that one chairman confessed that he had sanctioned several projects without knowing what 'IRR'(Internal Rate of Return) actually meant!

2.1 INTRODUCTION

There are several concepts and terminologies that are used in Business Research Methodology (BRM). These exclusive terminologies and concepts have been described in this chapter. This is on the premise that when these are mentioned while discussing the detailed aspects and process of conducting a research study later, these will be easily appreciated and integrated with the corresponding text. In addition to these, there are two tools namely PERT/CPM and creativity that are immensely useful in enhancing the quality of a research project as also its planning and implementing in such a way that the resources, including money and time, are utilised in the most optimum manner.

2.2 RESEARCH METHODOLOGY AND RESEARCH METHODS

Research Methodology covers a wide gamut of research concepts and activities. It is appropriately summarised with the help of following definitions:

"It is the analysis of the **principles of methods**, rules and postulates used in a field of study."

"It encompasses the **systematic study** of methods that are useful in a field of study".

It implies a particular procedure or **set of procedures**.

In fact, Research Methodology could be considered encompassing the following components relating to any research study:

- Collection of theories and practices
- Features of various methods
- Evaluation of various methods

Thus, methodology is more than the set of methods. It is illustrated through some examples below:

- **Teaching/Training Methodology**
  It includes various training methods such as
  - Classroom Lectures

- Case Studies
- Presentations
- Role Play
- Workshops and Seminars

It also includes training evaluation methods.

- **Selection Methodology**

  It includes:
  - Written tests of various types for testing verbal and non-verbal communication
  - Group Discussion
  - Interview
- **Research Methodology**

  It includes:
  - Listing of appropriate Research Methods
  - Logic behind the selection of methods used
  - Context and objective of each method
  - Scope and limitation of each method

As a broader philosophy, research methodology can be viewed as seeking answers to the questions:

- Why?
- How?
- When?
- Who?
- What?
- Where?
- Which?

For example,

In a research study conducted to improve productivity and efficiency of employees of an organisation, the above types of questions, could be worded as follows:

- Why is the study being conducted?
- How to measure the productivity of the employees?
- When the productivity started declining?
- What are the factors that led to declining productivity?
- What are the different types of skills that are desired for different categories of employees?
- Who are the employees needing immediate training?
- Where the training will be conducted?
- Which is the set of persons – internal or external – who will conduct the study and training?

As an illustration, the Director of a management institute desires to know the percentage of MBA students with more than one year of work experience before joining MBA. In such a situation, the methodology of seeking the requisite information could be obtained by any one of the following methods:

- Browsing through the application forms of the students and noting the number of students with the suitable experience.
- Scanning the computerised records of all the students.

- Personally contacting a sample of students, selected randomly, and enquiring about their work experience. Once the percentage of students with more than one year of experience is found, it can be used as an estimate of the percentage of such students in the population i.e. all the MBA students at the institute.

In general, people tend to make little distinction between methodology and methods in the context of research studies; and the two terms are used often without any distinction.

2.2.1 Selection of Appropriate Research Methods

The overall criterion in selecting an appropriate research method is to select the method which entails gathering the most useful information in the most cost-effective manner that leads to the most effective decision-making. The issues that have to be borne in mind for selecting appropriate methods are in the following:

- The information that is needed to facilitate the decision. In this connection, it may be prudent to quote the 'three laws of information' which are described at the end of this section in the box
- The methodology for collecting the desired information
- The methods that ensure credible and authentic information
- The appropriate methods for analysing and presenting the information

In general, a researcher uses a combination of data collection methods, e.g.

- A questionnaire to quickly collect a great deal of information from a lot of people
- Interview to get in-depth information from certain respondents to the questionnaire.

Even in case study method, described in Section 2.7 of this chapter, could then be used for more in-depth analysis of unique and exclusive information.

Incidentally, we would like to mention the three laws of information that have been enunciated which may be taken in a light manner or may be taken seriously to examine whether these are relevant and true in the work environment!

'Three Laws of Information'

- The information that we have is not what we want!
- The information we want is not what we need!!
- The information we need is not available!!!

2.3 CONSTRUCTS AND CONCEPTS

Each and every living being, say human being, and every non-living body or entity like a business organisation has certain characteristics that vary from individual to individual or from one entity to another.

In the case of human beings, there are certain physical or/and quantitative characteristics like height, weight, complexion, etc., there are certain abstract or qualitative characteristics like intelligence, integrity, creativity, etc.

While scientists have evolved scales to measure physical characteristics, psychologists and sociologists have evolved scales for measuring abstract characteristics. Intelligence is one such characteristic that is measured by Intelligence Coefficient (I.Q.). As discussed in Chapter 5 and Chapter

14 (Factor Analysis), scales have also been developed for various unobservable characteristics or **constructs** such as attitudes, sales aptitude, image, personality and patriotism.

Through scaling techniques discussed in Chapter 5, one could measure the above **constructs**.

One may wonder as to how these are related to business entities. One simple logic is that any business is run by the people (employees), for the people (customers) and it is essential to study the above constructs for successful running of a business.

Like human beings, a business organisation has also physical characteristics like: employees, sales, offices, etc. Being physical in nature, these are easily measurable. However, there are certain abstract characteristics (constructs), like reputation, image of the entity, motivation, work culture, commitment, customer's perception and trust. Some other examples are—truth, honesty, intelligence, happiness, motivation, achievement, satisfaction, personality, achievement, ambience, décor, beauty, justice, values, etc.

All these perceptions and feelings of customers are extremely important because they help the company to stay afloat and grow. Therefore, it is essential for the companies to consider the above constructs relating to employees and customers.

A construct is based on **'concepts'**, or can be thought of as a conceptual model that has measurable aspects. This allows a researcher to "measure" the concept and have a common acceptable platform when other researchers do a similar research. For example, measuring advertising effectiveness is a construct, and concepts related would be brand awareness and consumer behaviour. Quality of a TV is a construct, while picture, sound, contrast ratio, etc. would be concepts, that could be measured to define quality. In general, concepts are mental representations and are typically based on experience, and relate to real phenomena (students, customers).

In general, constructs are abstract and concepts are components of constructs and are concrete, and are, therefore, measurable.

Some examples are:

Construct	*Concepts*
Job Competence	Knowledge Skills Attitude
Mental Ability	Memory Analytical Ability Logical Power
Language Skill	Vocabulary Syntax Spelling

Operational Definition of Concept Sometimes, it may be possible to operationally define a concept to make it measurable. For example, prosperity of an individual is a concept. We may define 'income' as a variable measuring prosperity. Further, its operational definition could be 'Annual Income'.

Human Behaviour and Preferences In addition to the above types of variables, certain types of studies involve unobservable variables like intelligence, motivation, creativity, honesty, satisfaction, conformity, etc. Such variables are also called **'concepts'**.

Such traits or characteristics are not directly observed but get reflected in observed human behaviour. For instance, intelligence is reflected by marks obtained in verbal, analytical and logical tests.

Life Skills Assessment of Students

It is reported that there is a proposal that starting this year, CBSE students in India will be subjected to '**Life Skills Assessment**'. This could be treated as '**Construct**' in BRM terminology, and the following skills:

- Thinking Skills
- Social Skills
- Emotional Skills

specified for assessment could be considered as **'Concepts'**. The students will be graded on these skills on a five-point scale viz. **A+, A, B+, B and C.**

The '**Dimensions**' of each of the concept that are proposed to be used for overall assessment in a skill are given in the following Table.

Thinking Skills	*Social Skills*	*Emotional Skills*
Student demonstrates the ability to: • Be original, flexible and imaginative • Raise Questions, identify and analyse problems • Implement a well-thought out decision and take responsibility • Generate new ideas with fluency • Elaborate/Build on new ideas	Student demonstrates the ability to: • Get along with others • Take criticism positively • Listen actively • Communicate using appropriate words, intonation as well as body language	Student demonstrates the ability to: • Identify one's own strengths and weaknesses • Be comfortable with one's own self and overcome weaknesses for positive self concept • Identify causes of and effects of stress on oneself

It is understood that software has been developed for assessing life skills, which will generate a grade based on feedback developed by teachers.

2.4 VARIABLES

A business research study, invariably, involves study of characteristic(s) of an individual/item/unit/entity, etc. These characteristic(s) are represented by variables. As the name suggests, a variable changes values for different individual/item at the same time (e.g. income of individuals for the year 2009–10, prices of stocks on a day) or for the same individual/item at different time (income of an individual, sales of a company).

For example, the income of an individual is a quantitative variable, gender is a qualitative variable.

In a study, data is, generally, collected for relevant variables. These are classified in five categories as follows:

(i) Independent Variable
(ii) Dependent Variable

(iii) Moderating Variable
(iv) Intervening Variables
(v) Extraneous Variables

Brief descriptions of these variables are given below:

(i) Independent Variable

Independent variable, also known as explanatory variable, is a variable which influences or explains the variation in the other variables, under consideration, in the study. The value of this variable can be decided or controlled by the researcher. A researcher can increase or decrease the value of independent variable to assess its impact. For example, the use of fertiliser (independent variable) influences the yield of a crop. The researcher might increase the use of fertiliser by a certain amount to evaluate the corresponding increase in the yield of the crop. As another example, if a company increases the number of its branches, it will also increase the business as well as manpower in the company. The number of branches in this study is termed as independent variable. Similarly, if the increase in advertisement leads to increase in sales, then the advertisement expenses is considered as an independent variable.

In the research relating to studying the relationship between two variables, the independent variable influencing the dependent variable is also called **'causal'**(as it causes change in dependent variable) or **'explanatory'** variable (as it explains the change in the dependent variable).

(ii) Dependent Variable

A dependent variable is one which depends on an independent variable defined in a study. For example, 'Expenditure on R&D' could be taken as independent variable, in pharmaceutical firms, and sales could be termed as dependent variable (on expenditure on R&D). Incidentally, the value of dependent variable is not manipulated or controlled in a research study; it changes due to the change in the independent variable.

(iii) Moderating Variable

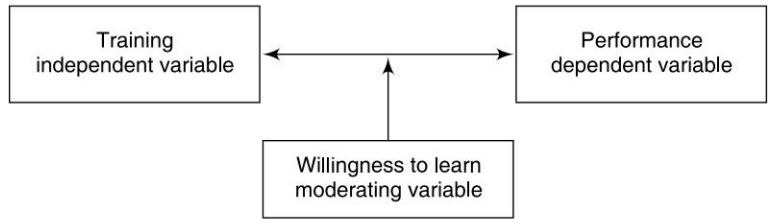
In a study involving an independent variable and a dependent variable, a relationship could be established through a variable. However, we may come across a third variable, which is not an independent variable but forms strong contingent/contextual effect on the relationship of the independent and dependent variables. We explain this with the help of the following example.

Let us consider a relationship between training and performance in an organisation. We may consider training as an independent variable and performance as a dependent variable. The variable 'willingness to learn' is neither an independent variable nor a dependent variable, but has strong effect on the relationship between training and performance. Those employees who are more willing to learn, will grasp the training effectively, and in turn will improve the performance. But if the employees are not willing to learn, even if they are trained, it may not improve their performance.

In this case, 'willingness to learn' becomes the moderating variable.

If we want to study the impact of training on performance, we must consider 'willingness to learn' as moderating variable, and conduct experiment, in such a manner that the moderating variable will not have an adverse effect on the independent–dependent variable relationship.

As another example, we may consider, that rigorous study by a student has impact on his/her marks. Here, the time spent in studying is an independent variable and marks, a dependent variable.



But the grasping power of a student has strong impact on the relationship. Two students having different grasping power may study for the same time, but may not get the same marks. In this case, 'grasping power' becomes the moderating variable. Some more examples are as follows:

Independent	*Dependent*	*Moderating*
Quality of teaching in a classroom	Performance of students in exams	Motivation for sitting in competitive exams
Percentage of discount in a store	Sales	Aptitude of Sales Personnel

(iv) Intervening Variables

In a study involving independent and dependent variables, there could be a variable/factor which might affect the dependent variable, but it cannot be directly observed or measured. For example, the sales of a retail store might increase with increasing discounts (e.g. 2%, 3%.... 5%, etc.), and a relationship could be established. Suppose, a scheme is introduced that monthly prizes will be given to randomly selected customers, this might increase the sales of a store but it cannot be measured; only its impact can be observed. Thus, monthly prizes scheme is an intervening factor. If the total amount of prizes offered (say Rs.10 lakh) is indicated, and the amount is a variable, then this could be considered as intervening variable, and one could study its impact.

(v) Extraneous Variables

Extraneous variable is one that is outside or external to the situation under study, and its impact on dependent variable is beyond the scope of the study. For example, the family income of students could be taken as extraneous variable while studying performance of MBA students. The value of extraneous variable may be 'controlled' to remove/neutralise its impact. For example, if we wish to compare the performance of MBA students with commerce and engineering background, we may include those students of both streams in the study who have almost the same financial background.

In general, the likely impact of an independent variable which is extraneous to the study could be eliminated by choosing those experimental units which are as homogeneous as possible. Another method to deal with the situation is to select the experimental units in a random manner so that the impact of extraneous variable is neutralised. For example, in the previous example of MBA students, the students from both the streams could be selected at random so that the impact of financial backgrounds is evened out.

In addition to the above variables, there is another variable called **control** variable. This variable is held constant in order to assess or clarify the relationship between two other variables.

Incidentally, the 'controlled variable' in Design of Experiments, has a different connotation. It is an alternative term for independent variable.

The various aspects like measurement and scaling relating to variables are described in detail in Chapter 5.

2.5 DEDUCTIVE AND INDUCTIVE LOGIC

The concepts of deductive and inductive logic are closely imbedded in a research study. It is, therefore, necessary to have thorough understanding of these basic concepts.

The words '**Deduction**', '**Deductive Logic'** and '**Deductive Approach'** imply the same and are used interchangeably. Similarly, the words '**Induction**', '**Inductive Logic'** and '**Inductive Approach'** imply the same and are used interchangeably.

2.5.1 Deduction

The basic concept in deduction is from

'Many to One'

or

'Population to Sample'

In this type of logic, we are given information about a population, and we deduce the information about a sample or just one unit.

A few examples/illustrations of deduction are given in Table 2.1.

Table 2.1 Examples/Illustrations of Deduction

(i)	Premise	Most MBA graduates are extremely intelligent	Validity of Deduction
(ii)	Given Information	Simran is an MBA student	
	Deduction/Conclusion	Therefore, Simran is extremely intelligent	Not valid. The premise is not universally true. There could be some MBAs who may not be so.
	Premise	All the MBA graduates recruited in a company through a rigorous selection process have proved to be very innovative and effective	
	Given Information	Sajay, an MBA student has been recruited in the company	
	Deduction/Conclusion	Sajay will prove to be very innovative and effective	Valid. The premise has been true for the organisation.
(iii)	Premise	Due to strict quality control through Six Sigma process, all the cars manufactured by 'Reliable' Company do not develop any defect for at least three years	
	Given Information	A given car has been manufactured by 'Reliable' company	
	Deduction/Conclusion	The car will perform hassle free for three years	Valid. From earlier reputation of the company followed by proven record in the past.

(*Contd*)

(*Contd*)

(iv) Premise (Law)	According to Newton's law of gravitation, any object thrown up will come down	
Given Information	Saluni throws a ball in the air	
Deduction/Conclusion	The ball will come down	Valid. Proven law.
(v) Premise	For all companies manufacturing retail products, advertisements do have favourable impact on their sales	
Given Information	ALPHA is a company manufacturing mobile phones	
Deduction/Conclusion	Therefore, the advertisement by 'ALPHA' company will improve its sales	Valid. Proven past record for all companies.

From the above examples, we may note that, Deduction reasoning works from the **'General to the Specific'**. It may also be termed as **'top-down'** approach.

It may be noted that it is analogous to **'Brand Image'** wherein, conclusions are drawn just by the name of the brand. That is how the brand image works. We infer about the quality of an individual/product/service depending on the image of the company where an individual works or the name of the company, where the product was produced or the name of the company which provides the service. In the case of an individual, we may draw inference about him/her even from country or race to which he/she belongs.

2.5.2 Induction

The basic concept of induction is from:

One to Many

or

Sample to Population

A few examples/illustrations of induction are given in Table 2.2.

Table 2.2 Examples/Illustrations of Induction

(i) Observation	The average work experience of a sample of MBA students at a management institute is 18 months.
Induction/Conclusion	We induce that the average work experience of all the MBA students at the institute is 18 months.
(ii) Observation	While cooking rice, a cook picks one piece of rice and finds if it is cooked
Induction/Conclusion	All the rice pieces are cooked.
(iii) Observation	One biscuit from a packet is stale
Induction/Conclusion	All the biscuits in the packet are stale
(iv) Observation	In GMAT 2009, the average score of Indian students was 562 as compared to 539 for the world

(*Contd*)

(*Contd*)

Induction/Conclusion	Indians are very competitive and hardworking It may be noted that, this inductive conclusion is not valid for the entire population of Indians. If at all one may say that "Indian students appearing in GMAT are very competitive and hardworking"

Induction, in simple terms, could also refer to '**Generalisation**' from what we observe or know. Induction involves reasoning about the future from the past, but in a broad sense, it involves reaching conclusions about unobserved things on the basis of what is actually observed.

Induction starts from **'Specific'** observations or set of observations to **Generalised** Theory or Law. It could be termed as **'bottom-up'** approach. A classical example is when Newton observed an apple falling from a tree—he generalised it to 'Gravitational Theory'.

Induction can also be considered as divergent thinking. It is used when nothing or little is known, and we wish to expand our knowledge. The following example illustrates such a process:

- A company is approached by a new management institute, to select its graduates.
- The company, through a written test and interview, selects 5 students.
- The company finds those students rendering excellent service to the company.
- The company selects 5 students again in the next year, by the same process.
- Once the company repeats such selection for some years, it forms an opinion or rather induces that the students of the institute, selected through the same process, would prove to be highly useful for the company. In fact, in due course of time, the company might do away with the written test, and select students just on the basis of interview.
- Based on the experience of this company, the other companies also start recruiting from the institute, and that is how, the image (brand) of the institute is created.

2.5.3 Deduction versus Induction

Inductive research is a model in which theories are developed from specific observations.

In deductive research, the specific expectations of a hypothesis are developed on the basis of general principle: we start from existing theory and then find its proof.

For instance, in Chennai, a social researcher observes that in a restaurant people from north India prefer to take tea over coffee. He extrapolates or uses inductive logic to conclude that all north Indians prefer tea over coffee.

In deductive logic, a researcher starts from the hypothesis that north Indians prefer tea over coffee, and then starts collecting observations to prove or disprove this hypothesis.

Inductive reasoning is open-ended and exploratory especially in the beginning. Deductive reasoning is specific in nature and is concerned with testing or confirming hypothesis.

In fact, all the researches that have taken place in various fields is a continuous cycling of induction and deduction approaches.

The difference in deduction and induction may also be appreciated by referring to statistical inference, described in detail in Chapter 11 wherein, estimation may be taken as inductive and testing of hypothesis may be considered deductive.

The explanation for treating deduction as 'Top-down' approach and induction as 'Bottom-up' approach, may be depicted as follows:

↓ Deduction	↑ Induction
Theory	Theory
Hypothesis	Hypothesis
Observation	Pattern
Confirmation	Observation

This has been explained by Illustrations 2.1 and 2.2.

These can be better appreciated after studying Chapters 10 and 11.

Illustration 2.1

Deduction

Theory : Advertising increases Sales

Hypotheses (Linear) : Sales = a + b Advertising

Data for a period of 'n' years

Year	*Sales*	*Advertising Expenses*
1	S_1	A_1
2	S_2	A_2
:	:	:
n	S_n	A_n

The regression equation can be derived as follows (Chapter 10)

$$\text{Sales} = (\hat{a}) + (\hat{b}) \text{ Advertising}$$

where $\hat{a}$ & $\hat{b}$ are estimates of a and b.

Test the significance of the regression equation: Here, we have started with a theory that 'Advertising increases sales' and collected sample to prove this hypothesis. If significant – theory is proved or confirmed.

Induction

Observations:

Year	*Sales*	*Advertising Expenses*
1	S_1	A_1
2	S_2	A_2
:	:	:
n	S_n	A_n

Hypothesis : Linear Relationship as evidenced by the above graph

Theory : Advertising increases sales

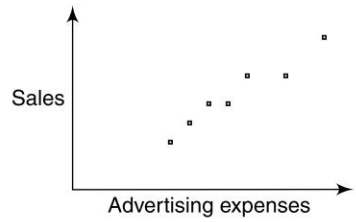


Illustration 2.2

Deduction

Daily Rates of Return on Sensex

Theory	Rates of Return follow symmetrical distribution with reference to some mean
Hypothesis	The rates of return follow normal distribution (Hypothesis is the statement which can be tested)
Observations	Collect **Data**
Confirmation	Fit a normal distribution to the observed data, calculate expected frequencies for various intervals, within the range of rate of return, confirm the hypothesis by chi-square test of significance, described in Chapter 11.

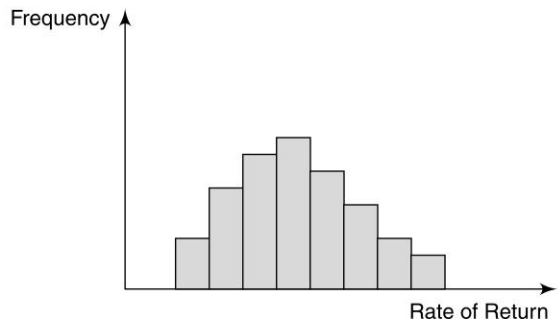
$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

Induction

Example:

For studying daily rates of return on Sensex, we may collect observations, say, for a year.

Making a histogram (discussed in Chapter 8) of these returns, we may observe a pattern as follows:



From the above pattern, we may develop the hypothesis that the distribution of rates of return follows normal distribution.

From this hypothesis, we may develop a theory that rates of return on Sensex follow a symmetrical pattern with respect to some mean. So, here we are going from observations in terms of histogram to come to one theory; therefore, it is an induction.

2.6 QUANTITATIVE AND QUALITATIVE RESEARCH

Business research is usually classified in two categories viz. **'Quantitative'** and **'Qualitative'.** We discuss these two types of researches in three different sections. The first one deals with Quantitative research, the second one describes Qualitative research, and the third section brings out the comparison of these two researches. However, we may add that the two types of researches are complementary—one supplementing the other. In most of the research studies, both the researches are conducted to draw conclusions of practical significance. An illustration is as follows:

The CEO of a company manufacturing BEST Vision brand of televisions sponsored a study to analyse the growth of various models vis-a-vis the growth of the other competitors. Such study is quantitative in nature, and would definitely give an idea of the changing market share—both in total number as also in various models. However, to draw a strategy for future, he would also like to have an idea of the following:

- Changing preference for various models by the customers—existing as well as potential, along with the possible reasons as also feedback and suggestions for sustained/improved growth
- Suggestions from the sales and service personnel for incorporating the features that could increase the demand for the company's televisions

For these purposes, the CEO would have to sponsor a qualitative research study.

Qualitative research is not a substitute for quantitative research; it is only to supplement the quantitative research. In fact, it is the combination of both qualitative and quantitative researches that are desirable in most of the situations.

If one were to study the past:

- What happened (Quantitative)
- Why and How it happened (Qualitative)

If one were to study the future

- What will happen (Quantitative)
- What should or ought to happen (Quantitative)
- How to make it happen (Qualitative)

A pragmatic approach to conduct a research study is to start the research as a qualitative, and conclude it as a quantitative. The qualitative method is used to explore and identify the ideas, develop hypotheses and variables, of interest to the researcher. This would be done through direct observations, interpreting or focus groups, described in Chapter 6. The concepts derived from the qualitative portion of the study are then studied through the use of quantitative methods. Further, the hypothesis tested through quantitative research gets more credibility.

2.6.1 Quantitative Research

We study such research in Statistics, which we may recall is based on quantitative or qualitative data relating to measurement, counting and frequency of occurrences. Such research originated from economic theory, and agricultural and medical experiments. Now, it is considered indispensible in any decision-making process that is based on facts and figures, especially in an uncertain environment. In business research, quantification *inter alia* relates to measuring consumer behaviour, knowledge and attitudes. In fact, quantitative research provides quite accurate and reliable measures—through well-developed statistical methods. For a detailed account of the basics and use of statistical methods, the reader may refer to the book titled **"Statistics for Management"** by the same authors,

published by Tata McGraw-Hill. However, the salient aspects of statistics relevant for BRM are given in Chapter 8 to Chapter 13.

Quantitative research answers the questions that need a quantitative answer like:

- 'How Much' (Measurement: How much growth has taken place in the retail outlets during the year 2010)
- 'How Many' (Counting: How many MBA students in an institute have work experience more than 2 years)
- 'How Often' (Frequency of Occurrence: How many times the production process went out of control during January 2010?)

It is also useful for testing any assumption which is

- Descriptive (Average increase in the equity prices of SENSEX companies during 2009 is 22%), or
- Causative (Coaching helps in better performance in CAT), or
- Associative (The salaries offered in campus placement and the final grade in MBA are correlated).

As a matter of recapitulation, all the topics listed next, that we study in statistics are studied through Quantitative Research.

- Measures of Location and Dispersion (Mean, Variance, etc.)
- Statistical Distributions (Normal, Binomial, etc.)
- Correlation and Regression Analysis (Correlation Coefficient and Regression Equation, etc.)
- Statistical Inference—Estimation and Testing of Hypothesis
- ANOVA and Design of Experiments
- Time Series Analysis
- Forecasting
- Decision Theory

Incidentally, it may be mentioned that the discussions in this book are predominantly on Quantitative Research Methodology, which generally connotes Research Methodology.

2.6.2 Qualitative Research

In simplistic terms, qualitative research may be defined as any research that is not amenable to be conducted by merely using quantitative research.

Qualitative research relates to inquiry/study, investigation of human behaviour and the explanation for the same. It is, therefore, applicable in investigations encompassing all fields. It proposes field-based theories i.e. theories that emerge through induction, based on observations in the field studies.

Finally, qualitative research is formally defined as based on:

"Researcher immersion in the phenomenon to be studied, gathering data which provide a detailed description of events, situations and interaction between people and things, providing depth and detail."

Qualitative Research can give some indication as to 'Why', 'How' or 'What' of the happenings of a phenomenon.

For example,

Why	Why the market share of the company has been declining?
How	How to meet the competition due to the opening of another store in the neighbourhood?
What	What are the factors that motivate the employees in an organisation?

Qualitative research was first used in social sciences, and later on its use spread to other fields like education, psychology, communication and even management studies, especially those relating to consumer behaviour, for new products and services. It investigates the 'why' and 'how' of decision-making in addition to what, where and when. Therefore, qualitative research relies on smaller but focused samples rather than large and random samples.

The Qualitative Research is broadly classified into the studies relating to:

(i) Reaction and Feelings
(ii) Learning (improving attitudes, knowledge)
(iii) Changes in Skills (improving effectiveness of lectures/workshop, etc.)
(iv) Effectiveness (improved performance attributed to improved behaviours)
(v) Response from existing and potential customers

Some of the situations needing qualitative research are:

- CEO of a company wishes to assess the reaction and feelings of employees for the health care scheme announced by the company
- Director (HRD) of a company wishes to assess the impact of the two-week training for middle-level executives, conducted by a consulting agency
- A consultant during his assignment relating to organisation structure, in a company, observes the need for improving behavioural skills of young MBA executives, and communication skills of old executives
- The Board of a company wants to have an assessment of the impact of the endorsement of its main product by a celebrity.
- A car manufacturer wishes to revise the prices of its different models based on the likely response by the new and potential buyers

2.6.2.1 Sources of Data for Qualitative Research Following are sources of data for conducting a qualitative research:

- People—individual or groups.
- Organisations—an individual or a representative of a group of organisations
- External environmental factors like regulatory, economic, social or technological
- Internal environment in an organisation encompassing work culture, incentive systems for hiring and retention, etc.
- Texts (Published or Internet)

2.6.2.2 Terminologies Used in Qualitative Research Qualitative research uses several terminologies which are specific to such research. Since the list is quite exhaustive, we have included it in the glossary.

2.6.2.3 Applications of Qualitative Research The following are some of the applications of Qualitative Research:

(i) Used for exploratory purposes or to investigate 'how' and 'why' of happening of a situation.

(ii) Used for pilot testing to design quantitative surveys of large scale.
(iii) Used for more diversity in responses as also flexibility to adapt to new developments or issues during the research process.
(iv) Used for better and deeper understanding of the situation/phenomenon. For example, how a customer goes about selecting a mobile phone and a service provider.
(v) Often used for policy and programme evaluation research since it can answer certain important questions more efficiently and effectively.
(vi) When a set of qualitative data is difficult to graph or display, pictorially. However, it categorises data into patterns which provide the basis for analysing and drawing conclusions.
(vii) Used for explaining and interpreting quantitative results like in Factor Analysis described in Chapter 14.

Some of the above applications are reflected in the following case:

CASE 2.1 QUALITATIVE RESEARCH FOR BUSINESS GROWTH

This case relates to the situation when growth in bank deposits was related to several non-financial factors like location (including building in which located), ambience in the branch, customer service, etc.

A bank hypothsesised that the growth in deposits, at its branches in posh residential areas, would be more than its branches in ordinary residential areas. However, the use of quantitative deductive approach at 10 branches in posh areas revealed that it was not so. For understanding the reasons for this phenomenon, a qualitative research was conducted at a couple of branches. The research involved observation of customers' behaviour, and interaction with them as well as the staff of the branches. It revealed that the customers were not too happy with the internal ambience, old building which housed these branches, arrangement for sitting, design of counters, customer service, etc. The residents of the area were, therefore, shifting or opening new accounts in the branches with a 'modern' look—pleasant ambience inside the branches, faster customer service with latest technology, ATM in the branch, etc. It may be added that while quantitative research found unsatisfactory growth in deposits, the qualitative research revealed or explained the reasons for the same.

The following case illustrates yet another situation involving the use of qualitative research:

CASE 2.2 QUALITATIVE RESEARCH FOR SETTLEMENT OF GRIEVANCES

In the departmental canteen of an organisation, owned and operated by the organisation itself, there were about 100 staff members including cooks, waiters, etc. There was perennial discontentment among the staff of the organisation about the quality of food and service. It went on for a couple of years, and the management took notice of it only when one day the staff of the organisation refused to go to the canteen for lunch. A quick review of the situation led the management to believe that the salaries of canteen employees were not being raised at par with the salaries of other staff. Analysis of the salary structures of the clerical/secretarial staff and the canteen staff revealed that the gap between their average salaries had increased from 10% to 20% over a period

of 3 years. Accordingly, the salaries of canteen employees were suitably revised. However, there was no perceptible change in the quality of food and service. The General Manager in charge of the administration then had personal interaction with the representatives of the canteen staff, and called a meeting of the canteen employees in the dining hall. He was profusely welcomed and offered a bouquet by the canteen staff. The General Manager asked for their free and frank comments and suggestions. The gist of the deliberations was that they perceived differential attitude of management with the canteen employees and the other staff of the organisation. For instance,

- They were not invited to any function organised by the management
- Their meritorious children were not awarded scholarships like children of other staff of the organisation
- The canteen was not air-conditioned like the office

They also gave some suggestions for bringing about a better menu without much cost implications.

The General Manager agreed to these suggestions. Later on, it was found that both the categories of the staff had developed cordial relations.

2.6.2.4 Use of Qualitative Research in Exploratory Research Qualitative research is most suited if there is not much information available about the phenomena/topic of interest. If reliable information is available, then one can easily form theoretical framework, develop hypothesis, collect data using questionnaire, analyse data and arrive at conclusions; which incidentally form the steps for a quantitative study. To form the theoretical framework, one needs to know the phenomena in depth. If one does not understand the phenomena, one can use qualitative approach in the form of open-ended questions, in-depth interviews, focus group discussions, etc., described in Chapter 6 to explore the phenomena in detail, and understand from the respondents. In fact, the qualitative research helps in fulfilling the basic objectives of exploratory research viz. diagnosis, developing alternatives and discovery of new ideas.

For example, if one wants to study the phenomena of customer dissatisfaction with a mobile service provider, one may not have any information about the same. Hence, the service provider may want to explore the reasons for dissatisfaction from a small group of its customers, by interviewing them or forming a small focused group discussion and noting down the views. This would facilitate to conduct a full scale study by suitably designing a questionnaire.

2.6.2.5 Qualitative Research—Data Collection, Analysis and Validation We describe, in this section, the three important aspects of qualitative research. These are: Data Collection, Data Analysis and Data Validation.

Data Collection:

The most important aspect of data to be collected in qualitative research is that the methodology is not fixed; it may keep on changing as the data collection process continues.

The data is collected through the following methods:

- Participant Observation
  Here, the researcher becomes a part of the group or plays the role of a participant
- Non-Participant Observations
  Here, the researcher does not become a part of the group
- Field notes—i.e. notes while observing a phenomenon, event, action, etc.

- Structured interviews—as per fixed set of points/questions, etc.
- Unstructured interviews—flexible set of points/questions which the researcher could vary from participant to participant as per the responses
- Focus Group technique (described in detail in Section 6.3.4)—It involves a moderator facilitating discussion within or interview with a small group of selected individuals who are well-informed or concerned with a particular topic. This system of getting data/information is quite useful in market research specially getting opinion/views about testing new initiative in the form of a product, service or a system.

Data Analysis:

The data collected by qualitative research may be referred as Qualitative Data. It is usually collected through unstructured/semi-structured interviews, focus group discussions (in-depth interviews), as mentioned above and described in detail in Chapter 6.

The most common analysis of qualitative data is observer's impression. First the observers examine the data, form an impression and report their impressions in a structured and sometimes quantitative form. These impressions can be the final conclusion of the analysis or some quantitative characteristics of the data to be further analysed using quantitative methods.

The analysis of qualitative data can be quite challenging. Since the data is unstructured, one is likely to miss out important factors. For avoiding this, there are software available for analysing the qualitative data. Computer Assisted Qualitative Data Analysis Software (CAQDAS) is a computer software designed to facilitate the analysis of qualitative data collected from groups, interviews or other qualitative textual sources. It is widely used in academic qualitative research but has not generally been adopted by business researchers.

Data Validation:

An issue which assumes importance is the validation i.e. credibility and reliability of data. One general way of establishing credibility is that the researcher should code the data, discern and document the whole methodology in a consistent and reliable way. The validation process continues along with data collection. In fact, the collection and validation process are done simultaneously.

2.6.2.6 Limitations of Qualitative Research

(i) In qualitative research, the sample sizes are generally smaller. Further, the sampling resorted is purposive or judgmental and not random.

(ii) The role of the researcher is quite critical and quite sensitive to the quality of a research study. It is utmost important that the researcher is competent and absolutely 'neutral' or 'unbiased'. This can be reasonably ensured through
   - Selection of the researcher who is known to be '**neutral**' to the topic for study, and, impressing upon him/her the importance of his/her role in the quality of the entire research process
   - Imparting the requisite training and guidance to the researcher
   - Testing the researcher's competence and credibility through a pilot study in a simulated environment and matching his/her findings with the known facts

(iii) The data collection process has far greater impact on the results as compared to a quantitative research. It is, therefore, imperative that the whole methodology of collecting and analysing data is well-documented to avoid ambiguity for the researcher and ensure credibility for users of research.

(iv) Though the sample size is smaller, the technique of data collection is more elaborate leading to the researcher's personal involvement, generally requiring more time and money.

2.6.3 Quantitative and Qualitative Research in Social Science

Social Science uses both 'Quantitative' and 'Qualitative' methods. Quantitative methods attempt to capture social behaviour or phenomenon, collect numerical data and focus on the links among a smaller set of attributes across many cases/individuals. Qualitative methods emphasise on personal experiences and interpretations, and are more concerned with understanding the meaning of social phenomenon and focus on links among a larger number of attributes across relatively few cases.

While quantitative and qualitative researchers/approaches are different in many aspects, both involve a systematic interaction between theories and data or between ideas and evidences. Ideas make social researchers draw conclusions from evidences and use evidences to extend, revise and test ideas. Social research, thus, attempts to create or validate theories through data collection and analysis, with the ultimate objective of exploring, describing and explaining.

2.6.4 Quantitative and Qualitative Research in Business Research—An Illustration

The relevance of both quantitative and qualitative researches, as an integral part of business research, may be explained by the following situation wherein, an entrepreneur wants to set up a plant to manufacture a product, may be explained by the following situation:

The entrepreneur would like to have an idea of the

- Size of the market
- Total revenue of all the existing companies manufacturing the product
- Average growth rate of the existing companies, in the past
- Futuristic demand of the product

This would require the use of quantitative research.

However, for ensuring success of the venture, the entrepreneur would also like to have an idea of the

- Customers' perception of the product and its growth
- Fulfilment of the needs/expectations from the product—scope for improvement
- Emerging technological advancement related to the product or production process
- Economic/social environment or statutory regulations, etc. which could impact the demand of the product
- Appropriate marketing strategy

All these would require the use of qualitative research.

2.6.5 Comparison of Quantitative and Qualitative Research

Both quantitative and qualitative researches have their own advantages and disadvantages depending on several factors like

- Objective of the research
- Problem or issue to be resolved
- Context in which it is used
- Resources made available
- Accuracy desired and margin of error to be tolerated (it is difficult to specify these aspects in qualitative research but some idea can be provided).

Table 2.3 contains salient features of comparison between the two types of research.

Table 2.3 Comparison between Quantitative and Qualitative Research

	Quantitative	*Qualitative*
Objective	Describe, explain and predict	Understand and interpret
Researcher's Involvement	Limited: Controlled to prevent bias	High: Researcher is participant or catalyst
Research Purpose	Describe or Predict: Build and test theory	In-depth understanding: Theory building
Research Design	Rigid design and framework: • Determined before commencing the project • Uses single method or mixed methods • Consistency is critical—involves either a cross-sectional or a longitudinal approach	Flexible design or framework: • May evolve or adjust during the course of the project • Often uses multiple methods simultaneously or sequentially • Consistency is not expected—involves longitudinal approach
Desired Sample Design	Probability	Non-probability: Purposive
Sample Size	Large	Small
Participant's Preparation	No preparation desired to avoid biasing the participant	Pre-tasking is common
Data Type and Preparation	Verbal descriptions reduced to numerical codes for computerised analysis	Verbal or pictorial descriptions reduced to verbal codes (sometimes with computer assistance)
Data Analysis	• Computerised analysis—statistical and mathematical methods dominate • Analysis may be ongoing during the project • Maintains clear distinction between facts and judgements	• Human analysis following computer or human coding; primarily non-quantitative • Forces researcher to see the contextual framework of the phenomenon being measured • Distinction between facts and judgements less clear • Always ongoing during the project
Insights and Meaning	Limited by the opportunity to probe respondents and the quality of the original data collection instrument; insights follow data collection and data entry, with limited ability to re-interview participants	Deeper level of understanding is the norm; determined by type and quantity of free-response questions. Researcher's participation in data collection allows insights to form and be tested during the process
Feedback Turnaround	• Larger sample sizes lengthen data collection; Internet methodologies are shortening turnaround but inappropriate for many studies • Insight development follows data collection and entry, lengthening research process; interviewing software permits some tallying of responses as data collection progresses	• Smaller sample sizes make data collection faster for shorter possible turnaround • Insights are developed as the research progresses, shortening data analysis

(*Contd*)

(*Contd*)

Conclusions	Generalising from sample to population is a valid subject to well-specified accuracy and judgemental error.	Generalisation from small-sized research has to be done with caution because of the subjectivity involved in data collection

This table is based on the Table 8.2 of the book titled "Business Research Methods" by David R Coopers and Pamela S Schindler, published by Tata McGraw-Hill.

2.7 CASE STUDY METHOD OF RESEARCH

A case study is one of the ways of conducting business research. Rather than using samples and examine limited number of variables, a case study involves in-depth study of several variables related to a single unit, instance, event, etc.

The use of word 'case' started in medical research. It was referred to a single individual who had some unique symptoms or who was subjected to some special treatment. The word 'case history' refers to the past medical record of a patient. It contains, *inter alia*, the symptoms and treatment given to the patient, in the past. Such cases were found useful in medical education and research.

Subsequently, the word 'case' was referred to legal suits which were of a typical nature—either the event/situation, suits itself, the arguments or the judgement.

A case study helps in understanding **'Why?'** and **'How?'** of the situations under study. It is a qualitative study but could also be quantitative or a combination of both.

A case study implies in-depth study about an individual unit of a population. For example, we might select

- a branch of a commercial bank
- a retail outlet out of several outlets of a company
- one management institute out of a group of several management institutes
- the mutual fund which declared maximum dividend on tax-save units, during the year 2009.
- a company that is declared the 'best company to work for'

The unit is selected for case study based on one or more of the following:

- exemplary performance or informal feedback
- very poor performance or informal feedback
- special features introduced at the unit
- experimenting new ideas/innovations.

In social and behavioural sciences, case studies have been used to understand human perceptions, behaviour, attitude, etc.

In business research, case studies have been conducted for product development, promotions and sales. Understanding customers' preferences is one of the major areas for study.

2.7.1 Objectives and Advantages

The main objectives of a case study are to:

- Build upon existing theory and practice
- Generate new theory and practice
- Develop certain 'learning' points which could be used for policy formulation for all units
- Answer the management/research questions, considered relevant
- Multiply the success achieved by the innovative strategies in one group/office to other groups/offices

- Understand and explain the causes that could have brought about an adverse situation in one group/office to avoid similar situation in other groups/offices
- Test new products, services, systems or strategies in a couple of groups/offices, and if successful, use in others.

The advantages of a case study are:

- It is a cost-effective methodology for learning through a real life set-up without the need for simulation
- It helps in better understanding of a complex or an unusual situation/phenomenon or issue
- It is an effective tool and mechanism for generating lively discussion and subsequent ideas to improve upon the offered suggestions or practical system.

For deriving the previously discussed advantages, a case study is developed by collecting information through observation, recording, interacting and suitably analysing it to pose issues for discussion. The details of solutions, if any, not included in the case, are sought from the selected group. All the suggestions/solutions are thoroughly discussed to arrive at a consensus. Such consensus, along with the earlier solution(s), if any, is deliberated further to reach to a much broader consensus. Deliberation and consideration on multiple solution help to analyse the situation from different perspective and hence, it can be used for either 'learning' or 'testing' in future.

Sometimes, one may conduct a number of case studies to have a better understanding of the issues involved by covering a number of units. For example, the branches of a commercial bank can be divided in four categories viz. metropolitan, urban, semi-urban and rural. These branches are quite different from one another. In such a situation, if the objective is to study behaviour of defaulting borrowers, one may select one branch from each of the four categories. Case studies at these branches might reveal some aspects which could be valid for the bank as a whole.

Similarly, if the objective of the bank was to study the pattern of car loans in Metro cities, one may select a branch each from the four Metros, and conduct four case studies at these branches.

We now describe two live case studies that were conducted primarily to bring out the relatively new concepts and practices introduced by some individuals who ultimately reached the top level in their careers.

CASE 2.3 A CASE STUDY ON MOTIVATING STAFF AND INVOLVING CUSTOMERS

A commercial bank's branch in a prosperous area was not doing as well as anticipated by the management of the bank. The management decided to post a relatively young but dynamic executive as the manager of that branch. The manager used the posting as a challenge to try some unconventional progressive ideas which, according to him, were bound to bring out favourable results. And, he was right. There was a remarkable increase in the business at the branch. In 2 years, the branch's business increased from 7 crore to 20 crore (in 1980s). That was when it was decided to undertake a study at the branch.

It was noted that the manager started the endeavour by introducing flexible timings for staff, according to their peculiar needs. It was used by the branch manager to open the branch 30 minutes earlier than the scheduled time and also close the branch 30 minutes later than the scheduled time. However, the branch manager was available to the customers 30 minutes before the opening as also after the scheduled closure of branch. This move was highly appreciated by the customers

and attracted new customers. The staff was highly appreciative of the branch manager as they felt here was one person who was sensitive to their personal needs, and was leading from the front.

The study involved interaction with the staff and customers as also the analysis of business records at the branch to ensure that the bank's interest was well-protected. The study revealed the following strategies used by the branch manager to develop excellent business at the branch:

- Set up consultative committee comprising high-valued customers
- Conduct weekly meetings of staff
- Conduct monthly meetings of customers
- Establish personal rapport with the staff to ensure support and commitment of staff
- Flexible timing at branch-led by example
- Communication channel for customers
- Meeting genuine and urgent demands of customers without sacrificing the interest of the bank
- By goodwill measures and personal interaction like birthday celebration at the branch, and
- Inviting families of staff for tea at residence

The following illustration is one more case where an unconventional approach solved the problem which was in the process of being resolved through conventional approach.

CASE 2.4 A CASE STUDY ON CUSTOMER SERVICE

A conventional research was on to increase the number of counters at a bank branch to meet the tremendous rush of customers at the branch just for about 45 minutes. The branch was situated just outside the premises of a big factory employing about 3000 employees. The salary of these employees was credited in their accounts at the branch. All these employees had just half an hour of time during lunch period to transact business at the branch. Obviously, two counters at the branch were not sufficient to serve more than 100 customers. (The number was much more in the first week).

The new branch manager adopted a novel approach to solve the problem. He took advantage of the fact that the branch had only about 1000 customers other than the 3000 accounts of the factory employees. Thus, the work load on the branch was very limited except during the lunch time of factory workers. The manager thought of 'taking the branch' to the workplaces of the employees. He deputed one officer and one clerk to visit the various sections of the factory in the morning to collect various instruments, such as withdrawal slips, cheques for deposit, requests for draft, etc., duly signed by them.

Before the lunch time, the bank staff kept the cash ready in envelopes. The employees would go to the branch, sign on the back of the withdrawal slip/cheque and take their envelopes. Similarly, drafts and other items like cheque book, statement of accounts, etc. were given to them as per their request.

2.8 GOAL SETTING FOR A RESEARCH PROJECT

In a wider perspective, every organisation sets up certain goals for itself as also for its various groups/offices. In general, the goal setting exercise comprises the answers to the following questions:

- Where are we?
- How we reached here?
- Where will we reach?
- Where we want to reach? Goal?
- How to reach there?

The term 'Goal Setting' in business research methodology, has a different connotation as it relates to conducting a research project. A research project is usually a team effort. It is, therefore, desirable, or rather necessary that the team should be made aware of the overall objective of the project along with the clarity about the role of each member of the team including the co-ordination among them. All this should be decided by discussing the plan of the study with the entire team members thus ensuring their commitment to the project and their '**ownership**' of the project. The ideal state would be if they could feel that their own prestige is associated with the success of the project. In this chapter, we have discussed the topic not only from students' perspective but also from the point of view of their role as a future executive.

2.8.1 Importance and Advantages of Goal Setting

We have listed the following factors that indicate the importance and advantages of goal setting by the process indicated in the previous paragraph.

(i) It facilitates focusing of sustained efforts by team members, in the desired direction.
(ii) It stimulates and motivates the team members to put in their best for achieving their own objective and goal. It has been noted that higher goals lead to higher motivation.
(iii) It provides immense satisfaction to the individuals when they achieve or exceed the goal which was set up by involving them, and became their own.
(iv) Helps in monitoring the performance and take appropriate action or even modify the strategy.
(v) Goals focus attention towards goal-relevant activities and away from goal-irrelevant activities.

Goal setting also requires motivation. Simply setting a target by the top management may lead to progress in the desired direction, but understanding why the target is desired encourages personal involvement into the achievement of the goal.

In order to achieve these characteristics in a goal, it is recommended to go for a participative approach of group members while deciding on the goal. Participative management theory positively affects goal and goal accomplishments because group members are clear about the goal, and feel accountable for accomplishment of the same. Hence, it would lead to intrinsic motivation and better performance.

There is ample evidence to conclude that the teams which work in this spirit succeed in smoother manner. In fact, several teams have achieved phenomenal success with this approach. A live case indicating the importance of participative management theory for business growth is described in the following:

CASE 2.5 CASE STUDY ON INVOLVEMENT OF STAFF FOR BUSINESS GROWTH

The top management of a medium-sized company having branches and regional offices all over India wanted a consultant to motivate his middle and senior level executives and help the company in accelerating growth. The consultant decided to do so by conducting a seminar for two and a half days.

A representative group of 30 executives of various levels and playing various roles at head office, regional and branch level was selected. During the first day of the seminar, the consultant gave an objective analysis and assessment of the developments at global and Indian level, in general and four companies—two relatively bigger than the company and two relatively smaller than the company. Thereafter, he divided the executives in four groups, and asked each one of them to draw the blueprint for their company for the next 3-year period. All the executives worked so enthusiastically that they had little sleep on the first day and almost no sleep on the second day. On the third day, each group was asked to present their suggestions. The consultant prepared the draft by consolidating their views, and presented before the board of directors, who were rather thrilled to have a plan that exceeded their own expectations. The company reaped the benefits of this exercise of involving the senior and middle-level executives, and it became the annual exercise in the subsequent years.

While bringing out the advantages of goal setting in a collaborative manner, here is a word of caution. This system does wonder as cited above, is ideal only if one has majority of employees who belongs to 'Y' category as enunciated by Mc Gregor. According to this theory, the employees may be divided into two categories; those who belong to category 'Y' are self-motivated to achieve higher goals but those who belong to 'X' category, are to be continuously guided and supervised and sometimes even coaxed to put in their best. One has to use the goal setting philosophy judiciously to take into account both categories.

2.9 PERT/CPM

(A Tool for Time Management and Optimum Utilisation of Resources)

PERT and **CPM** are acronyms.

PERT stands for 'Program Evaluation and Review Technique', and

CPM stands for 'Critical Path Method'.

There is a subtle difference, to be pointed out later, between the two but for our purpose we treat them as the same. It is so, because, both these techniques tell us:

"How to do a work in a systematic and scientific manner?"

The work may be conducting a study, organising admission process for MBA, writing a book, building a mall, etc. Contrary to the popular notion that these techniques are useful only for large-scale projects, we would like to emphasise that both these techniques are simple enough to be useful for any work where there is a concern for completing it on a given time and at a given cost.

Both techniques were developed almost at the same time in the year 1958, in different projects. While PERT was developed for the project relating to Polaris missile system in USA, CPM was

developed for the projects of overhauling of chemical and electricity generation plants in USA and UK, respectively.

While the emphasis in PERT was to reduce **'time'** for development of the system, and the emphasis in CPM, used in the other two cases, mentioned above, was to reduce time but with the ultimate objective of reducing **'cost'**. It may be appreciated that reducing time for any work, almost invariably, leads to reduction in cost.

These techniques were soon adopted for management of big projects to cope up with the problem of **'over run'** on time and cost which were an integral part of almost all the projects. The significant advantage of using these techniques propagated their use in almost all the spheres of office work and industrial activities. In fact, in India, L&T is known for using these techniques effectively to complete their projects like building Nehru Football Stadium in Chennai in stipulated time. The stadium was made in just 9 months, 9 days ahead of schedule! The first author of this book has successfully used this technique for publishing a book and organising annual board meetings.

2.9.1 Terminologies

In the context of managing projects, PERT is now more popular as 'Project Evaluation and Review Technique'. Project is defined as '**any work which has a definite beginning and a definite end, and which consumes resources**'; and as such PERT can be used for managing any 'project' like:

- Conducting a research study
- Building a bridge
- Selection process for MBA students at a management institute
- Finalising of annual accounts of a company
- Launching a new product or service

In the context of BRM, the use of PERT/CPM helps in managing the conduct of a research study in lesser time with optimum use of given resources without compromising on the quality of the research.

2.9.1.1 Phases of a Project Any project undergoes the following phases:

- Planning
- Scheduling
- Implementing
- Controlling

These are briefly described as follows:

Planning

The planning of a project involves the following:

- Listing of all the activities that are required to be completed for completing a project.
- Specifying resources for all the activities. Incidentally, there are six types of resources called six Ms of industry as follows:
  Men, Machine, Material, Money, Minutes (time) and Metres (space)
- Specifying the sequencing of and co-ordination among all the activities
- Estimating the realistic completion times for all the activities

These sound so simple, yet we leave it to the readers to recall having done this exercise before the start of a project! In fact, if this is done as described above, one could derive a good percentage of the advantages that is accrued by using PERT/CPM.

Scheduling

It implies indicating the starting and completion times for all the activities. It is a very important managerial part of the project management, and it decides the extent to which, one could derive the benefits of PERT/CPM, for any project.

Implementing

This is the physical part of carrying out the project as per the schedule.

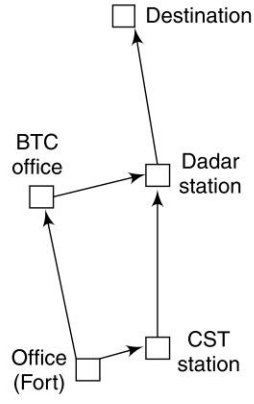
Controlling

This is the managerial part of overseeing that the project is completed as per the schedule, and if something does not go as per schedule, then to take appropriate corrective action. The objective is to complete the project within the time and cost budgeted for the project. However, if due to unavoidable reasons, there is overrun on time and cost, it is to be kept at minimum.

A point of caution may be noted. Mere use of PERT/CPM does not ensure that there will be no overrun on time and cost. The overrun can be minimised only to the extent it is humanly possible.

2.9.2 Illustration of a PERT Chart

The above four phases of a project as well as the other complicated aspects of a project can be explained with the help of a simple illustration given below.



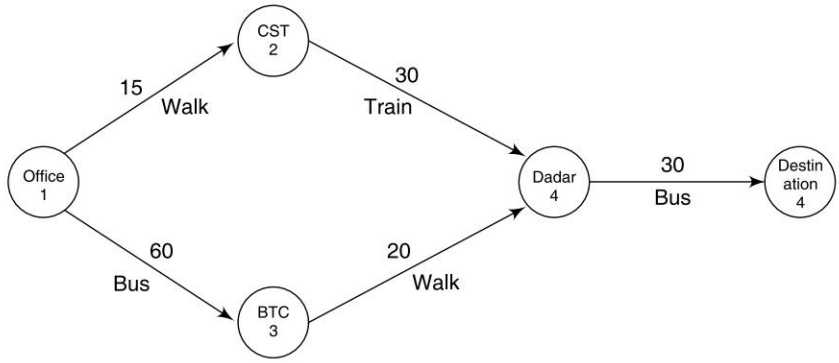
The manager of an office in Fort area (Mumbai) sends out two persons, calls them as 'A' and 'B' for some office work. The following diagram indicates the locations where these people are required to go, in connection with the assigned jobs.

'A' goes to BTC for collecting some documents, and from there goes to Dadar Station in Mumbai. 'B' goes to Dadar Station, and together, they go to the destination. In literal sense, the 'project' starts with 'A' leaving the office and both 'A' and 'B' reaching the destination. The **activities** or **jobs**, for our purpose are as follows:

- 'A' going from office to CST
- 'A' going from CST to Dadar Station

- 'B' going from office to BTC
- 'B' going from BTC to Dadar Station
- 'A' and 'B' going from Dadar Station to the destination

The sequencing and co-ordination of various activities or jobs are shown in the following:



The above diagram is known as PERT chart. The timings given above the arrows are in minutes.

It comprises circles and arrows. While arrows indicate the activities, circles represent **'Events'** or milestones. Every **activity** is bound by two events – one the starting and the other as the completion or finish. All the events are numbered from 1 to 5, and the names of activities are indicated above the arrow. The times taken for activities are indicated below the arrow. It may be noted that, the length of an arrow is not proportional to time but if one so desires, the length of an arrow can be made proportional to time. We explain the phases of a project through the above example as follows:

The first step in planning is to list the activities. The manager has identified the jobs as:

For 'A':

- Going from office to BTC
- Doing the assigned job at BTC
- Going from BTC to Dadar Station

For 'B':

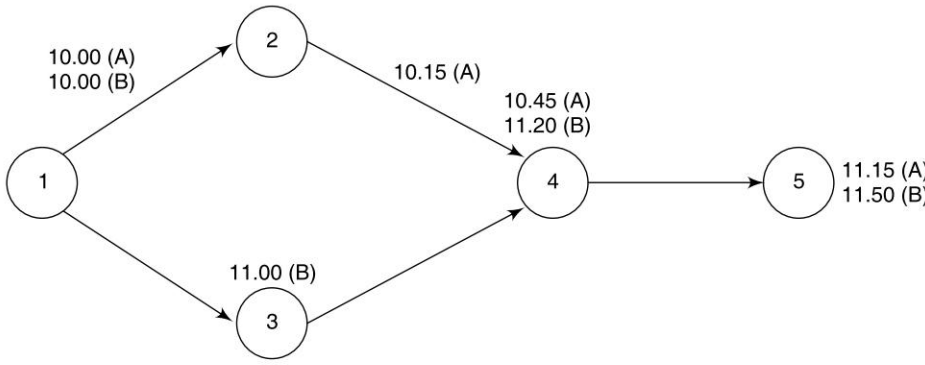
- Going from office to CST
- Going from CST to Dadar Station
- Going together with 'A from Dadar to the destination

Thus, the project comprises all the above activities. For the sake of keeping the PERT chart simple, we have ignored the job "Doing the assigned job at BTC" in the chart. The sequencing of activities is clearly indicated. The co-ordination is specified in the form of both 'A' and 'B' going together from Dadar to the destination.

As regards scheduling, we may consider the scheduling as shown in the chart as on next page:

The chart shows the times required for completion of each activity as also times for start and finish times for various jobs.

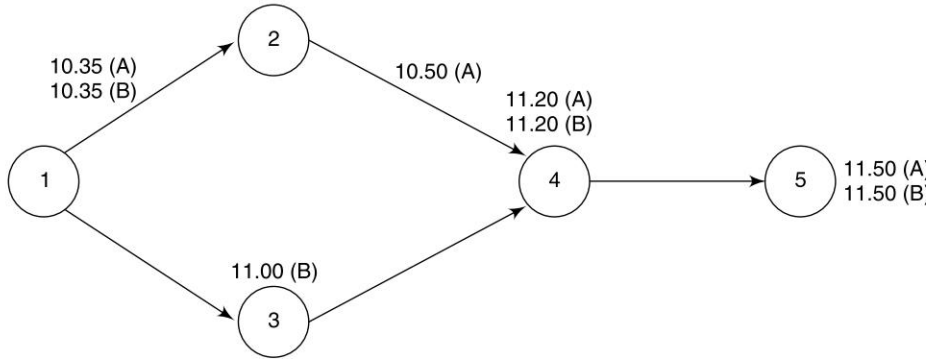
At the first look at the chart, one may think that the scheduling is defective in the sense that 'A' is wasting his time to the tune of 35 minutes as he has to wait for 'B' to arrive at Dadar. But there is a positive aspect also. There is a cushion for 'A'. Even if he/she is stuck with non-availability of train or whiles away his/her time, the manager will not be so much concerned about his movement.



He/She will concentrate on 'B', and might even tell him/her to phone when he/she reaches Dadar Station.

Further, if due to some reasons, after the project starts, he/she is required to expedite the project by 15 minutes, the manager may authorise taxi fare to 'B' to go from BTC to Dadar Station, and reach in 5 minutes. He/She need not incur any additional expenditure on 'A'.

However, if we accept the point of view that 'A' is wasting his time, he/she may be asked to start at 10.35 so that both will reach Dadar at the same time i.e. 11.20 as shown in the following:



However, now the manager will have to control the movements of both 'A' and 'B'. If any one of the two is delayed even by 5 minutes, the project will be delayed. Further, if he/she is required to expedite the project by 15 minutes; either he/she has to expedite the movement of both 'A' and 'B' or he/she expedites the movement from Dadar to the destination by telling them to go by taxi rather than bus. Thus, there are both pros and cons for the two schedules indicated above. There is a possibility of yet another scheduling, and that is to ask 'A' to leave office at 10.15. It provides 20 minutes of cushion to 'A'. In fact, 'A' could be asked to leave any time between 10.00 to 10.35. That is why, scheduling is considered as a managerial activity rather than a routine job. One can only guess what could be the scheduling concerns in a bigger project!

2.9.3 Further Definitions Relating to a PERT Chart

Some definitions relating to a PERT chart are as follows:

Independent and Dependent Jobs/Activities:
A job which can be started on its own is called 'independent' job; like 1–2. Further, two jobs are said to be independent of each other, if both the jobs can be done independently or simultaneously. For example, job 1–3 is independent of job 1–2.

But a job which can be started only when the earlier job is completed, is called 'dependent' on the earlier one. For example, the job, 3–4 is dependent on 1–3. The job 4–5 is dependent on both 2–3 and 3–4, and cannot be started unless both are completed.

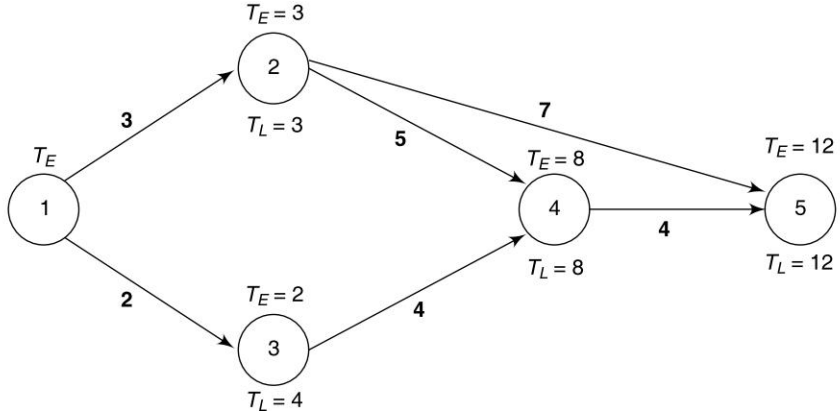
Simultaneous Jobs:
Two jobs which are independent of each other can be done simultaneously. For example, jobs 1–2 and 1–3 can be done simultaneously.

In fact, one of the biggest advantages of PERT chart is that it indicates independent jobs which can be done **simultaneously**, and, thus, **saves time**. **Incidentally, the biggest secret of reducing time for a project, through PERT chart is that independent jobs can be done simultaneously, and thus, saves time.**

Event:
An event is said to be completed if all the jobs leading to it are completed. For example, event '4' is said to be completed only when both the jobs 2–4 and 3–4 are completed as then only the job 4–5 can start. Incidentally, the event '4' indicates reaching that milestone as also the beginning for the next milestone.

2.9.4 Analysis of a PERT Chart

The analysis of a PERT chart involves calculation of certain quantities which help in the management of a project. These are described in the following with the help of a PERT chart of a project:



The times taken for a job, indicated above the arrows and for T_E and T_L are given in weeks.

Earliest Expected Time (T_E):
It relates to an event. It is the earliest time by which an event is expected to happen. An event is said to have happened when all the activities that are leading to it are completed. The T_Es for various events are given above the circle representing the event. The explanation for the T_E of event 4 is as follows:

As mentioned earlier, this event is said to have happened when both the jobs 2–4 and 3–4 are completed. The job 3–4 will get completed at the earliest by 8 weeks, and the job 2–4 will get completed at the earliest by 6 weeks. However, T_E for event 4 is 8 weeks, as the activity starting from event 4 i.e. 4–5 can be started only after 8 weeks.

Latest Allowable Time (T_L):

It also relates to an event. It is the latest time by which an event must be completed if the project is not to be delayed. For the last event, $\boldsymbol{T_L}$ indicates the project completion time.

For the last event, $\boldsymbol{T_E}$ and $\boldsymbol{T_L}$ are equal.

For event 5, the $\boldsymbol{T_L}$ is 12 weeks.

For event 4, it is 12 – 4 = 8 weeks.

For event 2, $\boldsymbol{T_L}$ for 4 is 8 weeks, and the job 2–4 takes 4 weeks, so $\boldsymbol{T_L}$ is 4 weeks.

Critical Path

Starting from event 1, we can reach the last event 5 via various paths. The path which consumes maximum time should really be called 'longest' path. But it is called **'Critical Path'** because on this path, even if there is slightest delay in any job, the project will get delayed. It is generally indicated by red ink, as it demands maximum attention. In the given earlier chart, the path 1-2-4-5 is the critical path. The term, Critical Path Method implies managing a project based on analysis using this path.

It may be noted that on critical path, the earliest expected times and latest available times are equal.

Slack or Cushion Time for Jobs:

The slack or cushion time for a job is calculated by subtracting the $\boldsymbol{T_E}$ for the previous (tail) event plus the expected time for the job from the $\boldsymbol{T_L}$ for the next (head) event. For example, for the job 3-4, $\boldsymbol{T_L}$ is 8 weeks. From this, we subtract 2 ($\boldsymbol{T_E}$ for event 2) and the time taken for job 3–4 i.e. 4 weeks; thus, getting 2 weeks as the cushion time for activity 3–4. The calculation is shown below:

8 ($\boldsymbol{T_L}$ for event 4) – 2 ($\boldsymbol{T_E}$ for event 2) – 4 (time for job 3–4) = 2

It may be noted that there is no cushion time for jobs lying on the critical path. These jobs are called **'critical jobs'**. Similarly, the events lying on the critical path are termed as **'critical events'**.

2.9.5 Drawing of a PERT Chart

For illustrating the drawing up of PERT chart for a project, we discuss the project of preparation of the budget in a manufacturing firm. For making it simplistic, we assume the budgeting exercise to comprise the following activities:

A: Forecast sales

B: Forecast revenue (based on commission, discount, etc.)

C: Draw production schedule

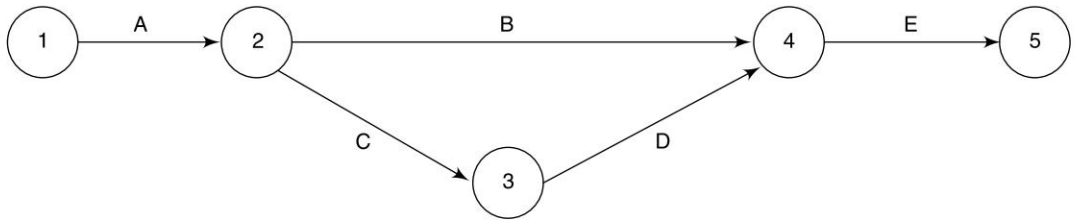
D: Cost the production schedule

E: Prepare budget

We prepare the following table indicating sequencing and interdependence of various activities:

Activity	*Depends on*
A	---
B	A
C	A
D	C
E	B and D

Using this table, the chart is drawn as follows:



It may be mentioned that a PERT chart is only a statement of logic showing sequencing and co-ordination, and its shape does not matter e.g. whether the arrows are slanting or are horizontal/ vertical.

2.9.6 Advantages of a PERT Chart

Some of the advantages of a PERT chart are given below:

(i) It provides an overview of the project with a single diagram on one page.

(ii) It helps in arriving at realistic estimates of time and cost in a scientific manner, as it is much easier to estimate time and cost of individual activities rather than the project as a whole.

(iii) One can easily comprehend and ensure that all the important activities are included in the chart as also their sequencing and co-ordination is logical.

(iv) It helps to exercise what is known as management by exception as one can pay more attention to the critical activities.

(v) The impact of delay in one activity can easily be assessed, and corrective action of expediting the relevant subsequent activities can easily be taken so as to complete the project on time.

(vi) The chart being only statement of logic does not change with time; one can only keep on revising the earliest and latest allowable times as also the critical path.

(vii) It helps in optimum utilisation of resources through levelling of resources i.e. transferring resources from one activity to another activity depending on the contingencies that arise during the completion of the project.

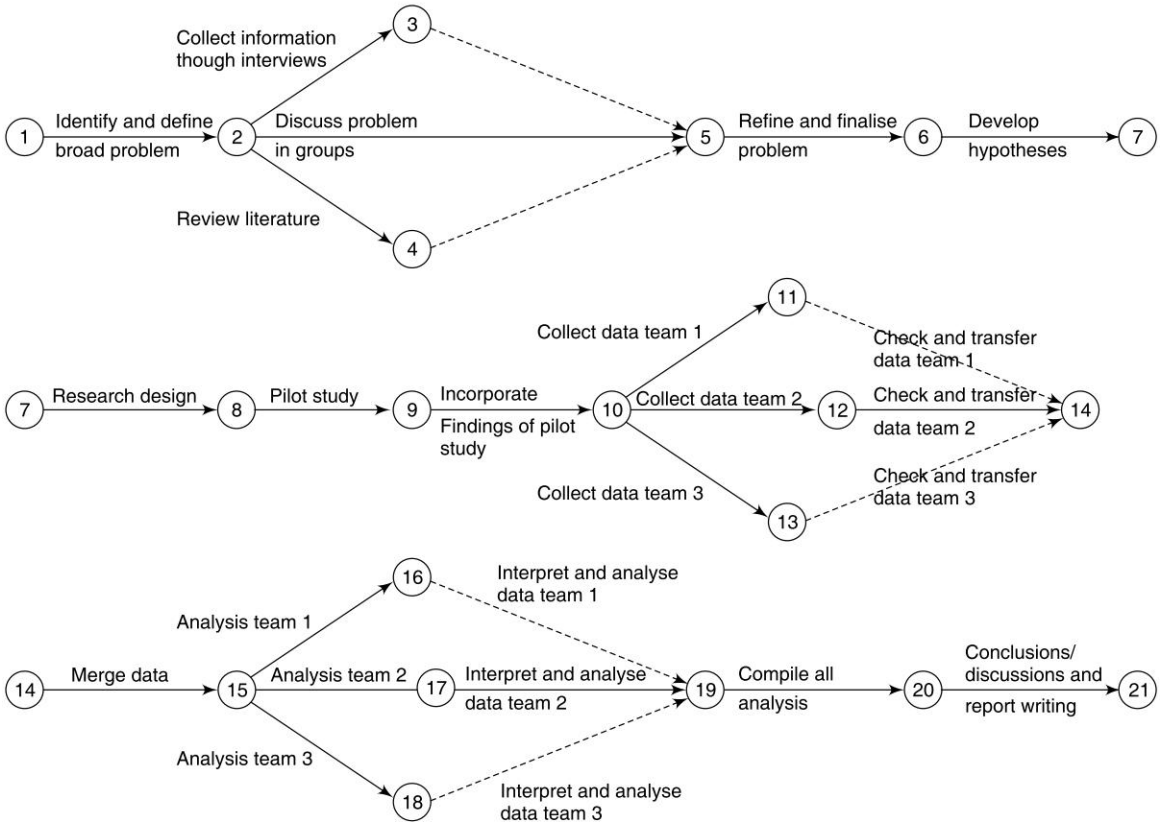
2.9.7 PERT and CPM

As regards the difference between PERT and CPM, for our limited purpose, we may mention that while PERT is event-based, CPM is activity-based.

2.9.8 Relevance of PERT Chart for Conducting a Research Study

The following PERT chart explains the sequencing and co-ordination of all the activities relating to conduct of a research study.

It may be noted that the level of complexity depends on size and type of a project. This PERT chart is only a guideline for preparing the PERT chart of an actual study.



It may also be mentioned that in a PERT chart, the emphasis is on management of time of various activities.

The dotted arrows are called **'dummy'** activities. These are used just to connect events.

The above PERT chart for the research projects indicates as to how the project can be conducted in teams. The project starts with identification of goals and defining problem. In the next stage, teams can be made and allocated different tasks like collecting information through interviews, discussing problem in groups, and reviewing literature, among themselves. The outcomes of all the activities mentioned above can be combined to refine and finalise the problem, and developing the hypothesis, research design and conducting a pilot study. After incorporating the findings of the pilot study, the team again may split into sub-teams for collecting data. After checking the data of different teams, the data can be compiled together and the analysis and interpretation can be distributed again in different teams. Finally, the entire analysis is compiled together in the form of a research report.

A PERT chart is a simple yet powerful device for managing a project in a systematic and scientific manner. It is a universal technique for managing projects in any field right from making **tea** to setting up a **nuclear plant**. Of course, the shape of the chart will go on getting complicated depending on the complexity of the project. The decision for using a PERT chart for a project does not depend on the size of the project; it depends on how serious one is to minimise overrun on time and cost and make optimum use of resources.

Apparently, it may appear that the real use of PERT chart is only when certain activities are parallel i.e. they can be done simultaneously. It is true to some extent, but the 'Planning' phase of PERT chart is useful in all the situations. In the case of BRM, collection and analysis of data is one activity which can be subdivided in suitably number of parts as shown above, each of which can be done simultaneously.

2.10 CREATIVITY AND RESEARCH IN AN ORGANISATION

There is no force as powerful as an idea whose time has come.

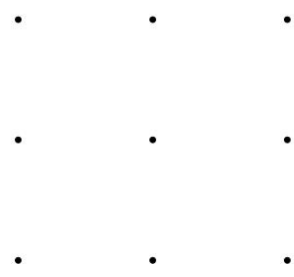
—Victor Hugo

Quoted by Dr. Manmohan Singh in his budget speech of 1991

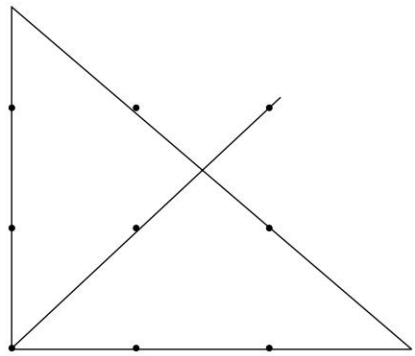
Creativity, in simple terms, is the trait of a person to think, create or do something new. Creativity is also defined as the ability to generate new ideas or concepts. As such both research and creativity are highly interrelated. In the context of application in business organisations, creativity is described as a process of developing and expressing novel ideas for solving problems, in general, and meeting specific needs of an organisation, in particular.

Normally, we associate creativity with art, music and literature but it may be appreciated that it is also associated with innovation and inventions. Thus, it is equally important in professions like business, architecture, industrial design, advertising, software engineering, video gaming, etc. It is due to this lack of appreciation that creativity, as a subject, is yet to get it's due place in the academic and in business industry.

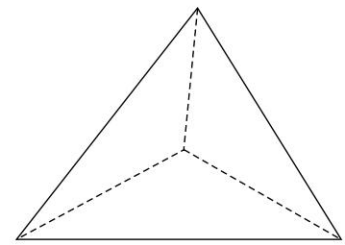
In the context of business environment, as mentioned earlier, it is highly useful in solving problems arising at various stages like planning, analysing and controlling. Sometimes, given a problem, creativity helps in looking at the problem from an angle, which is quite different from the normal or routine, and thus solve a problem. This concept is best illustrated with the help of the following two classical examples. The first example relates to joining the nine dots arranged in a square shape on a paper as follows:



The conditions for joining the nine dots are that the dots should be joined by drawing four straight lines by a pen 'but without lifting the pen from the paper'. The conventional approach for solving the problem is to try joining the dots while remaining within the square. The solution is, however, easily obtained when one goes outside the square, and joins the dots as follows:



The second problem relates to making four equilateral triangles with the help of six matchsticks; the sides of triangles being equal to the length of the matchsticks. Here, again, conventionally, one would keep the matchsticks on the surface of a table, and keep on trying to make four equilateral triangles by moving the sticks on the two-dimensional surface of the table. However, the solution is easily obtained when one ventures in the third dimension as shown in the following, and arranges the sticks in the form of a prism.



While creativity is relevant and an asset for each and every human being, it is all the more relevant for an organisation, as explained below.

The humans and organisations both go through the following life cycle:

Human	*Organisation*
Born	Born
Grow	Grow
Decay	Decay?

There is, however, a difference between the two, and the difference is attributed to creativity. While a human's decay cannot be postponed indefinitely, an organisation's decay can be postponed indefinitely, through creativity, if it keeps on innovating new products, services, processes and systems. A classical example is 'The Times of India' newspaper which is now more than 170 years old, and still dominates the circulation of English newspapers in India. However, many publications started after 'The Times of India' have become history. Another example is the Punjab National Bank which is more than 100 years old, and is still growing stronger steadily, and is now the second largest public sector bank in India. Several other banks which started much later do not exist; some of them are: New Bank of India, Hindustan Commercial Bank and Global Trust Bank.

Running a Small Business

The only son of a big businessman parted company with his father's business. He approached a banker and asked for his advice to run a small business. The banker replied, "When your father hands over the business to you, run it the way your father had been running in the past, without any change, you will have a small business to run!"

2.10.1 Role of Creativity in Research

Creativity plays a very important role in research—the prime reason being that while defining a problem for research, one has to critically examine that the real or perceived problems is really to be defined that way or is there another way to look at the problem. As we shall discuss in Chapter 3, many a times, we may have to redefine a problem to arrive at the solution.

CASE 2.6 REDEFINING A PROBLEM

A classical example of redefining the problem relates to a training institute. The institutes and hostel buildings in the same campus are shown in the following diagram:

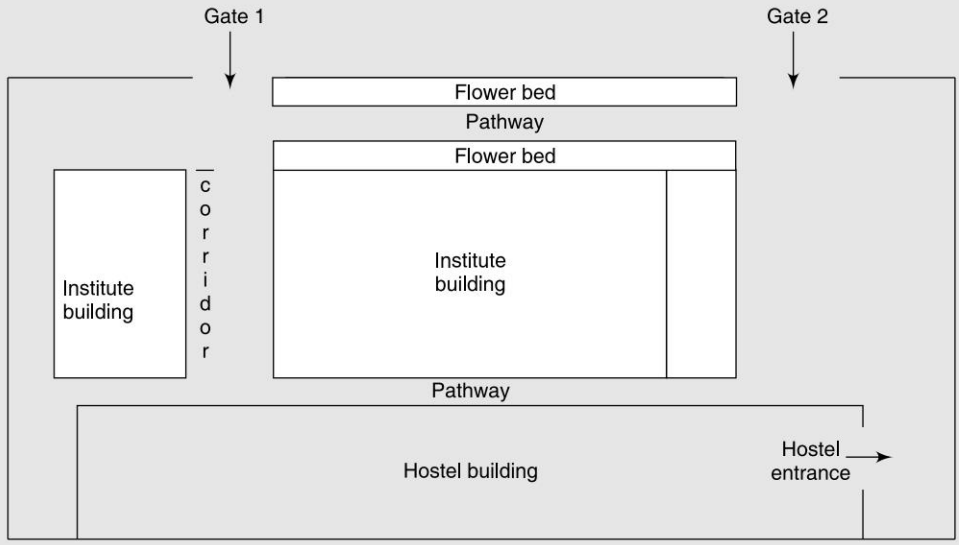


Fig. 2.1 *Institute Campus*

There are two gates—one for entry to the institute and the other for entry to the hostel building. Both gates are manned by one security guard each. As a part of the economy drive, the management wanted to reduce the number of security guards. After a study of the pattern of usage of the gates, it was decided to open only one gate at a time; and the timings for each gate were announced. This led to a lot of inconvenience for staff of the institute and hostel residents, especially for those who used cars or taxis for entry to the two buildings. Those hostel residents who came by car and had to enter through the institute gate could not reach the entrance of the hostel

building causing inconvenience in rainy season or having to carry luggage for a good distance of about 30 metres. On the other hand, the faculty and staff who came by car, entering through the hostel gate could not park their cars in the institute campus and had to park cars outside the campus due to limited parking capacity in the hostel compound. This went on for about a year. Thereafter, a new security officer joined the institute. He suggested the reduction of the width of the flower beds 'A' and 'B' in front of the institute's building, by 1 foot and 2 feet respectively. This increased the width of the pathway from 5 feet to 8 feet. Thus, a car could pass through the pathway. With this simple solution, the hostel gate was closed (except for utility vehicles) and the institute's gate was kept open for 24 hours. Thus, the hostel residents could come any time, enter through the institute's gate and reach the hostel building. Thus, the ideal solution was evolved by changing the original problem of deciding times for opening/closing of the two gates.

Yet another classical example of changing the perspective of a problem needing solution was during the Second World War when UK's merchant ships were being attacked and destroyed/damaged by German bombers in the Mediterranean Sea. A debate was going on in the British navy about fitting of ships with antiaircraft (AA) guns which were available only in limited number, and were required for protecting many civil/military installations in the country. Accordingly, only limited number of ships were fitted with AA guns. However, these guns could not shoot down sufficient number of bombers, and thus their utility was considered as doubtful by the military authorities. At this junction, U.K. government appointed a committee headed by Noble laureate Prof. P.M.S. Blackett; to look into the matter and offer its recommendations. The team collected and analysed the data. The analysis revealed that even though the guns fitted on ships did not destroy the bombers, their firing at the bombers did scare them and affected their accuracy of hitting the target with the result that many bombs did not hit the ships and even if they did hit, many ships were not sunk, and suffered only some damage that could be repaired on the sea itself. Prof. Blackett pointed out that **the criterion for judging the efficacy of installing AA guns on ships should not be the destruction of the bombers by the guns but the extent to which they were able to affect the bombers' accuracy of hitting the ships and thus protecting the ships. This view was accepted by the U.K. Government, and the guns continued to be fitted on ships. Thus, the perspective of the problem was changed to arrive at the solution. We could also say that the originally perceived problem was redefined.**

Incidentally, creativity has been associated with right or forehead brain activity or with lateral thinking, and the left brain is associated with memory and analytical ability.

Creative ideas are often generated when one discards preconceived notions and tries to follow a new approach which could be unthinkable or even 'laughable' for others! Incidentally, the present design of famous Tower Bridge in London (proposed by Sir W. G. Armstrong Mitchell & Company), which has stood the test of time was not approved at the first instance—it was approved only after discarding the suggestions that were received later.

2.10.2 Creativity in Business Organisation

It has been observed that a creative process begins with what is called '**divergent thinking**' that is quite different from usual ways of doing and observing. It helps to develop insights and new ideas. It is followed by '**convergent thinking**' which analyses divergent thinking into a specific product, service or system.

In the context of a business organisation, the term, **innovation**—for various components of the entire process—is used to refer to generating creative ideas and converting them to viable products, services and systems and the term, **creativity** is referred to generating new ideas by individuals or groups, as a necessary step within the innovation process. This leads to the saying, "**Innovation begins with creative ideas**".

Innovative staff has to be creative to continue to be relevant and remain fit to compete. Creativity can be nurtured in the individuals through practical training as described in the following:

Creativity is best practised for day-to-day decision making through well known '**Brainstorming**" technique, introduced by Alex Osborn. Another strategy for promoting creativity in the corporate world is setting up of '**Think Tanks'**. In fact, both the techniques of 'brainstorming' and 'think tanks' described in the following are instrumental in deriving the above mentioned benefits for an organisation.

2.10.3 Brainstorming

Brainstorming is a group creativity technique designed to generate a large number of ideas for the solution of a problem introduced in 1930s by Alex Osborn.

It is especially useful when the problem involves co-ordination and mutual support among the members of a team. It encourages innovative ideas in a group setting and generates a sense of involvement among the group/team members. While it leads to a sense of collective ownership of the emerging solution(s), it also ensures commitments by the participating members.

One such example was the meeting convened by the Director of a management institute of the faculty members to discuss the support system required for developing and using more case studies as teaching pedagogy by faculty members.

Another example is the General Manager of the credit card division, in a bank, called a meeting of the executives to discuss the different avenues to increase the profitability of the division, using this technique.

In fact, without naming so all the meetings that are conducted in an organisation to solve any specific problem could be termed as brainstorming sessions, if the following tenets of brainstorming are followed. These relate to maximising the number of ideas, remove inhibition among people to offer ideas, stimulate expression of the wildest idea and deriving synergy of ideas. These are as follows:

- Emphasise on quantity as per maxim 'quantity contains quality'
- No criticism of any idea
- Welcome all ideas even if they appear weird or extrinsic
- Combine and improve—combination of two or more ideas might lead to a new idea that could be better than the original idea

The steps in the conduct of a typical brainstorming session are outlined as follows:

- Specifying the problem
- Preparing and circulating an agenda containing the specific problem
- Selecting the participants—preferably in the range of 6 to 10, in number
- Conducting the sessions—encouraging each participant to offer suggestions without any inhibition, and recoding the same on a board
- Stimulating the idea generation process, if needed
- Evaluating the suggestions subsequently, and reaching consensus

2.10.4 Think Tank

A think tank is an embodiment of persons engaged in strategic or policy-level thinking over some issues with the objective of arriving at suitable action plan or course of action to resolve the issue.

The embodiment of persons could be as small as a group of persons in an organisation or as big as a professional institution like consultancy firm. The issues might relate to any one or more than one of the following illustrative (not comprehensive) sectors.

- Industry level (general)
- Business (specific)
- Economy
- Technology
- Defence
- Education

The term 'Think Tank' was used to refer to the rooms wherein war time related issues were discussed during Second World War. The term is now used for any group, organisation and corporation engaged in resolving policy/strategic issues relating to an individual, organisation, industry or even a country. It is reported that there are thousands of think tanks in USA and other countries.

2.10.5 Nurturing Creativity for Research in an Organisation

Incidentally, while creativity is the 'ability', research is 'action', and, thus, creativity is essential for carrying out research successfully. In fact, it is needed at each and every step of a research process, described in brief in Section 1.7 Chapter 1, and in detail in Chapter 4. It is especially useful in the initial stage of defining the problem.

The philosophy that '**there is always a better way**' has to be imbibed in the culture of an organisation. This could be achieved by adopting the following strategies in an organisation:

(i) Building basic skills leading to creative thinking and action
(ii) Promoting self-confidence and culture of taking risks within certain limits
(iii) Encouraging freedom for generating choices and opportunities for experimenting them
(iv) Stimulating and rewarding curiosity and exploration
(v) Creating culture of tolerance to bonafide mistakes or losses on account of trying new ideas
(vi) Demonstrating appreciation and reward for creative ideas and innovations

CASE 2.7 ENCOURAGING CREATIVITY IN AN ORGANISATION

The chairman of a middle-sized bank who himself brought about several innovative features to take bank to much greater heights, introduced a novel approach to encourage creativity among the senior and the top executives. An **'idea room'** was set up with just one table and a chair. The executives were requested to visit the room every quarter. They were given a sheet of paper, on which they were requested to offer new ideas and suggestions while sitting in the room. The sheets were submitted directly to the chairman for his perusal and taking necessary action. If an executive gave a blank sheet, it was placed in his personal file. This system brought out several innovations in the bank, and consequently it reached the level of top banks within a few years.

It may be useful to discuss the difference between creativity and innovation which is so commonly used in business organisations. While creativity is associated with generation of new ideas, innovation relates to translation of ideas into physical forms like products and services. In fact, innovation begins with creative ideas or in other words, creativity is the starting point for innovation. However, it may be noted that creativity is necessary but not a sufficient condition for innovation.

Intelligence and creativity are the hallmarks of successful personalities. While, behavioural scientists have evolved a measure of intelligence called I.Q (Intelligent Quotient), there is no measure of creativity as of now. However, both are correlated to some extent.

2.10.6 Creativity in Practice

It is argued that to enhance creativity in business, three components are needed:

- Expertise (technical, procedural and intellectual knowledge),
- Creative thinking skills (how flexibly and imaginatively people approach problems), and
- Motivation (especially intrinsic motivation)

A study of several successful Japanese companies revealed that creativity and knowledge creation played an important role in the success of organisations.

Following are some examples of a few innovations which have been evolved as path-breaking creative solutions.

Product Design	Apple: IPod
Manufacturing	Toyota Production Systems: 5 Whys
Sales and Distribution	Amazon.com: On-line Seller
Financial Products	Derivatives—Options and Futures: Black Scholes

2.10.7 Creativity in USA

Creativity as a subject, seems to be given maximum importance in USA where there is an exclusive **Master's** course in creativity conducted by **International Center for Studies in Creativity, Buffalo State**, and where international creativity festival is organised annually. Incidentally, now the creativity festivals are organised in other countries also. In fact, if one single factor is to be attributed to the phenomenal growth of USA, it is perhaps the creativity or innovative approach in various spheres of business, specially IT, telecommunication, aeronautics, pharmaceuticals, education, entertainment, etc.

Incidentally, some decades back when a famous scientist Norbert Wienner was asked to differentiate between philosophies of Americans and British people, in the context of rapid growth of USA as compared to that of Britain, he was reported to have replied,

> **"British give more importance to what they *do* than what they *think*,**
> **while**
> **Americans give more importance to what they *think* than what they *do*."**

It is because of the importance attached to creativity, that USA is the No. 1 economy leading in pharmaceuticals, health care products and services, entertainment, information technology, among several others.

2.10.8 Creativity in India

India is blessed with a vast number of younger, educated workforce with strong analytical skills. It has a tremendous potential for creativity ideas in the following emerging fields:

- Information Technology
- Biotechnology
- Space Research
- Pharmaceuticals
- Entertainment industry – Animation
- Education
- Hospitality
- Healthcare
- Agri-farming
- Finance and Banking

While it is difficult to mention thousands and thousands of innovations in different companies, we mention only some of them namely Infosys, TCS, Tata Motors and ICICI Bank.

Infosys, TCS and Wipro bear testimony to the role played by ideas and creativity. Our young IT professionals have provided solutions for many complex problems of developed countries, and have made India an IT giant in the process.

Tata Motor's 'Nano' is one of the most creative innovations of all times, in the entire global auto industry.

ICICI Bank gave a revolutionary approach to lending to Small and Medium Enterprise (SME), which has received praise from all over the world. The World Bank has selected it as one of the three banks (at global level) to be included in its special publication on SME. ICICI Bank has been declared as the best bank (SME) in Asia for the year 2009, by Asian Banker.

If our vast pool of young and energetic talents is encouraged to use their creative ability and innovate, it can change not only the face of India but also of the world to a significant extent.

SUMMARY

The difference between Research methodology and Research Methods is summed up as:

"The term 'methodology' connotes a much wider concept. It encompasses a complete gamut of activities required to complete a research study including research methods that are used in the study."

The criterion for selection of appropriate research method is to select the method which entails gathering most useful information in the most cost-effective manner that leads to most cost-effective decision-making.

The terms constructs and concepts used in the context of business research have been explained with suitable examples.

The chapter describes various types of variables that are the raw material for a research study. These are:

Independent Variable
Dependent Variable
Moderating Variable

Intervening Variable
Extraneous Variable

While the basic concept in deductive logic is 'Many to One' or 'Population to Sample', the basic concept in inductive logic is 'One to Many' or 'Sample to Population'.

Collaborative approach is considered essential when a research study involves team work.

Creativity is important for exploring research areas and in conducting research activities.

PERT/CPM is an indispensible tool for completing any assignment including conduct of a research study in a given amount of time and cost.

DISCUSSION QUESTIONS

1. Explain the basics of Constructs and Concepts with five examples.
2. Discuss the concepts of Deductive and Inductive logic with suitable illustrations.
3. Bring out the difference between Quantitative and Qualitative researches. For each of these, describe three situations from business environment.
4. Discuss comparative features of Quantitative and Qualitative researches.
5. Describe Case Study method of research with three suitable illustrations.
6. Describe the use of PERT/CPM in general, and in conducting a research study, in particular.
7. Discuss the role of creativity in an organisation, in general, and in conducting a research study, in particular.
8. Write short notes on the following:
   (i) Research Methodology and Research Methods
   (ii) General Applications of Qualitative Research
   (iii) Applications of Qualitative Research in Exploratory Research
   (iv) Goal Setting

Research Process

3

Contents

LEARNING OBJECTIVES

The main purpose of this chapter is to provide an exhaustive and comprehensive view of the various steps of a research process from problem identification to hypothesis development i.e. developing the statement to be tested for acceptance or rejection. This facilitates conduct of further study involving collection of data, carrying out relevant analysis, etc. and ultimately resolving the problem.

Further, the objective is also to facilitate preparing a research proposal, after the hypothesis is formulated, keeping in view the criteria for conducting a good business research.

Subsequent to finalisation of the research proposal, the issue of conducting the research either in-house i.e. by a team of researchers from within the organisation or getting it conducted by some outside consulting agency, has been deliberated in detail with pros and cons of both the approaches.

Relevance

National Dairy Products Ltd. is a small scale enterprise having two dairies in Kolhapur district of Maharashtra, started by Mr Vasant, from a modest small dairy. The business has been growing rapidly. The setup that was adequate for the small business, was not able to cope up with the growth. Ms Charu, an enthusiastic MBA from reputed B-School, joined her father's business with an ambition to take the business to new heights.

After joining the business, she immediately felt the need for in-house research department. She got her father's approval for the same, on a condition that she would utilise only internal talent for the department. She chose some of the talented staff from the other departments to set up the research department.

She felt the need for orientation of these people towards the research process.

She conducted a seminar to lay the foundation of research methodology for the new team. In the seminar, she explained the importance of research for the company, in general, and the importance of a systematic process for a conducting research. She also explained each and every step involved in research, and its importance.

At the end of the seminar, the selected employees who initially had a low morale, were highly motivated for conducting the research projects.

It generated great deal of satisfaction to Ms Charu, as this was the indication that the seminar was successful.

Ms Charu finally succeeded getting the approval for setting up of an internal research department in National Dairy Products Ltd.

3.1 INTRODUCTION

In Chapter 2, we have discussed certain terminologies/topics which are used extensively in business research methodology. In this chapter, we shall discuss the initial steps of the research process viz. defining/refining/redefining i.e. formulation of the problem to be solved, formulating issues/questions that are relevant for various hierarchal levels viz. managerial, researcher, investigator and measurement, development of hypothesis i.e. the statement to be tested to resolve the problem.

As mentioned in Chapter 1, conducting research follows a well-structured process and it comprises following broad steps:

(i) Specifying the area and the objective of the study
(ii) Defining and Refining a Problem
  - Defining Problem:
  - Refining/Redefining the Problem through
    - Literature Review
    - Interviewing relevant people
    - Group discussion with relevant people

(iii) Hypotheses Development
(iv) Preparing Research Design
(v) Collection of data
(vi) Analysing the data

(vii) Interpreting the results and drawing conclusions based on data
(viii) Report writing, stating the genesis of the problem, formulating the problem to be resolved, methodology used in collecting data, analysis, interpretation and conclusions/recommendations.

3.2 IDENTIFYING RESEARCH PROBLEM

"Every problem can be traced back to failure of management and leadership" Abhishek Gupta, Senior MD & Chairman, Blackstone India"

—Business Today - 11 Jan. 2009

With due regard to the above statement, we would like to add that many problems can be traced back to earlier success and solutions evolved because today's 'success' could be termed 'failure' later on, and what we consider as 'solutions' to the present problem lead to some other problems later on. We would also like to add that, in general, there is an endless queue of problems.

The moment, we solve one problem, the next in the queue—either by turn or by jumping the queue—pops up. That is how it is said:

"There is no finish line in problem solving"

A classical example of this aspect is the credit card business of banks in India. About 5 to 6 years ago, the banks went on rather recklessly about issuing cards to one and all without due verification of credit worthiness. Due to competition among banks to excel each others' number, the cards were even made free for lifetime. But now the banks are feeling the pinch of it in varying degrees. The 'success' achieved in increasing the number of credit card customers proved to be 'failure' in a limited sense that the banks are burdened with lots of money locked in overdues.

A research problem is a problem that requires a researcher to pursue it with a view to find out the best solution in a scientific manner.

Even though the discussions relate to the problems in an organisation, these are equally valid for any problem relating to an individual or group of individuals or a department of an organisation.

It may be prudent to take the following points into consideration for identification of the problem:

(i) Genesis of the Problem
(ii) Impact of the Problem

Brief discussions are as follows:

(i) Genesis of the Problem

A problem usually starts with some vague and unstructured thinking. Subsequently it goes through a series of refinements. Sometimes, it leads to even redefining the problem that may be quite different from the original problem.

The problem may be caused either due to the developments/decisions within an organisation or due to developments outside the organisation.

In addition to the factors leading to a problem that need to be resolved through research, other factors that induce a corporate body to conduct research as mentioned in Section 1.5.2 of Chapter 1, are listed here for recapitulation:

(i) Self-motivation
(ii) Regulatory

(iii) Competition
(iv) Customer Driven
(v) Failure
(vi) Technological Innovations
(vii) Environmental Considerations
(viii) Social
(ix) Economic
(x) Infrastructure
(xi) Operations / Process Driven
(xii) Coping with Changes

A problem might also arise due to

- Experiences in the field or operations
- Study of related literature. This might lead to people trying out similar experiments like expanding the scope of the problem
- Study of 'Request for Proposals', detailed in Section 3.7 or similar matter appearing in newspapers, magazines, etc.
- New visible opportunities

(ii) Impact of the Problem

If the problem is not solved, what would be the consequences?

The consequences could be:

- increase in costs/expenditure
- loss of revenue
- missing out on the opportunity for expansion or diversification or, could be as serious as the uncertainty over existence of the company e.g. due to non-compliance of regulatory directives
- loss of reputation
- erosion in USP (Unique Selling Points).

3.3 FORMULATING RESEARCH PROBLEM

The formulation of a research problem is the crystallisation of the thinking and deliberations about a research problem that is ultimately taken up for research study. It comprises the following steps:

- Defining a problem
- Refining or redefining the problem through
  - Literature Review
  - Individual and Group Discussions (Using Creativity)

3.3.1 Defining a Problem

The problem definition is based upon the problem faced or posed by the decision-maker. For example, a decision problem may be whether to start a new service. The corresponding research problem could be to assess whether the service would be found 'acceptable' by the intended users.

The problem should be stated in a general manner—without making it too specific or too vague.

3.3.2 Refining a Research Problem

In this stage, the problem is crystallised through

- Discussions by using creativity techniques like–Brainstorming. Think tanks, described in Section 1.8 of Chapter 1.
- Review of the relevant literature. This is one of the most important initial steps in a research project. This is also one of the most humbling experiences that one could have. In all probability, one is likely to find out that just about any worthwhile 'new' idea one has, is actually not so 'new', and someone else had thought of it before, at least to some degree. Literature review is essential in order to identify all the related research and also to set the current research project within a conceptual and theoretical context. Brief guidelines for literature review are as follows:
  - Look for similar topics in reputed sources of information including internet
  - Seek answer to the question/problem using multimedia and internet
  - Include all important relevant constructs/variables in the study
  - Facilitate selection of appropriate research design, including sampling, designing of questionnaire, measurement instruments, etc.
  - Anticipate common problems that might have been faced in earlier studies, and thus adopt strategies to overcome them

This might result in the modification of the original problem or even leading to a new problem. A researcher should refine the problem in such a way that it is amenable to research. Just like a doctor listens to all the symptoms indicated by a patient, examines his body, and prescribes tests, and on the basis of test reports diagnoses the problem and recommends a course of treatment. Similarly, the researcher has to diagnose the problem in a broader perspective and also at various research hierarchal levels, mentioned below. For example, if the managerial problem in a company is that of declining sales, the researcher could redefine or convert this problem to determine factors that contribute to decline in sales, and evaluate their impact on various aspects of sales. He/she could even extend the scope of the study to explore the ways in which sales could be increased.

While refining or redefining the problem, one may consider the following questions/issues:

In the context of conducting a research study in an organisation, there are different issues or questions to be addressed or answered at various hierarchical levels viz.

- Management
- Research
- Investigative
- Measurement

Management Question/Issues The management dilemma gets translated into management questions. The management questions convert the dilemma into question form. For example, the management dilemma could be declining sales in a retail outlet. Then the management questions that could be asked will be:

- Why the profits are declining?
- What should be done to improve sales?
- What should be done to increase customer satisfaction?
- What should be done to increase footfall in the retail outlet?

These questions are usually general in nature. The questions may, in turn, get narrowed down to specific sub-questions. In the above example, question "what should be done to improve sales?" can further be divided into:

- How to improve efficiencies to reduce cost?
- How to increase sales to increase profit?

Management questions give directions to the research. These are useful in exploring the system to find the root cause of the problem.

Research Questions/Issues After having clear understanding of the objective and the management questions, a researcher translates the management question into research question. Management questions generally explore all the possibilities for the solution of the problem. Some of the possibilities may not be relevant for the problem under consideration. Out of the several management questions, only few are selected by the researcher for further analysis. There should be understanding between the management and the researcher (if they are different) while selecting the research questions for further investigations. The questions chosen should address the management dilemma and achieve the objective.

Research questions are more specific than the management questions. The selected management questions are examined in-depth for further investigation.

In the above example, for the management question "How to increase efficiency by reducing costs", the research questions could be:

- How can we reduce employee cost without affecting the output?
- How can we reduce the logistic cost without affecting the functions of retail outlet?
- How can we reduce infrastructure cost?

Investigative Questions/Issues The next level of the questions' hierarchy is the Investigative questions. These questions disclose specific information that is useful to answer the research question. Also, these are more specific than the research questions. Further, these questions need to be satisfactorily answered to arrive at conclusions and lead to selecting appropriate research design.

In the above example, the research question, "How can we reduce employee cost without affecting the output?" can be divided into specific investigative questions like,

- How to have appropriate balance of permanent and temporary/casual employees to improve efficiencies?
- How to improve output of the employees without paying overtime/bonus, etc.?
- What is the best compensation policy that can be implemented?

These questions are the building blocks for the hypothesis development, discussed in the next section.

Measurement Questions/Issues These questions allow the researcher to collect specific information required for the research study. For each investigative question, the measurement questions are asked. These questions form the building blocks for designing the questionnaire.

For the above example, the measurement questions could be:

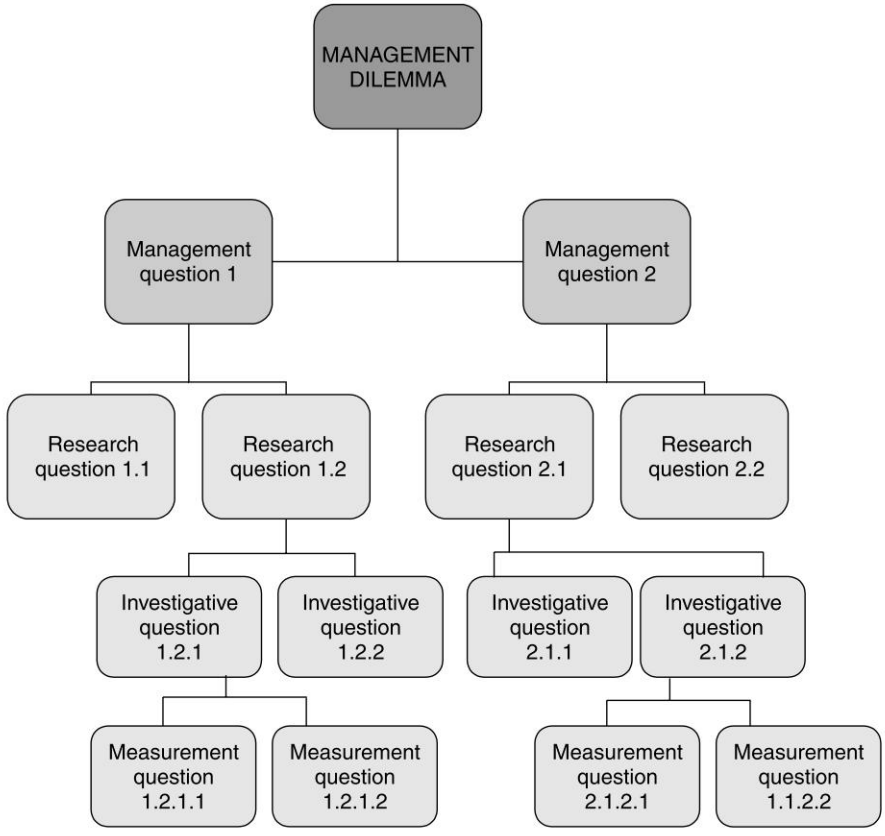
- How to measure efficiency of employees?
- How to measure the output of employee?
- How to measure productivity?
- What information about the employees is required?

The hierarchy of questions could be represented in the form of a diagram given on the next page.

3.4 HYPOTHESIS DEVELOPMENT

Developing or setting up a hypothesis for a research study is quite critical as the ultimate conclusions that are drawn would depend upon hypothesis that was set up for the study.

Questions' Hierarchy



3.4.1 Definition and Wording of a Hypothesis

The Oxford dictionary meaning of hypothesis is "A proposition or suggestion made as the basis for reasoning or investigation".

In the context of research methodology, it implies:

"Educated or informed guess about the answer to a question framed in a particular study."

An important pre-requisite before finalising a hypothesis is that it should be thoroughly deliberated upon so as to meet the objectives of the research study.

The hypotheses should be worded in a manner that it can be tested. Some of the examples are:

- Training improves the performance of sales personnel
- Advertisements improve the sales of a product/service
- Average I.Q. of a batch of MBA students at a management institute is 110.
- Batteries of brand A are better than brand B
- The price of a stock is an indication of the company's turnover
- The value of dollar is dependent on the stock market index.
- The Mumbai Stock Market (BSE) and New York Stock Market (NYSE) move in tandem

In the context of BRM, hypothesis could be defined as a proposition formulated for empirical testing of a descriptive statement that describes

(i) The value of a parameter representing a variable
(ii) Relationship between two or more variables or attributes

Hypotheses are an integral part of many of the research studies. Their role in a study is to:

- Authenticate or strengthen the confidence in any conclusion that could be drawn from the study
- Help in deciding upon the type of research design to be used
- Help in drawing up a sampling plan
- Help in deciding the issue relating to collection of data

Before, we deliberate further on hypothesis, we would like to explain the concept of 'population' in the context of BRM.

A population is a set of individuals, items, units, entities, etc.

Some examples are:

- Students at a management institute
- Cars manufactured in a factory
- Houses in a colony
- Companies e.g. those included in SENSEX of Mumbai Stock Exchange

A population is characterised by one or more number of variables. For example, human population is characterised by income, religion, nationality, etc. An item like electric bulb is characterised by 'life', wattage, etc. Students in a management institute are characterised by age, work experience, area of specialisation, etc. Companies are characterised by market capitalisation, P/E (price to earning) ratio, etc. Each characteristic is represented by a variable.

In business research, each characteristic like sales, profit, EPS (earnings per shares), etc. is represented by variables which are amenable to measurement.

In consonance with the dictionary meaning, hypothesis in the context of research methodology is an assumption or statement about

- A characteristic of a population
- Two or more characteristics (like background and specialisation of MBA students) of a population
- The same characteristics of two or more populations
- The same variable(s) of two or more population which can be tested to be true or false.

3.4.2 Types of Hypotheses

There are various types of hypotheses depending on the type and purpose of a research study. These are as follows:

- Descriptive
- Relational
  - Correlation
  - Causal or Explanatory
- Directional
- Non-directional

A brief description of all the above types of hypotheses is given below.

3.4.3 Descriptive Hypotheses

Such hypotheses relate to an assumption or statement relating to a population. For example, average life of light bulbs manufactured by a factory is 2500 hrs. The percentage of new batch of students, joining a management institute, having more than one year of work experience is 20%. Average mileage of a brand of car is 15 km per litre.

Such hypotheses also postulate either difference between a variable of two different populations or of the same populations before and after some factor causes a change. For example:

"The average salary packages offered to the students of two management institutes are same"

or

"Has the performance of the employees improved after introducing a new wage revision policy."

It may be noted that descriptive hypotheses and descriptive studies are different parts of a research study.

3.4.4 Relational Hypotheses

Such hypotheses are concerned with studying or analysing relationship or correlation between two variables (characteristics) or more than two variables (characteristics) of a population. Thus, such hypotheses are of two types viz. Causal or Correlational. These are explained as follows:

3.4.4.1 Causal or Explanatory Hypotheses Such hypotheses encompasses situations where we study the impact or influence of one factor (represented by a variable) on some other factor (represented by another variable). The variable which causes or influences change is called independent, causal or explanatory variable and the other variable which gets influenced is called dependent variable. For example, use of fertilizer influences yield of a crop; therefore, fertilizer is the independent variable and yield is the dependent variable. Similarly, we may study impact of advertising expense on sales turnover of a company, and draw conclusion like "Change in sales turnover is caused or explained by change in advertising expenses.

3.4.4.2 Correlational Hypotheses Such hypotheses are used when we want to test whether there is any correlation between two variables. Some examples are:

- Return on a stock and return on BSE Sensex/NIFTY
- Marks in entrance examination and final MBA grade
- Insurance amount and family size of a policy holder
- Return on stocks and return on investment in gold
- I.Q. and marks obtained in entrance/final MBA exam

We may add that correlational hypothesis is also used to test association between two factors like academic background (B.Sc./B.Tech., etc.) and the area of specialisation opted by MBA students at a management institute. Another example could be to test whether the performance of cricketers (one day matches) is associated with the days of a week.

One example relates to a behavioural scientist who was conducting a survey (to determine if the financial benefits, in terms of salary, influences the level of satisfaction of employees, or whether there are other factors such as work environment which are more important than salary in influencing employee satisfaction). A random sample of 300 employees is given a test to determine their level of satisfaction. Their salary levels are also recorded. The information is tabulated as follows:

Level of Satisfaction	*Annual Salary (Rs. in Lakhs)*			
	Up to 5	5 – 10	More than 10	Total
High	10	10	10	40
Medium	50	45	15	110
Low	40	15	5	50
Total	120	60	20	200

3.5 DEVELOPING OR SETTING UP HYPOTHESES

Suppose, we want to test the statement that the average height of Class X students in a school is 165 cms. The given statement or the statement to be tested is designated as **Null Hypothesis**. The word Null is used because the nature of testing is such that we try our best to nullify this hypothesis on the basis of the sample collected, and if we do not find sufficient evidence from the sample to do so then we have no alternative but not to reject (to accept) it. The null hypothesis is represented by $\boldsymbol{H_0}$, H with zero as subscript.

This is analogous to a situation when the Police presents an 'accused' before the judge for prosecution. The judge starts with the presumption that the accused is innocent. The police collect and present relevant facts and evidences to 'nullify' the assumption of the judge. But if the police is not able to do so, they have no alternative but to accept the presumption of the judge that the accused is innocent.

Incidentally, both the judicial system and statistics concentrate on disproving the null hypothesis rather than proving the alternative hypotheses. While attempt is made to reject the null hypothesis on the basis of evidence but if the evidence is not sufficient to reject the null hypothesis, the null hypothesis is accepted.

After setting up the null hypothesis, one has to set up an **alternative** hypothesis i.e. a statement which is intended to be accepted if the null hypothesis is rejected. It is denoted by $\boldsymbol{H_1}$. In the above case relating to the height of students, the alternative statement could be that the average height is not 165 cms; these hypotheses could be written as

Null Hypothesis : $H_0 : m = 165$ cms.
Alternative Hypothesis : $H_1 : m \neq 165$ cms.

Obviously, one of the two hypotheses is to be accepted based on the calculations from the values obtained through the sample.

In general, the null and alternative hypotheses are set up as follows:

(i) Setting up the Null Hypothesis

It is in the form

$$H_0 : m = m_0$$

where m_0 is the value which is assumed or claimed for the population characteristic. It is the reference point against which the alternative hypothesis is set up, as explained in the next step.

However, sometimes, H_1 is set up first, and the form of H_0 is decided accordingly. This is explained below.

(ii) Setting up the Alternative Hypothesis

It is in one of the following forms:

$H_1: m \neq m_0$

or, $H_1 : m > m_0$

or, $H_1 : m < m_0$

One has to choose from the above three forms depending on the situation posed, as explained below.

In the example relating to the heights of students discussed in Section 11.5.3, the situation involved only testing the statement made about the average height of Class X students, and therefore, the alternative was of the form $m \neq m_0$.

In a situation relating to the life of TV tubes, a Production Manager wanted to test whether the average life was more than 10,000 hrs., and therefore the alternative hypothesis was of the form $m > m_0$.

In yet another example, the manufacturer of a reputed brand of cigarettes ordered that the nicotine content in a cigarette should not exceed the stipulated level of 30 mgs. In order to check this, he selected a sample of 200 cigarettes from the lot to be packed, and found that the average nicotine content was 28.5 mgs. Could he be reasonably sure that the stipulations laid down by him were being met? In this case, the manufacturer wanted the nicotine content to be less than a particular value, and therefore, the alternative hypothesis is of the form $m < m_0$.

Incidentally, it may be noted that the null hypothesis is in the form of an equation like $m = 10$, or inequality like $m \leq 10$ or $m \geq 10$. The alternative hypothesis, can, however, be in the form of either

not equal to ($\neq$), less than (<), or greater than (>).

The following table explains the difference between hypotheses developed for quantitative and qualitative data:

Table 3.1 Hypotheses for Quantitative and Qualitative Data

For Quantitative Data	*For Qualitative Data*
Average Waiting Time at a counter = 5 mins.	How do the customers perceive the waiting time: Acceptable or Irritating
What is the average service time at a counter	How do the customers perceive the waiting time: Good, Average or Poor
Average service time for the same service at Counter 'A' is more than that at counter 'B'	Does the staff at branch 'A' lack experience, training and motivation ?
What is the difference in the service times at counter 'A' and counter 'B'?	Is the difference in the service times at the two counters 'A' and 'B' is due to difference in types of customers?
The percentage of defectives in the first shift is less than in the second shift.	Does the performance in the second shift (3.00 to 11.00 p.m.) get affected due to fatigue as the day passes?
	Is the skill of the workers less or there is ineffective supervision in the second shift?

(*Contd*)

(*Contd*)

	Is the level of motivation among the workers less in the second shift?
Training method 'A' is better than the training method 'B' as reflected in improved sales performance of the salesmen trained by the two methods.	What makes training method 'A' superior to training method 'B'?

We have illustrated all the concepts associated with hypotheses development through a live illustration given below.

Illustration 3.1 Hypothesis Development

It was the last day of the interviews of MBA candidates of a management institute. At the end of the day all the faculty and external members of the interview panels had assembled in the conference room, and were eagerly waiting for the registrar to signal that all the interview related papers were in order so that they could disperse for the day.

While waiting, they shared some of their experiences of the interviews. Following are the remarks made by some of them.

"The quality of this year's students is very good"

"Yes I agree. This year, the performance has been better than last year"

"Yes I also feel, the quality of applicants is much better than last year"

"I feel, this year the students were well prepared"

"It could also be that now more and more talented students are opting for MBA rather than other professional courses"

"Or may be that many of them were engineers"

"But the engineers that we get are not from reputed colleges so how they could be better than B.Sc. and B.Com. students?"

"This year, the girls have fared much better than boys"

"The better quality could be due to the fact that many more students have work experience and thus they fare better in interviews"

"Yes, this year the work experience seems to be more"

"The number of engineering graduates is much more than last year"

At this time, the registrar remarked "A quick analysis shows that the candidates in the morning batches have scored better than the students in the afternoon batches"

He added "I overheard some candidates commenting that lady members of interview panel were more considerate than male members"

	Statement	*Hypothesis*	*Comment*
1	The quality of this year's students is very good	H_0: There is no difference in the quality of this year students and last year students H_1: This year students are of better quality than last year students	If quality can be quantified (marks in the graduation/experience, etc.), it is testable. Directional Hypothesis
2	This year, the performance had been better than last year	H_0: There is no difference in the performance of this year students and last year students H_1: This year students have performance better than last year students	If performance can be quantified (marks in GD/PI written test, etc.), it is testable. Directional Hypothesis

(*Contd*)

(*Contd*)

3	The quality of applicants is much better than last year	Same as 1	
4	This year the students were well prepared	H_0: There is no difference in the preparedness of this year students and last year students H_1: This year students are better prepared than last year students	Cannot be directly tested unless data is collected on preparedness but could be tested if relevant data is available
5	More talented students are opting for MBA rather than other professional courses	H_0: No difference in the students that are opting for MBA H_1: More talented students are opting for MBA	Cannot be tested unless 'talent' is defined in terms of some measurable variables and data is collected on the same.
6	Many of them were engineers	H_0: There is no difference in the proportions of engineers and non-engineers H_1: The proportion of engineers is more than non-engineers	Testable directly Directional Hypothesis
7	How engineers could be better than B.Sc. and B.Com. students?	H_0: Engineers are same as non-engineers H_1: Engineers better than non-engineers	Cannot be tested since 'Better' is vague unless defined properly and quantified, and hypothesis cannot be tested
8	This year, girls have fared much better than boys	H_0: There is no difference in the performance of girls and boys H_1: Girls have performed better than boys	If performance can be quantified (marks in GD/PI written test, etc.), it is testable. Directional Hypothesis
9	The better quality could be due to more and more students having work experience and so they face interview better	(a) H_0: There is no relationship between experience of students and quality H_1: More the experience better the quality (b) H_0: There is no relationship between experience of students and their performance in the interview H_1: More the experience better the performance in the interview	(a) If quality can be quantified, it is testable. Directional Hypothesis (b) Testable directly Directional Hypothesis
10	Yes, this year the work experience seems to be more	H_0: There is no difference in the work experience of students of this year and last year H_1: This year students have more work experience than last year	Testable directly Directional Hypothesis
11	The candidates in the morning batches have scored better than the students in the afternoon batches	H_0: There is no difference in marks of the students of morning batch and the evening batch H_1: Candidates in the morning batches have scored better than the students in the afternoon batches	Testable directly Directional Hypothesis

(*Contd*)

(*Contd*)

12	Female members of interview panel were more considerate than male members	H_0: There is no difference in female and male panel members H_1: Female panelists are more considerate than males	Cannot be tested since 'Considerate' is vague unless defined properly and quantified, and hypothesis cannot be tested

3.6 RESEARCH PROPOSAL

The next stage after developing the hypothesis is to prepare a research proposal for submission to the management. The management, on its part, may decide to conduct the research either within the organisation or engage the services of an individual consultant or a consultancy firm. However, before we discuss about the management's role, we discuss the various aspects of a research proposal.

A research proposal is an individual's offer to the company or a company's offer to some outside agency like Government, to investigate the issues relating to product(s), service(s) or any other aspect relating to human resources, systems or organisations. Broadly speaking, a research proposal encompasses the methodology of conducting the research to solve the formulated research problem. The broad structure of a research proposal is as follows:

(i) Executive Summary

This is to serve the purpose of appreciation of the problem and the associated issue(s) by the management.

(ii) Problem Details

In this section, genesis of the problem, impact of the problem, risks and benefits, etc. are regenerated with managerial perspective. A reference should also be included about a similar problem either faced earlier in the same company or in another company and then explain how the company resolved the problem. Reference to relevant existing literature on the problem may be listed.

(iii) Strategy for Solving the Problem

As outlined earlier, various strategies for solving the problem are, suggesting varied approaches to solve the problem, estimating the resources (this is more relevant if the study is to be conducted within the organisation), recommending actions for the respective options of conducting the research within the organisation or engaging an external agency, conducting comparative evaluations including cost-benefit of various options, etc.

(iv) Details of Research Design

Various components of research design including sampling design, data collection, measurement instruments, and type of analysis are to be indicated.

(v) Allocation of Resources

Resources required/needed in terms of money, men, machine and material should be indicated. Resources of special nature may be specified.

(vi) Time Schedule

Time schedule based on PERT Chart may be drawn.

(vii) Cost-Benefit Organisation

Cost-benefit of the proposal including intangible costs/benefits should be evaluated.

3.6.1 Objective/Purpose of Research Proposal

The objectives for preparing and submitting a research proposal are as follows:

(i) Specifies research question/issue and indicates its importance
(ii) Helps in understanding the salient aspects of the research project without going into details.
(iii) Indicates the broad methodology and plan to conduct the study.
(iv) Indicates the type of interpretation that might be made from the study and thus assessing the importance as well as efficacy of the study.

3.6.2 Preparing a Research Proposal

While referring to the past research, one has to be totally objective and eschew any iota of past bias. Only facts and figures are to be provided without any comment. Based on these facts and figures and the developments that have taken place or those that are being envisaged, one has to justify the relevance and the benefit of the proposal. The emphasis has to be on 'Improvement' that is visualised in future rather than 'Criticising' what happened in the past. In fact, one could endeavour to appreciate the past efforts and results, while indicating scope for further improvement in quality or expanding and creating new products/services/systems. Use of such appreciative phrases makes the proposal as 'constructive'. One has to remember that the strategies, their relevance and effectiveness keep on changing with time; therefore the emphasis has to be on improvement rather than on criticism.

It is said that a research is as good as one's proposal.

A good quality proposal, in addition to the increased chances of acceptance by the concerned authorities also creates a good impression as well as establishes credibility of the researcher. It is, therefore, necessary to put in best efforts to ensure high degree of acceptance of the proposal and its smooth execution.

A research proposal serves the purpose of convincing that the research is worthwhile and the researcher has the requisite competence and ability to complete the project as per schedule.

It should reflect good grasp of various issues related to the topic supported by survey of relevant literature. Accordingly, it should answer the following questions:

- What is the objective to be achieved?
- What is its relevance and importance?
- What is the methodology to be used?
- What is the plan and schedule of completion?
- What are the scope and limitations?
- What is the extent to which the objectives might be achieved?

3.6.3 Components of a Research Proposal

The components or items of a proposal are as follows:

(i) Title
(ii) Preamble or Summary
(iii) Introduction including Literature Review
(iv) Suggested strategy/approach

(v) Methodology and resources
(vi) Discussion on the findings and their applicability
(viii) Epilogue–scope and limitation

A brief write-up on the above items is given as follows:

3.6.3.1 Title The title should indicate the gist or theme of the research. However, it should be catchy, and should instantaneously arouse the curiosity and interest of the reader. This would induce further reading of the proposal with favourable disposition towards the proposal.

3.6.3.2 Preamble or Summary It should be stimulating and engaging the mind of the reader. It should describe in brief the salient aspects of the entire proposal, including the strategy and methodology. Various components or items of the methodology, including design, collection of data, instruments to be used, data analytical tools to be used should be mentioned along with their justification.

3.6.3.3 Introduction It should start with the reference to the previous researches carried out on the subject, indicating the need for carrying out the research further to take care of their limitations or increase their scope. Of course, sometimes, the research topic may be path breaking and might attempt towards totally new dimension. In general, the following items should be included in this section:

(i) Literature Review
(ii) Relevance and Need for Further Study
(iii) Issues and the research problem for study
(iv) Variables included in the study
(v) Hypothesis or theory
(vi) Type of research study like exploratory, descriptive, etc.

3.6.3.4 Methodology and Resources It should outline the details of the entire process of carrying out the study. The justification for using a particular method should be provided. It should generate a sense of assurance that the methods are adequate and in consonance with the purpose of the study. For quantitative studies, the typical components of the methodology are:

- Design
- Sampling Design
- Instruments (like Questionnaire, Interview, etc.)
- Data Collection
- Statistical Analysis

Based on the aforementioned components, the resources are to be estimated and provided in the proposal.

3.6.3.5 Discussion on the Findings and Their Applicability While discussing the findings, their specific applicability to the issues raised in the research problem should be indicated.

3.6.3.6 Epilogue The highlights of conclusions along with the scope and limitations could be included. An outline of further research on the topic could also be indicated.

3.6.4 Research Proposal – A Sample

A sample research proposal submitted by a consulting agency to an organisation is given in the Appendix in this chapter. It may be appreciated that the consultants while following typical textbook approach, develop their own approaches depending upon the assignment.

3.7 REQUEST FOR PROPOSAL (RFP)

The need for floating Request for Proposal for a project arises when the promoter or sponsor of the project wishes to seek certain details from interested vendors. It is usually floated for big projects like the 'Sea Link', in Mumbai, or for projects which are quite technical in nature.

We may add that the system for RFP started from setting of big projects by government organisations. Before proceeding further, we would like to demystify this term by reproducing two RPF advertisements inserted by Government of India in the newspaper. One RFP relates to International Crafts Complex at New Delhi, and the other relates to Chicken Dressing Plant at Delhi.

Request for Proposal (RFP) for International Crafts Complex at New Delhi

In continuation to earlier Request for Proposal [RFP] called for International Crafts Complex published in the same newspaper dated 15.10.2008 & 23.1.2009, Revised Bids are being called, from reputed developers of Design, Finance, Build. Operate, Manage & Maintain state-of-the-art International Crafts Complex (ICC) at Vasant Kunj (near Nelson Mandela Marg and Hotel Grand), New Delhi, An area of **7153** sq. mtrs. has been earmarked for the project located at Vasant Kunj, New Delhi.

The selected bidder will enter into a Concession Agreement which will specify the terms and conditions governing the implementation of the Project. The successful bidder will have the obligation to construct, operate and maintain the ICC within the applicable development controls.

Project Configuration

The Project Site will be provided for the purposes of the development of the International Crafts Complex. This would be set up with the overall objective of providing facilities and platform for the promotion of the Handicrafts sector (including hand-knotted carpets) besides promoting and supporting artisans and entrepreneurs working for this sector. The Concession agreement would include:

- Hosting handicraft promotion exhibitions, and measures for promotion of handicraft artisans.
- Hosting handicraft promotion exhibitions, and measures for promotion of handicraft artisans.
- Providing facilities of international standards for handicraft goods oriented shops.
- Providing spaces of international standards for handicraft exporters and organizations related to the handicrafts sector.
- Providing facilities of international standards to member of the general public/tourists for shopping for handicraft products and
- Interacting/networking with the handicraft industry and providing all associated amenities (including restaurants, adequate parking, etc.) necessary for making such a complex and attractive, foreign tourists and members of the general public.

Bidding Process

The bidding process provides for two steps for selection of successful bidder, first of eligibility determination and second of techno-financial evaluation of the proposal submitted by the bidder. Detailed terms and conditions of the bid process are specified in the RFP document.

Issue and Receipt of Bids

RFP document can be obtained from EPCH's office at EPCH House, Pocket 6 & 7, Sector-C, LSC, Vasant Kunj, New Delhi-110070 on all working days from **7.4.2009** [**TUESDAY**] **from 11.00 am to 5.00 pm on** payment of Rs. 10,000/-. The payment shall be in form of Demand Draft from any Nationalised/Scheduled Band drawn in favour of 'Export Promotions Council for Handicrafts' payable at New Delhi. The RFP document can also be downloaded from www.epch.com subject to condition that the payment for the RFP document shall be made along with the bid. **The bidders those have paid earlier vide advertisement dated 15.10.2008 & 23.1.2009, can obtain the REVISED RFP documents on presentation of receipt earlier issued.**

The sealed envelope containing the bids should be superscribed as 'Bid for Development of International Crafts Complex at New Delhi' Bidders' Proposals in response **to RFP should reach EPCH at the address given below on or before 1400 hours (IST) on 28.4.2009 [TUESDAY].**

A pre-bid conference will be held on 17.4.2009 [FRIDAY] at 11.30 am hours in the committee Room of the Office of the Development Commissioner Handicrafts at West Block, 7, R. K. Puram, New Delhi-110 066.

These bids are being invited on behalf of Ministry of Textile, Govt. of India.

Address of submission of Bids:
Mr. Rakesh Kumar, Executive Director
Export Promotion Council for Handicrafts
EPCH House, Pocket 6 & 7, Sector-C, LSC, Vasant Kunj, New Delhi 110 070
Tel.: +91-11-26135256 Fax +91-11-26135518/19
Email: epch@vsnl.com www.epch.com

Source: *Economic Times,* Mumbai, Apr 4, 2009; Section: Business and IT, Page 8.

In the context of research projects, RFP describes the problem, the context in which it arose, the recommended approach for solving the problem, the amount to be paid for the assignment, etc. The formal document for the project is prepared, and issued for competitive bidding by the interested agencies. A typical format of the document, called 'Request for Proposal', is as follows:

(i) **Background**: Overview of the organisation
(ii) **Research Project (Problem)**: Overview of the research project and general expectations like time schedule, etc.
(iii) **Vendor Information**:
  (a) General profile of the bidding company
  (b) Record of previous experience in conducting such type of research

Request for Proposal for Setting up of
CHICKEN DRESSING PLANT
on Design-Build-Operate-Transfer (DBOT) Basis

DELHI AGRICULTURAL MARKETING BOARD (GOVT. OF NCT OF DELHI)
proposes to set up a chicken dressing plant at Poultry Market Gazipur on National Highway 24 to make available hygienic poultry meat to the population and to minimize human handling of birds. Capacity of plant varies from 40,000 birds in a single shift to 80,000 birds in double shift and can be expanded to 80,000 birds in single shift and 1,60,000 birds in double shift.

SALIENT FEATURES OF THE PROJECT

- **Most eco-friendly plant in the country. CNG node behind the plant site.**
- **Arrivals at mandi exceed 1,20,000 live chicken per day.**
- **Current demand in Delhi/NCR region approx 3,50,000 chicken/day.**
- **Supply of hygienically dressed chicken in existing institutional markets (50,000 eating joints, 70 four & five star hotels, Airlines and railways etc.)**
- **MCD Law prohibits sale of live and slaughter of chicken at shops in Delhi region. Chicken meat from authorized slaughter plant can only be sold from licensed meat shops.**
- **Grant of Rs. 15 crores from Ministry of Food Processing Industries to successful bidder is also available.**

The successful bidder will run the first shift as service slaughter facility for traders & buyers at predetermined tariff of Rs. 9 per bird and the second shift for himself to recover investment & to make profit.

Delhi Agricultural Marketing Board invites Request for Proposal (RFP) from agencies/companies/firms/ consortia/JV on Design Build Operate and Transfer (DBOT) basis for a period of 21.5 years (i/c 18 months of construction period)

For the above proposal DAMB will provide: -

(i) About 7.0 acres of land in the existing Poultry Market at NH-24 near Delhi-UP Border on mutually agreed terms.

(ii) Assistance in getting approval of drawings from DDA, DUAC & Local authority.

The Details of the eligibility criteria, Pre-bid meeting date and other terms and conditions are available on the website and in the RFP document which may be obtained from the office of Project Engineer II, at the address given above from 8.4.09 to 21.4.09 during working hours. The RFP document may also be downloaded from the websites: **www.delhigovt.nic.in** and **www.delagrimarket.org.**

The last date for submission of RFP 16.06.09 upto 3.00 p.m.

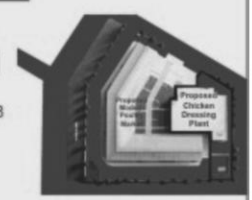
Delhi Agricultural Marketing Board (DAMB) reserves the right to reject any or all RFP without assigning any reason thereof.

Estimated Cost	:	Rs. 73.86 crores
Earnest Money	:	Rs. 1.48 crores
Cost of RFP Document	:	Rs. 5000.00

Project Engineer-II

Delhi Agricultural Marketing Board

(Govt. of NCT of Delhi)
9, Institutional Area, Pankha Road, Janak puri, New Delhi-110058
Tel:- 011-28522281, 28525232 Fax:-011-28521871
E-Mail: ambdl@nic.in, pe2damb@yahoo.co.in
Website: www.delagrimarket.org



NIT No. 13-DAMB/Engg/2009

www.delagrimarket.org/103-2009

www.delhigovt.nic.in/103-2009

(c) Proposed strategy for conducting the research includes research design, human resources to be deployed, time schedule, etc. It may also include seeking support from within the organisation in terms of manpower, internal team composition, infrastructure support, etc.

(d) Implementation Schedule in case of acceptance of the bid. For example, if some specific information system is to be set up in the organisation, then how much time it would take?

(e) Cost Proposal

(f) Statutory and Regulatory commitments or/and obligations for the vendor.

In a typical RFP, the format of information to be supplied by the company that floats the proposal, and the format in which the vendor is to supply the information are as follows:

From the Interested Company

- An Overview of the Company
- Details of the Project
- Details of All Approvals and Clearances from Regulatory Authorities
- Estimated Cost of the Project, including Hardware and Software for Information Technology, content of the project – with clauses for installation and overrun
- Ethical Issues containing *inter alia* the clauses for confidentiality and privacy

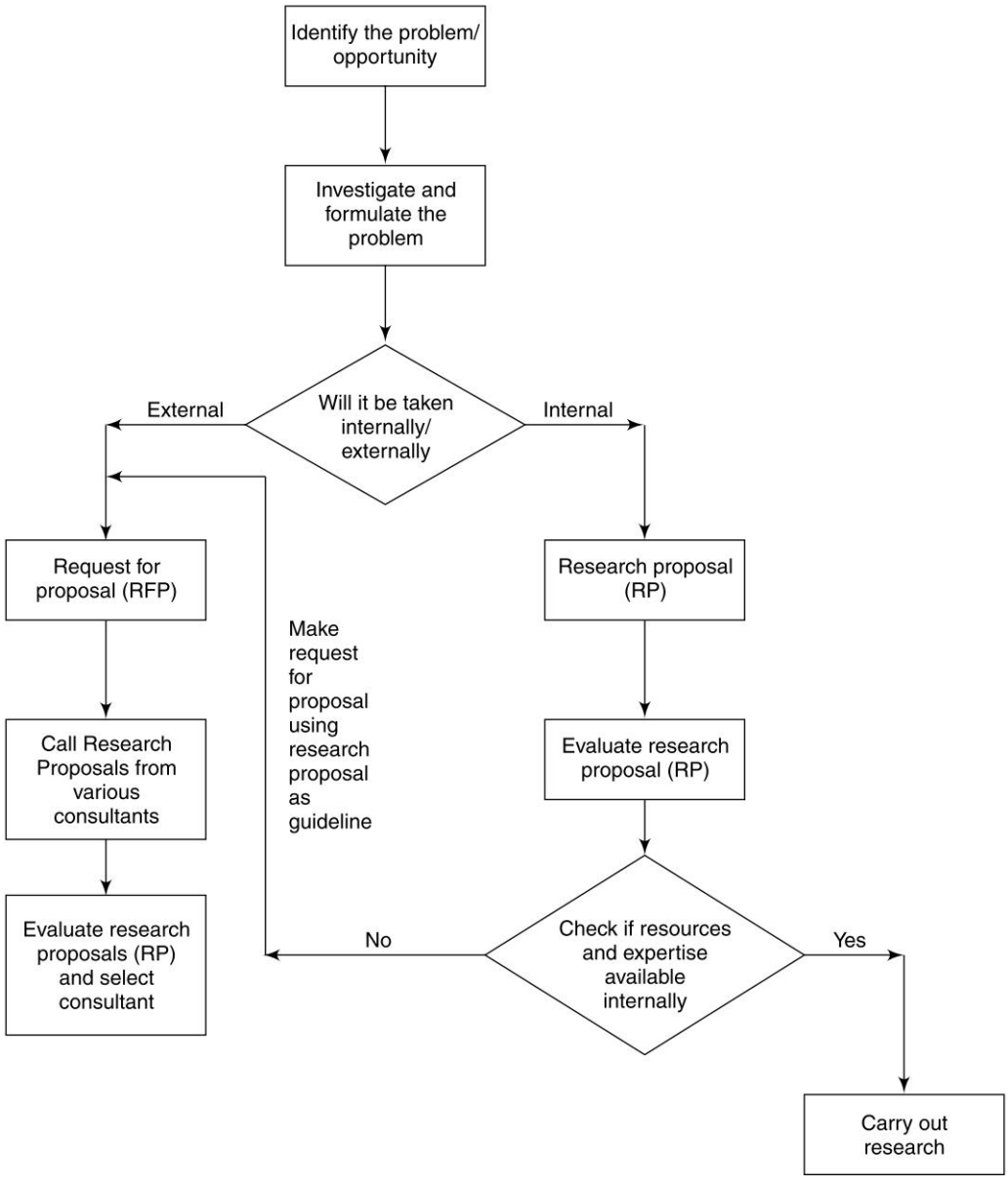
From the Vendor Side

- Profile of the Company
- History and Description
- Legal Status and Constraints, if any
- Partners and Alliances to Participate in the Project
- Proposed Approach/Solution
- Services and Support
- Cost Proposal–Pricing of Resources including men, material and machines
- Time Schedule (PERT CHART)
- Contractual Terms and Conditions for Implementation and Maintenance

So far, we have discussed conduct of business research in an organisation. However, due to liberalisation policy and global integration of economies, an organisation is no longer insulated from other organisations, especially those of the same sector like automobile, pharmaceutical, etc. Therefore, while conducting research in an aspect, we also have to conduct research for that sector. What is happening in the sector as a whole is bound to impact other organisations. This leads us to the need for discussing certain salient aspects of research both at corporate (organisation) and sectoral levels.

3.8 FLOW CHART FOR CONDUCTING RESEARCH

Before deliberating on various facets of internal and external research, a flow chart explaining the decision process relating to conduct a research study is given as follows:



3.9 INTERNAL AND EXTERNAL RESEARCH

Internal research is conducted by a team of in-house experts of the organisation. External research is conducted for an organisation by an outside agency like a consultant, consultancy firm or a professional like a faculty in an academic institution.

Whenever in an organisation a problem needs to be solved or a research study is considered desirable, decision needs to be taken on whether the research should be conducted internally or through

an external agency. Both strategies have certain advantages and disadvantages. In this section we deliberate on the pros and cons and other relevant issues.

3.9.1 Internal Research

Even before taking a decision on engaging an external agency, it may be advisable to first give it to an internal team for preliminary investigation and offering suggestions. This approach is especially suitable if the nature of problem or study is internal i.e. specific to a particular department and needs urgent attention. The advantages of internal research are as follows:

- The internal team is well-versed with the environment, systems and procedures, culture, etc. in the organisation.
- The team may get a quick grasp and comprehend the various aspects of the problem/study.
- The problem is solved quickly. Of course it is assumed that the requisite, competent and skilled team is available within the organisation.

An essential criterion for the success of the internal team is their credibility and acceptability. The acceptability criterion is all the more relevant if the problem/study involves issues relating to staff and systems and procedures or some other internal matter.

The following case study from an organisation illustrates this point:

CASE 3.1 CREDIBILITY OF INTERNAL TEAM

An all India level company had a department called 'Organisations and Methods'. The role of this department was to bring out or facilitate suitable changes within the company to function efficiently in the ever changing environment. The department was doing reasonably well and had established some credibility within the company. However, all changed when management filled the vacancy of the Head of the Department, with a senior executive who had been overlooked for promotion for the last three years. This had resulted in lowering of his image in the organisation. This move reduced the credibility of the department as any of its move/suggestion was perceived as that of the executive whose reputation had gone down considerably because of being overlooked for promotion. Subsequently, he offered some highly useful suggestions relating to improvement of systems and procedures within the organisation, with informal guidance of one of his friends who happened to be a professor of management in a management institute: but they were all ignored. The repercussions were soon noticed in the company. To get over this impasse, the Chairman adopted the strategy of handing over the additional responsibility of this department to one of the highly successful General Managers. However, he retained the ignored executive as number 2 in the department and conveyed a strong message to him that it was his last chance to justify his existence in the company. It so happened that the General Manager who was individually appreciative of the suggestions offered by the ignored executive, reworded those suggestions with some modifications, and circulated them under his signature. Surprisingly, most of these suggestions were accepted and even implemented yielding encouraging results.

The above case leads to the importance of having qualified and competent team of researchers who have credibility in the organisation and command respect for their professional approach. In the organisations where formal organisational setup exists for researchers, two approaches are followed viz.

- Centralised Research
- Decenteralised Research

3.9.1.1 Centralised Approach In such a setup there is an exclusive department devoted to research activities. All the problems that are perceived or arrive at any branch, region or corporate level are referred to this department which is generally equipped with adequate expertise to undertake the assignments. The researchers in this department must have multidisciplinary competence with the exception of some specialists in relevant areas.

3.9.1.2 Advantages of Centralised Research

(i) Research is not duplicated. The research done for one department/region/branch could be used for others with suitable modification.
(ii) Limited number of experts are required in the central department.
(iii) Experience gained by researchers in one department can be put to use in the other departments.
(iv) Researchers follow unbiased approach to all departments.
(v) The staff of the department can have free and frank interaction with the researchers.

3.9.1.3 Disadvantages of Centralised Research

(i) Researchers are far from the scene of action and may not be able to comprehend a true picture of the situation.
(ii) Researchers may not have thorough knowledge of systems and procedures of all the departments.

3.9.1.4 Decentralised Approach

In such a system, each department has its own research team to take care of its research requirements that arise from time to time.

3.9.1.5 Advantages of Decentralised Research

(i) Researchers in an individual department are well-versed with the systems and procedures of the department as also the behavioural aspects of the staff, so chances of success of the research increases.
(ii) Researchers are nearer to the problem area and could have direct feel of the situation.
(iii) Researchers can easily try out various solutions before selecting the most suitable alternative.
(iv) It is easier to develop expertise in one particular area than to develop expertise in several areas.

3.9.1.6 Disadvantages of Decentralised Research

(i) The researchers may not have a detached approach. It may instead be in consensus with various other sections of the department and its personnel, and at times may have certain preconceived notions.
(ii) Researchers may overlook shortcomings of an individual or a system in the department for the sake of its reputation.
(iii) Researchers may have only limited exposure, and thus the solution may be limited in approach.
(iv) Employees may hesitate in having free and frank interaction, especially if their views are not shared by their bosses.
(v) A department may try to camouflage anything going wrong within the department, and might influence its researchers. Similarly, a claim made by it may not be easily verifiable.

3.9.2 External Research

The external research is conducted either due to necessity or as a matter of choice. This research is suitable for smaller companies which do not have the requisite expertise with reference to the problem faced by them or exploit the opportunity available to them.

Sometimes, even a part of the entire project could be given to an external agency with exclusive expertise which is relevant for that particular module/portion.

If sometimes there are conflicting views in the organisation, as a strategy to resolve conflict within an organisation, the management may award a project to an outside agency.

3.9.2.1 Advantages of External Research Some of the specific advantages of having external research are described as follows:

(i) Availability of Expertise

These research agencies have general expertise and also expertise in various aspects of business like, Finance, Marketing, HRD, Organisational Structure. They are well-groomed for conducting research in a professional manner.

(ii) Pool of Expertise

They have a pool of expertise, and experience of successful and also not so successful companies. Such experience goes on increasing as they conduct research for more and more companies. They also know the problems faced in conducting various types of researches and how to overcome or even avoid the hurdles. Such type of experience is usually not available within individual companies.

(iii) Fresh Perspective and Thinking

Many a times, individuals or teams of persons working in an organisation, develop certain concepts and notions or even mental blocks, and they continue to look at a problem in the same manner; their vision just does not go beyond a certain limit. This is a natural phenomenon, and the individuals and team members should not to be blamed. It has happened with even those with the best of caliber. However, when an outsider comes and looks at the same problem, he is able to develop a fresh new perspective of the problem or a new approach to solve the same problem.

(iv) Acceptability

Sometimes, a company's personnel have the solution(s) to problem(s) but either they are hesitant to offer the same as it may not be palatable to the management or even if they offer, the management may not pay the due attention. It is ironical but it does happen that if the same solution is offered by the outside agency, in a different way with its own logic, it is accepted by the management.

(v) Quality

Usually, the external agency's representatives discuss various issues with the company's staff, and thus are able to get their frank, free and considered views which help the representatives to frame their own views in a much better perspective.

(vi) Broaden the Scope

A distinguishing feature of external consultants is that, sometimes, they help a company to broaden the scope of the perceived problem to include other relevant aspects.

3.9.2.2 Limitations of External Research Along with the advantages, there are certain limitations of external research, which are as follows:

(i) The external team may not be well-versed with the environment, systems and procedures, culture, etc. prevalent in the organisation.
(ii) They may take time to grasp and comprehend the various aspects of the problem/study.
(iii) The selection of external agency plays a crucial role, and has to be done carefully to ensure credibility and acceptability of their research in the organisation.

Therefore, a comprehensive list of criteria for the selection process is as follows.

3.9.2.3 Criteria for Selection of an External Research Agency The following criteria may serve as a guide to select an external agency for assigning a research assignment.

(i) Reputation—Overall
The overall reputation of an agency in the market is an important criterion for selection.
(ii) Record of completing assignments in time
This information is also easily available. However, it has to be ensured that delays, if any, were preliminarily caused by the agency. Sometimes the delays take place because of non-cooperation from the clients or because of the factors on which the agency had no control.
(iii) Credibility in maintaining ethical standards
The agency should have the reputation of maintaining privacy and confidentiality of the assignments undertaken by them.
(iv) The agency should have such flexibility and this is dependent on its competence and experience in diversified areas. This ability develops over a period of time, having varied experience and developing maturity.
(v) Quality of past assignments
This can be verified by interacting with the earlier clients of the agency.
(vi) Experience – Overall experience and/or experience in similar assignments
This is very important for ensuring the quality of report by the agency.
(vii) Quality of Staff
The agency should have relevant qualifications and experience/expertise including communication skills.
(viii) Sharing of ideology and value systems
Through intense interaction, it has to be ensured that the company and the external agency share similar ideology and value system.

3.10 SPONSORING RESEARCH

Now, we shall look upon the research assignments/projects from another perspective i.e. the sponsoring aspect. All research studies are sponsored in one way or the other. Various types of sponsorships are mentioned below.

3.10.1 Management Sponsored Research

Such research is sponsored by either the top management of an organisation or by the Head of a Group/Department

- To be carried out within the organisation
- To be carried out by an outside agency

3.10.2 Individual/Group Sponsored Research

Such research is sponsored by an individual researcher or by a group of researchers and presented to the management outlining the genesis, and thus justifying the cost vis-a-vis benefit. Such research is also called solicited research.

3.10.3 Faculty Sponsored Research (Applicable for Management Institute)

Such research can be of two forms:

(i) Faculty may suggest the research study to the Institute, which it wants to pursue.
(ii) Faculty may assign research assignments to be carried out by the students.

3.10.4 Government/Corporate Sponsored Research

It is the research awarded to academic institutions by the Government bodies or corporate entities. In USA, most of the research needed in Government bodies like Defence, and corporate bodies like Boeing is sponsored to the universities and other academic institutions. This trend has also set in India.

3.11 RESEARCH AT CORPORATE AND SECTORAL LEVELS

Research can be conducted either for the issues related to an individual organisation/corporate entity/enterprise like Reliance Communication, ICICI Bank, Hindustan Lever, etc., or it can be conducted for a sector like Banking, Telecom, Retail, etc. comprising a group of companies carrying out similar business.

The studies conducted for an individual entity are specific to that organisation, and are, in general, narrow in scope limited to the needs of that organisation. For example, for a bank, the study might relate to finding ways and means for making home loan portfolio more profitable. In the case of sectoral study, the issue for study is more general. For example, Reserve Bank of India might like to study the issues relating to home loan with the objective of making home loan cheaper by all the banks. For this purpose, the Reserve Bank might like to study cost of funds i.e. rates for deposits, risk factor i.e. non-payment of dues by the borrowers, national housing policy, etc.

Following are the typical steps or stages for conducting a sectoral study:

(i) Collection and analysis of national and international level information about the sector as well as any other sector that might have impact on the sector under study
(ii) Review of legal provisions which might be relevant for futuristic growth/development of the sector
(iii) Interaction with the leading companies in the sector
(iv) Online survey and interview—on the basis of questionnaire/schedule developed for the purpose
(v) Analysis of data–both qualitative and quantitative
(vi) Prepare a futuristic scenario for the sector, critical review of the past developments and performance, discerning trends and relationships, etc.

Incidentally, the list of sectors in India is as follows:

Automobile	Banking and Finance	Insurance
Reality	IT	Travel and Tourism
Pharmaceutical	Hospitality	Cement
Retail	Oil and Gas	Coal
Electricity Generation	Mining	Gems and Jewellary
Garment	Aeronautics/Airline	Health care
Shipping	Telecom	Transport

3.12 GUIDANCE FOR GOOD BUSINESS RESEARCH

There are no specific criteria laid down for evaluating the quality of a research study. Basically, the quality is assessed by the user of the research report to the extent that his objective is served. However, the following guidelines are generally used for assessing the quality of a research.

Table 3.2 Guidance for Good Business Research

Objectives are clearly defined	Brief statement of business objective that a researcher is seeking to accomplish (e.g. increase revenue, reduce cost, improve customer satisfaction, etc.)
Translation of objectives to hierarchical questions	On the basis of genesis of a problem, the researcher identifies manager's perception of the problem, and formulates the research problem
Research process detailed	The researcher defines the complete research process with inputs/outputs at each stage and where appropriate, an illustration of the process
Research design thoroughly planned	• Need for the exploration is assessed and type of exploration described • Exploratory procedures are outlined • Sample unit is clearly described along with sampling methodology • Questionnaire is designed in accordance with the required data and desired analysis requirement • Data collection procedures are collected and designed
Deviations revealed	• Desired procedure is compared with actual procedure in report • Desired sample is compared with actual sample in the report • Impact on findings and conclusions is detailed • If the deviations are likely to significantly reduce confidence in the conclusions, then additional steps are taken to remove those limitations
Findings presented unambiguously	• Findings are clearly presented in words, tables and graphs • Findings are logically organised to facilitate reaching a decision about the manager's problem in line with the original objectives of the research project • Executive summary of conclusions is outlined
Conclusions justified	Decision-based conclusions are matched with detailed findings
Researcher's experience reflected	The researcher ensures reflection of expertise through the quality of the report

Source: Based on the table given in the book titled "Business Research Methods" by David R Coopers and Pamela S Schindler, published by Tata Mc Graw-Hill.

3.13 A CONSULTANT'S APPROACH TO PROBLEM SOLVING

The intelligence of a consultant is like the power of a car engine but the problem-solving approach is the steering wheel that guides the car in the desired direction.

We conclude this chapter by describing an approach followed by some Strategy Consultants. The reason for including this in the book is that many times, a corporation's business managers have to play this role while carrying out an assignment given to them by the senior management. We use the word 'client' for the department, section, or group leader who has given the assignment. However, we would like to mention that this approach should not be interpreted as the only approach followed by consultants—a consultant starts with the textbook approach and adds his or her own ideas and experience to evolve the approach to suit the business problem.

In Strategy Consulting, problem-solving requires a structured way for thinking and communicating about problems. The approach can be defined as a series of key steps.

3.13.1 Define the Problem with a Crisp, Clear and Concise Statement

The problem statement should be phrased in such a manner that it is thought-provoking and specific. It should not be phrased in a manner that re-states an obvious fact or is beyond dispute or is too general.

Illustrative Examples:

Problem Statement	*Evaluation*
The company should grow revenue and become profitable.	Obviously!
The company should be managed differently to increase profitability.	Too general!
The company should grow revenue and profit by using distribution partners instead of its own salesmen.	**Ideal!**

3.13.2 Decompose Problem Statement into Key Issues

The problem should be decomposed into smaller problems or issues that are distinct from each other. To ensure that no key issue is missed, the issues should be **"Mutually Exclusive and Collectively Exhaustive"**. This is known as the **"MECE"** approach. Once the problem is decomposed, solving the key issues would lead to solving the main problem.

3.13.3 Analyse Key Issues and Outline Findings

(i) Breakdown Issues into Questions

Break issues into questions that need to be answered with fact gathering or data analysis.

(ii) Conduct Analysis for Each Question

Start collecting the facts or data to answer the question. In some cases, the question may involve framing a hypothesis that needs to be either proved or disproved. In subsequent chapters, we will delve much deeper into hypothesis testing. But for now, it is sufficient to understand that the analysis should help provide convincing rationale regarding the hypothesis. Not all hypotheses require analysing data–in some cases, it is sufficient to gather facts through interviews with key subject matter experts inside the company (e.g. business unit executives) or outside the company (e.g. industry analysts).

(iii) Analyse to Appropriate Level of Accuracy

When analysing data, it is easy to get trapped in what is often called "**Analysis Paralysis**", i.e. we keep trying to refine the analysis, get the data sample to be more comprehensive, etc. The risk

is that we may never finish the analysis in the timeframe expected for making a decision. So it is important to understand the level of accuracy needed in the analysis.

(iv) Understand Decision-Makers' Motivation

In conducting the analysis, it is also important to think about the *decision-makers* and their motivations. What might be their concerns and issues around the hypothesis? These could be based on past experience (e.g. they have tried changing their distributors before and have never been successful). Sometimes decision-makers are swayed by political concerns (e.g. Head of Marketing may not easily accept a solution that relies on putting more power in the hands of Head of Customer Care). It is important to recognise these issues and address them in the analysis.

3.13.4 Develop Recommendations

Once the hypotheses have been developed and validated with rigorous analysis, we should outline the management actions that need to be taken with well-defined recommendations. A good way to develop compelling recommendations that grab the decision-maker's attention is to use strong **action verbs** at the start of each recommendation.

3.13.5 Present Recommendations for Final Decision

Each recommendation should be supported with facts and analytical findings that provide the rationale for accepting the recommendation. For most complex business problems, there is no easy answer. If a final decision cannot be proposed, but the decision has been narrowed to set of two or three choices, it is prudent to present them all but with clearly outlined "pros" and "cons" for each. Another alternative is to define a set of criteria by which a decision should be made, review them in advance with the decision-makers and then present the recommendations with an assessment of each against the set criteria.

3.13.6 Illustration

Problem Statement:

Company should grow revenue and profit by using distribution partners instead of its own salesmen.

Key Issues/Key Questions:

(i) Using distribution partners has not been successful in the past.
   (a) Why did past attempts fail? What were the root causes? (e.g. lack of focus, bad selection of partners, poor negotiation, etc.)
   (b) Does the company know the distribution partners it would like to use?
   (c) Are distribution partners still receptive to working with the company? Which ones?

(ii) Distribution partners may take a long time to start producing the revenue growth needed.
   (a) What is typical 'ramp-up' time for a distribution partner? Can it be accelerated?
   (b) How much revenue growth can be expected from distribution partners?

(iii) Distribution deals can be expensive.
   (a) What is typical revenue sharing agreement in the industry?
   (b) Can the company negotiate deals that are better than those of its competitors?
   (c) What is the expense/revenue ratio that can be achieved with distribution partners and how does it compare to that of its own sales force?

(iv) The company has strong culture of growing revenue by growing sales force.
  (a) What is the productivity of sales force?
  (b) What would it cost to increase sales force to reach revenue growth target?
  (c) Would the increased sales expense fit allow the company to meet profit targets?

3.13.7 Recommendations/Key Supporting Facts and Analytics

(i) Start negotiating with distribution partners
  (a) Partners provide expense to revenue ratio of 4% vs. 12% from the company's sales force.
  (b) Distribution partners can cover 5 remote regions where the company's sales force has no presence.
  (c) Distribution partners have strong relationships with 10 customers that have not been penetrated as yet by the company's sales force.

(ii) Conduct a pilot project in 2 of the 5 regions
  (a) It will allow company to assess effectiveness of a new approach.
  (b) It will outline potential challenges that have not been foreseen till date.
  (c) It will enable improvements to the overall program before it is expanded to all regions.

APPENDIX

Proposal for Business Process Improvement at AGROTECH India Limited

Background AGROTECH India Limited has been a pioneer in the biotechnology research-based agro products. The organisation has been serving the farming community in India over the past two decades and has made an impressive growth from Rs. 6 cr in 1996 to Rs. 200 cr in 2007. However, for the last 2 years the top line has seen a downward trend and was at Rs. 170 cr in 2009 though the overall industry has seen a healthy growth in the same period. The bottom line has also eroded and the PBT is down from a level of Rs. 20 cr in 2007 to Rs. 10 cr in 2009. The management has taken several initiatives in the current year including setting up the strategic agenda for the organisation and aspires to be a $150 million company by 2012.

The organisation, now wants to align its key business processes such as sales and operations planning, demand generation, and annual budgeting with the overall business objective. Visionary Consultants (VCs) have been partnering with AGROTECH India Limited over the last one year in several initiatives in HR space and has been requested to submit a proposal for streamlining the key business processes using its operations consulting expertise.

Our Understanding of the Current Situation A team from VC visited the corporate office and interacted with key people from HR, Finance, Supply Chain and Planning and Sales including few Territory Managers. The following were the key observations:

- The actual revenues were 30% less than the budgeted revenues for FY08-09 and are about 15% less for the current year till Oct-09.
- In the current year, the actual monthly sales has been only about 60% of the forecast.
- There is a heavy skew in sales towards the end of the month – 50% of the sales happen on the last day.
- The products are highly seasonal and there is a very short selling season.

The team subsequently studied a few key business processes and identified the following issues/ concerns:

Issues/Concerns in Sales and Operations Planning Process The entire Sales and Operations Planning Process is not aligned to the Manufacturing and Procurement Lead Times.

- The plan for every month is finalised by 7th and the last dispatch needs to be done by 27th/28th –this leaves the supply side with only about 20 to 22 days to cope up with the entire months requirement.
- The dispatches of goods from all manufacturing locations generally begin only from mid of the month. About 50% of FG is received by the C&FAs on the last 2 days of the month. This may result in loss of sales as the selling window is very small.
- The planning for procurement of domestic RM & PM is based upon the forecast given for the next month. However, this forecast is made on an *ad-hoc* basis leading to a variation of more than ± 50% between forecast and actual requirement. This results in either RM/PM not available or pile of non-moving stock.
- The planning of imported RM (LT more than 60 days) is based on the corporate budget, while the actual requirement is far different from the corporate budget.
- The metrics used for reviewing performance indicates more than 95% availability (based on Rs dispatched vs Rs planned) while for more than 3/4th of the month the required SKU may not be available at the C&FA.
- Availability of stocks at C&FA is a no man's territory
- Inventory norms for FG/RM/PM have not been scientifically arrived at.

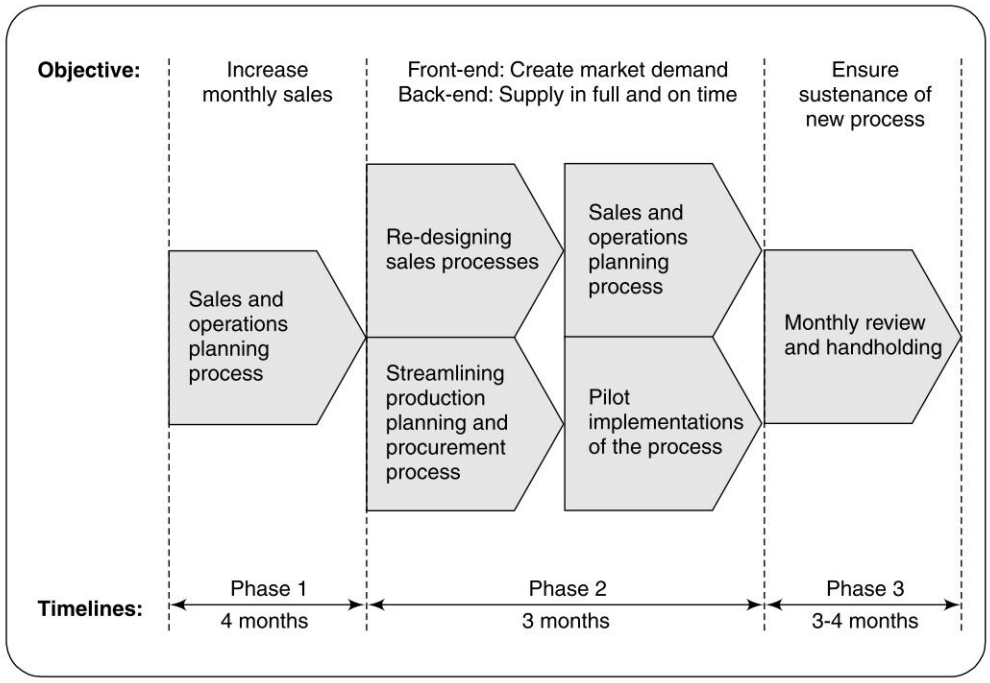
Issues/Concerns in Sales Processes

- There is a variation in the way the sales is forecasted and there is also a high variation in forecast vs. actual sales across the 4 units.
- The focus is higher on lag indicators such as sales in Rs lakh collections, outstandings, etc. while a few lead indicators such as details of farmer meets, number of dealers visited, dealer-wise secondary sales, which are key for generating demand are not adequately covered.
- The processes related to capturing market intelligence such as share of competitors at dealer level, product level, crop level need to be further strengthened.
- There is no structured review and standardised review mechanism across all units.
- The sales return at 30% is significantly higher than the industry average.

Issues/Concerns in the Annual Budgeting Process

- The budget focuses more on financial measures while performance indicators such as number/ details of new dealers to be appointed, new villages to be targeted, new crops to be targeted, etc. need to be a part of the budgeting exercise also.
- The final annual budget is not deployed in a structured manner across the organisation. Thus, actions/initiatives to ensure achievement of budget are not identified.
- The budget review mechanism is inclined more on top line and bottom line rather than status of actions/initiatives for achievement of the same.

Suggested Approach for Improving the Business Processes AGROTECH should re-look at some of its key business processes in a phased manner as under:



Phase 1 The 1st phase will be for duration of 4 months and the objective will be to redesign and implement the Sales and Operations Planning Process, which will result in

- Bridging the gap between the forecast and actual sales;
- Reducing the skew towards month-end in both dispatches as well as sales;
- Ensuring that the right stock keeping units (SKUs) are available throughout the month.

Phase 2 The 2nd phase will be for duration of 3 months with the following two objectives:

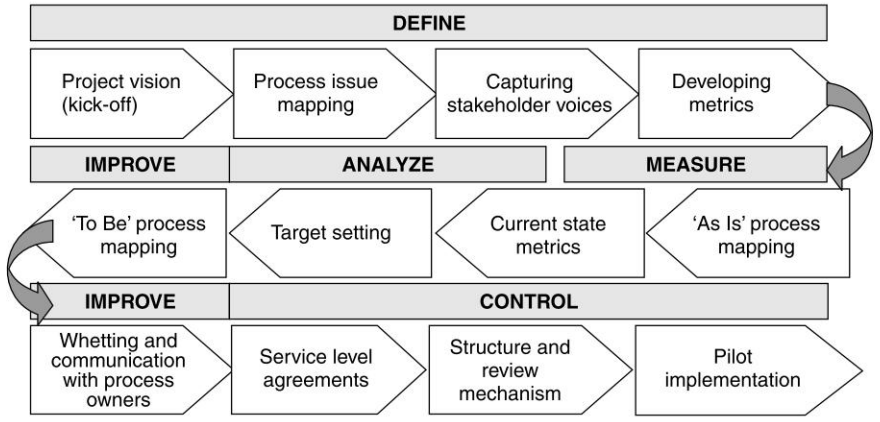
- The first objective will be to redesign all sales processes such as demand generation, daily order fulfillment, territory planning, market intelligence gathering, managing promotion activities, etc. Subsequently all the entire sales force will be trained on the new process and related standards/formats/policies.
- The second objective will be redesign and implement Production Planning and Procurement Planning Process at all the three manufacturing locations.

Phase 3 The 3rd phase will be for duration of 3 to 4 months where a senior member from VC will provide handholding and coaching on a part-time basis and participate in monthly reviews.

Engagement Objective for the 1st Phase The engagement objective for the 1st phase will be to design and implement the Sales and Operations Planning Process for the Agri and Pesticide division.

VC Methodology VC would form a cross-functional team comprising of members from Supply Chain, Sales, Production, Marketing and Finance. This team will use a very structured methodology as stated below to redesign and implement the Sales and Operations Planning Process.

The team would work rigorously over a span of 4 months along with the VC team and would focus on the following:



- Analysing the existing product portfolio to identify runners/repeaters/strangers
- Define the planning philosophy for runners/repeaters/strangers
- Fix the point of commit in the supply chain and scientifically arrive at the inventory norms for FG, RM and PM
- Identify the demand time fence for all key activities in the to-be process
- Design, communicate and train the concerned members on all new formats/standards
- Define Service Level Agreements between key related functions such as Production, Sales, Marketing and Supply Chain
- Define the metrics and review mechanism for Sales and Operations Planning Process
- Execute the new process for 1 full cycle

Project Plan The broad project plan for the above engagement will be as under:

Activity	M1		M2		M3		M4	
	F1	F2	F1	F2	F1	F2	F1	F2
Project visioning								
Process issue mapping								
Capturing stakeholder voice								
Developing metrics								
As-is process mapping								
Current state metrics								
Target setting								
To-be process designing								
Whetting & communicating with process owners								
Finalizing service level agreements								
Structure & review mechanism								
Pilot implementation								
SC review		▲		▲		▲		▲

Resource Requirements

Resource Deployment from VC Limited VC will deploy of a team of three resources as under:

- A full-time consultant with appropriate skill sets will be deployed for a duration of 3 months who will work along with the team from AGROTECH for redesigning and implementing the Sales and Operations Planning Process
- A Principal Consultant will be deployed on a part-time basis for the entire duration of 4 months. He would mastermind the initiative, guide the teams and participate in all reviews.
- A Chief Consultant will provide the necessary thought leadership to all teams including AGROTECH Top Management.

Resource Requirement from AGROTECH AGROTECH will have to commit the following resources:

- One of the Senior Managers will have to champion the entire initiative and will have to commit about 20% of his/her time on this initiative.
- A middle-level manager, who will work with the VC Consultant, will have to be committed on a full-time basis for the entire duration of 4 months.
- A cross-functional team comprising of members from Supply Chain, Production, Marketing and Sales will have to commit about 15% of their time of the next 4 months.

Benefits and Deliverables

The following would be the deliverables from the engagement:

- Redesigned Sales and Operations Planning Process
- Implementation of the new to-be process for 1 cycle with subsequent fine-tuning, if necessary
- Inventory norms for FG, RM and PM
- Metrics and Review mechanism for the standard and operating process (Sales and Operations Planning Process)
- Service Level Agreements across key related functions

The following would be the benefits from the above engagement:

- Bridging the gap between the forecast and actual sales
- Reducing the skew towards month-end in both dispatches as well as sales
- Ensuring better availability of SKU's across C&FAs

Professional Fees

VC would charge a professional fee of Rs. 24 lakhs (Rs. Twenty Four Lakhs Only/-) for the first phase for a duration of 4 months. The total Professional Fee will be paid as per the following schedule:

- Rs. 5 lakh will be paid as Mobilisation fee before the start of the assignment.
- Rs. 5 lakh will be paid as monthly fee for the first three months.
- Rs. 4 lakh will be paid as monthly fee for the fourth month.

Service tax, as per prevailing government rules (presently 12.24%), will be payable on the invoice amount. Any new tax, applicable in future, will be similarly treated.

All out-of-pocket expenses such as local travel, travel to various plants/regional offices, lodging and boarding at plants/regional offices, communication expenses will be payable by AGROTECH India Limited.

All invoices and debit notes will be payable within 7 days from the date of receipt.

SUMMARY

The first three components of a research process are as follows:

- Specifying area and objective
- Defining and refining/redefining a problem
- Hypothesis development

The three issues relating to identifying the problem are.

- Genesis of the problem
- Impact of the problem
- Dissecting cause of the problem and objective to be achieved

After the formulation of the problem, the next stage is development and formulation of various types of hypothesis to be tested.

After formulating the hypothesis, the next step is preparing a research proposal, containing inter alia strategy, methodology and resources required for conducting the research study.

Once the research proposal is finalised, the next decision to be taken by the management is that whether the study should be conducted in-house i.e. by the internal team of experts or by an external agency.

Sometimes, when the research study involves a major project like setting up a communication network, ERP, a management institute, etc., then the organisation floats what is called RFP. RFP facilitates better understanding of the expectations from the vendor.

DISCUSSION QUESTIONS

1. Discuss various steps in conducting research study, giving a suitable example.
2. Write short notes on
   (a) Formulating a research problem
   (b) Criteria for good business research
3. Discuss the process of formulating a problem, with an example.
4. "Solution of a problem leads to another problem". Discuss this with two illustrations.
5. While conducting a research study, what are the issues to be discussed at various hierarchal levels? Illustrate with an example for each level.
6. What are the various types of hypotheses? Give an example for each of the hypotheses.
7. Describe the process of setting up or developing a hypothesis, with an illustration.
8. Describe the objectives and contents of a research proposal.
9. Describe the advantages and disadvantages of external and internal research.

EXERCISE

1. Think of some problem in any area, and proceed to discuss the various steps that are required to refine or redefine the problem, and resolve it in a systematic manner.

Research Design

4

Contents

LEARNING OBJECTIVES

The main purpose of this chapter is to provide a comprehensive knowledge about the various experimental and research designs, and their applications in business environment.

In addition to conventional designs and sampling schemes, the objective is also to acquaint with two other types of designs that are highly useful in research environment to provide complete knowledge about conduct of research studies.

One type of studies relates to conduct of research of a particular phenomenon for a cross-section of entities like companies, institutions, at one point of time or over a period of time.

The other type of studies relates to an environment where it is advisable to simulate or generate data in a futuristic scenario, and use it for managerial decisions.

Relevance

Modern Electronics, a leading seller of electronic products has a chain of several stores all over India. It has been growing at about 20% for the last 3 years. However, the opening of stores of a multinational brand of electronic products has affected the sales of Modern Electronics considerably during the last six months. The management of the company has been toying with various options to deal with this unprecedented situation. The CEO thought of seeking answers to the following questions before drawing out a strategy for future:

- Have the sales of all products been affected equally?
- Have the sales of the six zones been affected equally?
- Is there any interaction between the decline of sales of products and the zones i.e. whether any particular products have been affected more in some zones than others?
- What is the perception of customers towards Modern Electronics, and what are their suggestions for greater acceptability of its products?
- What are the employees' suggestions to boost up the sales?

The CEO engaged the services of a consultant for this purpose. The consultant used the research designs and experimental designs discussed in this chapter to provide answers to the above questions which helped the CEO to draw up a competitive strategy to successfully compete with other multinational companies.

4.1 INTRODUCTION: DEFINITIONS AND OBJECTIVES

Research Design is a comprehensive plan of the sequence of operations that a researcher intends to carry out to achieve the objectives of a research study. It provides the conceptual structure or blueprint for the conduct of the research study. It could also be considered a planned sequence of the entire process involved in conducting a research.

A research design comprises the blueprint of methodology for

- Collection of data
- Measurement of data elements/units
- Analysis of data

The basic objective of a research design is to ensure maximum information that is needed for decision-making through a research study, with minimum resources. For a specific problem, a research design covers the following issues:

(i) Objective of the research study
(ii) Methodology for obtaining or collecting information
(iii) Resources including Men, Money and Minutes (Time), and their allocation for various segments of the research study.

While selecting an approach from the possible approaches, one could strike a balance between objectives and resources including availability of data. For instance, once a study, on urgent basis, was to be undertaken to explore linkage between amounts of loans sanctioned and disbursed (outstanding) by financial institutions and certain economic and banking indicators in India. However, when the study started, it was found that the financial institutions had data about outstanding amount of sanctioned loan at the end of every month but no such data about loans actually outstanding at the end of every month. It took about two months to collect the required data from all financial

institutions in India. Although the information was useful and led to setting up an appropriate information system, it was not available at the time when urgently needed for policy formulation.

4.2 ADVANTAGES OF RESEARCH DESIGN

Just as in the architecture's design for building a house, it is necessary to ensure that all the specified requirements are met without any exception. The research design helps the researcher to ensure fulfillment of the objective of the study.

It may be noted that defining the objective, formulation of the problem, etc. are conceptual part of a research process, and involve only mental thinking, the research design involves physical work!

4.3 TYPES OF RESEARCH DESIGNS

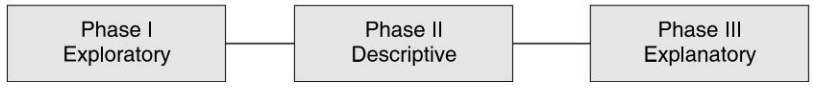
There are three basic types of research designs depending on the three types of studies:

- Exploratory
- Descriptive
- Explanatory/Causal/Relational

These can be considered as the different phases of the research study. The research passes through each phase. In the initial phase when the research topic is new and not much is known/researched about the topic, the exploratory study is undertaken. Though exploratory study can be qualitative as well as quantitative, more emphasis is given for the qualitative approach. The phase mainly concentrates on understanding the topic and identifying variables.

The next phase of research is descriptive; in this phase, the variables identified are collected using a tool/instrument (like interview, questionnaire, etc.) and descriptive analysis is performed. This helps in further understanding of the variables.

The next phase of research is explanatory/causal. Once there is a clear understanding of the identified variables, the next step is to try to find relationship among the variables. In this phase, the variables and the relationships are identified, like independent variable, dependent variable, moderating variable, etc. This research can also be done by conducting field experiments, discussed in Section 4.5.



We consider different designs for these three different types of studies as there are different requirements of design for the three phases. For example, flexible design is suited to exploratory studies, and rigid design is suited for explanatory studies. The descriptive studies may have a mixed approach. A flexible design is amenable to the required changes in the research. In the case of exploratory study, not much is known about the research topic, and therefore frequent changes are required. The design should allow such changes. In the case of explanatory research study, the research is generally matured through the previous two stages and the design should be rigid enough to avoid any bias that might crop in the study.

We will discuss these phases, in brief, in the next section.

4.3.1 Exploratory

We have discussed the exploratory studies in Chapter 1. Thus, we limit our discussion on exploratory studies as a type of research design.

Exploratory studies are generally carried out:

- When not much is known about the situation – prevailing or encountered, and yet we want to have some assessment;
- When we want to solve a problem but no information is available as to how same or similar problem was solved in the past.

Exploratory study is conducted to explore a problem, at its preliminary stage, to get some basic idea about the solution at the preliminary stage of a research study. It is usually conducted when there is no earlier theory or model to guide us or when we wish to have some preliminary ideas to understand the problem to be studied, as also the approach towards arriving at the solution. It might help in modifying the original objective of the study or might even lead to a new perspective rather than the earlier perceived problem.

4.3.2 Descriptive

Such studies deal with:

- Description of a phenomenon like accidents in a city
- Describing a variable like revenue, life of an item, return on investment etc., representing a characteristic of a population under study
- Estimation of the proportion of the population having certain characteristic(s) like colour preference of cars, specified qualification or experience, etc.

The main goal of this type of study is to describe the data and characteristics of what is being studied. The idea behind conducting descriptive study is to study frequencies, averages and other statistical calculations. **Although this study is highly accurate, it does not explain the causes behind a situation.**

Unlike exploratory research, descriptive research may be more analytic. It often focuses on a particular variable or factor.

4.3.3 Explanatory/Causal/Relational

Such studies involve studying the impact of one variable on the other and also the relationship between two variables.

The relevance of causal study arises only when there exists correlation between two variables. For example, if there is correlation between two variables, say sales and advertising expenses, one may like to study which of the two is the **'cause'** and which is the **'effect'**. In this case, advertising expenses is the cause (called independent variable) and sales (called dependent variable) is the effect. Incidentally, causal variable is also called **'explanatory'** variable as it explains the effect or impact on the dependent variable. Similarly, if one finds existence of association between two factors, one may investigate which of the two factors is the 'cause' and which one is the 'effect'. But the study cannot proceed further. However, in the case of correlation study relating two variables, once we know the 'cause' and the 'effect', the study can proceed further to find out the relationship between the two variables.

When we talk about association or correlation, we refer to association or correlation between two factors or variables. A factor is an attribute or characteristic like colour, qualification, area of

specialisation in MBA, and is usually not measurable. A variable is a measurement of some characteristic like income, age, etc.

While association is used in the context of studying factors, correlation is used in the context of studying variables.

Some examples of association study are:

Association between

- Academic background (like commerce, science, engineering, etc.) of an MBA student and his/her option of area like marketing, finance, operations, etc.
- Motivation and Performance of Salesmen

Some examples of correlation study are:

Correlation between

- Income and the insurance cover by individuals
- Expenditure on advertisement and sales turnover of a company

In the case of association, we can conclude only when

- There is any association, or
- There is no association

However, in the case of correlation, we can also find whether the correlation is positive or negative. If the correlation is positive, both variables move in the same direction like sales and advertising expenses. If the correlation is negative, both variables move in the opposite direction like availability of a commodity and its price.

Once causal study establishes cause and effect relationship between two or more variables, the relational study attempts to measure the relationship in a mathematical equation derived by using statistical methodology. The relationship could be in the form

$$\mathbf{y}\ (\text{sales}) = 25 + 10\ \mathbf{x}\ (\text{Advertising Expenses})$$

It may be interpreted as follows:

For a unit (say, Rs. in crore) change in the value of x, y changes by 10 units (Rs. in crore). It implies that, on an average, an increase of Rs. 1 crore in advertising expenses causes an increase of Rs. 10 crore in sales.

4.4 VALIDITY OF A RESEARCH DESIGN

Validity refers to the strength and the accuracy of a research design. We consider two basic types of validity in a research design viz.

- Internal Validity
- External Validity

Campbell and Stanley (1963) have defined **Internal validity** as the basic requirement for an experiment to be interpretable, and **External validity** as the criterion for generalisation.

These are briefly described as follows.

4.4.1 Internal Validity

Internal validity describes the ability of the research design to unambiguously test the research hypothesis. In other words, internal validity refers to the extent to which one can accurately state that the independent variable is responsible for the observed effect in the dependent variable and no other variable is responsible for the effect. We may explain it in a simplified form as follows:

If the effect on dependent variable is only due to variation in the independent variable, then we may conclude that the internal validity is achieved.

To achieve high internal validity, a researcher must consider all the factors that affect the dependent variable and control them appropriately. It ensures that these factors do not interfere with the results.

It may be noted that internal validity is only relevant in studies that aim at establishing causal relationship. It is not relevant in most of the exploratory, observational or descriptive studies.

Some of the threats to internal validity are:

- History
- Maturation
- Testing
- Selection of Respondents
- Statistical Regression
- Experimental Mortality

A brief discussion on each of these threats is given below.

History

History refers to the events that are beyond the control of the experiment. These events may change the attitude of the respondents irrespective of whether the independent variable is changed or not. Thus, it is impossible to determine whether any change on the dependent variable is due to the independent variable, or the historical event.

Maturation

The respondents may not have the same level of responses at the later part of experiment as in the beginning of the study. The permanent changes, such as physical growth and temporary changes such as fatigue, may provide alternative explanations; thus, they may change the way a respondent would react to the independent variable. So upon completion of the study, the researcher may not be able to determine if the cause of the discrepancy is due to time or the independent variable.

Testing

When the respondents are subjected to the repeated test, i.e. in the experiments where the respondents are tested more than once, the bias could crop in as the respondent may remember what they are being tested for. The mental ability tests in MBA selection process is an example. The respondents may learn the techniques by practicing and may get higher scores than before. This may not be attributed to the independent variable, thus leading to bias.

Selection of Respondents

The inappropriate selection of respondents may lead to bias in experimental design. If the selected respondents are not uniform, inadvertent randomisation may take place leading to bias.

Statistical Regression

The statistical regression refers to the bias that may crop in due to some respondents giving extreme responses. This bias is known as error sum of squares in statistical regression analysis.

Experimental Mortality

This can occur when the respondents drop out during the experiment especially in the experiment involving pre-test and post-test. The same respondents who take up the pre-test may not be available for the post-test. This results in excluding the entire pre-test data from the analysis for the dropped-out respondents.

4.4.2 External Validity

External validity is related to generalisability of the findings/results. It refers to the degree of generalisability of the conclusions to other situations. In other words, external validity is the degree to which the conclusions in the study for a given population could be made applicable to other populations or other situations.

It may be noted that more the internal validity for an experiment lesser will be its external validity. It is because when steps are taken to increase the degree of internal validity, it leads to controlling of many other variables (also known as extraneous variables, see Chapter 2), which are not under the scope of study but can still influence the causal relationship between variables under study. The end-result is that it limits the generalisability of findings.

4.5 EXPERIMENTAL DESIGNS

In this section, we shall discuss various experimental designs that are used in the conduct of experimental research. In such research, data is generated through experimentation or by conducting experiment whose basic objective is to make a discovery or

- To test hypothesis like Customers prefer "One + One Free" rather than "50% discount" or "I.Q.s of students get improved after doing MBA"
- Demonstrate a belief: Performances of male and female students are same in MBA.

4.5.1 Relevance and Historical Development

Historically, the subject of experimental designs started with the analysis of data relating to agricultural experiments. These experiments were conducted, and are being conducted even now due to continuous quest by the human being for better and better productivity of crops. Such experiments are also known as **'Design of Experiments (DOE)'**. It is so because the experiment is first designed in consonance with the objective of study. **An important feature of such studies is that data is generated through an experiment.**

Even though these experiments were evolved in the context of agriculture, they were soon picked up for use in pharmaceutical and medical research wherein there is continuing need for better and better medicines and treatments. This objective is achieved through designing experiments, collecting experimental data and drawing conclusions from the same.

Evaluations of educational and training systems are also dependant largely on conducting and evaluating systems and tools using such experiments.

In fact, design of experiments plays an important role in any situation where the objective is to compare two or more treatments or compare two or more levels of '**treatments**'. Just like in medical language, the word treatment is used for the medicines and their dosages, in agricultural experiments, the same word was used to evaluate impact of fertilisers and seeds on yields of different crops. This word continues to be used for all the applications in sociology, educational system, marketing, finance, operations, etc. We shall describe this term as well as other terms used in agricultural experiments that continue being used in all experiments in all fields.

The word 'treatment' and its levels may be used in experiments as follows:

Area	Possible Treatments for Experiments
Agriculture	Fertilisers—Types or different amounts of fertilisers per unit of land.
Education	Training systems—Types with respect to contents and methodology or the same training inputs but covered in different time periods like 1 week, 2 weeks, etc.
Marketing	Strategies, Discounts, Packing size, etc.
Finance	Evaluation and Comparison of Schemes, Products, etc.
Operations	Machines, Inputs, Processes, etc.

Before we proceed further, we would like to define the following terms used in Design of Experiments, so that their usage will be understood and appreciated in the subsequent paragraphs.

Table 4.1 Explanation of Various Terms in Experimental Designs

Experimental Unit	It is the object on which the experiment is to be conducted. Examples: Plots (of land), Students, Salesmen, Patients
Response	It is the dependent variable of interest. Examples: Yield (of plot), Return on investment, Performance scores (like marks, grades, sales), Quality of product/service
Treatment or Factor	Those independent variables whose effect on the response is of interest to the experimenter
Quantitative Factor	Measured on a numerical score like discount in price
Qualitative Factor	Not measured numerically like gender, colour, location
Levels of a Factor	Different values of a factor. Values of the factor that are used like 10 gms, 12 gms and 15 gms per sq. metre of land, dosages of medicine, duration of training, percentages of discounts (5%, 10%, 15%)
Types of Factors	Different varieties of fertiliser, different medicines, same training by different institutes, different mutual funds or different schemes of mutual fund, different makers of machines used in a factory
Experimental Designs	The experimenter/researcher controls the specification of treatments and the method of assigning experimental units to each treatment
Completely Randomised Experiment	Herein, the experiment is concerned with the study of **only one** factor. Each treatment/factor is assigned or applied to the experimental units without any consideration
Block	Each block (analogy with agricultural plots) comprises of same number of experimental units as the number of treatments under experimentation
Randomised Block Design	Herein, the experiment involves study of **two or more** factors. One experimental unit from each of the blocks, say **'n'** in number, is assigned to each of the, say, '**m**' treatments. Thus, 'n' blocks have 'm' treatments in each block. For example, suppose, the experiment involves comparing I.Q.s of students (experimental units) of each of the three areas ('treatments') viz. Marketing, Finance and Operations
Blocking	Blocking implies control of factors which are either not of interest or their effect is removed/filtered/averaged out.

(*Contd*)

(*Contd*)

Control Group	This term is explained with the help of an example. One may like to study the productivity of a particular fertiliser. This can be studied by taking some, say 12, plots of similar type. While the fertiliser could be used only on 6 plots, the other 6 plots could be cultivated without the fertiliser. Thereafter, the yields of the two sets of plots may be compared to see the effect of the use of the fertiliser. In Experimental Designs terminology, the fertiliser is called **'Treatment'** and the plots are called '**experimental units**' or simply **'Units'** which are subjected to treatment for generating the desired data. The plots which are not treated with the fertiliser are called **'control'** group. They are used to evaluate impact of the treatment.
Replication and Randomisation	In addition to the concept of control, explained above, two other concepts viz. **Replication and Randomisation** are very relevant and important for designing of experiments. Replication, as the name implies, means repeating. Obviously, no conclusion can be drawn by conducting experiment just on one unit. Just like, in statistics, no worthwhile or meaningful conclusion can be drawn just by taking one observation in the sample; similar is the case in DOE. Just like in a statistical study wherein the units on which the observations are to be recorded must be selected **randomly**, similarly, the units must be selected **randomly** as also the treatments under consideration should be applied to the units in a random manner. Formally a **randomisation is defined as random assignment of treatments to experimental units**.

We may reiterate that **control**, **replication** and **randomisation** play a very important role, and, in fact, are at the core of Design of Experiments.

Following are some of the types of experiments which can be conducted for studying:

- Productivity of two or more types of fertilisers, seeds, irrigation systems
- Productivity of different amounts of the same fertilisers per unit of land
- Effectiveness of two or more types of medicines
- Effectiveness of two or more doses of the same medicine
- Effectiveness of two or more types of training systems
- Impact of various marketing strategies
- Impact of various incentives for improving sales
- Impact of responses to products and services in different cities, regions, etc.
- Evaluation of different returns on various stocks or market indices

The statistical technique used for analysing the data collected for conduct of experiments is ANOVA, discussed in Chapter 12.

We have used the concepts of 'Design of Experiments' with the help of agricultural experiments, as these were the first such experiments, and more important, they are simple to understand.

4.5.2 Types of Experimental Designs

Some of the most popular experimental designs described in this section are:

- One-Factor Experiment
- Two-Factor Experiments
- Two-Factor Experiments with Interaction
- Latin Square Design
- Factorial Designs
- Quasi-Experimental Design
- Ex Post Facto Design

4.5.2.1 One-Factor Experiment This can be better explained by the following example:
The yield of a crop depends on several factors like fertilisers, variety and quality of seeds, type and quality of soil, methods of cultivation, amount of water made available, climate including temperature, humidity etc., method of harvesting, etc. Out of the several factors contributing to the yield of a crop, say rice, suppose we are interested in any one factor, say varieties of rice.

In such cases, the type of treatment is one, e.g. variety of rice. Let there be three varieties of rice. The data is collected about the yield of rice for each variety on a number of plots of equal size and similar type of soil. The care to be taken while designing the experiment is that the yield from plot to plot should vary only due to variety and not due to other factors. A typical table giving data collected through an experiment is given as follows.

(Yield in Quintals per Unit of Plot)

Plots	Variety of Rice		
	A	*B*	*C*
1	6	7	6
2	5	6	5
3	5	7	4
4	4	8	5
5		6	6
6		5	
Total (T_1)	20	39	26
Average	5	6.5	5.2

Using the appropriate ANOVA technique, it can be concluded that there is significant difference among the yields of the three varieties of rice. This example is solved in Chapter 12, Section 12.5.1.

Illustration 4.1: One Factor
Three groups of five salesmen each, were imparted training related to marketing of consumer products by three Management Institutes. The amount of sales made by each of the salesmen, during the first month after training, is recorded and is given in Table 4.2.

Table 4.2 Amount of Sales by Salesmen

		Salesmen				
		1	2	3	4	5
Institutes	1	65	68	64	70	71
	2	73	68	73	69	64
	3	61	64	64	66	69

The problem posed here is to ascertain whether the three institutes' training programmes are equally effective in improving the performance of trainees. If m_1, m_2 and m_3 denote the mean effectiveness of the programmes of the three institutes, then, statistically, the problem gets reduced to test the following null hypothesis, i.e.

$$H_0: m_1 = m_2 = m_3$$

against the alternative hypothesis that it is not so i.e.

$$H_1: \text{All means are not equal}$$

This situation is resolved by using ANOVA – One Factor, explained in Chapter 12. The conclusion reached is that all the training institutes are **not** equal with respect to the training programmes conducted by them.

4.5.2.2 Two-Factor Experiments Suppose it is claimed that the yield of any variety of rice depends not only on the variety itself, but also on the type of fertilisers used. Let there be three types of fertilisers under consideration. Thus, we would also like to test as to whether yields due to all the three fertilisers are equal. Such experiment is called two-factor experiment, and the data collected is in the following format.

		Fertilisers			
		I	II	III	IV
Varieties	A	6	4	8	6
	B	7	6	6	9
	C	8	5	10	9
	Total	21	15	24	24

If we take the totals of variety 'A' for all fertilisers i.e. 24, it removes the effect of varieties, and indicates yield of only variety 'A'. Thus, the totals of the three varieties i.e. 24, 28, and 32 indicate only the differences among 'A', 'B' and 'C'. The impact of fertilisers has been averaged out. We can also say that the impact of the factor fertiliser has been **'blocked'.**

Similarly, the totals of the four fertilisers, i.e. 21, 15, 24 and 24 indicate only the differences in fertilisers. The impact of varieties has been '**averaged out**' or '**blocked**'.

The analysis for two-factor experiment is the same as the analysis for two-way ANOVA, and is given in Chapter 12. Using two-way ANOVA, it can be concluded that there is no significant difference among the three varieties of rice as also among the four varieties of fertilisers.

Illustration 4.2

The following table gives the number of subscribers added by four major telecom players in India in the months of August, September, October and November 2005. The data are given in 000's and are rounded off to the nearest 100, and are thus in lakhs.

Additions to Subscribers

(*In lakhs*)

Months	Company			
	Bharti	*BSNL*	*Tata Indicom*	*Reliance*
August	6	6	2	5
September	7	6	2	3
October	7	6	6	4
November	7	8	7	4

(*Source*: Indiainfoline.com on India Mobile Industry)

It is required to test:

(i) If the four companies significantly differ in their performance?

(ii) Is there significant difference between the months?

Using the appropriate ANOVA technique, described in Chapter 12, following would be concluded:

(i) There is significant difference among the four companies in terms of adding subscribers.

(ii) There is no significant difference among the months with respect to adding subscribers.

Illustration 4.3

Several studies have been conducted to ascertain whether the entire range of indices on BSE move in the same direction.

This study relates to four important indices – BSE 30, BSE 100, BSE 200 and BSE 500, of Mumbai Stock Exchange. The idea was to test whether the average percentage change in these four indices were similar or not. Data was recorded for the last 12 months.

The movements in the closing value of four indexes have been provided below:

Month	*BSE 30*	*BSE 100*	*BSE 200*	*BSE 500*
January 2005	6,555.94	3,521.71	2,726.49	2,726.49
February 2005	6,713.86	3,611.90	2,825.65	2,825.65
March 2005	6,492.82	3,481.86	2,734.66	2,734.66
April 2005	6,154.44	3,313.45	2,610.50	2,610.50
May 2005	6,715.11	3,601.73	2,829.20	2,829.20
June 2005	7,193.85	3,800.24	2,928.31	2,928.31
July 2005	7,635.42	4,072.15	3,124.78	3,124.78
August 2005	7,805.43	4,184.83	3,273.00	3,273.00
September 2005	8,634.48	4,566.63	3,521.83	3,521.83
October 2005	7,892.32	4,159.59	3,198.69	3,198.69
November 2005	8,788.81	4,649.87	3,568.37	3,568.37
December 2005	9,397.93	4,953.28	3,795.96	3,795.96

ANOVA technique resulted in drawing the inference that:
All the four indices have similar average % changes.

4.5.2.3 Two-Factor Experiments with Interaction

While dealing with two factors, as above, one could argue that the yield of rice depends not only on variety of rice and type of fertiliser but also on the interaction between variety of rice and type of fertiliser. It could happen one particular type of fertiliser gives much more yield for one variety of rice as compared to the other variety. Thus, the yield of rice could also depend on the combination of a variety of rice and a type of fertiliser. If *V*1, *V*2 and *V*3 indicate the varieties of rice and *F*1, *F*2 and *F*3 indicate the types of fertilisers, then various interactions on combinations are nine as follows:

V1F1 V1F2 V1F3
V2F1 V2F2 V3F3
V3F1 V3F2 V3F3

Illustration 4.4
The following data refers to the yields of rice on two plots each with each combination of the variety of rice and type of fertiliser.

		Fertilisers			
		F1	*F2*	*F3*	*F4*
Varieties	*V*1	6	4	8	6
		5	5	6	4
	*V*2	7	6	6	9
		6	7	7	8
	*V*3	8	5	10	9
		7	5	9	10

It may be noted that in the data given in Table in Section 4.5.2.2, there is only one value available for each of these combinations analysing variation due to any combination. However, for isolating the interaction factor, we should have minimum of two values for each combination. Accordingly, the above table gives data with minimum of two yields, obtained on two different plots of the same type, for each of the combinations.

The analysis of such data is carried out with the ANOVA analysis for two-factor interaction, as will be explained in Chapter 12. It can be concluded that:

- The yields of three varieties of rice are significantly different.
- The yields of four types of fertilisers are significantly different.
- The interactions among varieties of rice and types of fertilisers are significantly different.

Illustration 4.5: Two Factors with Interaction
It has been observed that there are variations in the pay packages offered to MBA students. These variations could either be due to specialisation in a field or due to the institute wherein they study. The variation could also occur due to **interaction** between the institute and the field of specialisation.

For example, it could happen that the marketing specialisation at one institute might fetch better pay package rather than marketing at the other institute. These presumptions could be tested by collecting following type of data for a number of students with different specialisations and different institutes. However, for the sake of simplicity of calculations and illustration, we have taken only two students each for each interaction between the institute and field of specialisation.
The data is presented below in a tabular format:

	Institute A	*Institute B*	*Institute C*
Marketing	8	9	9
	10	10	8
Finance	9	10	6
	11	11	7
HRD	9	8	6
	7	6	6

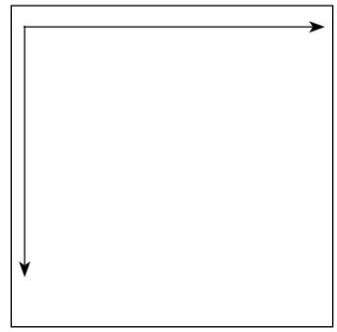
Here, the test of hypotheses will be
For Institute:
H_0 : Average pay packages for all the three institutes are **equal**
H_1 : Average pay packages for all the three institutes are **not** equal
For Specialisation:
H_0 : Average pay packages for all the three specialisations are equal
H_1 : Average pay packages for all the three specialisations are **not** equal
For Interaction:
H_0 : Average pay packages for all the nine interactions are equal
H_1 : Average pay packages for all the nine interactions are **not** equal
As worked out in Chapter 12, the conclusions drawn are as follows:
- Average pay packages for all the three institutes are different
- Average pay packages for all the three specialisations are equal
- Average pay packages for all the interactions are **not** equal

4.5.2.4 Latin Square Design A Latin Square is an $n \times n$ table filled with n different symbols in such a way that each symbol occurs exactly once in each row and exactly once in each column.

A	B	C
B	C	A
C	A	B

With the Latin Square design, a researcher is able to control variation in two directions. It may be noted that:
- Treatments are arranged in rows and columns.
- Each row contains every treatment but only once.
- Each column contains every treatment but only once.

Latin square designs were developed in the context of agricultural experiments. Suppose there is a big agricultural plot available for experimentation. This plot is to be divided into several smaller plots (experimental units) for experimenting to compare the yield of different fertilisers.



If the plot is such that its fertility changes along with its length as well as its breadth, then allocation of fertilisers has to be done in such a way that variations are averaged out in both directions. If there are 4 levels of fertilisers, the plot is divided into 4×4 (=16) plots and the different types of fertilisers are assigned to different plots as follows:

F_1	F_2	F_3	F_4
F_2	F_3	F_4	F_1
F_3	F_4	F_1	F_2
F_4	F_1	F_2	F_3

It is better if the shape is square. The square given above may be made smaller to accommodate. Such design is called Latin Square design.

It may be noted that:

- Every treatment is used in all rows and columns
- One treatment is used only once in each row
- One treatment is used only once in each column

Because of such allocation, the variation in fertility along length and breadth does not matter for comparing the yields of different fertilisers.

The above allocation is only one of several possible allocations satisfying the above criterion.

4.5.2.5 Factorial Designs A **factorial experiment** is an experiment whose design consists of two or more factors, each with discrete possible values or 'levels', and whose experimental units take on all possible combinations of these levels across all such factors. Such an experiment allows studying the effect of each factor on the response variable, as well as the effects of interactions between factors on the response variable.

Factorial experiments allow for investigation of the interaction of two or more factors or independent variables. A factorial design allows for testing of two or more treatments (factors) at various levels, and also their interaction. For this reason, they are more efficient i.e. providing more information with lesser resources.

Suppose, one wants to answer the following:

(i) What is the effect of different salaries offered to MBA graduates of an institute for posting in the corporate office or field offices?

(ii) What is the effect of varying design (plain or checks) for two colours of shirts (white and blue)

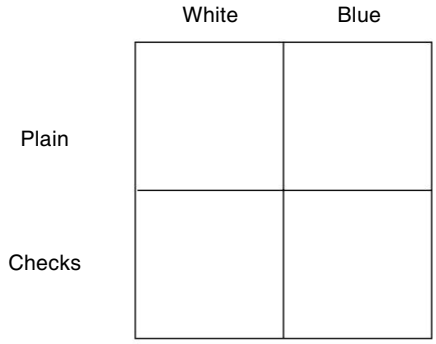
In the first example, independent variables or factors are corporate office and field postings.

In the second example, independent variables or factors are design and colour of shirts.

In the second example, suppose one wish to study the number of customers buying two different designs for each of the two colours.

The factorial experiment would consist of four experimental units: plain white, plain blue, checks white, checks blue. Each combination of a single level selected from every factor is present once.

Sales for each of the four combinations of the two factors viz. discount and packing size.



In general, for the vast majority of factorial experiments, each factor has only two levels. For example, with two factors A and B each taking two levels viz. A_1 and A_2, and B_1 and B_2, a factorial experiment would have four treatment combinations in total, and is called a 2×2 factorial design.

A_1B_1	A_1B_2
A_2B_1	A_2B_2

A factorial experiment can be analysed using ANOVA or regression analysis.

It is relatively easy to estimate the main effect of a factor. To compute the main effect of a factor "A", subtract the average response of all experimental runs for which A was at its low (or first) level i.e. A_1 from the average response of all experimental runs for which A was at its high (or second) level i.e. A_2.

4.5.2.6 Quasi-Experimental Design A quasi-experiment is a scientific research method primarily used in social sciences. "Quasi" means likeness or resembling, and therefore quasi-experiments share characteristics of true experiments which seek interventions or treatments. The key difference in this empirical approach is the lack of random assignment. Another unique element often involved in this experimentation

method is use of time series analysis to ascertain whether the impact of any factor has undergone change over a period of time. Experiments designed in this manner are referred to as having quasi-experimental design.

One such example is that if we want to examine whether the annual rate of returns on stocks of Infosys, TCS and WIPRO are the same over a period of last 5 years. The data is available in an organised form, and as such no randomisation is required.

Since quasi-experimental designs are used when randomisation is impossible and/or impractical, they are easier to set up than true experimental designs; it takes much less effort to study and compare subjects or groups of subjects that are already naturally organised than to have to conduct random assignment of subjects. Additionally, utilising quasi-experimental designs minimises threats to external validity as natural environments do not suffer the same problems of artificiality as compared to well-controlled laboratory settings. Since quasi-experiments are natural experiments, findings in one study may be applied to other subjects and settings. Also, this experimentation method is useful in longitudinal research that involves longer time periods which can be followed in different environments.

Illustration 4.6

A consulting agency has been helping an investment company to recruit about 20 MBA students as business analyst executives from two management institutes, each year. The company offers lucrative compensation package aimed to attract the best talent.

One year, the consulting agency thought of organising an online competition between the two management institutes from where they were recruiting the executives. It was prompted because of the debate in the academic circles about the superiority of one over the other.

They picked up 20 top students, based on their latest academic scores, from each of the institute, who agreed to participate in the competition. We may notice that it was **not a random selection**. They divided the 20 students in 5 groups each having 4 students. The groups comprised students with ranks 1 to 20, as follows:

Institute A:

Group 1	1	6	11	16
Group 2	2	7	12	17
Group 3	3	8	13	18
Group 4	4	9	14	19
Group 5	5	10	15	20

Institute B:

Group 1	1	6	11	16
Group 2	2	7	12	17
Group 3	3	8	13	18
Group 4	4	9	14	19
Group 5	5	10	15	20

It may be noted that the assignment of the students to the groups is not done in a random manner.

The business investment game was given to each of the two corresponding groups from each institute. Group 1 of institute A competed with group 1 of B, Group 2 of A with Group 2 of B and so on, and the scores recorded for all the groups. The institute whose three or more groups recorded wins was declared as the winner.

4.5.2.7 Ex Post Facto Design The literal meaning of expost facto is "from what is done afterwards" or "after the fact". It means something done or something occurring after an event with a retroactive effect on the event.

In an experimental approach an investigator has direct control, or can manipulate at least one independent variable. He can choose his experimental units at random and assign treatments to groups at random.

In ex post facto approach, one cannot control the independent variable or variables because they have already occurred, and cannot assign subjects or treatments at random. In this situation, an investigator must take things as they are and do his best in trying to sense and disentangle them. In variables' language, ex post facto research means that an investigator starts by observing a dependent variable(s), and the possible causes for it i.e. independent variable(s), and then he studies the independent variable(s) retrospectively for its possible effect on the dependent variable(s). For example, if the sales (dependent variable) of a product have declined then one may like to study as to whether it was due to change in price (independent variable) or change in quality (independent variable) or some other factors.

Ex post facto method has been used in all fields of social sciences, dealing with problems which do not lend themselves to experimental inquiry. As a matter of fact a large number of researches in sociology, education and psychology are ex post facto. The method, in effect, has offered a valuable tool to:

- sociologists, who, for instance, wanted to study the cause of crime, drug addiction, delinquency, family breakdown, and many other social ills that afflicted every society;
- psychologists who wanted to study individual and group behaviour, roots of adult personality, racial discrimination, conflicts and disagreements, and child-rearing practices;
- educational scientists who wanted to study school achievement, teaching methods, intelligence, teacher personality, home environment, etc.

Many of the above studies could not have been made through the normal way of merely collecting the data and interpreting them, simply because they cannot be subjected to true experimentation.

4.6 CROSS-SECTIONAL STUDIES

These are the studies that are conducted over a group of companies or organisations over the same point of time. Such research makes observations at one and the same point of time for all the entities under study.

For example,

- Placement offers to MBA students of 2009 batch at all IIMs
- P/E (Profit to Earning) ratios of all automobile companies as on 31st March 2010.
- Conducting opinion poll on a particular day
- Percentage increase in net profits of a group of companies for the financial year 2009-10

Cross-sectional research study may be viewed as taking a '**Photograph**' or picture of a group of entities. For example, the following table gives the closing prices of some stocks on BSE and NSE on 18th November 2009.

Stock	*BSE*	*NSE*
Reliance	2102.45	2100.05
ONGC	1176.50	1179.85
Infosys Technologies	2433.60	2433.30
ICICI Bank	905.25	905.50
Bharti Airtel	296.95	296.00
L & T	1627.70	1627.85

Yet another example could be the study of highest or average marks obtained in the paper on Business Research Methodology by students in each of the areas of specialisation like marketing, finance, operations, IT and HR.

Cross-sectional study could also be made to study the relationship among several variables relating to an entity over the group of similar entities like all automobile manufacturers. The variables in this case could be sales, net profit, advertising expenses, etc. For example, one could study the changes in 'sales' and 'net profit' and their interrelationship in several similar companies in a particular sector like cement, in a particular year.

The major advantage of cross-sectional research is that data can be collected on many entities of different kinds in a **short span** of time.

Since the data is collected at one point of time, it can be easily collected at lower cost. For example, if we want to study the changes in share prices of some selected companies due to the announcement of budget, we may collect the data from just one newspaper of the next day. Similarly, if we wish to compare business parameters like deposits, credit, etc. of commercial banks for the year 2009-10 over the year 2008-09, we may refer to just one annual publication of Reserve Bank of India.

Thus, it may be noted that the cross-sectional data may give the position at one point of time, during the same period of time or it may indicate the change at one point of time over the previous day/week/year.

The main advantage of cross-sectional study is that it is cheaper and faster to conduct such a study. However, the main disadvantage of such study is that it reveals little as to how the changes occur.

4.7 LONGITUDINAL STUDIES

Longitudinal research studies are conducted over a period of time. An example of data in such studies is as follows:

Closing Prices of RIL on BSE and NSE

Date	*BSE*	*NSE*
16-11-2009	2147.75	2153.60
17-11-2009	2133.75	2132.35
18-11-2009	2102.45	2100.05
19-11-2009	2081.95	2083.80
20-11-2009	2125.15	2123.30

Longitudinal studies are quite popular in social and behavioural sciences, socio economic research, banking and finance, etc. These are conducted over a period of time. Such studies can be made to study changes in an individual or a group of individuals, a country or a group of countries over a period of time. Following are some more examples of longitudinal studies:

- Expenditure pattern over a period of time of an individual or a group of individuals
- 'Quality of Life' parameters of a state or a country
- NPAs (Non-performing Assets) of an individual bank or the banking industry
- R&D Expenditure by a sector of companies like pharmaceuticals
- Communication among people over a period of time in one or more regions/countries (postal, telephonic, e-mail, etc.)

It may be noted that all variables studied under longitudinal studies are measured over a period of time.

4.7.1 Advantages of Longitudinal Studies

The advantages of longitudinal studies are:

- Discover trends and patterns of change
- Locate the times when the trend or pattern changed – It might lead to investigating the factors that caused the change

4.7.2 Problems Associated with Longitudinal Studies

Several problems or issues that might arise because of the studies being conducted over a long period of time are as follows:

- Expensive
  Since a study is conducted over a period of time, the cost is quite high for collection of data. However, if the entire data of the past is available at one place in a publication, then the cost factor is not important.
- High dropout rates or obsolescence
  If a study is to be conducted over a period of time on the same individual/unit/company, etc., then there is the risk that
  - an individual may not be available after some time
  - an individual's behaviour/attitude may change with age
  - the unit or the company may close
  - even if the unit/company does not close, it might undergo substantial changes either due to internal or external factors, so that the comparison of data with the past may not be relevant. This has become all the more relevant because of frequent acquisition and mergers taking place in the industrial world.

4.7.3 Examples of Longitudinal Studies

Some of the examples of longitudinal studies are as follows:

- A study about employees' behaviour and attitude was conducted before and after pay revision, in a company.
- During the economic turbulence in 2008, in the banking systems in India, many customers of some banks switched their accounts to some other banks. After the stability in the banking system, as also in the bank, a public relation exercise was conducted to study about the change in perception of the customers towards the bank.

4.8 ACTION RESEARCH

Kurt Lewin, who was a professor at MIT, first coined the term **"action research"** in about 1944. He described action research as "a comparative research on the conditions and effects of various forms of social action and research leading to social action" that uses "a spiral of steps, each of which is composed of a circle of planning, action, and fact-finding about the result of the action". It stemmed from the belief that the motivation to change was strongly related to action: if people are involved in decisions affecting them, they are more likely to adopt new ways.

In action research, a researcher works in close collaboration with a group of persons to discuss a problem in a particular setting. Therefore, a researcher has to have skills of group management and well-versed with the mechanics of group behaviour and understanding of group dynamics. This type of research is more popular in areas such as organisational management, education, etc. But it is equally useful in solving any problem or resolving any issue arising in any situation whether it relates to business or personnel.

Action research begins with a process of communication and agreement among persons who together are involved in bringing about a change. It is imperative that these persons are open to new ideas and willing to discuss these in a positive and collaborative spirit. The discussions undergo four phases:

- Planning
- Acting
- Observing
- Reflecting

Action research uses various research methods like questionnaires, interviews and focus groups, as described in Chapter 5.

Action research can also be undertaken by larger organisations or institutions, assisted or guided by professional researchers, with the aim of improving their strategies, practices and knowledge of the environments within which they practice.

We would like to conclude with formal definition of Action Research by Reason and Bradbury as

"Action research is an interactive inquiry process that balances problem-solving actions implemented in a collaborative context with data-driven collaborative analysis or research to understand underlying causes enabling future predictions about personal and organisational change."

4.9 SAMPLING SCHEMES

This section emphasises understanding of the role of sampling that provides reliable information in a lesser time, at lower cost and even with lesser manpower. However, the advantages accrue only when a sample is a true representative of the population. This advantage is invaluable while conducting business research studies as we arrive at more accurate and reliable conclusions with lesser resources. This is achieved through the use of some sampling schemes described here.

4.9.1 Relevance of Sampling

The use of sampling in making inferences about a population is possibly as old as the civilisation itself. One of the simplest examples from day-to-day life is that a cook, while cooking rice, draws inference about the whole pot of rice by taking only few or even one piece of rice.

In the context of BRM, while planning for collection of data, one has to take a decision whether one would like to study the entire population or only a sample drawn from the population. If the population size is up to 500, or so, studying the entire population may not pose many problems. But, if the population size is more, then studying the entire population may pose some problems. Truly speaking, there is no cut-off point like 500, as mentioned above. It is only an illustrative figure. Nevertheless, a decision about studying only a part of the population through sampling or studying the full population may be based on the following considerations. In general, studying the entire population consumes more resources including man-hours, time, money, etc. At the same time, the conclusions drawn on the basis of the study of full population may be more accurate. However, it has been proved empirically that the census (studying each and every member of the population) may not lead to more reliable results as compared to studying only a sample from the population, as described below.

One of such situation arose in the Defence Department, USA, which was required to inspect the items before accepting them for use by the armed forces. Since it was a war-like situation, it was thought that all the items should be inspected to ensure defect-free supplies to the armed forces. Accordingly, a team of inspectors carried out 100% inspection of items, and the percentage of defective items was worked out. In order to test the effectiveness of such inspection, a sample was taken from the lot and was inspected by another team of highly competent inspectors. They found out the percentage of defective items in the sample. It was different from the percentage of defectives found through 100% inspection by the earlier team. The differences in the percentages of defective items, in the same lot of items, reported by two teams led to 100% thorough inspection of all the items in the lot by a third team of inspectors. This team reported yet another percentage of defective items in the lot. However, it was noted that the percentage of defective items reported by the third team was much closer to the percentage of defective items reported by the second team as compared to the first team.

This led to the conclusion that a sample collected and analysed by efficient persons provides a better idea of the population than 100% inspection carried out by ordinary persons. This is primarily because of the following two factors:

- While sampling, one can entrust the job to a few qualified and competent persons. It may not be possible to find more persons of such type if the entire population is to be studied. Obviously, the data provided by the well-qualified and competent people will be more reliable.
- It has been observed that when a person is required to do a job, repetitively, a psychological factor called 'human fatigue' sets in which may affect the efficiency of the person. Obviously, chances of the setting in of human fatigue factor are lesser in sampling than in census.

One important point while sampling is to ensure that the sample is a true representative of the population, i.e. if one looks at the sample with a magnifying glass, the sample should appear like the population, or in other words, **the sample should be like the photo or image of the population**. If this care is not taken, then the results obtained by the sampling may be misleading to varying extent.

For example, in one of the opinion polls surveys conducted in a country during Presidential elections, an agency selected the prospective voters for ascertaining their opinion with the help of the telephone listings in the telephone directory. The agency did not realise that the persons listed in the telephone directory were not representative of the entire voting population. In fact, as could be well imagined, the listed persons belonged to the affluent section of the society whose percentage in the entire voting population could be small. Further, it is observed, in general, that the percentage of the affluent persons going for voting is lesser than the other classes of voters. Because of these factors, the prediction made by the agency about the poll results was totally off the mark. The agency suffered a big setback in its reputation, and had to eventually stop its publication.

If due care is taken in selecting a representative sample from the population, the results obtained will, generally, not only be more reliable and accurate but also will consume lesser resources in terms of manpower, time, money, etc.

4.9.2 Census Vs Sampling

The collection of data from a population can either be on sampling basis or census basis. Some of the comparative features of both the methods are described as follows:

4.9.2.1 Census: Advantages and Disadvantages

Advantages: Accurate and reliable – however, this advantage is a myth if the population is quite large.

Disadvantages:

(i) More resources in terms of men, money, time, etc.

(ii) If the test is destructive i.e. the item is destroyed while collecting the information about the item, this option is totally ruled out. Some examples are:
- Estimating the life of bulbs/tubes, etc.
- Testing the quality of bullets, fuses, etc.
- Testing the quality of food, etc.

4.9.2.2 Sampling: Advantages and Disadvantages

Advantages: The advantages of sampling are as follows:

(i) Less resources in terms of manpower, money, time, etc.

(ii) Highly qualified and skilled persons can be deployed for collection of data as the manpower requirement is relatively low. This aspect assumes greater significance when the collection of data requires special skills or knowledge.

(iii) Indispensable or a must if in the process of getting the desired information about the unit, it gets destroyed (items like, bullets, fuses) or gets consumed (e.g. fruits) or becomes useless (item like an electric bulb, tubelight) after its failure time or "life" is recorded.

Disadvantages: The disadvantages of sampling is that it is, less accurate and reliable because the sample may not be a true representative of the population. This disadvantage can be minimised by selecting a sample such that it is a true representative of the population, but it cannot be eliminated. However, if the population size is quite large and the collection of data needs more than normal knowledge and skills, this disadvantage could be eliminated.

On the whole, if the population size is large, sampling is much better option than census. It is in this context that, while discussing any methodology or analysing any data, data obtained through sampling is referred.

4.9.3 Size of a Sample

The size of a sample, i.e. the number of units to be selected in a sample, depends on several factors as indicated below:

- **Population Size** – Normally bigger the size of the population, bigger will be the sample size needed to draw meaningful conclusions from the sample.
- **Heterogeneity** in the population's concerned characteristic e.g. age or income in the case of human population or life of electric bulbs in the case of a physical item, or high school examination results in the case of the population of schools. More the heterogeneity in the data, more the size of the sample required. As mentioned earlier, in the case of rice being cooked, even a sample of one piece of rice is sufficient to draw conclusion about the extent of cooking.
- **Accuracy and Reliability**—In general, results obtained from a bigger size sample would be more accurate and reliable as compared to results obtained from a smaller size sample. Therefore, more the accuracy and reliability required, more would be the requirement of sample size.
- **Allocation of Resources**—The sample size depends on the resources allocated or made available. Obviously, more the resources in terms of manpower, money, time, etc. are made available, more the sample size can be increased.

Allocation of resources and accuracy/reliability desired are interrelated–if more accuracy/reliability is required, more the resources have to be allocated.

4.9.4 Sample Size Calculations

The calculations leading to the desired sample size based on the above considerations are given in Chapter 11 on Statistical Inference.

As an example, if we want to estimate the mean of a population characteristic like income, with the level of confidence as 99%, margin of error as 40 and standard deviation as 200, the minimum sample size required is 166.

The method of calculation is illustrated in Chapter 11 on Statistical Inference.

4.9.5 Some Terms and Definitions:

Before we proceed further, let us define and explain some terms in sampling design.

Unit: A unit is an element or a group of elements, living or non-living, on which observations can be made. For example, a person living in a city or a household in the city, or an account, or an employee or a branch, in a bank, etc.

Population (or Universe): The collection of all the units of a specified type at a particular point or period of time is called a population or universe. For example, the persons or households in a given city, or the accounts at a branch of a bank or branches of a bank, themselves could constitute a population. The total number of units, generally denoted by '*N*', in the population is called population size. A population is said to be a finite population or an infinite population, depending on whether its size is finite or infinite.

Sample: One or more units, selected from a population according to some specified procedure are said to constitute a sample. The number of units, selected in the sample, is called the sample size, and is usually denoted by "*n*".

Sampling Frame: We ordinarily select a sample of units by selecting the numbers that identify them, e.g. savings accounts at a bank's branch. Generally, we first, identify each unit of the population by giving them a distinct number, generally from 1, 2, 3, …, N where N is the population size.

Sampling With or Without Replacement: Let a population consist of N units. If a sample of size n units is obtained by first selecting one of the N units, replacing it, then making a second selection and replacing the unit before making a third selection, etc. until n selections are made, then the sample is said to be selected with replacement. Since there are N possible results in each of the n selection, the total number of possible samples of size is ${}^{N}P_{n}$. It is to be noted that a unit could appear more than once in the sample.

If a sample of n units is obtained by first selecting one of the N units, and, without replacing it, selecting one of the remaining $(n - 1)$ units, and without replacing the two selected units, and so on, so that at the nth selection, there are $(N - n + 1)$ units, then we say that the sample has been selected without replacement. It may also be obtained by the first method (i.e. with replacement) if the number of selections are continued till n distinct (different) units are selected and all repetitions are ignored. In this case, the total number of possible samples is ${}^{N}C_{n}$.

But these are unordered samples. Each of these ${}^{N}C_{n}$ samples has $n!$ ordered samples, and hence the total number of all possible ordered samples is $n!\ {}^{N}C_{n}$.

Illustration 4.7:
Let a, b and c be three units in the population, and we want to select a sample of 2 units, i.e. $N = 3$ and $n = 2$. The possible samples in two cases (i.e. with and without replacement) are given below:

(i) Sampling With Replacement
In this type of sampling,
Total Number of possible samples $= {}^{3}P_{2} = \dfrac{3!}{(3-2)!} = \dfrac{3!}{1!} = 6$
and all possible samples of size 2 are:
aa ab ac bb bc cc

(ii) **Sampling Without Replacement**
In this type of sampling,
Total number of possible samples $= {}^{3}C_{2} = \dfrac{3!}{2!(3-2)!} = \dfrac{6}{(2)(1)} = 3$
and all possible samples of size 2 are:
ab ac bc

Now, we proceed to discuss some commonly used sampling schemes with their salient features. These are:

- Simple random sampling
- Systematic sampling
- Stratified sampling
- Cluster sampling

These are described as follows.

4.9.6 Simple Random Sampling

Sample Selection Procedure (omitting with or without sampling)
Step 1: First of all, the sampling frame is prepared.
Step 2: Then the 'n' random numbers, where n is the sample size, are chosen from the Random Number Table 4.3 in the following manner.

First, we decide the digit of the random numbers to be chosen and it depends on the population size N. If N is a one-digit number, then one-digit random number is used, if N is a two-digit number, then two-digit random number is used, and so on. Examples of three-digit numbers are given in Section 4.9. Then n random numbers, each less than or equal to N are chosen from the random number table by starting, usually, from the first row of any one selected column of the table and observing each number of the selected column from its starting row. If the column is completely exhausted, and we do not find the required number n of random numbers from the column, then we use the next column in the continuation, and so on.

Step 3: Finally, the units corresponding to the selected n random numbers are listed separately which constitute the sample and further required information is collected from them according to the plan of the survey.

Estimation of the Population Mean

The estimate of the population is the sample mean calculated as

$$\bar{x} = \frac{\Sigma x_i}{n}$$

where x_i is the value of the characteristic of the ith unit in the sample ($i = 1, 2, 3, \ldots, n$), where n is the sample size.

The sample mean may change from sample to sample. The extent of variation in these sample means is measured by a quantity called **standard error (s.e.)** and is equal to

$$\text{Standard Error of } \bar{x} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

where n is the sample size
N is the population size
σ is the s.d. of population

The ratio of sample size to population size i.e. n/N is called **sampling fraction.**

If N is very large say, 10,000, and n is comparatively small, say 100, then $\sqrt{\frac{N-n}{N-1}}$ is $\sqrt{\frac{9900}{9999}}$, which is approximately equal to 1, and therefore,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Thus, if the population size is infinite or large, the s.d. of sample mean, also called **standard error**, is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

but if the population size is finite i.e. not large, then the s.d. of the sample mean is to be multiplied by the term

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

For this reason, the above term is called **finite population correction**.

4.9.7 Other Sampling Schemes

Now we shall discuss three other sampling schemes which are used in varying situations. These are:

- **Systematic Sampling**

- **Stratified Sampling**
- **Cluster Sampling**

These are described as follows.

4.9.7.1 Systematic Sampling In simple random sampling, the units in the sample are selected with the help of the random number table. However, there is another method of sampling in which only the first unit of the sample is selected with the help of the random number table, and the rest are selected automatically according to a pre-determined pattern. The method is known as Systematic Sampling.

For instance, there are 50 students in a class, and each of them has the roll number from 1 to 50. Suppose, we wish to select 10% i.e. 5 of the students for assessment of their views on the library facilities, then, we may select one random number from single digit random numbers varying from 0 to 9, say 5. Thus, the first student in the sample would be the one with student number 5. The number of the next student would be obtained by adding 10 to the first number 5, i.e. 15, and so on. Thus, the five students selected in the sample would be the students with students: 5, 15, 25, 35 and 45. If the first random number selected was 0, then the first student would have been the one with student number as 10, and the subsequent students in the sample would have been 20, 30, 40 and 50. The detailed procedure is outlined below:

Selection Procedure

Step 1: In systematic sampling also, first the sampling frame is prepared. However, the units are not just randomly identified like in simple random sampling, but they are first arranged in a fixed order with the help of some information and according to the purpose in mind. For example, in the illustration cited above, the roll number was used to arrange the students in a particular order.

Step 2: Depending on the sample size desired, sampling interval is worked out. The interval gives the difference among successive units to be selected in the sample. The sampling interval, say I, is the ratio N/n. If N/n is not an integer, its integral part could be treated as I for the purpose of selecting the sample. Then a random number, say R, is selected from the appropriate random number table such that $1 \leq R \leq I$.

Step 3: The units corresponding to the serial number R, $R + I$, $R + 2I$, …, $R + (n - 1)\ I$ would constitute the sample.

The above method of sampling is known as **Linear Systematic Sampling**.

If N/n is not an integer, the sample size will vary, being either n or $n + 1$, depending on the random number selected.

For example, if $N = 10$ and $n = 3$, then $N/n = 10/3 = 3.33$, and therefore $I = 3$ (integral part of 3.33). Now, a number is to be selected from 1 to 3 . Suppose the number is 3, then the sample to be selected would be the units corresponding to 3, 6 and 9. However, if the random number selected between 1 and 3 is 1, then the units in the sample would be the units corresponding to 1, 4, 7 and 10 i.e. a sample of size 4.

One way of avoiding the difficulty of varying sample size is to select the sample through **Circular Systematic Sampling**, described below.

Select a random number R from 1 to N. Start at the unit corresponding to this number, and thereafter, select cyclically every Ith unit (I is the integer nearest to N/n) until n units are chosen for the sample. The cyclic selection involves assigning the number $N + 1$ to the first unit in the sampling frame, $N + 2$ to the second unit, and so on. This is done to continue sampling even when the nth unit has been reached. For example, in the above illustration when the population size was 10 and

the sample size required was 3, *I* worked out to be 3. Now, if the random number selected between 1 and 10 is, say 6, then the numbers to be included in the sample would be 6, 9 and 12. However, since the number of units in the population is 10, the unit corresponding to the number 12 would be 2 (= 12 – 10), and this would be included in the sample. Thus, the sample would comprise the units 2, 6 and 9. This procedure is known as Circular Systematic Sampling.

In the above example of 50 students in the class, suppose the sample size desired is 8 students. Then the sampling interval *I* is equal to the integral part of 50/8 i.e. 6. Instead of taking a random number between 1 and 6 as in linear systematic sampling, to start the circular systematic sampling, one takes a random number between 1 and 50. Suppose it is 30. We divide 50 by 30 to get the quotient as 1 and the remainder as 20. Now, we start sampling with this number 20, and get subsequent numbers in the sample as 26 (20 + 6), 32 (26 + 6), 38 (32 + 6), 44 (38 + 6), 50 (44 + 6), 56 i.e. 6 (50 + 6), and 12 (6 + 6). Thus, the sample of eight students would comprise employees with numbers 6, 12, 20, 26, 32, 38, 44 and 50.

The arithmetic mean is an estimator of the population mean.

4.9.7.2 Stratified Sampling Stratified sampling involves classifying the population into a certain number of non-overlapping homogeneous groups called strata, and then selecting samples independently from each stratum (singular form of strata). For example, in India the entire population can be classified into two strata as rural and urban, in three strata as 'lower income group', 'middle income group', and 'higher income group' or in three strata as 'children' (up to 18 years of age), 'adults' (more than 18 years of age) but up to 60 years of age, and 'senior citizens' (more than 60 years of age).

The strata should be as homogeneous as possible within each stratum, and as heterogeneous as possible among various strata.

The main advantage of using the stratified sampling is to increase the efficiency (concept explained in Chapter 11) of the estimators estimating the population characteristics. The other advantages are as follows:

(i) When estimates are required with given precision not only for the population as a whole but also for the various strata, stratified sampling is used.

(ii) When an organisation has field offices in various zones in which the country may have been divided for administrative purposes, it might be desirable to treat zones as strata for facilitating the organisation of fieldwork.

(iii) When an organisation has field offices in various zones in which the country may have been divided for administrative purposes, it might be desirable to treat zones as strata for facilitating the organisation of fieldwork.

(iv) When there is some periodical variations in the population, like hill stations, religious places, etc., the stratification is useful for reducing the chances of getting bad samples.

Further, stratification is highly useful when the population is skewed e.g. income of individuals, business at branches of a commercial bank, productivity of rice (yield per acre) in various parts of the country, etc.

Selection Procedure

Step 1: First, the population of *N* units is divided into '*k*' strata with the help of the prior knowledge, intuition, etc. such that the strata are as homogeneous as possible. Let the number of units in the *i*th stratum be N_i; *i* varying from 1 to *k*. Thus,

$$N = \Sigma N_i$$

Step 2: The total sample size n is allocated to each stratum in such a way so as to provide an estimate of the population mean with maximum precision for a given cost. Such allocation is referred to as the principle of Optimum Allocation.

But due to practical difficulties in arriving at these allocations, mostly Proportional Allocation or sometimes Equal Allocation is used in practice. Let n_i denote the sample size allocated to the ith stratum, then

$$n = \Sigma n_i$$

Step 3: After the strata and the sample sizes in each stratum are determined, the samples are drawn independently from various strata following any of the above mentioned sampling scheme viz. Simple Random, Systematic, etc., which is most efficient or/and convenient to the stratum concerned.

One of the biggest advantages of stratification is that one can use different sampling schemes in different stratum.

Estimation of Population Mean

Let $\overline{X}_i$ denote an unbiased estimator of the mean of the ith stratum, then an estimator of the population mean is given by:

$$\overline{X}\text{st} = \Sigma W_i \overline{X}_i$$

Where,

$$W_i = N_i/N$$

4.9.7.3 Cluster Sampling Sometimes, the entire population is grouped into clusters which comprise a number of units each. For example, the entire rural area in the country or in a district could be divided into villages – the villages constituting the clusters.

In cluster sampling, the sampling unit is a cluster. Thus, cluster sampling involves formation of suitable clusters of units, and then selecting a sample of clusters treating them as units by an appropriate sampling scheme. It is to be noted that all the units of each selected cluster are enumerated. The advantage of cluster sampling from the point of view of cost arises mainly due to the fact that collection of data for nearby units is easier, faster, cheaper and more convenient than observing units scattered over a wide area.

As an illustration, a bank could be divided among its branches—branches constituting the clusters, if the intended study is to exclude regional offices and the head office. For example, a bank wanted to assess the computer training needs of its employees working in branches. There were three types of branches—manually operated, partially computerised and totally computerised. Their numbers were 560, 250 and 75. The bank decided to follow cluster sampling approach i.e. selecting samples of branches of each type and then ascertaining the needs of all the employees at those branches rather than taking samples of individual staff members from those posted at all the branches.

4.9.7.3.1 Multi-stage Sampling It is to be noted that if the number of units is fixed because of resource constraints and the size of clusters is large, then only a limited number of clusters could be selected for 100% observations in those clusters. For example, suppose the number of units, which could be included in the sample for observation, is 500. Further, assume that the size of clusters is about 100 each, and in all there are 40 clusters. In such a case only 5 clusters could be selected for recording observations on each unit in the clusters, and there will be no information available for the rest 35 clusters.

It can be proved mathematically that for a given number of units, distributing the units over a large number of clusters leads to greater precision than by taking a small number of clusters and completely enumerating them.

In consonance with the above, a modification of cluster sampling called Two-Stage sampling or Sub-sampling was evolved. Such a scheme envisages first selecting clusters and then choosing specified number of units from each selected cluster. This implies that more clusters are included in the sample, but instead of recording observation for each and every unit in these clusters, only samples are selected.

The clusters which are selected randomly at the first stage are called the First Stage units, and the units or the groups of units within clusters which are selected subsequently are called **Second Stage units**. Such a scheme could be extended to three or more stages and is termed Multi-stage sampling.

Two-stage sampling, used in the above illustration relating to 40 clusters, could be like this: First, random sampling is resorted to select, say 10 clusters out of 40 clusters. Then, from each cluster, samples of 50 units are selected at random.

Number of Clusters = 40, Number of units in each cluster = 100

Number of units in the population = 40 × 100 = 4000

Thus, in cluster sampling, a sample of 500 units was selected with all the100 units from each of the 5 clusters selected at random out of the 40 clusters. In Two-stage sampling, the number of clusters was increased to 10 to get better representation of clusters, but the sample size from each cluster was reduced from 100 to 50, thus maintaining the overall sample size to 500. This can also be illustrated through a tabular presentation given below.

Cluster Sampling	Two-stage Sampling
Number of Clusters selected = 5	Number of Clusters selected = 10
Number of units selected from each of the 5 clusters = 100 Sample size = 500	Number of units selected from each of the 10 clusters = 50 Sample size = 500

4.9.8 Non-probability Sampling

At this stage, it may be mentioned that all the sampling schemes described above are categorised as **Probability Sampling** wherein **each** unit of population has some predefined probability of being selected in a sample. However, there is another category of sampling schemes which are referred to as **Non-probability Sampling.** In this type of sampling scheme, the selection of units is subjective and not based on any probability considerations. Each of the schemes is suitable for a particular situation, and in general, is better than the schemes described earlier in the chapter. These schemes are described below.

Haphazard Sampling

This method of selecting a sample is in total contrast to random sampling. In such sampling, the units from the population are selected without any set criteria. They are selected based on the preference, prejudice or bias of the person(s) selecting the sample. This method is usually followed when one wants to know the opinions of people in the crowd coming out after watching a film or a play, etc.

Purposive Sampling

As the name implies, under this type of sampling, units of the population are selected according to the relevance and the nature of representativeness of sampled units. For example, if one wants to assess the likely reaction of employees to certain new measures contemplated in an organisation, it might be better to include those employees in the sample who are likely to influence on the thinking and actions of a vast majority of the employees. Incidentally, the sample size in such cases is not fixed. We may terminate the sampling i.e. recording of information when we feel that no further information or suggestion is being obtained.

Quota Sampling

Such sampling is, sometimes, considered a type of purposive sampling. It is usually resorted when some quota about the number of units to be included in the sample is fixed. The quota is fixed due to constraints on availability of time or/and cost. Within the quota stipulated, one has to select a sample which is representative of the entire population. For example, within the overall quota of interviewing 100 persons for some opinion poll, one may contact some persons from various categories like college students, housewives, shopkeepers, office-goers, daily wage earners, etc. Similarly, in an organisation, one might include persons from all categories of staff cadre-wise as well as function-wise, department-wise, etc.

Judgement Sampling

In such type of sampling, the selection of units, to be included in the sample, depends on the judgement or assessment of the person(s) collecting the sample. The sample is selected based on their judgement/assessment as to what would constitute a representative sample. This is specially useful when the sample size is small, and if random sampling is adopted, then the units which are more important and critical to the objective of the study might not get included in the sample.

For example, in a training institute, the teaching staff was 30. However, for urgent academic or administrative matters, the Director used to get opinion of one particular faculty as he was known to have balanced views, did not belong to any group, and was frank enough to express his views. Thus, the Director used to rely on a sample of size one.

Convenience Sampling

Such sampling is dictated by the needs of convenience rather than any other consideration.

For example, one may select some persons from a telephone directory, for getting their opinion on some issue provided the views of those who own phones are relevant to the issue. For instance, their views on TV programmes might be relevant but their views on some party or a person in a general election may not be much relevant as they represent only relatively affluent class of people.

Similarly, one could select a sample of persons from the list of credit card holders.

Another example relates to opinion poll when one may find it easier to get the opinion of those in the shops or restaurants or walking on pavement rather than going from house to house.

Snowball Sampling

Snowball sampling—also known as **chain referral sampling**—is considered a type of purposive sampling. In such sampling, the sampling units are not fixed in advance but are decided as the sampling proceeds. We may move to sample the units one after the other depending on the response received from the previous units. If the units are human beings, one individual might refer to other individual who, in turn, might refer to some other individuals. That is how it is called as "chain

referral' sampling. In this method, participants or informants with whom contact has already been made use their influence/social networks to refer the researcher to other people who could potentially participate in or contribute to the study. Snowball sampling is often used to find and recruit 'hidden populations', that is, groups not easily accessible to researchers through other sampling strategies.

Inverse Sampling

In normal sampling, we take a sample of units and estimate about the characteristics of the units in the population. However, if the proportion of units of a certain type is very small like fake notes in circulation, then the method may not work. For instance, if we do not find any fake note in a sample of 1000 pieces, examined at random, can we say that the proportion of fake notes is zero? **In such cases, inverse sampling could be used.**

Inverse sampling is a method of sampling which requires that drawings of random samples shall be continued until certain specified conditions dependent on the results of the earlier drawings have been fulfilled, e.g. until a given number of units of specified type have been found.

Thus, referring to the above problem of estimating fake notes, we may continue taking samples till we reach a certain number, say 10, of, say, Rs. 100 denomination notes. Suppose, we find 10 fake notes in a total sample of 10 lakhs pieces, it implies that the chance of a note being fake is 10/10 lakhs i.e. 1 in 1 lakh. Thus, if the number of Rs. 100 notes in circulation are 100 crores i.e. 10,000 lakhs, then about 10,000 lakhs × (1/1,00,000) = 10,000 notes are fake notes. It is only a rough approximation but better than pure guess.

In the case of human beings, this method may be used to estimate the number of persons with some rare characteristic like say, having 6 fingers on a palm, or some rare disease.

For illustrating the process of inverse sampling, a simplistic approach for estimating fake notes is described as follows:

Starting from a day, say, 1st January, a daily record could be kept of the number of notes (coming back to the Reserve Bank after circulation) examined at random, and number of fake notes detected. When the number of fake notes reaches 5, this number 5 divided by the number of notes examined would provide a quick estimate of the proportion of fake notes. Selection of number 5 is arbitrary, and is based on the assumption that proportion of fake notes in the system is very small. When the number of fake notes reaches 10, it will provide a better estimate. The estimate will keep on improving, and might stabilise at certain level which could be considered a fair estimate of the proportion of fake notes which have successfully passed through the system and have come back to the Reserve Bank.

Of course, there are certain assumptions implicit in this approach, and it can be improved by taking into consideration several other factors which are beyond the scope of this book.

4.9.9 Comparison of Probability Sampling Schemes

Type	*Description*	*Advantages*
Simple Random	Each population unit has an equal chance of being selected into the sample. Sample is drawn using random number table	• Easy to implement

(*Contd*)

(*Contd*)

Systematic	• Selects an element of the population at a beginning with a random start and following the sampling fraction (n/N) selects every Ith element, here I (I integral part of N/n) is the sampling interval • Linear (random no. R selected between 1 and I); the sample comprises R, $R + I$, $R + 2I$,… • Circular (random no. R selected between 1 and N; the elements to be selected as R, $R + I$,… I are the nearest integer to N/n)	• Simple to design • Easier to use than the simple random • Less expensive than simple
Stratified	Divide population into subpopulations or strata and use simple random sampling for each strata. Results may be weighted and combined.	• Sample size can be allocated to each stratum according to criteria of size, cost, etc. • Increased statistical efficiency • Provides data to represent and analyse subgroups • Enables use of different methods in different strata
Cluster	Population is divided into appropriate subgroups. Some are randomly selected for further study either in full or part (two stage)	• Provides an unbiased estimate of population parameters if properly done • Economically more efficient than simple random • Lowest cost per sample, especially with geographic clusters • Easy to do without a population list

Table 4.3 Random Number Table

298	997	555	848
061	205	180	020
206	587	281	154
099	254	441	484
568	563	616	495
669	984	048	455
875	408	324	700
949	258	699	948
113	428	802	882
229	800	399	961

Illustration for Using the Table for Selection on Random Basis

Suppose, a random sample of 10 students is to be selected from a class of 50 students. We may number the students from 1 to 50. Then, we may select the first two digits from the above random numbers in any pre-defined order. For example, we may select the second column so that the digits are

99, 20, 58, 25, 56, 98, 40, 25, 42 and 80

For numbers greater than 50, we may divide the number by 50 and take the remainder. Thus, the numbers are

49, 20, 8, 25, 6, 48, 40, 25, 42 and 30

Since one number 25 is repeated, we may take the first number in the third column i.e. 55 (first two digits of 555). Dividing this number by 50, we get the remainder as 5. Thus, the sample of 10 students would be

5, 6, 8, 20, 25, 30, 40, 42, 48 and 49.

4.10 SIMULATION

So far, we have discussed three types of data by:

- Observing e.g. model of a car, colour of a shirt, gender of a person, etc.
- Recording e.g. income of an individual, salary offered to MBA students in campus placements, sales turnover of a company, price of a stock, etc.
- Soliciting (in writing) e.g. through questionnaire/schedule filled by a respondent
- Experimenting and recording e.g. impact of a medicine on a patient, yield of a variety of rice by using a particular fertiliser, marks obtained by students through examination, etc.

However, there is yet another method of collecting data, and it is through simulation.

In the context of BRM, simulation is used to generate data for carrying out the desired research study without actually observing, recording, collecting or conducting an experiment.

For example, if we want to assess the likely impact of variation in 'sale price' per unit of an item on the profit of a company manufacturing that product and if the company were to fix a certain price, then traditional approach would be to record its sale after fixing that certain price, increase the price further by 5%, record its sale, again increase the sale by 10%, and again record its sale. But this will be a time-consuming process. One of the simple ways is to set up a mathematical formula relating 'sale price' (S.P.) to the profit. Let the formula be

$$\text{Profit} = n \times \text{S.P} - n \times \text{C.P.} - 200$$

where n is the number of units sold (say, 100), C.P. is the cost per unit, say Rs. 2, and 200 is the fixed cost. Let S.P. be Rs. 10 per unit.

Using the above formula, we can see the impact of varying S.P. by increasing S.P. to 12, 15 and 20, in the following table:

S.P.	*Profit*
10	600
12	800
15	1100
18	1400

It may be noted that the above profit values have been **'generated'** by using the given formula, and help us to assess the impact of S.P. on the profit of the company. What we have described here is a very simplistic situation just for illustration. In real life, simulation is more complex.

Incidentally, we may recall or note that, in the above case, we also get an idea as to how sensitive is the profit to change in S.P. This is called **'Sensitivity Analysis'**. However, there is a difference between sensitivity analysis and simulation. As we shall see later, in simulation, the independent variables like S.P. are selected randomly rather than mathematically, as we have done.

When we use the word simulation, we refer to any analytical method meant to imitate a real-life system, especially when other analyses are too mathematically complex or too difficult to reproduce.

Simulation is defined by T. H. Taylor as "A numerical technique for conducting experiments on a digital computer, which involves certain types of mathematical and logical relationships necessary to describe the behaviour and structure of a complex real word system over extended period of time."

Even though simulation is a very sophisticated technique, as indicated by its formal definitions given above, some simple aspects of simulation are described in this section.

4.10.1 Types of Simulation

There are two main types of simulation:

- **System Simulation**

This technique is applicable to situations in which the business or operating environment is reproduced with all its complexity and the impact of alternative management actions is analysed. System Simulations are commonly used in engineering and technological applications.

A classical example is that of flight simulators used for training of pilots.

- **Monte Carlo Simulation**

Monte Carlo simulation is named after the place Monte Carlo in Monaco, where the primary attractions are casinos containing games of chance. It may be noted the games of chance, such as roulette wheels, dice and slot machines, are based on random factor or matter of chance.

This technique uses modelling of key variables with defined random distributions to cover potential values in solving analytical problems. This technique is more popular in business applications because it generally involves less complexity and can be implemented more quickly and at a lower cost than System Simulation. **Therefore, we shall discuss only Monte Carlo simulation**.

4.10.2 Monte Carlo Simulation

Monte Carlo (MC) methods are stochastic techniques—meaning they are based on the use of random numbers and probability statistics to investigate problems. We can find MC methods used in everything from economics to nuclear physics to regulating the flow of traffic. The use of MC methods to model physical problems allows us to examine more complex systems than we otherwise can. Solving equations which describe the interactions between two atoms is fairly simple; solving the same equations for hundreds or thousands of atoms is impossible. With MC methods, a large system can be sampled in a number of random configurations, and that data can be used to describe the system as a whole.

4.10.3 Steps for Monte Carlo Simulation

The steps for Monte Carlo Simulation are as follows:

Step 1: Create a parametric model
Step 2: Generate a set of random inputs
Step 3: Evaluate the model and store the results
Step 4: Repeat steps 2 and 3 for $i = 1$ to n.

Step 5: Analyse the results using histograms, summary statistics, confidence intervals
This can be explained by the following example:
Reliance Company wants to know how profitable it will be to market their new gadget, realising there are many uncertainties associated with market size, expenses and revenue. It is planning to use the Monte Carlo simulation to estimate profit and evaluate risk. The simulation exercise comprises the following steps:

Step 1: Creating the Model
We use a top-down approach to create the sales forecast model, starting with:
Profit = Income – Expenses

Both income and expenses are uncertain parameters, but we aren't going to stop here, because one of the purposes of developing a model is to try to break the problem down into more fundamental quantities. Ideally, we want all the inputs to be independent. Does income depend on expenses? If so, our model needs to take this into account somehow.

We assume that income comes solely from the number of sales (S) multiplied by the profit per sale (P) resulting from an individual purchase of a gadget, so that

$$\text{Income} = S*P$$

The profit per sale takes into account the sale price, the initial cost to manufacture or purchase the product wholesale, and other transaction fees (bank credit, shipping, etc.). For our purposes, we assume that P may fluctuate between Rs. 2350 and Rs. 2650.

We could just leave the number of sales as one of the primary variables, but for this example, the company generates sales through purchasing leads. The number of sales per month is the number of leads per month (L) multiplied by the conversion rate (R) (the percentage of leads that result in sales). So our final equation for income is:

$$\textbf{Income} = \boldsymbol{L*R*P}$$

We assume the expenses to be a combination of fixed overhead (H) plus the total cost of the leads. For this model, the cost of a single lead (C) varies between Rs. 10 and Rs. 40. Based upon some market research, the Reliance Company expects the number of leads per month (L) to vary between 1200 and 1800. Thus, the final model for the Reliance Company sales forecast is:

$$\textbf{Profit} = \boldsymbol{L*R*P - (H + L*C)}$$

It may be noted that H is also a part of the equation, but for illustration purpose it may be treated as a constant. The **inputs** to the Monte Carlo simulation are just the uncertain parameters L, C, R and P.

Normally, all the inputs in the model should be independent variables. In this case, even though L is related to income as well as expenses, and thus, income and expenses are not independent. However, for the sake of simplicity, we shall assume that L, R, P, H, and C are all independent.

Step 2: Generating Random Inputs
The key to Monte Carlo simulation is generating the set of random inputs. As with any modelling and prediction method, the "garbage in equals garbage out" applies equally to this method.

Input Values

	Nominal	*Min*	*Max*
Leads per Month (*L*)	1500	1200	1800
Cost per Lead (*C*)	25	10	40
Conversion Rate (*R*)	3.0%	1.0%	5.0%
Profit per Sale (*P*)	2500	2350	2650
Overhead per Month (*H*)	40000		

The table above uses "Min" and "Max" to indicate the uncertainty in *L*, *C*, *R*, and *P*. To generate a random number between "Min" and "Max", we use the following formula in Excel (Replacing "min" and "max" with cell references):

= min + RAND()*(max – min)

We can also use the Random Number Generation tool in Excel's Analysis ToolPak Add-In to generate a bunch of static random numbers for a few distributions. However, in this example we make use of Excel's RAND() formula so that every time the worksheet recalculates, a new random number is generated.

Suppose, we want to run $n = 5000$ evaluations of our model. Incidentally, this is a fairly moderate number when it comes to Monte Carlo simulation.

A very convenient way to organise the data in Excel is to make a column for each variable as shown in the following Excel snapshot.

Cell A2 contains the formula:

=Model!$F$14+RAND()*(Model!$G$14-Model!$F$14)

Note that the reference Model!$F$14 refers to the corresponding Min value for the variable *L* on the Model worksheet, as shown in Figure 4.1.

To generate 5000 random numbers for *L*, one can simply copy the formula down 5000 rows. We repeat the process for the other variables (except for *H*, which is constant).

Step 3: Evaluating the Model

Since our model is very simple, all we need to do, to evaluate the model for each run of the simulation, is to put the equation in another column next to the inputs, as shown in Figure 4.1 (the Profit column).

Cell G2 contains the formula:

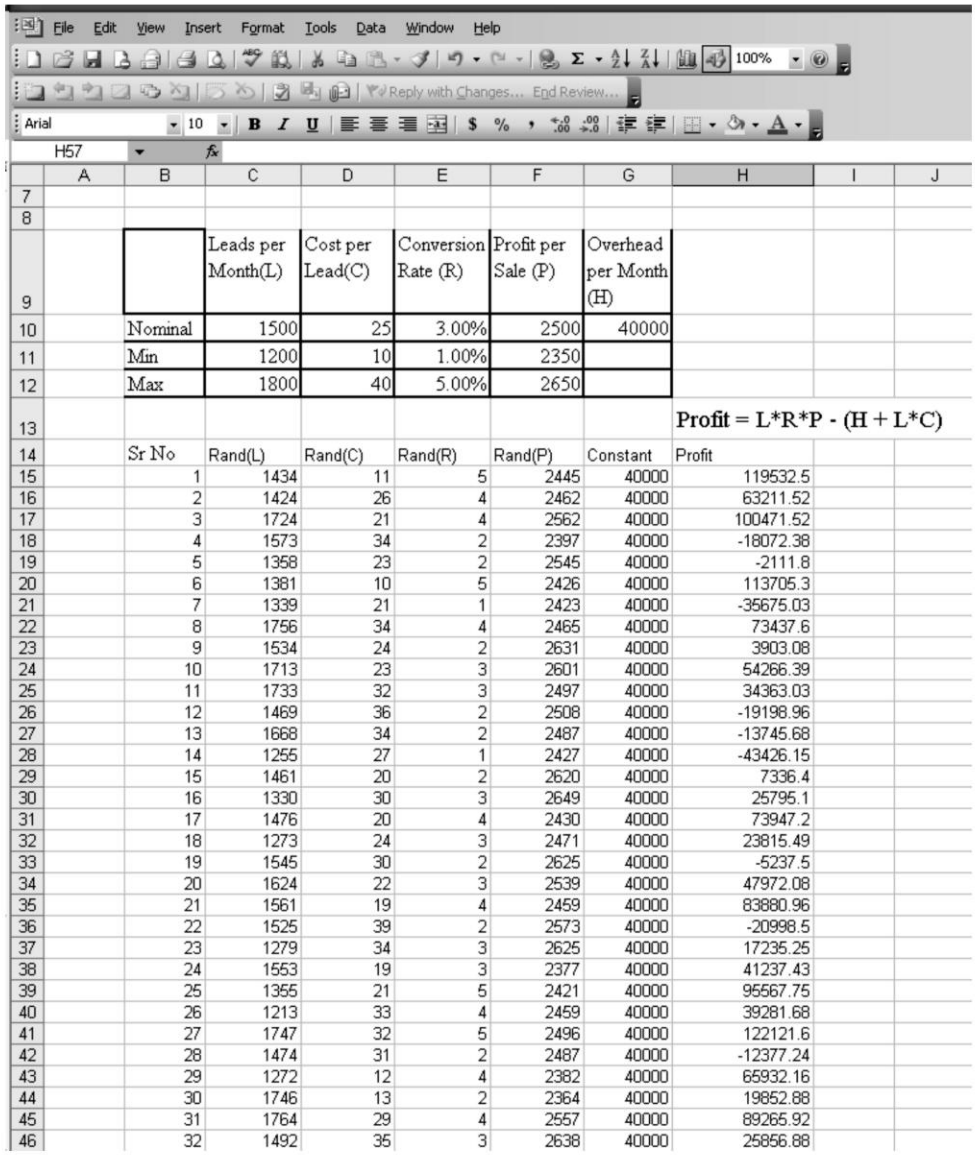
=A2*C2*D2-(E2+A2*B2)

Step 4: Run the Simulation

We do not need to write a macro for this example in order to iteratively evaluate our model. We simply copy the formula for profit down 5000 rows, making sure that we use relative references in the formula

Rerun the Simulation: F9

Although we still need to analyse the data, we have essentially completed a Monte Carlo simulation. Because we have used the volatile RAND() formula, to re-run the simulation all we have to do is recalculate the worksheet (F9 is the shortcut).



	Leads per Month(L)	Cost per Lead(C)	Conversion Rate (R)	Profit per Sale (P)	Overhead per Month (H)
Nominal	1500	25	3.00%	2500	40000
Min	1200	10	1.00%	2350	
Max	1800	40	5.00%	2650	

$$\text{Profit} = L*R*P - (H + L*C)$$

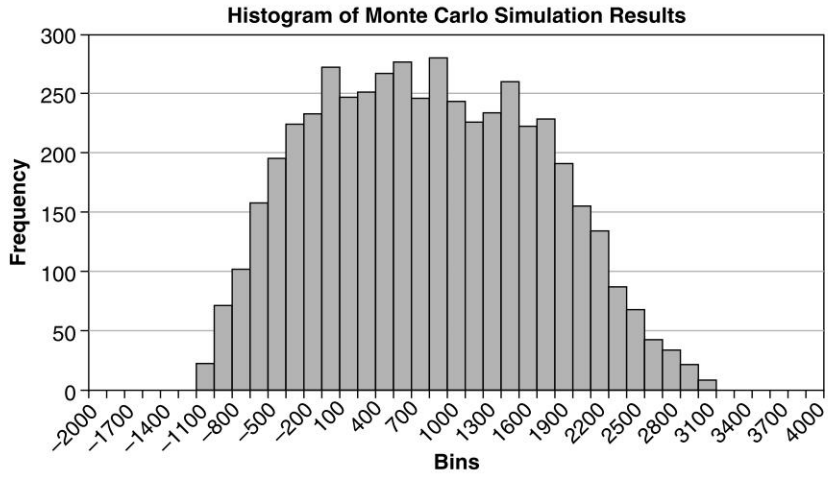
Sr No	Rand(L)	Rand(C)	Rand(R)	Rand(P)	Constant	Profit
1	1434	11	5	2445	40000	119532.5
2	1424	26	4	2462	40000	63211.52
3	1724	21	4	2562	40000	100471.52
4	1573	34	2	2397	40000	-18072.38
5	1358	23	2	2545	40000	-2111.8
6	1381	10	5	2426	40000	113705.3
7	1339	21	1	2423	40000	-35675.03
8	1756	34	4	2465	40000	73437.6
9	1534	24	2	2631	40000	3903.08
10	1713	23	3	2601	40000	54266.39
11	1733	32	3	2497	40000	34363.03
12	1469	36	2	2508	40000	-19198.96
13	1668	34	2	2487	40000	-13745.68
14	1255	27	1	2427	40000	-43426.15
15	1461	20	2	2620	40000	7336.4
16	1330	30	3	2649	40000	25795.1
17	1476	20	4	2430	40000	73947.2
18	1273	24	3	2471	40000	23815.49
19	1545	30	2	2625	40000	-5237.5
20	1624	22	3	2539	40000	47972.08
21	1561	19	4	2459	40000	83880.96
22	1525	39	2	2573	40000	-20998.5
23	1279	34	3	2625	40000	17235.25
24	1553	19	3	2377	40000	41237.43
25	1355	21	5	2421	40000	95567.75
26	1213	33	4	2459	40000	39281.68
27	1747	32	5	2496	40000	122121.6
28	1474	31	2	2487	40000	-12377.24
29	1272	12	4	2382	40000	65932.16
30	1746	13	2	2364	40000	19852.88
31	1764	29	4	2557	40000	89265.92
32	1492	35	3	2638	40000	25856.88

Figure 4.1 *Screen capture from the example sales forecast spreadsheet.*

In Part II of this Monte Carlo Simulation example, we completed the actual simulation. We ended up with a column of 5000 possible values (observations) for our single response variable, profit. The last step is to analyse the results. We create a histogram in Excel, a graphical method for visualising the results.

We can draw many conclusions from this histogram:

- Profit will be positive, most of the time.



- The uncertainty is quite large, varying between –1000 to 3400.
- The distribution does not look like a perfect normal distribution.
- There doesn't appear to be outliers, truncation, multiple modes, etc.
- We can estimate the probability of profit being below a given value or above a given value, or between a set of two values.

Some Other Applications:

(i) Evaluation of Investment

Many business firms invest large sums to expand capacity, reduce cost of production, etc. There is considerable risk associated with each investment plan, and this risk can be minimised, if one could visualise the impact of several factors by evaluating the alternative courses of action. If these interactive alternatives involve many parameters and large volume of data, it becomes difficult for the human mind to digest and analyse all the relevant information. This inability of human mind to process large amount of data is called as bounded rationality. Simulation offers a great deal of help by reducing the complexities in such cases and hence mitigating the issues of bounded rationality.

(ii) Evaluation of Net Present Value (NPV)

A telephone company, in USA, was considering the sale of its Calling Card Platform (CCP) that was facing declining usage due to the rising popularity of cellular phones. They asked an investment bank to calculate the likely value of the CCP. The bankers built a financial model that calculated the Net Present Value (NPV) of the CCP. Since several key variables in the NPV calculation could not be known with certainty, they assumed them to be uniformly distributed within a range of likely maximum and minimum values.

The range for key variables used in the NPV calculation was:

S.No.	*Key Variable*	*Minimum*	*Maximum*
1	Annual decline in call volume	2%	4%
2	Annual decline in price per minute	3%	5%
3	Gross margin of calling card business over time	45%	55%
4	Cost of capital for calling card business	5%	10%
5	Terminal value as multiple of final year cash flow	5	10

The result of performing a Monte Carlo Simulation with 5,000 iterations across 5 variables is shown below:

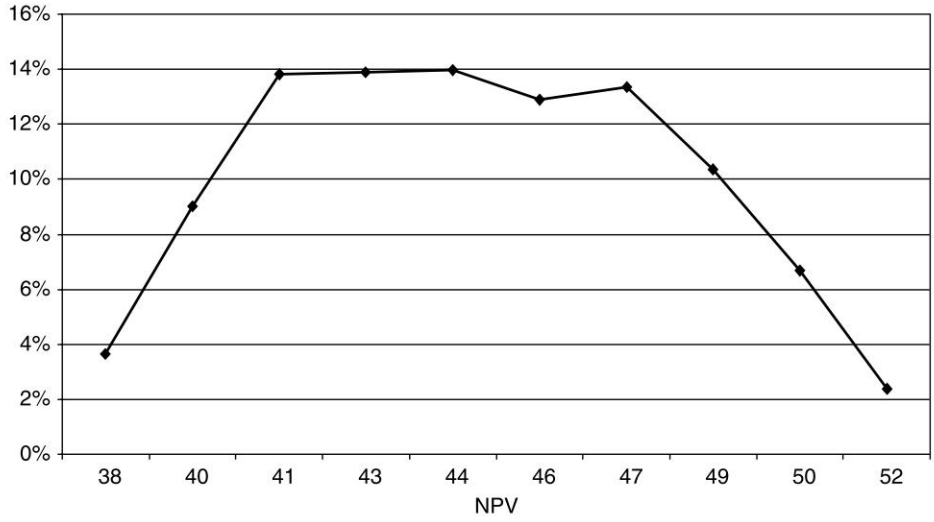


Figure 4.2 *Relative Probabilities for NPV (in $Million)*

Expected NPV Average:	$44.0
Standard Deviation:	$3.4
Maximum:	$51.8
Minimum:	$37.0

Thus, the Monte Carlo Simulation provided the management of the telephone company with more information for negotiating with prospective buyers than a single NPV value based on assumptions that may have been inaccurate.

(iii) Other Applications in Finance

Monte Carlo methods are used in finance to value and analyse financial instruments, portfolios and investments by simulating the various sources of uncertainty affecting their value, and then determining their average value over the range of resultant outcomes.

In finance, the Monte Carlo method is used to simulate the various sources of uncertainty that affect the value of the instrument, portfolio or investment in question, and then calculate a representative value given these possible values of the underlying inputs.

Some other applications are

- Portfolio evaluation
- Personal financial planning
- Corporate Finance
- Project finance

Monte Carlo Methods are also used by financial analysts who wish to construct "stochastic" or probabilistic financial models as opposed to the traditional static and deterministic models. Here, in

order to analyse the characteristics of a project's net present value (NPV), the cash flow components that are heavily impacted by uncertainty are modelled, mathematically reflecting their "random characteristics". Then, these results are combined in a histogram of NPV (i.e. the project's probability distribution), and the average NPV of the potential investment–as well as its volatility and other sensitivities is observed. This distribution allows, for example, for an estimate of the probability that the project has a net present value greater than zero (or any other value).

Some other areas where simulation is used are:

(iii) Assembly Line and Maintenance Scheduling
(iv) Bank Counters' Allocation and Scheduling
(v) City Bus Scheduling
(vi) Telephone Traffic Routing
(vii) Consumer Behaviour Forecasting
(viii) Brand Selection and Sales Promotion
(ix) Warehouse Location

4.10.4 Advantages of Simulation

The important advantages of the simulation techniques are as follows:

(i) It is useful in solving problems where all values of the variables are either not known, or partially known.
(ii) In situations where it is difficult to predict or identify bottlenecks, simulation is used to foresee these unknown difficulties.

4.10.5 Use of Monte Carlo Simulation in Companies

It is reported that several reputed companies use this method. Some of these are as follows:

(i) General Motors (GM), Procter and Gamble, and Eli Lilly (a pharmaceutical company in USA) use simulation to estimate both the average return and the risk associated with new products. At GM, this information is used to determine the products that come to the market.
(ii) GM uses simulation for activities such as forecasting net income for the corporation, predicting structural costs and purchasing costs, and determining its susceptibility to different kinds of risk (such as interest rate changes and exchange rate fluctuations).
(iii) Lilly uses simulation to determine the optimal plant capacity that should be built for each drug.
(iv) Wall Street firms use simulation to price complex financial derivatives and determine the Value at RISK (VAR) of their investment portfolios.
(v) Procter and Gamble uses simulation to model and optimally hedge foreign exchange risk.
(vi) Sears (a departmental store in USA) uses simulation to determine how many units of each product line should be ordered from suppliers, each year.

4.10.6 Use of Random Number Table for Simulation

Perhaps, the simplest example of simulation with the help of a random number table is to simulate the experiment of tossing a coin.

Table 4.4 Random Number Table

298	997	555	848
061	205	180	020
206	587	281	154
099	254	441	484
568	563	616	495
669	984	048	455
875	408	324	700
949	258	699	948
113	428	802	882
229	800	399	961

Instead of actually tossing a coin, and observing head (H) or tail (T), we can, for convenience, simulate the experiment of tossing a coin by selecting a one-digit number from any column of the table, reproduced here.

Thus, the ten numbers are

2, 0, 2, 0, 5, 6, 8, 9, 1 and 2

If a number is odd, it may be taken as H, and if it is even, it may be taken as T.

Thus, the simulated sequence of H and T is

T, T, T, T, H, T, T, H, H

This sequence may be taken as the simulated result of the tosses of a coin.

Another example of simulation is to simulate the experiment of throwing a dice, and noting the number on the dice. We can select, two-digit numbers from the first column of the above random number table. Thus, the numbers are

29, 06, 20, 09, 56, 66, 87, 94, 11 and 22

Dividing each of these numbers by 6, and matching the remainder 0 with number 1 on the dice, remainder 1 with number 2 on the dice, and so on. Finally, the remainder 5 is matched with number 6 on the dice. Thus, the sequence of numbers is

6, 1, 3, 4, 3, 1, 4, 5, 6 and 5

This sequence may be taken as the **simulated result of the experiment of throwing a dice.**

The above two examples illustrate **how a Random Number Table can help in simulating an experiment without actually conducting it.**

We can also generate random observations from any distribution or pattern which can be described mathematically.

SUMMARY

The research designs are:

- Exploratory

- Descriptive
- Explanatory/Causal and Association

The experimental designs are:

- One-factor Experiments
- Two-factor Experiments
- Two-factor Experiments with Interaction
- Latin Square Design
- Factorial Design
- Quasi-experimental Design
- Ex Post Facto Designs

In addition to conventional designs and sampling schemes, two other types of designs are highly useful in research environment to provide complete knowledge about conduct of research studies.

One type of studies relates to conduct of research of a particular phenomenon for a cross-section of entities like companies, institutions, at one point of time or over a period of time.

The other type of studies relate to an environment where it is advisable to simulate or generate data in a futuristic scenario, and use it for managerial decisions rather than limiting to available actual data—which might be irrelevant!

DISCUSSION QUESTIONS

1. Describe various types of research designs with suitable examples.
2. Describe relevance and historical development of experimental designs.
3. Describe a two-factor experiment with interaction in a business environment.
4. Write short notes on
   (i) Latin Square Design
   (ii) Factorial Design
   (iii) Quasi-experimental Designs
   (iv) Ex Post Facto Design
   (v) Action Research

EXERCISES

1. A research firm wants to conduct "A study of the effect of tips given by brokers to retail investors on stock investments".
   (a) Write two objectives of this study.
   (b) Identify major variables of the study.
   (c) Suggest appropriate design for the study giving justification.
2. A company wants to study the effect of managerial control on the company's performance.
   (a) Identify and classify the variables in the study.
   (b) Identify major variables of the study.
   (c) Suggest suitable design for the study.

Measurement Scales

5

When you can measure what you are speaking about and express in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind.

—Lord Kelvin

Contents

10. Non-Comparative Scaling Techniques
    (a) Continuous Rating Scale
    (b) Itemised Rating Scale
        (i) Likert Scale
        (ii) Semantic Differential Scale
        (iii) Staple Scales
    (c) Simple/Multiple Category Scales
        (i) Multiple choice, Single-response scale
        (ii) Multiple choice, Multiple-response scale
        (iii) Verbal Frequency Scale
11. Guidelines for Deciding Scales
    (a) Data Properties
    (b) Number of Dimensions
    (c) Number of Scale Categories
    (d) Balanced versus Unbalanced
    (e) Odd or Even Number of Categories
    (f) Forced versus Unforced Scales

LEARNING OBJECTIVES

- Providing a comprehensive understanding of the four types of measurement scales
- Creating an awareness of the measurement errors at the four hierarchical levels of a study
- Enabling to distinguish between comparative and non-comparative measurement scales, and their different sub types
- Equipping with basic tool-kit of scales for research

Relevance

Mr. Jitendar and Mr. Jay, students of a reputed B-School, joined a rapidly growing mobile service provider company as summer interns. After completing the initial formalities, they were asked to meet their corporate guide, Mr. Dixit. He accorded a warm welcome to them, and explained the project assigned to them. It related to conducting a research study relating to the penetration of the GPRS system in Indian Mobile Phone User Segment.

Mr. Dixit assured them that adequate resources would be provided to them, and the concerned staff would provide them full co-operation and support. Mr. Dixit, however, smilingly told them that the management had great expectations, and hoped they would enhance reputation of their institute.

Jitendar and Jay started the project with great excitement. However, the excitement waned off when they were asked to do two more additional projects that were accorded a higher priority by the company.

After completion of the other two projects, when they reverted to the first project, they realised that there was hardly any time available for the project. This resulted in hurryingly designing of the requisite questionnaire. The data was collected from 350 participants. After the data was coded and they started with the analysis, they realised the blunder they had

committed while using scales in the questionnaire. Most of the questions asked were purely qualitative and the scales used were inappropriate to carry out any quantitative analysis, they were planning to carry out.

At this stage, they realised that they should have given more time and thought while defining the measurements and scales in the questionnaire appropriate for the project. As a result, they could not fulfill the envisaged objectives.

After the presentation of the project report, Jitendar and Jay had mixed feelings. Though the presentation was not impressive, they could salvage the situation as they had established good reputation in the company, thanks to their two successful projects. But the lesson they learnt from the failed project was much more than what they learnt from the successful ones.

5.1 INTRODUCTION

The characteristics of individuals and business entities vary from individual to individual and from entity to entity.

In the case of human beings, there are certain physical and/or quantitative characteristics like height, weight, complexion, etc., and there are certain abstract or qualitative characteristics like intelligence, integrity, creativity, attitude, etc.

Like human beings, a business organisation has also physical characteristics like: employees, sales, offices, etc. Being physical in nature, these are easily measurable. However, there are certain abstract characteristics (known as constructs) like reputation, image of the entity, motivation, work culture, commitment, customer's perception and trust.

All these perceptions and feelings of customers and employees are extremely important because they help the company to stay afloat and grow. Therefore, it is essential for the companies to consider the above constructs relating to employees and customers.

As mentioned in Chapter 2, constructs are abstract and concepts are components of construct that are concrete, and, therefore, measurable.

Some concepts that are normally relevant for understanding the psychology of employees and customers are:

- Achievement
- Aptitude
- Attitude
- Intelligence
- Personality

It may be appreciated that even abstract characteristics or concepts such as mentioned earlier have to be measured for their meaningful assessment. It is reflected in the quotation of Lord Kelvin, given at the beginning of this chapter.

Accordingly, behavioural scientists have evolved the following instruments to measure the above concepts:

- Achievement Tests
- Aptitude Tests
- Attitude Scales
- Intelligence Coefficient
- Personality Profiling

Measurement can be defined as a process of associating numbers or symbols to observations obtained in a research study. This can be done through some scales like hours, metres, grams, etc. For example, if one has to measure motivation, it is quite difficult to measure or quantify the same. This can be done by assigning some number to motivation and forming a scale.

In this chapter, we will discuss the different types of measurement scaling techniques with their advantages and limitations.

This chapter discusses the various types of measurement scales that have been evolved which led to development of the above instruments. The advantages and limitations of the scales have also been elaborated.

The variables associated with a study are classified in two basic categories viz.

- Quantitative/Numeric/Metric
- Qualitative/Categorical/Non-Metric

This forms the basis for the classification of the measurement scales. These have been elaborated in the next section.

5.2 QUANTITATIVE AND QUALITATIVE MEASURES

The measures/variables can be divided into two basic types, namely

- Quantitative/Numeric/Metric
- Qualitative/Categorical/Non-Metric

Incidentally, only quantitative variables can be measured and qualitative variables can only be counted. This distinction of the two variables is quite evident while analysing the data, as most analysis can be done using quantitative data. For example, if the data is collected in quantitative form, then in Descriptive statistics one can find mean, standard deviation, etc. as discussed in Chapter 9. The qualitative variable has limitations for been subjected to such analysis. One can only count such variable; mean, standard deviation, etc. cannot be computed. It is, therefore, very important to make detailed plan of what analysis the research project requires and set up appropriate hypothesis accordingly **before the data collection stage**. The researcher must have a clear understanding of the type of variables to be used, and appropriate analysis to be performed on them.

The examples of categorical random variables are responses such as "yes" or "no" to a question; gender of a newborn child i.e. "male" or "female"; the result of students in an examination viz. "pass" or "fail"; type of medals in a sports event viz. gold, silver and bronze, etc.

The examples of numerical random variables are height, weight, income of individuals, marks obtained by students, number of children in families, number of runs scored by a cricketer, etc.

The variables can also be divided into two different categories based on different criteria viz.

- Continuous
- Discrete

The continuous variable arises in situations when some sort of measurement is involved e.g. height, weight, life of an electric bulb, waiting time for customers at a bank's counter, etc. In such cases, the variable assumes all possible values in its range.

The variable is said to be discrete if it assumes only some specified values in a given range. The discrete variable arises in situations when counting is involved, e.g. number of children in a family, number of credit cards held by an individual, number of customers visiting a branch, number of defective items in boxes of 100 items, etc.

5.3 CLASSIFICATION OR TYPES OF MEASUREMENT SCALES

Stevens (1946) postulated that all measurement scales can be classified into the following four categories:

(i) **Nominal**
(ii) **Ordinal**
(iii) **Interval**
(iv) **Ratio**

5.3.1 Properties of Scales

Scales possess typically the following properties:

- **Distinctive classification**
- **Order**
- **Equal distance**
- **Fixed origin**

These are briefly described as follows:

Distinctive Classification

A measure that can be used to classify objects or their characteristics into distinctive classes/categories is said to have this property. This is a minimum requirement for any measure. For example, gender classifies the individuals into two distinctive groups, males and females. The individuals may also be classified on the basis of their occupation, like student, salaried, businessman, etc. Similarly, the qualification of an individual could be used to classify individuals into various categories such as undergraduate, postgraduate, professional, etc.

Order

A measure is said to have an order if the objects or their characteristics can be arranged in a meaningful order. For example, marks of a student can be arranged in an ascending or a descending order. As another example, a consumer may rank four telecom service providers on connectivity; the result will be the order of companies as 1, 2, 3 and 4.

It may be noted that all quantitative measures have implied order. Qualitative measures may also have order. The first example described earlier is a quantitative measure and the second is a qualitative measure.

Equal Distance

If, for a measure, the difference between any two consecutive categories (generally termed as values for numeric variables) of a measured attribute, are equal, then the measure is said to have equal distance. For example, the time difference between 2.00 pm to 3.00 pm is same as the difference between 3.00 pm and 4.00 pm i.e. 1 hour. Another example could be the temperature as a measure; the difference between 40^{o}C and 50^{o}C is same as between 60^{o}C and 70^{o}C.

All numeric measures satisfy this property.

Fixed Origin

A measurement scale for measuring a characteristic is said to have a fixed origin if there is a meaningful zero or 'absence' of the characteristic. Examples are: income of an individual, sales of a company, etc. These scales have a meaningful zero or 'absence' of the characteristic; zero income signifies no income or absence of income, and zero sales signifies no sales or absence of sales.

Before we describe these details, we would like to discuss various properties that a scale possesses.

5.3.2 Types of Scales

Depending on the presence or absence of the above properties, the four scales are classified as nominal, ordinal, interval and ratio. These are described below.

Nominal Scale

A qualitative scale without order is called **nominal scale**. This scale can only be categorised and do not satisfy other three properties described above. It is termed as 'nominal' as, though one may represent the categories using numbers, the numbers are just 'nominal' or namesake, they do not carry any value or order or meaning. For example, the colour of bikes is a nominal measure. The different possible answers to the question 'Which colour will you prefer for a bike?'—could be blue, black, red, etc. One may number these colours as 1, 2, 3 or 4 or 100, 200, 300, 400, in any sequence i.e. this scale neither has any specific order nor it has any value.

The nominal scale involves classification of measure objects into various categories such as 'Yes' or 'No', 'Pass' or 'Fail', type of population group viz. metropolitan, urban, semi-urban, vehicle used for going to office, bus, car, motor cycle, etc. Numeric value is assigned to these classified categories. Such numbers are used for identifying individuals.

The data collected through a nominal measure scale is called nominal data.

The data obtained through a nominal scale is of a type that can be classified into categories or groups, and given labels to describe them. Examples are: house number, telephone number, car number, roll number (of a student), model (of a TV), etc.

Sometimes, instead of numbers, codes are used for classification like STD codes for cities, bar codes for items in departmental stores, codes for various subjects in a university, codes for books in a library, blood group of individuals, etc.

Ordinal

Ordinal scale is a scale that does not measure values of the characteristic(s) but indicates only the order or rank like 1^{st}, 2^{nd}, 3^{rd}, etc.–like in a beauty competition. Even when objects or their characteristics are measured quantitatively, the scale converts them into ranks like students in a class.

As another definition, a qualitative scale with order is called an ordinal scale.

This scale possesses first two of the four properties of the scales, viz. the properties of distinctive classification as well as order.

Rank as a measure is always considered as ordinal. The difference between any two ranks is not necessarily equal. The difference between first and second rank does not connote the same differential. For example, if in a class of students, the highest mark is 95, next is 85 and the next is 84, converting marks to ranks will lead to 1, 2 and 3. Incidentally, it may be noted that the difference

in the performance of the 1st ranker and 2nd ranker is not the same as the 2nd ranker and 3rd ranker. Thus, one can only conclude that 1st ranker has performed better than 2nd ranker and 2nd ranker better than 3rd ranker.

The data obtained using ordinal scale is termed as ordinal data.

Ordinal data is essentially the same as nominal data, except that there is now an order within the groups into which the data is classified. However, as we are dealing with qualitative data, we are unable to say by how much they differ from each other.

Some examples are:

- Ratings of hotels, restaurants and movies. We can say a 5 star hotel is better than a 4 star hotel, but we cannot say that a 4 star hotel is twice as good as a 2 star hotel
- Class of travel in a train or an aeroplane
- Grades of students in a class

Interval

A measurement scale whose successive values represent equal value or amount of the characteristic that is being measured, and whose base value is not fixed, is called an interval scale.

This is a quantitative scale of measure without a fixed or true zero.

The data obtained from an interval scale is termed as interval data.

Interval data is quantitative data that can be measured on a numerical scale. However, the zero point does not mean the absence of the characteristic being measured. Some examples are: temperature, time, longitude, latitude, etc.

Ratio

Ratio scales are quantitative measures with **fixed** or true zero. Ratio scale has all the four properties of scales that are described in the Section 5.3.1.

The data obtained from ratio scales are referred to as ratio data.

Ratio is also a quantitative data that can be measured on a numerical scale but, here, the zero point is fixed and implies the absence of what is being measured. In fact, if a scale has all the features of an interval scale, and there is a true zero point, then it is called a ratio scale. For example, a weighing scale is a ratio scale. Some other examples are: height, life, price, length, sales, revenue, etc. In all these cases, zero implies absence of that characteristic.

The table given in the following depicts the four properties followed by different types of scales:

Type of Scale	*Category/ Distinctive Classification*	*Order*	*Distance*	*Origin*
Nominal	Yes	No	Not fixed	Not fixed
Ordinal	Yes	Yes	Not fixed	Not fixed
Interval	Yes	Yes	Fixed	Not fixed
Ratio	Yes	Yes	Fixed	Fixed

It may be noted again that the ratio scale follows all the four properties of a scale.

Some more examples of the measurement scales/data are as follows:

Nominal	*Ordinal*	*Interval*	*Ratio*
• Gender • Name • Roll number • Serial number • Bank account number • Phone number • PAN number • Codes like bar codes • Number of jersey of a player	• Ranks • Feedback (very effective, somewhat effective, neutral, somewhat non-effective, not effective) • Rating of hotels (given in terms of stars) • Rating of movies (given in terms of stars) • Class of travel	• Temperature • Calendar dates • Rating scale • Longitudes • Latitudes	• Weight • Height • Salary • Number of times one visits a bank • Frequency of buying a product

5.4 STATISTICAL ANALYSIS BASED ON SCALES

Depending on the property of the scales, there is a limitation on the descriptive statistics one can perform on the scales. This is discussed in more detail in Chapter 9.

The following table summarises the Descriptive Statistics that can be used on the types of scales:

Type of Scale	Mode	Median	Arithmetic Mean	Geometric Mean	Standard Deviation, Coefficient of Variation
Nominal	√	×	×	×	×
Ordinal	√	√	×	×	×
Interval	√	√	√	×	√
Ratio	√	√	√	√	√

5.5 CHARACTERISTICS OR GOODNESS OF INSTRUMENTS/ MEASUREMENT SCALES

A measurement scale has to have certain desirable characteristics or criteria to judge its **'goodness'** so that one could have faith or trust in the scale that it will measure what it is intended to measure, and measure it consistently and accurately in an economical manner. These are as follows:

(i) Accuracy and Precision
(ii) Reliability
(iii) Validity
(iv) Practicality

These characteristics are described in the following:

(i) Accuracy and Precision

The accuracy of a measurement scale implies that it should lead to the true value. In popular language, it means that if a person's weight is 60 kg., the measurement scale should indicate 60 kg. on the scale (in a dietitian's clinic, the weighing scale might show 62!)

The precision, however, means the power to discriminate/distinguish, and indicate the extent of accuracy that can be achieved with the measurement scale. A simple example is the one-foot scale used by students. It has two scales viz. 'inch' and 'centimetres' on its edges. Both scales measure length accurately; but the 'cms.' scale is more precise because it can measure accurately up to one

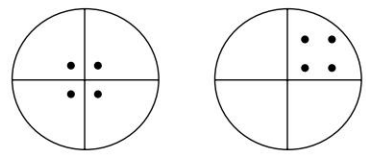
twenty-fifth of an inch (1 inch = 25 mm) while the 'inch' scale can measure precisely only up to one-eighth of an inch. Similarly, the Fahrenheit scale of measuring temperature is more precise than the Celsius scale. Incidentally, those using eyelenses wish that the precision of lenses could be more than one-quarter i.e. 0.25!

An analogy could be had with the examination conducted to measure the knowledge and understanding of the students as also to distinguish the students from one another. The marks scored out of, say 100, would provide better accuracy and precision than simply grading the students A^+, A, B^+, B and C. However, this grading system would provide more accuracy and precision than grading the students merely as A, B and C.

(ii) Reliability

Reliability indicates the confidence one could have in the measurement obtained with a scale. It tests how **consistently** a measuring instrument measures a given characteristic or concept i.e. if the same object/characteristic/attitude, etc. is measured again and again, it would lead to about the same conclusion.

However, it may be emphasised that reliability does not necessarily imply that the measuring instrument is also accurate. All it means is consistency in drawing conclusion. The following drawing diagram explains it clearly:



The first diagram exhibits reliability as well as accuracy. In the second diagram, however, the measuring instrument is consistent but not accurate.

(iii) Validity

The validity of a measuring instrument indicates the extent to which an instrument/scale tests or measures what it is intended to measure. For example, if we intend to measure intelligence, the instrument, say question paper, ought to be such that it results in measuring true intelligence; if the paper tests only general knowledge, the instrument is not valid. As another example, if the reward system for measuring performance of salesmen is based only on the sales figures, irrespective of the territory, the reward system may not be valid, if the territories do impact sales.

Types of Validity

There are three types of validity. These are:

- **Content** Validity
- **Criterion** Validity
- **Construct** Validity

A brief description of these is given below:

(a) Content Validity

It indicates the extent to which it provides adequate coverage of the issues that are under study.

(b) Criterion Validity

These are of two types. One indicates the success of the measuring instrument used for predicting. The other, also called **concurrent** validity, is used to estimate the present status.

(c) Construct Validity

It is one of the most significant research in development of measurement theory and practice. It links psychometric notions and practices to theoretical notions. It attempts to explain the variation observed in tests conducted on several individuals. For example, if a test of intelligence is conducted on individuals, and if the test scores obtained by a measurement scale vary from individual to individual, one would like to know the factors or constructs behind this phenomenon. For illustration, when customers have differing assessment about the quality of a TV, it could be due to their perception/preference about the constructs/features like picture quality, sound quality and contrast ratio.

Incidentally, the construct validity is closely related to Factor Analysis, a subject that is discussed, in detail, in Chapter 14.

The classifications of validity into three types (just discussed) was prepared jointly by the following bodies:

- American Psychological Association
- American Educational Research Association
- National Council on Measurements (used in education).

Thus, the classification is rather better suited in the field of educational research. Accordingly, the detailed discussion on the three types of validity, in consonance with the objectives of this book, is beyond the scope of this book.

(iv) Practicality

From theoretical viewpoint, a measure ought to be reliable and valid. However, from practical viewpoint, the measure should be

- Economical
- Convenient
- Interpretable

The economic considerations lead to a compromise between the idealism and availability of budget for a study. Thus, the measuring instrument has to take cognisance of this aspect and designed accordingly.

The convenience implies the ease with which an instrument like questionnaire could be easily administered to the subjects/respondents/participants. This poses more challenge in the situations wherein the concepts and constructs are rather difficult to understand.

The interpretability of an instrument, like questionnaire, is the ease with which the researcher is able to interpret the responses from the subjects/respondents/participants.

5.6 ERRORS IN MEASUREMENTS

So far, we have discussed the properties of a scale as also the characteristics of a good scale. However, we may mention that however good a scale might be like time, age, income, etc., when it is used for measuring the objects or their characteristics, some errors do creep in. We may only minimise such errors by a suitable research design, it may not be possible to eliminate such errors. In this section, we have discussed the different sources of errors that could creep in while collecting data.

Incidentally, the accuracy of any study solely depends on the accuracy of the data. Research studies often follow the GIGO principle commonly used in information technology. The GIGO principle states 'Garbage In Garbage Out' i.e. if the data itself is inaccurate, then however appropriate or accurate analysis is performed on it, the analysis will still yield inappropriate results.

The errors could be classified as:

- **Systematic Error**
- **Random Error**

Systematic errors generally occur due to bias involved in a study at various levels viz.

- **Researcher**
- **Participants/Subjects/Respondents**
- **Implementer/Measurer/Interviewer**
- **Tool/Instrument**

and may follow some pattern. On the other hand, the random error does not have any pattern or source, and hence difficult to control.

Sources of Errors

The major sources of systematic errors at various levels are as follows:

- **Researcher**
  Quantitative research is said to be more dependent on the researcher's understanding of the subject as, unlike in qualitative research, there is not much scope of learning from the participants and making changes in the design in the quantitative research. Thus, data collection can have a researcher's bias. Following are some of the typical researcher's errors that can affect the research study:
  - — It is possible that a researcher requires a particular information for a research study but the information sough is different. For example, if in a research study on credit card perceptions it is required to collect data on the dependent and earning family members separately, the researcher may consider the consolidated data of all the family members. This will introduce error in the study.
  - — The error may also be introduced due to inappropriate selection of scales. For example, if a study requires purely numerical data, and the researcher uses ordinal data, it will lead to error.
  - — The error may also be introduced due to inappropriate definition of the population by the researcher. This is a common problem faced by researchers while defining the population. For example, the definitions like "Literate", "Affluent", "Urban", "Rural" could be difficult to define, as the definitions may vary from researcher to researcher.
- **Participants/Subjects/Respondents**
  Such errors arise due to academic, economic, social, political, cultural and regional backgrounds of the respondents, as these factors affect the responses on various issues. The sampling design should allow for minimising the impact of such situations. For example, if a study is conducted to understand the risk taking abilities of managers, the study may have the participant's bias, if the sampling scheme is not designed properly. This may be avoided either by considering an appropriate cross-section of managers with the above mentioned backgrounds or by simply restricting the study for a type of population and concluding only about the same.
- **Implementer/Measurer/Interviewer**
  The interviewer may, inadvertently introduce error by changing the order of the questions, changing the form of the questions, rephrasing the questions or translating in simpler form so as to be understood by respondents. Body language of the interviewer such as nodding, smiling, etc. while seeking answer for a particular question, may also influence the response, and, thus, cause error. For example, for the study conducted to understand customer perception for

a particular brand of products, the interviewer may influence respondent by nodding or smiling at his preferred brand, thus, introducing error. While analysing data, inappropriate coding, wrong calculations, etc. can also introduce errors.

- **Tool/Instrument**
  The flaws in the instrument itself may cause errors. The flaws could be in the form of inappropriate words/language used to ask questions, ambiguous meanings, incorrect order of questions, incorrectly designed questionnaire, not giving enough choices for respondents also called response choice omission, etc. For example, if in the questionnaire it is asked "What is your income?" without mentioning monthly/yearly individual/family gross/net income from salary or total income from all sources, then each respondent might respond according to his definition of income, thus, leading to error.

  Poor printing, not providing enough space for answers, etc. are also considered as instrument errors.

In addition to the above four levels, there is yet another source of error viz. '**Circumstantial**'.

There could be errors due to circumstances like presence of someone, while answering the questions, which could influence the answers. For example, if the participant of a survey is an employee, then the presence of his/her boss or any senior may influence the responses, thus, adding error. Another example of circumstantial error is the error that could crop in, if a participant is not taken into confidence or not assured anonymity as he/she might give guarded responses with caution, suppressing the real responses.

All these errors needs to be controlled or neutralised by using appropriate design, understanding the subject under consideration in depth, training the interviewer, simplifying the tool, etc.

5.7 TYPES OF SCALES OR SCALING TECHNIQUES

We have discussed classification of scales, their properties and the characteristics of good scales. In this section, we shall discuss formats of several scales which have been developed to enable a researcher in collecting appropriate data for conducting a study.

The scales are broadly divided into two categories viz.

- **Conventional Scales**
- **Unconventional Scales**

The conventional scales are used in the questionnaire format and are most common. The unconventional scales are used for unconventional collection of data through games, quizzes, etc. However, the unconventional scales are beyond the scope of this book, and we shall discuss only the conventional scales.

The conventional scales are of two types viz.

- **Comparative Scaling Techniques**
- **Non-comparative Scaling Techniques**

and, are described below.

Comparative Scaling Techniques

The comparative scales involve direct comparison of the different objects. For example, in a study of consumer preferences for different airlines, a consumer may be asked to **rank** a list of factors that he/she would consider while choosing a particular airline out of the indicated factors like price,

punctuality, food, flying returns programme, etc. The consumer has to assign rank 1 to the most preferred factor and the last rank to the least preferred factor. A sample question is indicated in the box as follows:

Please rank the following factors in order of importance to you, while choosing an airline. Assign Rank 1 to most important and Rank 5 to least important factor. Do not repeat the ranks.

Factor	*Rank*
Price	
Punctuality	
Food	
Flying Returns Programme	
Seat Comfort	

The response will be indicative of the relative importance attached to different factors. Such comparative scales generally use ordinal scale and are interpreted only in relative terms, and as such they generate non-metric/non-numerical data.

Non-Comparative Scaling Techniques

The non-comparative scales involve scaling of each object independently of other objects. For example, in a study of consumer preferences for different airlines, a consumer may be asked to **rate** a list of factors that he/she would consider while choosing a particular airline for the indicated factors like price, punctuality, food, flying returns programme, etc. The consumer has to give rating of 5 to the most preferred factor and rating or 1 to the least preferred factor. The response will be rating of all the factors independent of one another; for example, a respondent may give rating of 5 to price as well as flying returns programme.

A sample question is indicated in the box as follows:

Please rate the following factors according to the importance attached by you, while choosing an airline. Indicate by √ against the appropriate rating. Rate 1 to 2 least important and Rate 7* to 2 most important factor.

Factor	*1–Least Important*	*2*	*3*	*4*	*5*	*6*	*7*–Most Important*
Price							
Punctuality							
Food							
Flying Returns Programme							
Seat Comfort							

* rating could be on any scale. In this illustration, we have used a 7-point scale.

The data generated through non-comparative scaling technique is usually in interval scale. The advantage of non-comparative scales is that they can be continuous, metric/numeric variables. This allows wide avenues for analysis as compared to non-numeric variables. This advantage makes it the most preferred scaling technique.

5.8 COMPARATIVE SCALING TECHNIQUES

There are several comparative scaling techniques. However, we shall confine our discussion only to the following three most commonly used Comparative Scaling techniques:

- **Paired Comparison**
- **Rank Order**
- **Constant Sum**

A brief description of each of these techniques is given below:

Paired Comparison

In paired comparison scales, the respondent is asked to select one object from the list of two objects, on the basis of some criteria. This forces the respondent to compulsorily select one of the two. Such scales are used when the study requires to distinguish between the two specified objects. An example is given below:

In a study of consumer preferences about two brands of Glucose biscuits viz. Parle-G and Tiger Glucose, the following paired comparisons were solicited on three characteristics.
Select only one of the two brands.
Which Glucose biscuits do you prefer on the basis of 'TASTE'?

☐ Parle – G ☐ Tiger

Which Glucose biscuits do you prefer on the basis of 'PRICE'?

☐ Parle – G ☐ Tiger

Which Glucose biscuits do you prefer on the basis of 'PACKAGING'?

☐ Parle – G ☐ Tiger

It may be noted that the data generated through this scaling technique is ordinal in nature.

This scaling technique is useful when the researcher wants to compare two or more objects. It may be noted that in the above example we have compared two brands over three factors, hence, the number of comparisons are three. If the number of objects to be compared are more, this can lead to too many comparisons requiring more time on the part of the respondent.

In general, if there are n objects to be compared on k factors, the total comparisons will have to be

$$k \times \frac{n \times (n-1)}{2}$$

which can be high number for large values of n and k. Hence, such scales are generally used only when number of objects to be compared are smaller in number.

For example, in the above case, instead of two brands if one considers five brands to be ranked on three factors, the total number of paired rankings to be given by the respondent is

$$3 \times \frac{5 \times (5-1)}{2} = 30$$

In general, the following factors may be taken into account while using Paired Comparisons:

- The method is most effective when actual choices in the real situation are always between two objects
- The researcher must be willing to forgo measurement of the distance between items in each pair, as this scale is ordinal and not numeric
- The number of objects to be compared may be limited to avoid making the respondent's task too difficult

Rank Order Scaling

This is also one of the commonly used comparative scaling techniques. It is also termed as **Forced Ranking** Scale. Unlike paired comparisons, rank order scaling technique prompts respondents to rank a given list of objects. This, generally, is less time consuming and less tedious on part of the respondents. The lists up to 10 to 12 objects can be ranked easily as opposed to paired comparison of 10 to 12 objects. A list with more than 12 objects can be tedious for respondents, and one may not get accurate responses in such situations.

An example of rank order scale is given in the following:

Rank the following services in the order of importance attached by you, while selecting a new mobile service provider. The most preferred can be ranked 1, the next as 2 and so on. The least preferred will have the last rank. Do not repeat the ranks.

Feature	Rank
1. Connectivity	_____
2. Minimum Call Drops	_____
3. Value Added Services	_____
4. SMS	_____
5. Roaming	_____
6. Ring tone/Caller tune	_____
7. Alerts	_____
8. Downloads	_____
9. Internet	_____

A sample response of a respondent is as follows:

Feature	Rank
1. Connectivity	__1___
2. Minimum Call Drops	__3___
3. Value Added Services	__5___
4. SMS	__4___

5. Roaming	__2___
6. Ring tone/Caller tune	__8___
7. Alerts	__7___
8. Downloads	__9___
9. Internet	__6___

An advantage of this technique compared to paired comparisons is that it requires only $(n - 1)$ ranking decisions for the respondents as compared to $\frac{n \times (n-1)}{2}$ ranking decisions in paired comparisons. Another advantage is that it is easier for the respondents to understand the rank order than the paired comparisons.

This technique typically creates ordinal data.

The following factors may be taken into account while using Rank Order Scaling:

- The number of objects to be ranked may be less than 10, to avoid making the respondent's task too difficult.
- The major focus may be on the relative standing of the entities, not their absolute position.
- The researcher must be willing to forgo measurement of the distance between ranks.
- As with other scales yielding ordinal data, an analysis is confined to a limited set of statistical procedures that do not require equal intervals.

Constant Sum Scaling

When it is required to assess the relative importance attached by a respondent to the objects in a list, the constant sum scaling technique is used. In this technique, a respondent is asked to allocate certain points, out of a fixed sum of points, for each object according to the importance attached by him/her to that object. If the object is not important, the respondent can allocate zero point, and if an object is most important he/she may allocate maximum points out of the fixed points. Generally, the total fixed points are 100 for simplicity but it may be taken as some other value depending on the study.

An illustrative example is as follows:

Allocate the amount you would like to spend on your birthday on the following items, out of total amount of Rs 5000/- (please note that total amount allocated should be exactly 5000)

Item	Amount
1. Cosmetics	_____
2. Clothes	_____
3. Accessories	_____
4. Jewellery	_____
5. Dinner	_____
6. Movie	_____
Total	5000

A sample response of a respondent is as follows:

Allocate the amount you would like to spend on your birthday on the following items, out of total amount of Rs 5000/-

Item	Amount
1. Cosmetics	_____
2. Clothes	1000_
3. Accessories	1500_
4. Jewellery	_____
5. Dinner	2000_
6. Movie	_500_
Total	5000

The data generated from this scaling technique can sometimes be considered as numeric data. The amount assigned to each of the objects in the list is purely numeric but generalisation of the amount beyond the list of objects is not possible. Due to this limitation, it is appropriate to consider the data as ordinal data. The advantage of this method is that it can distinguish the respondent's preference between objects, in lesser time than the other comparative scaling methods.

The following factors may be taken into account while using Constant Sum Scaling:

- The respondent may not assign exact total amount, they may assign either lesser or more than the specified amount. In such cases, the data may have to be discarded.
- Only up to 10 objects may be used in the list. Allocation over large number of objects may create confusion for the respondent leading to respondent's error.
- This cannot be used as a response strategy with children or uneducated people.

5.10 NON-COMPARATIVE SCALING TECHNIQUES

Some of the commonly used Non-Comparative Scaling techniques are:

- Continuous Rating Scale
- Itemised Rating Scale
  - — Likert Scale
  - — Semantic Differential Scale
  - — Stapel Scales
- Simple/Multiple Category Scale
- Verbal Frequency Scale

These are described as follows:

Continuous Rating Scale

This is also termed as a **Graphic Rating** scale. In this type of scale, the respondents indicate their rating by marking at appropriate distance on a continuous line. The line is labelled at both ends usually by two opposite words. For simplicity and understanding of the respondent, the line may contain points like 1 to 100. Alternatively, the scale can be written on a card and shown to the respondent during the interview.

An illustrative example is as follows:

Rate the following criteria, while choosing a LCD TV, on the basis of importance attached by you. (mark × at an appropriate distance)

1. Price	Most	_____	Least
2. Picture Quality	Most	_____	Least
3. Sound Quality	Most	_____	Least
4. Service	Most	_____	Least

Alternatively, the scale can also be used as following:

Rate following aspects, while choosing a LCD TV, on the basis of importance attached by you. (mark × at appropriate distance)

1. Price	Most	_____									Least
	100	90	80	70	60	50	40	30	20	10	**0**
2. Picture Quality	Most	_____									Least
	100	90	80	70	60	50	40	30	20	10	**0**
3. Sound Quality	Most	_____									Least
	100	90	80	70	60	50	40	30	20	10	**0**
4. Service	Most	_____									Least
	100	90	80	70	60	50	40	30	20	10	**0**

Theoretically, an infinite number of ratings are possible if the respondents are qualified enough to understand and accurately differentiate the objects. The accurate score can be measured by measuring the length of the mark from the either side.

The data generated from this scale can be treated as numeric and interval data.

The disadvantage of this scale is that it is more time consuming and difficult for editing, coding and analysis compared to the other rating scales. Graphic scales are generally used with children since they have limited vocabulary that prevents the use of scales dominated with words.

The following factors may be taken into while using continues rating scale:

- This method is most applicable where evaluative responses are to be arrayed on a single dimension.
- Scale extremes (both ends of the scale) may be labelled extremely to define the dimension, and the labels used must be bipolar opposites.
- In the vast majority of cases, the intermediate scale value should not be labelled with words, and only numbers spaced at equal intervals may be used.

Itemised Rating Scale

In an itemised scale, respondents are provided with a scale having numbers and/or brief descriptions associated with each category. The categories are usually ordered in terms of scale position.

The respondents are asked to select one of the categories, that best describes the product, brand, company or any other attribute being studied.

The commonly used itemised rating scales are:

- **Likert Scale**
- **Semantic Differential Scale**
- **Stapel Scales**

Likert Scale

The **Likert Scale** is the **most frequently** used variation of the summated rating scale commonly used in the studies relating to attitudes and perceptions.

Summated Rating Scales comprise statements that express either a favourable or an unfavourable attitude toward the object of interest on a 5 point, 7 point or on any other numerical value. The respondents are given a list of statements and asked to agree or disagree with each statement by marking against the numerical value that best fits their response. The scores may be summed-up to measure the respondent's overall attitude. It is not necessary to sum up the scores this scale can be used in isolation without summing up. The summing up may be misleading especially if there are statements designed to avoid leanings towards either side. The summed-up score in such cases does not interpret the actual attitude towards the objects.

The following illustration relates to a retail store:

Please rate the statements given below. 1 – Disagree …. 5 – Agree

Statement	*Disagree*				*Agree*
The ambience at this store is good	1	2	3	4	5
This store has clean, attractive and convenient public areas (restrooms, trial rooms)	1	2	3	4	5
This store has merchandise available when the customers want it.	1	2	3	4	5
Employees in this store have the knowledge to answer customers' questions.	1	2	3	4	5
The behaviour of employees in this store instills confidence in customers.	1	2	3	4	5
Customers feel safe in their transactions with this store.	1	2	3	4	5
Employees in this store give prompt service to customers.	1	2	3	4	5
Employees in this store are too busy to respond to customer's requests.	1	2	3	4	5
This store gives customers individual attention.	1	2	3	4	5
This store willingly handles returns and exchanges.	1	2	3	4	5
Employees of this store are able to handle customer complaints directly and immediately.	1	2	3	4	5
This store offers high quality merchandise.	1	2	3	4	5
This store accepts most major credit cards.	1	2	3	4	5

Likert Scale has several advantages that make it more popular. It is relatively easy and quick to compute. Further, it is more reliable and provides more data for a given amount of respondent's time, as compared to other scales.

The data gathered is interval data.

Incidentally, it may be mentioned that originally creation of a Likert Scale involved a procedure known as **item analysis**.

The following factors may be taken into account while using the Likert Scale:

- The Likert Scale may be used for several items, rather than just one or two, to achieve economy of scale.
- The items may be sufficient to capture a broad range of responses.
- The researcher may ensure to avoid a situation wherein most of the respondents do not tend to follow the middle path by indicating the response as '3' (on a 5-point scale, equivalent to 'neutral' opinion) in the above illustration.
- If a summated score is to be used, about half the statements may be inclined towards the positive side of the issue and half towards the negative side, to avoid respondent's inclination towards one side, irrespective of his/her true opinion. This avoids inappropriate interpretation of the summated ranks.

Semantic Differential Scale

This scale provides a measure to the psychological meaning of an attitude or an object, using bipolar adjectives. The respondents mark in the blank spaces provided between the two objects, indicating how they would best rate the object. Commonly, this is rated on 7-point scale. The Semantic Differential Scale is based on the proposition that an object can have several implied or suggestive meanings to an expressed opinion.

An illustrative example is as follows:

Rate the ATM you have just used in respect of the indicated parameters. Mark × at appropriate location that best suits your answer.

The ATM was _____ for operations

Easy :___:___:___:___:___:___:___: Difficult

The processing time was

Slow :___:___:___:___:___:___:___: Fast

The security person was

Cordial :___:___:___:___:___:___:___: Indifferent

The advantage of Semantic Differential Scale is that it is versatile and gives multidimensional advantage. It is widely used to compare image of brands, products, services and companies.

The data generated from this scale can be considered as numeric in some cases, and can be summed to arrive total scores. **If the objects reflect about a single store/product etc., the data is considered as ordinal**.

The following factors may be taken into account while using semantic differential scale:

- Adjectives must define a single dimension, and each pair must be bipolar opposites labelling the extremes.
- Precisely what the respondent is to rate must be clearly stated in the introductory instructions.

Stapel Scales

These scales are named after John Stapel who developed these scales. It is a unipolar rating scale with 10 categories numbered from –5 to +5 without a neutral point or zero. The respondents are asked to rate how each term describes the object. Positive rating indicates that the respondent describes the object accurately and negative rating indicates that the respondent inaccurately describes the object. Fewer response categories may also be used in certain cases. This is usually presented in vertical form as against other scales which are generally presented in horizontal form.

This scale is an alternative to semantic differentiation specially when it is difficult to find bipolar adjective that matches the question.

Example:

Rate the outlet on the following factors. +5 indicates that the factor is most accurate for you and – 5 indicates that the factor is most inaccurate for you.

+5	+5	+5
+4	+4	+4
+3	+3	+3
+2	+2	+2
+1	+1	+1
Good Ambience	Quality Products	Excellent Service
- 1	- 1	- 1
- 2	- 2	- 2
- 3	- 3	- 3
- 4	- 4	- 4
- 5	- 5	- 5

The data generated in staple scale is interval data.

The other advantage of this scale is that it can be used to collect data through telephonic interview.

The following factors may be taken into consideration while using Stapel Scale:

- This method is most applicable where evaluative responses are to be rated on a single dimension.
- The scale is most economical where several items are all to be rated on the same dimension.
- The method assumes discerning respondent. In the absence of such respondents, the technique may create errors in data.

Single/Multiple Category Scales

This scale is also termed as a dichotomous scale. It offers two mutually exclusive response choices, typically a 'Yes' or 'No' type of response. This response strategy is especially useful for demographic questions or where a dichotomous response is solicited.

Such scales are of two types viz.

- **Multiple choice, Single-response scale**
- **Multiple choice, Multiple-response scale**

In the first type, there are multiple options for the respondent, but only one answer can be chosen. The scale used is called the **multiple choice, single-response scale**.

In the second type, as a variation researcher may use **multiple choice, multiple-response scale** also termed as **Check List** wherein the respondent is given a list of multiple choices and can choose more than one choices from the list.

Example:

Single Category Scale:

1. Do you own a car?
   ○ Yes ○ No
2. Do you own a house?
   ○ Yes ○ No
3. Do you own a laptop/computer?
   ○ Yes ○ No
4. Do you own a club membership?
   ○ Yes ○ No

Example:

Multiple Categories – Single Response Scale:

1. Please indicate your Educational Qualification
   ○ Under Graduate ○ Graduate ○ Post Graduate
2. Please indicate your Current Occupation
   ○ Student ○ Salaried ○ Self-Employed
   ○ Professional ○ Retired ○ Home Maker

Simple attitude scales are easy to develop, inexpensive, can be highly specific, and provide useful information, if developed skillfully.

The following factors may be considered while using Single/Multiple category scale:

- The category names may define a set of discrete alternatives, with clear distinction in the minds of interviewers and/or respondents. The named categories should be mutually exclusive, so that a response does not fit into more than one of the categories.
- It may be ensued that the labeled alternatives capture majority (about 90%) of answers that are likely to be given by the respondents. As an abundant caution, "Others" category may be listed at the end to include any answers that do not fit into the named categories.

Example:

Multiple Categories – Multiple Response Scale (checklist)

Do you Own? (Please Mark × in the box for each object you own)

1. Car ☐
2. House ☐
3. Laptop/computer ☐
4. Club membership ☐

The checklist scale allows the respondent to select one or several alternatives. It is simple to understand and saves considerable time.

The following factors may be considered while Multiple Category Multiple Response or Checklist scale:

- The instructions and response task are quick and simple, and many options can be included.
- The scale yields data only in the form of discrete, nominal, dichotomous data.

The data gathered in single category scale is nominal.

Multiple category scale (single response) is either nominal or ordinal.

The checklist data formed is nominal, and each object in the list is coded as separate variable with 'Yes' and 'No' type ('Yes' means the object was ticked by the respondent. Any 'No' means the object was not ticked)

Verbal Frequency Scale

The verbal frequency scale is used when it is difficult for the respondent to answer in exact numbers.

Example:

How often do you watch Movies?
1. Always 2. Often 3. Sometimes 4. Rarely 5. Never

The following factors may be considered while using the verbal frequency scale:

- The scale is most appropriate when respondents are unable or unwilling to compute exact frequencies.
- This scale is used when only an approximation of frequency is desired.

This scale generates ordinal data.

5.11 GUIDELINES FOR DECIDING SCALES

A researcher must take into account the following factors before choosing any scales:

- Data properties
- Number of dimensions
- Number of scale categories
- Balanced versus unbalanced
- Odd or Even number of categories
- Forced versus Unforced scales

A brief description of each of the above factors is given as follows:

Data Properties

The properties of the data generated by scales is one of the important factors in deciding the scales. Each data type has certain property and can be used for only selected type of analysis. For example, nominal and ordinal data are not amenable to numerical analysis vide Section 5.4.

When the research design is planned, along with each variable, corresponding analysis is also decided. This aspect should be taken into consideration while deciding the scales.

Number of Dimensions

Measurement scales can be unidimensional or multidimensional. In unidimension scale, only one attribute of the participant or object is measured. For example, ambience of a retail store can be measured by a single measure like layout, or it may be measured as a combination of multiple measures in a single measure called 'store ambience'.

A **multidimensional scale** presumes that an object might be better described with several dimensions than a single dimension. In the above example, the store's ambiance can be defined on different dimensions like store design, store décor, and friendly environment, lighting, layout, lobby area, etc.

Number of Scale Categories

This is an important decision while choosing scales. It has trade-off between precision and simplicity. More the number of categories, better will be the precision. However, from a respondent's point of view, it may be more complicated and time consuming. Generally, the preferred number of categories is up to 10.

It may be pointed out that the number of scale points needed, to produce accuracy when using single-dimension versus multiple-dimension scales, is larger.

Balanced versus Unbalanced

A **balanced rating scale** has an equal number of categories above and below the midpoint.

It leads to odd number of responses if it has neutral option. Normally, rating scales should be balanced, with an equal number of favourable and unfavourable response choices. The balanced scale may or may not have the midpoint or neutral option.

An example for balanced scale is:
Strongly Agree – Agree – Neutral – Disagree – Strongly Disagree

An **unbalanced rating scale** has an unequal number of favourable and unfavourable response choices.

An example for unbalanced scale is:
Strongly Agree – Partially Agree – Agree – Neutral – Disagree – Strongly Disagree

The scale does not allow participants who are unfavourable to express the intensity of their attitude as much as for the respondents with the favourable viewpoint. An unbalanced rating scale is justified in studies where researchers know, in advance, that nearly all participants' responses will lean in one direction or the other.

When researchers know that one side of the scale is not likely to be used, they try to achieve precision on the side that will receive the participant's attention.

Odd or Even Number of Categories

In case of odd number of categories, the middle point is generally 'Neutral' option. The presence of neutral can have significant influence on the responses. For example, if we remove neutral category, the respondent will have to be inclined towards positive or negative side. This would force the respondent to give some opinion.

Forced versus Unforced Scales

An **unforced-choice rating scale** allows participants to express no opinion when they are unable to make a choice among the alternatives offered. A **forced-choice scale** requires that participants select one of the offered alternatives. Researchers often exclude "no opinion," "undecided," "don't know," "uncertain" or "neutral" when they know that most participants have a firm opinion on the topic. When many participants are undecided, and the scale does not allow them to express their uncertainty, the forced-choice scale may introduce bias in the results.
An example of Unforced Scale:

Indicate your level of agreement or disagreement with the following statements by placing a tick mark in the relevant grid (**5** = Strongly agree, **4** = Agree, **3** = Neutral, **2** = Disagree, **1** = Strongly disagree).

		Strongly Agree	*Agree*	*Neutral*	*Disagree*	*Strongly Disagree*
Price is of core importance	Undecided	☐	☐	☐	☐	☐
Opinions of family and friends are not important	○	☐	☐	☐	☐	☐
Advertisements accurately depict product feature	○	☐	☐	☐	☐	☐
Consumer are developing mental blocks to advertisements	○	☐	☐	☐	☐	☐
I try most of the new products in the market when they are launched.	○	☐	☐	☐	☐	☐

An example of Forced Scale:

Indicate your level of agreement or disagreement with the following statements by placing a tick mark in the relevant grid (**5** = Strongly agree, **4** = Agree, **3** = Neutral, **2** = Disagree, **1** = Strongly disagree).

	Strongly Agree	*Agree*	*Neutral*	*Disagree*	*Strongly Disagree*
Price is of core importance	Undecided	☐	☐	☐	☐
Opinions of family and friends are not important	☐	☐	☐	☐	☐

(*Contd*)

(*Contd*)

Advertisements accurately depict product feature	☐	☐	☐	☐	☐
Consumer are developing mental blocks to advertisements	☐	☐	☐	☐	☐
I try most of the new products in the market when they are launched.	☐	☐	☐	☐	☐

SUMMARY

There are two types of measures viz. qualitative and quantitative. Further, the four properties of scales are classification, order, equal distance and fixed origin.

The characteristics of a good measurement scale are accuracy, precision, reliability, validity and practicality.

Various types of measurement errors could creep in while conducting a study.

The four categories of measurement scales/data are nominal, ordinal, interval and ratio.

The measurement scales are of two types viz. conventional and unconventional. The conventional scales are further classified in two categories viz. comparative scaling techniques and non-comparative scaling techniques. Almost all important conventional scales are generally used in research studies are illustrated with suitable examples.

DISCUSSION QUESTIONS

1. Describe various aspects of quantitative and qualitative measures associated with a research study.
2. Discuss the four categories of measurement scales. Explain each of these scales with each of the three hypothetical situations.
3. Describe the characteristics of various measurement scales with suitable examples.
4. Describe various sources of errors in measurement as also ways to avoid or minimise such errors.
5. Discuss with examples the three comparative scaling techniques.
6. Describe the relevance and applications of Likert scale with an application.
7. Discuss the guidelines for deciding the use of scales in research studies.
8. Write short notes on:
   (i) Statistical analysis based on various scales
   (ii) Continuous Rating Scales
   (iii) Semantic differential Scale
   (iv) Stapel Scale

Primary and Secondary Data and Their Sources

6

Contents

LEARNING OBJECTIVES

- To explain primary and secondary types of data with respective advantages and limitations
- To acquaint with the various sources of primary and secondary data and the various methods to collect such data
- To guide web-based searches

Relevance

Mr. Anil, a senior consultant at ABC Consultant Ltd., had gone to meet one of its clients Mr. Arjun, the owner of a reputed luxury hotel chain, at his Nariman Point Building. Mr. Arjun was working on his new ambitious project of setting up a hotel with a budget of about Rs. 350 crore. He was looking for a firm that had a previous experience of research in the hotel business. ABC Consultant, having already worked for three reputed hotel chains successfully, was the obvious choice.

The meeting was arranged to discuss the detailed research design prepared by the firm and also to discuss the different sources for collecting the data required for the research. The firm,

in its research proposal had mentioned about the requirement of considering qualitative as well as quantitative methods of data collection. The qualitative method included observations, semi-structured or unstructured interviews; and the quantitative method included questionnaires and structured interviews. Since the study also required some economic parameters to assess the feasibility of the project, it was also felt necessary to consider some secondary data for the analysis.

After fruitful discussions in the meeting, the different strategies of data collection were finalised with the tentative dates of execution. Mr. Anil remarked in the end, "Most researchers find these matters trivial and do not invest enough time and thinking for such decisions, but this casual attitude towards such decisions may prove to be fatal at the later part of the research study."

Mr. Arjun agreed with the statement and nodded with appreciation.

6.1 INTRODUCTION

Data is the raw material for almost all research studies. The type of data and methodology of its collection vary according to the requirements of the study. It also depends upon the **'unit of study' i.e. an object, an individual, an entity,** etc. and the type of data that may be required for the study i.e. quantitative or qualitative, primary or secondary, etc. Incidentally, the concepts of primary and secondary data are explained in the next section. Further, the decisions on type and methodology of collection of data may also depend on the type of planned research design. This chapter deals with various methods of data collection like observations, focus groups, interview, survey, etc. and the various types of data sources like primary and secondary data sources.

6.2 PRIMARY AND SECONDARY DATA

Primary data is collected directly by the researcher for some specific purpose or study. It may be collected by methods such as personal investigation and/or questionnaires. A questionnaire is either filled up by an investigator based on the information given by the respondent or by the respondent himself/herself.

Secondary data is the data disseminated through some media like reports (external or internal), newspapers, hand books, magazines, websites, etc. Nowadays, it is also being disseminated by some agencies.

However, the source, which releases the secondary data, would have derived it either from the primary data collected by itself or from the secondary data released or published by some other source. This chain can go even further. However, the ultimate source of any secondary data has to be primary data. For example, primary data about prices of commodities is collected in various centres/cities in India. Government of India compiles this data and derives consumer price index and wholesale price index. The Government releases these indicies, which in turn get published in newspapers and magazines. The Reserve Bank of India also publishes these price indices in its publications. For the Reserve Bank, it is secondary data and so it is for anyone who uses the Reserve Bank's publications.

The data relating to banking in India is collected by the Reserve Bank of India, and so it is primary data for the Reserve Bank. However, the same data when published in the Reserve Bank publications becomes secondary for the person or organisation that uses the data.

The most common examples of secondary data are the data collected from published sources like newspapers, magazines/journals, books, reports, institutional publications like 'Handbook of Statistics on the Indian Economy' published by Reserve Bank of India, Economic Survey and Economic Census of Data, published by Government of India, etc. The data provided by the organisations on their websites could be primary data collected by them but for the visitor to the website, the data is secondary.

In subsequent sections, we shall discuss the advantages and limitations, and the different sources of primary and secondary data.

6.2.1 Advantages and Limitations of Primary and Secondary Data

Both the primary and secondary data have their own advantages and limitations. Research based on primary data has the following strengths:

Advantages of Primary Data

- The data is collected for a specific purpose/objective; hence, is most relevant for the research study.
- There is greater degree of control for the researcher. The control could be over the budget for the study, the size of the sample, the method of sampling, etc. The accuracy/error can be controlled by the researcher by increasing or decreasing the resources used for data collection. The accuracy of the data is generally determined by the nature and depth of the study. For example, a small study undertaken to understand the consumer preferences in a departmental store will require different accuracy level than the feasibility study of setting up a hotel. The resource allocation for the former may be far lesser than the later. The population can be selected by the researcher suiting the objective. The method of collecting data like observing, interview, focus groups, etc. can also be selected by the researcher.
- Such data can give the researcher a holistic and realistic view as also better insight about the population under study. Thus, the data collected is exclusive to the study.
- There is a possibility to control bias in the information collected through primary study.

Limitations of Primary Data

- The process of primary data collection is time-consuming and expensive. The entire process of data collection has to go through different stages like questionnaire design, pilot study, data collection, coding, editing/checking etc. This is time consuming, and makes it rather difficult to conduct primary study if it has to be conducted within a small time frame.
- The collection of primary data requires expertise beyond plain analytical skills and cannot be done by unskilled persons. This increases the cost of the study. In many cases, the primary data is more expensive than the secondary data.
- Primary data collection may not be always feasible. In case, the data required is scattered over different companies, geographical areas, countries, etc., it may not be feasible to use this method. If data like GDP, inflation, business parameters of the banking system, financial parameters of corporate bodies/entities like banks, companies are required, it may not be feasible to use primary data collection method.

Advantages of Secondary Data

- The data available is quick and ready to use. Hence, for a study having limitation of time schedule, it is easier to use secondary data method than the primary data.
- If the source of the data collected is reliable, such data can avoid the errors that otherwise could creep in while collecting primary data.
- In many cases, the secondary data may be less expensive than the primary data as it does not involve any elaborate process.
- Such data may allow the researcher to cover a wider geographic, cross-sectional or temporal range. It also provides an opportunity for cross-cultural analysis with lesser resources like time and cost.
- Secondary sources of information can yield more accurate data than that obtained through primary research. This is not always true but where a government or international agency has undertaken a large-scale survey, or even a census, this is likely to yield far more accurate results than custom designed and executed surveys when these are based on relatively small sample sizes.

Limitations of Secondary Data

- The biggest challenge while using secondary data is that the data should be relevant, recent and reliable for the given objectives of a study.
- The research conducted by primary data collection methods is generally controlled by the researcher; however, if the data used is secondary or is not collected by the researcher, the origin of the information may be questionable if the data is not verified appropriately.
- Secondary data does not address the specific needs of the researchers. For example, if the researcher wants the demographic profile in specific category like income, the categories defined by the researcher may not match with the categories defined in the data.
- Since the data available may not be that is specifically required by the researcher, it can be considered as inefficient spending of resources for the secondary data. The quality control of the data is also not possible.
- Since the secondary data is available to many others, the exclusiveness of the information derived from the data is lost.
- The secondary researcher may have to compromise on the envisaged objectives and scope of the study depending on the availability of data that is available.

The choice between collecting primary data or using secondary data, depends on the objective, confidentiality/exclusiveness of the study, type of data required, reliability of the available data, etc. If the data required is readily available, in a Government or some statutory/authentic agency's publication, the obvious decision would be to use the available data. But if the study is exclusive or consumer related where the consumer preferences change rather rapidly, the old data may not be reliable and valid, and it would be preferable to conduct a **primary research** i.e. research based on primary data.

It may be noted that even when one decides to conduct primary research/study, the secondary data can play a substantial role in understanding the research perspective especially in the exploratory phase, while defining the objectives and setting up the hypothesis. Secondary data might give valuable insights about the background of the research, and could be extremely useful both in defining the population and in deciding the sample size.

6.3 PRIMARY DATA SOURCES

The primary data can be collected by using **Quantitative methods** or **Qualitative methods** of data collection.

Incidentally, in this chapter, we have discussed only the methods of data collection, the actual process of collection is explained in the next chapter.

Survey is the most popular method for quantitative method. A **survey** is a method for collecting quantitative information about items in a population. The information is collected either using human intervention viz. a highly structured interview or without human intervention. Surveys may have different approaches like personally administered survey, telephonic survey, mail survey or electronic survey. All surveys are conducted using a fixed format of questions, termed as a **questionnaire**.

Advantages of Quantitative Method:

- Well-known methodology. This is the most popular method of collecting data.
- The entire process is structured. The process can be planned in advance.
- The process is independent of the researcher. The analysis uses standard methods/software.
- Some of the process can be outsourced; hence, the researcher's involvement can be minimised. For example, the researcher may not personally administer a survey, he/she may outsource it to some other specialised agency.
- Since all the respondents are asked the same set of questions, the quantitative analysis of the data is possible. The general approach to statistical analysis is to find similarities and differences between the subsets of answers.

Limitations of Quantitative Method:

- Its structured approach gives this method less flexibility.
- The method solely depends on a researcher's understanding of the topic. If the researcher does not have adequate understanding of the issue/problem, the method might not deliver valid results.
- Only written responses are considered, the unwritten responses like body language, the tone of talking, etc. are not considered in this method.

Qualitative Methods of Data Collection

Qualitative methods of data collection include observations, focus groups, semi-structured and unstructured interviews, etc., described later in the section. These methods do not use questionnaires. The emphasis of such methods is to have a flexible or unstructured approach while collecting data. This method is used when the research is in the exploratory stage and due to lack of research conducted on the topic, the researcher does not have enough knowledge about the research topic. In such cases, it is very difficult for a researcher to design a questionnaire and follow a systematic approach. In all the cases that require exploring new insights from the respondents, the qualitative methods of data collection are most appropriate.

Limitations of Qualitative Methods

- These methods are generally time consuming and the amount of information one can get through these methods may not be commensurate with the time spent on these methods

- As the data collected is mostly in the form of notes, recorded tapes, etc. it requires a highly qualified and experienced people.

Most researchers follow a mixed approach, wherein initially the qualitative methods of data collection are used by a researcher to understand the problem/issue under consideration. Once the insights are obtained using qualitative methods, these could be included to prepare a comprehensive questionnaire and then surveys can be conducted. Most market researchers use this approach to understand and analyse the consumer preferences.

We shall discuss some of the quantitative as well as qualitative methods of data collection in the following sections.

6.3.1 Surveys

Survey is a method of data collection, usually on a large scale. This is a structured method of collection of data. A survey, generally, has a fixed questionnaire containing a set of specific questions that are close-ended, and the responses are analysed statistically.

A survey can be administered using the following methods:

- Personally
- Telephonic
- Mail
- Electronic Media

These are described as follows:

6.3.1.1 Personally Administered Survey or Structured Interview The set of designed questions are personally asked by the researcher or interviewer. In this method, either the questionnaire is handed over personally, and taken back after completion by the respondent, or he/she is asked the questions orally and the responses are noted down by the interviewer. The first method is easier for the researcher and takes lesser time, and many questionnaires can be filled in a limited time by different respondents. In fact, the presence of the researcher is only to clarify any doubts that might be raised by the respondent while answering the questions. The disadvantage is that the respondent may take longer time, and because of which one may not find enough persons to fill the questionnaires.

In the second method wherein the responses are asked orally, and the researcher notes down the answers, the survey is easier for the respondent who has to only speak. This method, also termed as **Structured Interview,** may be adopted in two ways. First, the interviewer may hand over the questionnaire to the respondent, and the respondent may speak out the answers orally which are noted by the interviewer. Second, the interviewer may hold the questionnaire, asks the question orally and also note down the respondents' answers.

The disadvantage of personally administered survey is that it requires personal involvement, on the part of the researcher, to conduct the process. This becomes a limitation if the samples are distributed over large geographic area or the sample size is very large.

6.3.1.2 Telephonic Survey Many a times, it is not possible to personally conduct survey for each unit in the sample. This drawback is overcome by conducting a telephonic survey. In this type of survey method, the data is collected through telephonic interaction. The questions are asked over phone by professional callers, and the responses are noted down by the callers. The advantage of this method is that it can cover a larger geographic area than can be covered by personally administered surveys. It takes lesser time for the interviewer and can be convenient for the respondents too. The high telephone penetration in

India has made this method more convenient. The other advantages are, this method gives considerable cost advantage than the personally administered survey method. The telephone responses may be immediately entered to save time and cost of the data entry. The time spent in telephone interview method is much lesser than most other methods. If necessary, one can conduct the telephone survey within a day's time, which is not possible in other survey methods. Interviewer bias caused by the physical appearance, body language and actions of the interviewer is reduced in telephonic survey.

Major Limitations

- The questionnaire should be specially designed for telephonic survey. It should be precise, clear and short. If the questions are not understood by the respondent, it may create bias. Certain type of scaling questions like rank order questions can make the interview difficult for the respondents.
- The interviewer has to be given proper training to conduct the interview.
- The respondent may not be willing to respond. One should respect the respondents' choice if they are not willing to co-operate. The conversion rates i.e. the proportion of willing respondents for telephonic interview is much lesser than for the personally administered survey.
- The length of the survey questionnaire is very important in telephone survey. If the survey is too long, the phone could be disconnected by the respondent even before completion of the survey.

6.3.1.3 Mail Survey This method is used when either the respondents are geographically dispersed and too far to call or the survey is detailed/extensive to be conducted on phone. In some cases especially in the rural areas, the telephones may not be there. In such cases, the questionnaire is mailed to the respondents and the detailed instructions about filling up the questionnaire, and sending it back are explained in the mail. This is also termed as **Self-Administered Survey.** The respondent has to voluntarily fill the questionnaire and send it back to the researcher. The postage is paid by the researcher.

Mail surveys are typically perceived as providing more anonymity than other communication modes, including other methods for collecting data.

Limitations

- The conversion rate is too low. Participants may not co-operate with a long and/or complex questionnaire unless they perceive a personal benefit.
- This could be more expensive than the electronic surveys, described later.
- This method is time consuming as there is no certainty as to when the response is received back.
- Due to lack of direct communication, if any questions are not understood by the respondent, either the questions will be left unanswered or would be answered inappropriately.

6.3.1.4 Survey Using Electronic Media The self-administered surveys discussed earlier can also be sent through electronic media. This gives maximum reach. Practically, any location on the globe can be reached through this media.

Electronic mail and Internet are common in most of the countries. The Information Communication Technology (ICT) reach is increasing at exponential rates. This format of data collection is most efficient and cost-effective compared to the other formats. The questionnaire is sent through electronic media in two ways:

- By sending a document file containing the questions through the e-mail services.
- Using **on-line survey services** and making an **on-line survey**, and sending the **link** of the survey to the respondents by e-mail.

In the first method, though the electronic media is used, the process is not automated and the compilation of data through different documents may still consume time and resources.

In the second method, the survey is made online. It is very convenient for the respondents to participate in online survey. It takes minimum time for responding, by the use of popup, drop down menu, checkboxes, etc.

It is also easy for the researcher as most survey providers give readymade data file either in Excel or in SPSS form. This saves data coding efforts and also maximises the accuracy as the process is automated.

Some of the sites that provide this facilities are:

www.**SurveyMonkey**.com
www.**surveypro**.com
www.e**surveyspro**.com
www.**survey**gizmo.com/

Limitations

- This method can give biased results as the people who willingly respond to the survey may be from a section of society thus restricting the sample to that section of society. For example, most of the students are willing to respond to such surveys. If a consumer study is conducted through this media, the majority will be only from a section of a society, and, thus, the results of the study may not be valid, in general.
- The surveys are limited to the educated people who have access to mail. In India, still majority of people do not have this privilege. This again restricts the use of this media. Especially if the study undertaken requires data from rural/semi-urban area, then this media may not be much useful.

6.3.2 Observation

Observation is a **qualitative method of data collection**. In this technique, the information is captured by observing objects, human behaviour, systems and processes, etc.

For example, in a study about a product, instead of asking consumers a fixed set of questions like in questionnaire, one can appoint a person in a store to observe and analyse their behaviour. This could be more expensive method than the questionnaire method, but the results obtained through this method are more reliable. The respondent error discussed in Chapter 5, could be eliminated through this process. The observer generally does not interfere in the process, and to that extent, the observer's bias is also eliminated. The observation techniques may reveal some critical information, that otherwise is not disclosed in the other forms of data collection. This method requires a qualified person for data collection. The most common use of this method is done to access a process, where a qualified person walks through the process, observes the process critically and notes down about the process.

Advantages

- The data collected through this method is original, first hand, accurate and authentic.

- One may capture information that the participant might ignore, if asked in any question in a questionnaire, assuming the information is too trivial or not important for the study.
- Certain type of information to be obtained by accessing a process, procedure or event can be collected only through this method.
- This method has the least intrusion for the participants.

Limitations

- This method requires physical presence of a qualified observer.
- The process is very slow, and the researcher may have to spend considerable time before acquiring any valuable information.
- The process is also more expensive than other methods.
- The information captured is solely dependent on the skills of the observer, and is subjective. The quantitative analysis of the information is generally not possible.
- It is difficult to find the logic or rationale behind the behaviour from this method. The method only accesses the behaviour and not the rationale behind the behaviour. This enables the researcher to conclude about the results, but not about the reasons behind the results.

It may be noted that the selection of the observer is very critical to the success of this method. We shall, therefore, discuss some of the qualities desired in an observer.

Qualities Desired by the Observer

- The observer should have the ability to remember the details of the experiment/event/process or note down quickly while still observing.
- The observer should be able to concentrate and pick relevant information from the behaviour/event/process.
- The observer should also be able to remain unobtrusive.
- The observer should have patience to collect data as this is a slow process.

6.3.3 Interview

Interviewing is the most commonly used method of data collection. Interview could be of three types:

- **Structured Interviews**
- **Semi-Structured Interviews**
- **Unstructured Interviews**

Structured interviews are the quantitative method of data collection whereas semi-structured and unstructured interviews are qualitative methods of data collection.

Structured Interviews

A survey can be conducted by using a structured interview. This method is generally termed as personally administered survey. We have discussed this method in Section 6.3.1.1. This is a quantitative method of data collection and the analysis of the data collected is generally quantitative.

Semi-Structured Interviews

This method is used when the researcher asks the respondent some basic questions, and then lets the respondent answer, interfering whenever necessary. In this method, the interviewer sets some

guidelines for the questions to be asked. The succeeding questions are generally on the basis of the preceding questions.

The most common example for semi-structured interview is 'Job Interview'.

Unstructured Interviews

Unstructured interviews are those that allow the interviewer to get opinions and get a feel of general attitudes of the respondents. They are exposed to lesser degree of bias and based on their inputs and explanation, a researcher can get deeper insights. This can help researchers interpret the respondent's output better than in structured questions. Unstructured interviews are generally used in exploratory research, and are generally time consuming. The biggest challenge of this method is that the data generated is in unstructured format making it difficult for quantitative analysis. The coding for such data, is extensive to allow methodical analysis, and needs expertise. Though there are selective software tools like Computer Assisted Qualitative Data Analysis Software (CAQDAS) available for this purpose, their scope is limited.

Adequate attention has to be paid while interviewing respondents who may not be very adept at expressing themselves.

6.3.3.1 Focus Groups This is yet another variation of interviewing technique. Focus groups are small selected group of participants who are interviewed by a trained researcher. The participants are from a target research audience whose opinion is of interest to the researcher and the client. The discussions are generally in form of exchange of experiences, opinions on how they feel and their ideas on a specific topic.

The researcher generally guides the discussion in a direction that will lead the participants to opine on the relevant issue. The discussions during such interactions allow a free and open discussion, wherein the researcher might receive some tips on a varied line of thought that could be of advantage and was not previously perceived.

The selection of a focus group has to be given due importance. Smaller groups are preferred to achieve natural and well co-ordinated discussions. The participants are to be selected, as far as possible, from a similar economic, social and cultural background. This minimises any conflict that could arise within the group, and contribute towards achieving the set objectives.

The researcher's skills are very important in keeping the discussions alive and smooth without getting entangled by any controversy or bias. In fact, he/she has to play the role of a **moderato**r, and, thus, should have the ability to make the participants at ease. His/her timely intervention and probe to get more information in a congenial atmosphere is very important. The researcher should, therefore, have fair knowledge of the topic that is to be discussed and should be able to understand and effectively utilise the group dynamics or group behaviour.

In the context of market research, this method is used to elicit their perceptions, opinions, beliefs and attitudes towards a product, service, concept, advertisement, idea or packaging. In particular, focus groups allow companies wishing to develop, package, name or test market a new product, to discuss, view and/or test the new product before it is made available to the public.

In fact, focus groups could provide more reliable information, and could be less expensive than other forms of traditional marketing research.

Some other variants of focus groups are as follows:

- **Dual moderator focus group** – One moderator ensures the session progresses smoothly, while another ensures that all the topics are covered

- **Teleconference focus groups** – Telephone network is used for interaction
- **Online focus groups** – Computers connected via the internet are used for interaction

It is reported that United States Federal government makes extensive use of focus groups to assess public education materials and messages for their many programs.

6.3.3.2 Projective Techniques The projective technique is a technique of interviewing. It is an indirect method of questioning in which the interviewer directs the questioning to receive responses on questions not directly related to the respondent, and analysing the response behaviour to matters not directly related to him. With this technique, one is able to bring out the hidden or suppressed feelings. The interviewer is able to project the responses of the respondents which are more directly applicable to him/her. This method of data collection requires psychological skills and tools. One of the common uses of these techniques is in marketing research wherein consumer attitudes, motivations, beliefs, feelings, etc. are uncovered/revealed with reference to a product, service, system, etc.

Some of these techniques include:

Word Association	The respondents are asked to read a list of words and to recollect and indicate the first word that comes to their mind.
Pictures and Words Association	The respondents are given a number of words and pictures and are asked to choose those they associate with a brand or product and to explain their choice.
Sentence/Story Completion	Respondents are given an incomplete sentence, or a story, and asked to complete it.
Thematic Apperception Test (Cartoons or Empty Balloons)	Respondents are required to give opinions of other people's actions, feelings or attitudes. "Bubble" drawings or cartoon tests provide an opportunity to fill in the thought or speech bubbles of the characters depicted.
Expressive	Respondents are required to role-play, act, draw or paint a specific concept or situation.
Brand Mapping	Respondents are presented with different brands and are required to offer their perceptions with respect to certain criteria.

6.4 SECONDARY DATA SOURCES

We have discussed in Section 6.2 about secondary data, its advantages and limitations. The most important requisite about secondary data is the reliability of the source. We shall discuss in this section the various reliable sources of secondary data.

The major sources of secondary data are:

- **Publications**
- **Project and Research Reports**
- **ERP/Data Warehouses and Mining**
- **Internet/Web**

These are described as follows:

6.4.1 Publications

The material appearing in the print media can be labelled as a publication. Daily newspapers, magazines, encyclopaedias, textbooks, hand-books, reports, journals, etc. can be termed as publications.

These are also termed as the reference material, and these constitute widely available sources of data. Most researchers use this media at literature review stage of the research study for getting a wide perspective of the topic. The indexes, bibliographies, catalogues, etc. help search the topic in a systematic manner. Proper referencing should be given to identify the source appropriately. Some of suggested publications are:

- Government of India publications like Economic Surveys presented before the budget, budget speeches, etc.
- Statutory bodies like Reserve Bank of India publications like Annual Reports, Reports on Currency and Finance, Handbook of Statistics on Indian Economy, Trend and Progress of Banking in India, etc.
- CMIE publications encompassing a wide range of economic parameters relating to the Indian economy
- Publications of
  - — Associations of segments of industries like Indian Banks Association
  - — Confederation of Indian Industry (CII)
  - — All-India Association of Industries (AIAI)
  - — Association of Chambers of Commerce and Industries in India
  - — National Association of Software & Service Companies (NASSCOM)
  - — Telephone Regulatory Authority of India (TRAI), etc.

Incidentally, the Government and the statutory bodies are generally considered to be more authentic sources of secondary data.

Some reputed consultants like Gartner Group, McKinsey, etc. also bring out reports containing valuable data that could also be considered as secondary data sources.

"Dun & Bradstreet (D&B) is the leading provider of international including Indian business information. HOOVERS, also a leading data base company, provides data about companies at global level.

6.4.3 ERP/Data Warehouses and Mining

Data is stored in organisations in two major types of systems viz.

- Transaction Processing (TP) Systems or Enterprise Resource Planning (ERP) systems, and
- Data Warehouses

These are described in the following:

Transaction Processing (TP) Systems or Enterprise Resource Planning (ERP) Systems

These systems store the recent data on the business transactions. Depending on the size of the data, this system may store data from one day to one year or more. The data is well protected by different types of accounts (user IDs and passwords). Each account on the system is defined specific roles/functions, and is allowed access to specific data. For example, a purchase officer's account will have access to all the information relating to the decision of purchasing like different vendors, their products, prices, their history of past purchases, delivery time, quality control, etc. but he/she may not have information on the other aspects of the vendor like the bank account number of the vendor, or any other data that is not required by the purchase officer.

Relevant access to collect the data through ERP or TP systems

A researcher may need to collect this data if the research is about the process/transactions. The data collected from these systems is generally considered as primary data. In most other cases, the data for research is collected from the data warehouses, described in the following:

Data Warehouses and Mining

A **data warehouse** is a repository of an organisation's electronically stored data. Data warehouses are designed to facilitate reporting and analysis, based on data. Most organisations maintain historical data separately so that it is easy to analyse, and does not affect the speed of the transactions. These objectives are achieved by setting up a separate Data Warehouse.

The data collected through data warehouses is considered as secondary data. Data warehouses store huge amount of data making it difficult to analyse manually and, thus, requiring software to analyse the data. This is done by **Data Mining** software. Data Mining is a specialised branch encompassing Information Technology and Statistics. It uses statistical techniques such as Outlier analysis, Correlation and Regression analysis, Analysis of Variance, Discriminant analysis, Cluster analysis, Factor analysis, etc. for extracting valuable knowledge like patterns, associations, etc. from the data. The concept of Data Mining is used when the data is huge and is beyond human capabilities to analyse. The statistical techniques are used with the help of computer programs and preset algorithms.

There are many software companies provide mining tools. Excel Miner is a mining tool by Microsoft. Statistical packages like SPSS, SAS and SYSSTAT also have data mining features.

It may be noted that we have discussed most of the earlier-mentioned statistical techniques in the book. We have also used tools like SPSS or Excel for each of these techniques. But using merely a tool is not data mining. When one uses the tool, all the variable selections and other selections like method, type of tool, etc. is done by the user, whereas data mining is a completely automated process. The system decides which variables to use; in fact, it tries all the possible combinations of variables. It also uses the artificial intelligence to decide the method to be used and gives the desired output.

6.4.4 Internet/Web

One of the best ways to collect any secondary data is to search on the web. Specific related words could be entered to activate the search process. World Wide Web (WWW) is collection of millions of WebPages containing information about practically all the topics. The current trend is to store the publications in the web form. This may include directories, dictionaries, encyclopaedias, newspapers, journals, e-books, Government reports, etc. One may choose the relevant website from the vast list displayed on the screen. Some of the important websites are given in the next section.

Internet is also responsible for data explosion. The internet is flooded with data, and the challenging task is to identify or locate the relevant or useful data from the huge amount of data. This task is generally not easy. Though there are intelligent searches available like Google, Yahoo, etc., even then these searches yield thousands of options and one needs to be patient to locate the exact data one is looking for. To get over these limitations, one can also use online databases, like EBSCO, ECCH, CMIE, Captialine, manupatra, legalpundits, etc. These are specialised databases, generally subscribed with a fee. The searches in these databases can be easier and specific than on websites. There are few free online databases like Wikipedia which is most useful site for researchers. Online journals also form a major source of secondary data.

There are also some journal database providers like Elsevier, Emerald, JSTOR, ProQuest, sciencedirect, etc. These subscribe to different journals and provide a combined access to the subscribers. This saves trouble of separately subscribing to the different journals for an individual subscriber. This also saves money as one has to only subscribe to the journal databases than the individual journals.

6.4.4.1 Some of the Important Websites

Owner/Sponsored	*Site Address*	*Description*
Reserve Bank of India	www.rbi.org.in	Economic Data, Banking Data
Securities Exchange Board of India	www.sebi.gov.in	Data
Bombay Stock Exchange	www.bseindia.com	Data
National Stock Exchange	www.nseindia.com	Data
Finance Ministry	www.finmin.nic.in	Data/Publications
Glossary of Statistics*	www.wikipedia.com	Definitions and Brief details*
Princeton University	www.dss.princeton.edu	Data and Statistical Services
Statistics Help Us	statsdirect.com	Data and Statistical Services (Paid Service)
The Oecd's Online Library	http://caliban.sourceoecd.org	Statistical Databases, Books and Periodicals
Nationmaster	http://www.nationmaster.com/index.php	Sources as the CIA World Fact book
India Stat	www.indiastat.com	Statistical data and useful information on India
Federation of Indian Chambers of Commerce and Industry	http://www.ficci.com	Commerce and Economic data
Iassist	iassist**data**.org	Indian Publications
Bepress	www.bepress.com	Electronic journals (paid site)
International Telecommunication Union	http://www.itu.int/	Publications and cases studies
World Bank	www.worldbank.org	Data
World Health Organisation	Health topics - www.who.int/topics/en/ Countries - www.who.int/countries/en/ About WHO - www.who.int/about/en/ Publications - www.who.int/publications/en/	Data
ISI (Indian Statistical Institute)	www.isical.ac.in/~library/	Web library
Telecom Regulatory Authority of India	http://www.trai.gov.in	Data on telephone network

(*Contd*)

(*Contd*)

Centre for Monitoring Indian Economy	Industry Analysis Service - www.cmie.com/.../industry-analysis-service.htm Economic Intelligence Service - www.cmie.com/database/?service=database-p... Products - www.cmie.com/database/?service=database-products.htm	
Euro monitor	Euromonitor.com	International market intelligence on industries, countries and consumers
India Info line	http://www.indiainfoline.com/	Information of stocks and other financial services
EBSCO	http://web.ebscohost.com	RESEARCH DATABASES (paid)
Confederation of Indian Industry	http://www.ciionline.org/	Business data, cases, publications
ECCH (European Case Clearing House)	www.**ecch**.com	Case study database (paid)
National Informatics Centre, India	http://www.nic.in	Indian data
HOOVERS (A Database Company)	www.hoovers.com	Data about companies at global level
D&B - Dun & Bradstreet	www.dnb.co.in	Leading provider of international and Indian business information

*This is a very useful site for searching the definition and other details of statistical topics.

6.4.4.2 Searching Databases/WebPages Manually, the books/journals can be searched by the catalogs available in the library. This manual method has restriction, as the name of the book, journal, author, year of publication, etc. is required to be known before the search. Online databases are much easier to search. Since these are in the electronic form, one can search using various index tags like author, topic, date, publication type, etc. The databases may have full text papers or the abstracts. One can also choose to search in full text articles or in abstracts. The search options may vary from database to database. A researcher should follow the guidelines while searching the databases/websites.

Guidelines for Searching Databases/WebPages

- Select a specific topic to be searched. The topic should not be too general or too specific. Too general search will lead to thousands of searches and will be time consuming for the researcher to evaluate the searches and pick the relevant ones. Too specific topic may lead to too few searches, and inadequate information about the topic. For example, if one inputs 'telecommunication' as the topic of search, there will be too many searches to select from. On the other hand, if one inputs 'telecommunication policy for rural areas', there may be too few searches.
- Select a database that is appropriate to the topic selected.
- Construct a 'Query' on search databases, using different tags or indexes.

- Modify the query if it fails to give desirable result or it gives too many or too few results.
- It is always advisable to save the results at appropriate location with the detailed references, so that the researcher can trace them to the source and can give appropriate references at the end of the report.
- One could also repeat this process on the websites after searching the databases.

SUMMARY

The primary and secondary sources of data have advantages and limitations as well.

The two basic methods of data collection viz. qualitative and quantitative methods of data collection can be differentiated with respective advantages and limitations.

The following are the methods of conducting a survey:

- Personally Administered Survey
- Telephonic Survey
- Mail Survey
- Survey Using Electronic Media

The other methods of collecting primary data are observation, interview, projective techniques and focus groups. There are guidelines to collect the data from various sources such as:

- Publications
- Project and Research Reports
- ERP/Data Warehouses and Mining
- Internet/Web

DISCUSSION QUESTIONS

1. Discuss the advantages and limitations of primary and secondary data.
2. Give a brief description of the sources of primary data. Describe two situations where a particular source would be more appropriate, for each of the sources.
3. Describe various types of surveys used for collection of data, indicating their respective advantages and limitations.
4. Discuss the issues relevant to 'observation' as a method for collection of data.
5. Describe various facets of the 'interview' method of data collection.
6. Describe the various sources of secondary data. Provide guidelines for collecting data through secondary sources in ten points.
7. Name five websites indicating the type of data available on those sites.

Collection and Preparation of Data

7

Contents

LEARNING OBJECTIVES

- To explain the process of collecting primary and secondary data
- To provide requisite knowledge about various aspects associated with designing a questionnaire for collection of primary data
- To outline the steps for preparation of data
- To provide guidance in collecting/recording data from secondary sources

Relevance

The XYZ Consultants had bagged a research project from one of its clients, a leading player in telecom sector. The project related to assessing the impact of the Mobile Number Portability to be introduced by TRAI for the Indian telecom sector.

Mr. Uday, the chief consultant, met the client's representatives for discussing various strategies for collection of data, and to finalise the questionnaire for data collection.

The discussion was important as the study was to be conducted across a wide geographical area and covering a wide range of telecom users. The span of the study was gigantic. It was to cover rural and urban areas, different states, different users speaking different languages, etc.

There were many issues to be discussed in the meeting such as: What is going to be the unit i.e. object, entity, system, etc. of the study?
How will the data be collected i.e. personally administered, telephonic, mail, or e-mail survey or mix of these methods?
What questions to be asked?
What should be the ideal length of the questionnaire?
What should be the sequencing of the questions?
How should the questions be framed?
Is there any requirement of multiple questionnaires?
What could be the advantages and limitations of having multiple questionnaires? etc.

After the fruitful discussion, the team jointly decided to go ahead with a mixed approach of data collection. It was also decided to have multiple questionnaires, different for rural and urban customers, and analyse them separately for the two groups.

7.1 INTRODUCTION

We have discussed in Chapter 6, the different sources of data and their respective advantages, limitations and also various methods of collection of data. In this chapter, we shall discuss the methodology of collecting the data using various tools/instruments or schedules. We shall also discuss preparing the data for presentation and analysis.

This has been done separately for primary data and secondary data.

7.2 COLLECTION OF PRIMARY DATA

Different methods of collecting primary data like questionnaire, observation, interview, projective techniques, focus groups, etc. have been earlier discussed. In this section, we have discussed

collecting, coding and preparing the data for subsequent presentation and analysis for each of the methods.

7.2.1 Questionnaire

Most research studies use questionnaires. Surveys use questionnaires to collect data in the systematic format. A questionnaire is a set of questions asked to individuals to obtain useful information about a given topic of interest. If properly constructed and responsibly administered, questionnaires can become a vital instrument by which inferences can be made about specific groups or people or entire populations. The advantage of questionnaire is its ease of use, getting answers 'to the point' and the ease at which the statistical analysis can be carried out. In this method of collecting data, a wide range of information can be collected from a large number of individuals.

The designing of the questionnaire forms the most important part of a survey. Adequate questionnaire design is critical to the success of the survey. Any errors at the design stage of the questionnaire could prove to be fatal at a later stage. The questionnaire is at the centre of any quantitative research study, and, therefore, due attention should be paid to design the questionnaire. Inappropriately asked questions, inappropriate order of questions, format, scaling, etc. could generate errors in the research study as the responses may not actually reflect the opinions of the respondents. Most of these errors can be avoided by conducting a **pilot study** or **pretesting** the questionnaire. After the questionnaire is constructed, it is tested on a set of people on various aspects like completeness, time taken to answer, clarity of questions, etc. If executed appropriately, pretesting can reduce design errors that otherwise could have made the research study worthless. The questionnaire is generally revised after incorporating the learning's from the pretesting.

It may be noted that even if the pretested responses is retained as such in the final study, the pretested questionnaire is retained as such.

7.2.1.1 Questionnaire Modes of Responses

Different methods by which the questionnaire can be administered have been earlier discussed. We will discuss in this section the guidelines for designing the questionnaire, for different modes.

Personally Administered Survey

Since the trained personal is present at the time of collecting data, a questionnaire designed for personally administered surveys could be extensive seeking in depth information. Incidentally, in-depth studies can be possible only through this format. The questionnaire could be seen by the respondent (unlike telephonic interview). This could be another advantage of this format. There might be separate guidelines for interviewer and the interviewee. Two samples of personally administered surveys are displayed at the end of the chapter as Sample Questionnaires 1 and 2.

Telephonic Survey

In case of a telephonic survey, the interviewer interacts with the respondents, but is not personally present. The questionnaire is also not seen by the respondent. These form the limitations for collection of data using this method. These limitations restrict the questionnaire design for this method of data collection. The questionnaire should not be too long. It should not have questions that take long time on phone. The instructions for answering questions should be simple so that the respondent understands them easily. There are restrictions on using some of the scales like fixed sum scales,

diagrammatic scales, semantic differential scales, etc. The interviewer should be trained for the survey. There should be a mechanism to verify if the survey really took place or not. The data can be directly entered in a form or an Excel sheet to save the coding time.

Mail Survey

Mail surveys are self-administered. Though the questionnaire can be seen by the respondent, there is no interaction with the interviewer. This limitation should be considered while designing the survey. The questionnaire must have clear instructions, clearly stated questions, etc. Since this is self-administered, the questions should be so designed so as to hold the interest till the end of the questionnaire. The questionnaire should not be too lengthy that it leads to no response. The flow of the questions should follow a pattern like, easy questions in the beginning, followed by difficult questions, followed by easy questions at the end. This ensures maximum responses. This pattern assumes that the concentration level of an individual is less in the beginning, increases over time and reduces after some time. Time spent while answering questions is very important factor for these surveys. Lesser the time spent, more will be the responses. This factor is primarily considered during the pretesting of the questionnaire.

E-mail Survey

As earlier discussed, the survey through electronic media can be done in two ways—sending the soft form through e-mail, or using online survey services. In the first method, i.e. sending the soft version of the form through e-mail, there is no difference in the questionnaire design than used in mail surveys, as only the medium of collecting data is different. In the later case, the survey is constructed using HTML format and tools. It makes the questionnaire easy to construct, easy to answer, (by clicking answers), and easy to code. But this method adds some limitations in the type of questions. The online survey providers do not provide all types of scales—for example, scales like semantic differential scale, graphic rating scale, etc.

Open-ended questions are also difficult to answer and code. A sample of E-mail survey is displayed at the end of the chapter as Sample Questionnaire 3.

7.2.1.2 Designing of a Questionnaire We have discussed in the previous section, the considerations specific to method of survey, while designing a survey. We shall discuss in this section some general guidelines for designing a questionnaire.

The questionnaire design often faces trade-off between the depth of the study to be conducted and the likely receipt of the responses. The more the depth of the study, lesser could be the responses received. The researcher should consider this aspect and design the questionnaire that has adequate balance between these two factors. There are many other factors to be considered for improvement in the responses. These are discussed in brief in section 7.2.1.3.

The questionnaire for a research study is designed based on the questions hierarchy discussed in Chapter 3.These questions form the building blocks for questionnaire design. The survey strategy is designed only after understanding connection between investigating questions and possible measurement questions.

Some of the strategies are:

- Deciding the type of scale needed
- The method of data collection to be used

- Dividing questionnaire into different sections
- Order of the questions to be asked
- Wording of the questions
- Length of the questionnaire
- Types of questions to be used like structured/unstructured
- Approach of the questionnaire—disguised/undisguised

We have discussed each of these, in brief, as follows:

Decision about the Type of Scale

Various measurement scales have been earlier discussed, viz. comparative scales like paired comparison, rank order, constant sum, etc., non-comparative scales like continuous rating scale, itemised rating scale, likert scale, semantic differential scale, staple scales, simple/multiple category scale, verbal frequency scale, etc., their advantages and limitations.

The decision of using appropriate scale is taken, considering the measurement questions, respective hypothesis and intended analysis to be done. Certain scales have limitations on the type of analysis to be carried out as brought out in Chapter 5.

Method of Data Collection

Different modes of questionnaire responses have been described in Chapter 6. The choice of method of data collection like personally, telephonic, web-based, etc. is an important strategy decision depending on the type of design used, the depth of the study, etc.

Dividing Questionnaire into Sections

The questionnaire is generally divided into different sections, to sustain interest, to allow logical sequence and to keep similar types of questions together. The decision about number of sections should be balanced. While too many sections may create negative impact on the respondent, too less sections (and eventually too many questions in a section) may affect responses due to losing interest by the respondents. **The economy of a questionnaire is defined as the time spent by a respondent to answer the questionnaire.** This can be achieved by keeping similar scale questions together so that the respondents do not waste time in reading the instructions about the scales again and again.

The sections should have clear instructions to indicate the change of scale, if any.

Order of the Questions

The order of questions to be asked is a very important aspect to be considered for avoiding errors. If certain questions are asked in wrong order, the respondent error may creep in.

Some general guidelines are given as follows:

It should sequence the questions:

- From general to particular.
- From easy questions to difficult ones.
- From factual to abstract.
- From closed format questions to open-ended.

Wordings of Questions

This is the most important aspect of designing a questionnaire. Many a times, the questions are not worded appropriately to get uniform or desired responses. It may result in different respondents interpreting the same question differently, or the respondents interpreting the question in a different way than meant by the designer of the questionnaire. This could result in the respondent's error. This problem may arise due to the different levels of vocabulary between the respondents and the questionnaire designer. This problem may also be due to long ambiguous questions. This error can be avoided by getting the questionnaire screened by others (other than the ones who design it).

Length of the Questionnaire

There are no universal agreements about the optimal length of questionnaires. It depends on the type of respondents, type of study, etc. However, short simple questionnaires usually attract higher response rates than long complex ones. The questionnaire should be lengthy only if it is necessary. The length of the questionnaire can also vary for different demographics. The time spent willingly by students and by executives is not the same. The topic of survey also plays an important role. People willingly participate in survey on interesting topics. The design of the survey also plays an important role. If the survey is designed carefully, people are willing to participate even in a lengthy survey.

Types of Questions—Structured/Unstructured

The questionnaire generally contains more structured questions than the unstructured. If the study requires extensive use of unstructured questions, the preferred method could be interviewing. The respondents generally avoid lengthy open-ended answers for unstructured questions.

There is always trade-off between the open-ended questions and the ease of analysis. More the open-ended or unstructured questions, more difficult will be the analysis using quantitative methods. The unstructured questions should be analysed separately to understand the pattern of responses.

Approach of the Questionnaire—Disguised/Undisguised

A disguised question is a question asked without revealing the actual purpose of the study. The study may always have some amount of disguise to obtain unbiased responses. For example, if a respondent comes to know about the company for which the study is done, the responses may be biased, either for the company or against the company, dependent on the experience the respondent had with the company. The objective of the study is generally disguised to avoid the errors. The information that the respondent might share willingly, need not be disguised. If the questions require to be disguised, the projective techniques may be considered.

7.2.1.3 General Guidelines of a Questionnaire We shall discuss the general guidelines to be considered by a researcher while designing a questionnaire.

There are two main objectives to be achieved while designing a questionnaire, viz.

- To maximise the response rate response rate i.e. the proportion of respondents answering the questionnaire
- To obtain accurate, relevant and necessary information for the survey.

To a certain extent, these two objectives are contradictory to each other. For example, to obtain the requisite information, a researcher may have to ask many questions that may lead to lengthy questionnaire affecting the response rate. Similarly, getting accurate and relevant information may

consume more time of the respondents. There should be an appropriate balance between these two objectives.

To maximise the response rate, the researcher may consider the following:

- One needs to consider carefully how to administer the questionnaire
- Establishing rapport goes a long way to increase response rate
- Explaining the purpose and the importance of the survey helps
- The researcher needs to remind those who have not responded
- The length of the questionnaire should be appropriate

To obtain accurate, relevant and necessary information, a researcher may consider the following aspects:

- Asking precise questions
- Using short and simple sentences is important for the respondent's understanding. Thus, only one piece of information should be asked at a time
- The questionnaire should be planned as to what questions be asked, and in what order, how to ask them, etc. according to the objectives of the study
- Sometimes, additional questions can be used to detect the consistency of the respondent's answers. For example, there may be a tendency for some to tick either "agree" or "disagree" to all the questions. Additional contradictory statements may be used to detect such tendencies
- Selecting the respondents who have the requisite knowledge

7.2.1.4 Types of Response Questions We have discussed in Chapter 5 various scales and types of questions that are appropriate for questionnaire design. The questions generally evolve from the questions' hierarchy, the hypothesis developed and the analysis planned for the study discussed in Chapter 3. Appropriate questions should be used keeping those aspects in mind.

7.2.2 Observation

The data collected through observation method requires to be transformed in appropriate form to perform requisite analysis on the data. The data collected using this method is generally in the form of notes by the observer or the sound clip/video clip of the recording of the process. This makes it difficult to analyse directly. Some artificial intelligent software can be used for compiling such data.

7.2.3 Interview

If the interview technique is structured, the data collected is in the similar form as the questionnaire, and can be easily compiled. In case of unstructured interviews, elaborate compilation skills are required to analyse the data.

7.2.3.1 Focus Groups The moderator of the focused group collects the data. The data may be collected in the form of notes, audio or video recordings of the discussion. This makes it difficult to analyse directly. Some artificial intelligent software can be used for compiling such data. The extraction of information from the data requires relevant expertise for compilation and interpretation of the data.

7.2.3.2 Projective Techniques Since these methods of data collection require psychological skills, the data needs to be collected by a skilled person. The compilation and analysis of the data also requires relevant expertise. These methods are generally used by the specialists and the data is collected and compiled by the experts.

7.3 PREPARATION OF DATA

The data collected from the respondents is generally not in the form to be analysed directly. After the responses are recorded or received, the next stage is that of preparation of data i.e. to make the data amenable for appropriate analysis. This process encompasses the following stages:

- **Editing**
- **Coding**
- **Validation**

These are described as follows:

Editing

The purpose of editing is to ensure consistency in the responses, and to locate omission(s) of any response(s) as also to detect extreme responses, if any. Further, editing also checks legibility of all the responses and seeks clarifications in the responses wherever felt necessary.

Coding

The process of converting responses into numeric symbols is called codes. The purposes of coding are:

- Easy to input and store in computerised systems
- Easily amenable for computerised processing and analysis
- Easily amenable for sorting and tabulation

In fact, coding of data is decided according to the needs of inputting, storing, processing, sorting and analysis of data.

Incidentally, the codes could be alphanumeric also but for using software packages like Excel and SPSS, the codes should preferably be numeric.

Validation

After the data is coded, it is validated for data entry errors. The data is then used for further analysis. The purpose of validating the data is that it has been collected as per the specifications in the prescribed format or questionnaire.

For example, if the respondent is asked to rate a particular aspect on 1 to 7, then the obvious responses should be 1 or 2 ….., or 7. Any other inputted number is not considered as valid. In validation of the data, the above data will be restricted to the integers between 1 and 7. This minimises the errors. The other validations are age within a number like 100, dates such as birth dates, joining dates, etc. should not be future dates, etc.

Incidentally, while editing is done after the receipt of the responses, validation is done after inputting the responses after coding. The validation is especially used to reduce data entry errors.

7.4 COLLECTION OF SECONDARY DATA

We have discussed various secondary data sources in the previous chapter. In this chapter, we shall

confine the discussions to the aspects related to the collection of the secondary data that is authentic and relevant to the study being conducted.

The relevance of the secondary data is the most important factor to be considered. The second important factor is to ascertain the authenticity or reliability of the secondary data.

Authenticity of Data

While collecting or recording data from a secondary source for a researcher, it is to be ensured that the data has been published by the agency that is authorised to either collect the primary data or get it directly collected from relevant sources.

For example, Government of India is the competent authority to collect data relating to prices and production. Thus, the data published by Government can be taken as an authentic source for secondary data relating to prices and production. The same data can also be taken as authentic when published by some statutory body like Reserve Bank of India. Similarly, Reserve Bank is authorised to collect data about banking related parameters like deposit, advances, etc. and, therefore, the data published by Reserve Bank of India may be taken as authentic. In fact, Reserve Bank of India disseminates a variety of data relating to Indian Economy, through various publications (listed in Chapter 6) and all those data are considered as authentic secondary sources of data for any study about Indian economy.

The data collected and disseminated by associations and federations of various industries could also be taken as authentic secondary data by a researcher.

At the global level, authentic data related to various aspects of different countries—either in consolidated form or country/region wise is published by United Nations, World Bank, International Monetary Fund, etc. Such data could also be taken as authentic secondary data by a researcher.

The other reliable sources could be the database services, which regularly conduct surveys and elaborate the methodology of collecting data and the error control. Published company data can be also reliable to a certain extent. Other sources of data should be first verified for the reliability before considering the use of the data.

While referring to the data published by private bodies, consultants, newspapers, magazines, etc., it may be ensured that they refer to the official publications as mentioned above before accepting the data as authentic.

Some other relevant issues while referring to secondary data are the definitions used for various parameters like deposits and profit, the period for which the data is available, extent of coverage of data, etc. In fact, the footnotes given for a data should be read carefully to assess their impact, if any, on the research study.

7.5 SAMPLE QUESTIONNAIRES

In this section, we have given some sample questionnaires for reference purpose.

SAMPLE QUESTIONNAIRE I

Questionnaire on Social Networking

1. Name:

2. Age:
3. Gender • Male • Female
4. Occupation:
   • Student • Working/Business • Housewife • Others
5. What is the purpose for your using the Internet? (Please rate 5-maximum and 1-minimum)

Purpose	*5*	*4*	*3*	*2*	*1*
Work					
Entertainment					
I have a lot of free time at hand					
That's the trend					
Need for networking with friends					

6. Which networking sites have you registered on?
   • www.orkut.com
   • www.facebook.com
   • www.ibibo.com
   • www.hi5.com
   • www.myspace.com
   • www.bigadda.com
   • Others _____ (Please specify)
7. What made you join these sites? (Please rate 5-maximum and 1-minimum)

Reason	*5*	*4*	*3*	*2*	*1*
My friends asked me					
They are the trend					
Advertisement					
Just came across them while surfing the internet					
Need for networking to stay in touch with friends					
Need_____ (Any please specify)					

8. Which is your most favourite website? (Please rate 5-maximum and 1-minimum)

	Visited					*Preferred*					*Useful*				
Orkut	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
Facebook															
Ibibo															
Hi5															
Myspace															
Bigadda															
Others_____															

9. What is it that makes you like the website? (Please rate 5-maximum and 1-minimum)

Features	*5*	*4*	*3*	*2*	*1*
Interface					
Crowd					
Simplicity					
Speed					
Other features _____(Please specify)					

10. How many times do you visit your account approximately?
    - Daily
    - _____ times in a Week
    - _____ times in a Month
11. What is it that you enjoy the most? (Please rate 5-maximum and 1-minimum)

	5	*4*	*3*	*2*	*1*
Adding friends					
Putting up pictures					
Making a profile					
Scrapping/Chatting					
Looking around at what friends are doing					
Playing games					
Making groups/writing blogs					
Seeking relationships online					
Ahh I'm there but I hate these sites!!					
Others_____ (Please specify)					

12. Do you feel safe to upload your picture and disclose your true self?
    - Yes, absolutely
    - No, absolutely not
    - Maybe, if there is restriction on the viewers
13. Has your attitude towards these sites changed after the adnan patrawala case?
    - Yes
    - No
    - I don't know about it
14. What is your take on the number of social networking sites that have flooded the internet?
    - Yes, I'm all for it
    - No, I think they are shady
    - Well, each one to his own
15. Do you think these sites are short-lived? why?
    - Yes
    - No
    - I Don't Know

_____

16. Would you prefer a networking site with a purpose or some use apart from just networking and fun? what kind of purpose would you like it to solve?
    - Yes
    - No

networking with a purpose!! (please rate 5-maximum and 1-minimum)

	5	4	3	2	1
Business networking					
Intellectual networking i.e. meeting likeminded people					
Project sharing					
Purposive networking, ex- Marriage, Jobs					
Carpooling					
Meeting for common purposes like social work, animal activists					

- Others _____

_____

_____

(Please specify)

SAMPLE QUESTIONNAIRE 2

Consumer Study for Two Wheelers

1. Name:
2. Age:
   (a) 18-20
   (b) 21-25
   (c) 26-30
   (d) above 30
3. Income:
   (a) > 1 lakh
   (b) 1-2 lakhs
   (c) 2-3 lakhs
   (d) > 3 lakh
4. Rate the following on the basis of mode of your preference. (Use √ against number) (1-least preferred, 5-most preferred)

Mode	1	2	3	4	5
Auto					
Train					
Bus					
Car					
Two wheeler					
Other (Specify) _____					

5. If given an opportunity, would you like to switch your mode to scooty?
   (a) Yes
   (b) No
   (c) Maybe
   (d) Maybe not

   If your answer to the above question is NO, please answer from question 12.
6. Rate the following on the basis of your preference of using scooty.(Use √ against number) (1-least preferred, 5-most preferred)

Preference	*1*	*2*	*3*	*4*	*5*
Travel to work					
To be reliable					
To be independent					
Easy to park					
Value for money					
Other (Specify) _____					

Rate following statements 1-Fully disagree, 5-Fully agree

	1	*2*	*3*	*4*	*5*
7. You seek brand name if purchasing a scooty					
8. Colour have preference if purchasing a scooty					
9. You are satisfied with the average current mileage of a scooty					
10. The accessories available for riding safely are sufficient					

11. Rate following on the basis of initial attraction you seek in a scooty.(Use √ against number) (1-least preferred, 5-most preferred)

Initial Attraction	*1*	*2*	*3*	*4*	*5*
Colour					
Design					
Seat Shape					
Other (Specify) _____					

12. What add-on feature would you want in your scooty?
    (a) Extra storage
    (b) Foot rest distance
    (c) Mobile charging
    (d) Other if (Please specify)
13. What do you think about the salient features available to make riding safer?
    (a) Excellent
    (b) Good
    (c) Average
    (d) Bad

14. Rate the following on the basis of your preference of two-wheeler. (Use √ against number) (1- least preferred, 5-most preferred)

Initial Attraction	*1*	*2*	*3*	*4*	*5*
Small and Compact					
Low Maintenance Cost					
Affordable Price					
Other (Specify) _____					

15. How much are you ready to spend on a two-wheeler?
    (a) 15,000–25,000
    (b) 25,000–35,000
    (c) 35,000–50,000
    (d) above 50,000

SAMPLE QUESTIONNAIRE 3

Mobile Users Web-based Questionnaire

Telecom Survey

1. Age _____
2. Gender
   - Male
   - Female
3. Relationship status
   - Single
   - Committed
   - Married
4. Education Qualification
   - Under Graduate
   - Graduate
   - Post Graduate
5. Current Occupation
   - Student
   - Salaried
   - Self-employed
   - Professional
   - Retired
   - Home-maker
   - Others (Please Specify) _____
6. Salary/Income/Pension/Pocket money per month (in Rupees)
   - Less than Rs. 2000
   - Between 2001–5000
   - Between 5001–10000
   - Between 10001–20000
   - Between 20000–50000
   - Above Rs. 50000
7. Since how many years are you using the mobile phone?
8. Which service provider are you using? (In case you are having more than 1 phone, select the most used one)
   - Vodafone
   - Airtel
   - MTNL (Dolphin)
   - MTNL (Trump)

- Reliance CDMA
- Reliance GSM
- TATA CDMA
- Spice
- Virgin
- Idea
- BPL
- Aircel
- BSNL

Others (Please Specify) _____

9. Have you ever changed your service provider?
   - Yes
   - No
10. What type of connection do you use?
   - Post paid
   - Prepaid
   - Lifetime Prepaid
11. What is your average monthly cell phone expenditure? _____
12. How do you pay your phone bills?
   - Cash
   - Cheque
   - Credit Card
   - Other (Please Specify) _____
13. Who pays for your bills?
   - Self
   - Parents
   - Company
   - Children
   - Other (Please Specify)
14. Do you use a corporate plan?
   - Yes
   - No
15. Rate the following services on the basis of your preference (1 – Least preferred; 5 – Most preferred)

	1-Least Preferred	*2*	*3*	*4*	*5-Most Preferred*
SMS	1	2	3	4	5
Ring tones/Caller tunes	1	2	3	4	5
Alerts	1	2	3	4	5
Downloads	1	2	3	4	5
Internet	1	2	3	4	5

16. Rate the following depending on whom you call the most?
1-Least called 2 3 4 5-Most called

	1-Least	*2*	*3*	*4*	*5-Most*
Family	1	2	3	4	5
Friends	1	2	3	4	5
Work related	1	2	3	4	5
One Specific number	1	2	3	4	5
Making Enquiries	1	2	3	4	5

17. Major part of your bill is?
- Local
- STD
- ISD
- Other (Please Specify)

18. What is the call rate you are currently charged for local calls?
19. What is the call rate you are currently charged for STD calls?
20. Rate your current service provider
Poor Below Average Average Good Excellent Not Applicable

	Not Applicable	*1-poor*	*2*	*3*	*4*	*5- Excellent*
Call Rates	0	1	2	3	4	5
Tariff Plans	0	1	2	3	4	5
Network Coverage	0	1	2	3	4	5
Voice clarity	0	1	2	3	4	5
Roaming Facility	0	1	2	3	4	5
Value-added Services	0	1	2	3	4	5
Easy mode of payment/Easy availability of recharge	0	1	2	3	4	5
Customer Services	0	1	2	3	4	5

21. What factors do you consider while selecting a service provider? Please rank them (1-Least considered, 5-Most considered; 0 – if NA)

Parameter: Price

1-Least Considered 2 3 4 5-Most Considered 0-Not Applicable

	Not Applicable	*1-Poor*	*2*	*3*	*4*	*5- Excellent*
Local Calling Rates	0	1	2	3	4	5
STD Calling Rates	0	1	2	3	4	5
ISD Calling Rates	0	1	2	3	4	5
SMS Charges	0	1	2	3	4	5
Value-added Charges	0	1	2	3	4	5
Roaming Charges	0	1	2	3	4	5
Credit Limits	0	1	2	3	4	5
Payment options available	0	1	2	3	4	5

22. What factors do you consider while selecting a service provider? Please rank them (1- Least considered; 5- Most considered; 0 – if NA); PARAMETER : Connectivity/Performance
1-Least Considered 2 3 4 5-Most Considered 0-Not applicable

	Not Applicable	*1-Poor*	*2*	*3*	*4*	*5-Excellent*
Least call drops (Lesser the call drops, better the coverage)	0	1	2	3	4	5
Voice Clarity	0	1	2	3	4	5
Network Coverage	0	1	2	3	4	5
Roaming Network	0	1	2	3	4	5
Down Time	0	1	2	3	4	5

23. What factors do you consider while selecting a service provider? Please rank them (1-Least considered; 5-Most considered; 0 – if NA):
PARAMETER: Value-added Services (VAS)
1-Least Considered 2 3 4 5-Most Considered 0-Not applicable

	Not Applicable	*1-Poor*	*2*	*3*	*4*	*5-Excellent*
SMS	0	1	2	3	4	5
MMS	0	1	2	3	4	5
Voice SMS	0	1	2	3	4	5
Ring Tones	0	1	2	3	4	5
Caller Tunes	0	1	2	3	4	5
Internet	0	1	2	3	4	5
Voice Conferencing	0	1	2	3	4	5

24. What factors do you consider while selecting a service provider? Please rank them (1-Least considered; 5-Most considered; 0 – if NA)
PARAMETER : Customer Care

	Not Applicable	*1-Poor*	*2*	*3*	*4*	*5- Excellent*
Bill Payment Ease	0	1	2	3	4	5
Call Centre Efficiency	0	1	2	3	4	5
Recharge Availability	0	1	2	3	4	5
Branch Availability	0	1	2	3	4	5

25. What factors do you consider while selecting a service provider? Please rank them (1-Least considered; 5-Most considered; 0 – if NA) PARAMETER: Other Reasons
1-Least Considered 2 3 4 5-Most Considered 0-Not applicable

	Not Applicable	*1-Poor*	*2*	*3*	*4*	*5-Excellent*
Advertisements	0	1	2	3	4	5
Peer Effect	0	1	2	3	4	5
Prior Experience	0	1	2	3	4	5
Word of mouth publicity	0	1	2	3	4	5

26. Do you know about mobile number portability?
    - Yes
    - No
    - Comments/Others (Please Specify)
27. If given an option of changing your service provider while retaining your current number, will you do it?
    - Yes
    - No
    - Can't Say
28. Which service provider would you change to and why?
29. Any new facility you want to be introduced by service providers/cell phone companies?

SUMMARY

Primary data is collected directly by the researcher for some specific purpose or study. Secondary data is the data disseminated through some media like reports, newspapers, etc. The next stage after collection of data is the preparation of data. Various steps involved are editing, coding and validating.

DISCUSSION QUESTIONS

1. Describe the questionnaire modes of responses with illustration for each of the modes.
2. Describe the various issues to be considered for designing a questionnaire.
3. Describe the activities for preparation of data with illustrative examples.
4. Evaluate the three questionnaires critically, and offer suggestions for improvement in the same.
5. Write short notes on:
   (i) Focus Groups
   (ii) Projective Techniques
   (iii) Sources of Secondary Data
   (iv) General Guidelines for a Questionnaire Relating to Collection of Data

Presentation of Data

8

Contents

LEARNING OBJECTIVES

The main objective of this chapter is to provide the methodology and scope of various modes of presentation of data. This is intended to help in making effective presentation of data and conclusions relating to analysis and evaluation for an assignment or a project. Some of the charts and graphs discussed are:

- Bar Chart
- Pareto Chart
- Pie Chart
- Histogram
- Line Graph

Two distinguishing features of the chapter are:

- Use of live data to describe various charts and graphs and thus help in using the same for any live data in the work environment.
- Highlight use of Graphs as a management tool through a live case study so as to facilitate using the same or innovating another way to improve effectiveness of managerial functions.

8.1 RELEVANCE AND INTRODUCTION

It is said that a picture is equal to a thousand words. The same is true about graphs and charts that are used to present data in a form which can be easily comprehended. **About graphs and charts, it**

is said that instead of our looking at them they look at us! If fact, they reflect our performance. There is sufficient scope for making effective use of graphs and charts for managerial functions. We have illustrated this through a case study in Section 3.3 of this chapter.

Various aspects relating to classification and presentation of the data are described in this chapter with the help of live data taken from newspapers and magazines. Incidentally, this is referred to as **Exploratory Analysis** which comprises easy to construct tables and diagrams that summarises and describe the data.

8.2 CLASSIFICATION AND TABULATION

8.2.1 Ungrouped (Raw) Data

The following data gives value of equity holdings of 20 of the India's billionaires.

Name	*Equity Holdings (Millions of Rs.)*
Adi Godrej	5561
Ajay G. Piramal	4923
B. Ramalinga Raju, Rama Raju & Family	3862
Habil F. Khorakiwala	4187
K.K. Birla	3534
Karsanbhai K. Patel & Family	3144
Keshub Mahindra	4506
Kiran Mazumdar-Shaw	2717
M.V. Subbiah & Family	4784
Naresh Goyal (through Tail Winds)	6874
S.P. Hinduja	5071
Sashi Ruia	3527
Subhash Chandra	5424
The Kirloskar Family	4745
The Murthy Family	4310
The Nilekani Family	2796
The Punj Family	3098
Uday Kotak	5034
Vijay Mallya	6505
V.N. Dhoot	6707

Source: BUSINESS WORLD Special Issue of 4th September 2006.

The above equity holdings are tabulated below in Table 3.1.

Table 3.1 Equity Holdings of 20 Billionaires

(Millions of Rs.)

5561	4923	3862	4187	3534
3144	4506	2717	4784	6874
5071	3527	5424	4745	4310
2796	3098	5034	6707	6505

The above data will be referred to as 'Billionaires Data' in subsequent sections and chapters in this book.

8.2.2 Arrangement of Data

The above data could be arranged in many ways to get a better idea about the equity holdings of the group. For example, the data could be arranged in ascending order of equity holdings as follows:

(Millions of Rs)

2717	2796	3098	3144	3527
3534	3862	4186	4310	4506
4745	4784	4923	5034	5071
5424	5561	6505	6707	6874

It may be observed that the data arranged in ascending order is more meaningful and conveys more information than the raw data. For example, one can immediately comprehend that:

(i) The minimum equity holdings of the billionaires is Rs. 2,717 (Millions) and the maximum is Rs 6,874 (Millions)

(ii) The number of billionaires holding equity less than 3,000 are only two

(iii) The number of billionaires holding equity more than 6,000 are only three, and so on.

Similarly, the data can be arranged in descending order, and similar type of conclusions drawn.

Stem and Leaf Diagram A stem and leaf diagram presents a visual summary of a data. The diagram provides sorting of the data and helps in detecting the distributional pattern of the data.

Stem and leaf diagram for the billionaires' equity data is given below. While the stem units 2, 3, 4, 5 and 6 represents '000 in data values, the leaf values on the right side indicate '00 in data values. Thus, e.g., 2,796 comprises of 2 units as stem and 8 units (796 is approximated as 800) as leaf. Similarly, the last value, 6,874 has 6 as stem value and 9 (874 is approximated by 9) as leaf value.

Stem-and-Leaf Display

Stem unit: 1000

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

8.2.3 Classification and Tabulation

Classification is the first step in Tabulation. Classification implies bringing together the items which are similar in some respect(s). For example, students of a class may be grouped together with respect to their marks obtained in an examination, their age or area of specialisation, etc.

After classification, tabulation is done to condense the data in a compact form which can be easily comprehended.

Specific advantages or objectives of Tabulation are that it:

- Summarises data into rows and columns.
- Gives appropriate classification with number of data items into cells (intersection of rows and columns), subtotals of rows and columns, etc. This help in drawing useful interpretations about the data.
- Provides significant features of data including comparisons that are revealed.

For example, the following tabulation of marks of 200 students, classified at intervals of 20 marks each, reveals that the maximum number of students have obtained marks between 61 and 80.

Distribution of Marks among Students

Range of Marks	*Number of Students*
21–40	25
41–60	35
61–80	90
81–100	50
Total	200

If marks are also classified with respect to sex of students, the tabulation, given below, might reveal that the percentage of girls with marks more than 60% is more than percentage of boys with marks more than 60%, as

Distribution of Marks among Boy and Girl Students

Range of Marks	*Boys*	*Girls*
21–40	15	10
41–60	20	15
61–80	40	50
81–100	25	25
Total	100	100

Procedure for Classification and Tabulation The procedure for classification and tabulation of a data is illustrated in this section through the example of data on billionaires.

The first step in classification and tabulation is to group data into suitable number of class intervals. This has been done for the billionaires' data, as follows:

Class Intervals@	*Actual Observations*	*Tally Sheet**	*Number of Observations*
2000–3000	2717, 2796	||	2
3000–4000	3098, 3144, 3527, 3534, 3862	𝍸	5
4000–5000	4187, 4310, 4506, 4745, 4784, 4923	𝍸 |	6
5000–6000	5034, 5071, 5424	||||	4
6000–7000	6505, 6707, 6874	|||	3

@ Whenever the data is grouped in class intervals, some information is lost. In ungrouped data, we know the individual observations but, in grouped data we only know only the number of observations in an interval but not their individual values. However, this disadvantage is compensated by the advantage of comprehending the data in grouped form—and the ease of calculations if the number of observations is large. The number of class intervals is decided by considerations of both advantages and the disadvantages. If the number of intervals is large, the advantages get reduced, and if the number of intervals is small, the disadvantages increase. Therefore, one has to decide the number of intervals rather judiciously. In the above example, the number of intervals seem to be optimum. The intervals can be continuous or discontinuous. In continuous intervals, the upper value of the previous interval is the same as the lower point of the next interval—as in the above case. In such cases, it has to be stated clearly as to if an observation is exactly equal to the value of the lower/upper interval, in which group that observation is to be taken. For example, in the above case, it could be stated that when an observation is exactly equal to the upper value of one interval and lower value of the next interval, it will be included in the next interval. The only care to be taken, in such a case, is that no observation should be equal to the upper value of the last interval.

*One tally mark means one observation. Every fifth observation is represented by horizontal line-cutting through all the tally marks.

The width of the class intervals is generally, taken as equal, but it is not a must.

The intervals are sometimes open either for the lower-most interval or for the upper-most interval or for both the intervals. For example, in the case of the data relating to annual income of individuals, the lower interval could be less than Rs 50,000, and the upper class interval could be more than Rs 10 lakhs, as in the following table:

Annual Income of a Group of Individuals

Class Interval Frequency	*Number of Individuals*
Less than Rs 50,000	15
50,000–1,00,000	25
1,00,000–2,00,000	30
2,00,000–5,00,000	15
5,00,000–10,00,000	10
More than Rs 10,00,000	5
Total	100

The most commonly used presentation of grouped data is in the form of a frequency table given below for billionaire's data:

Frequency Table/Cumulative Frequency Table

Class Interval	*Frequency (Number of Observations)* (f_i)*	*Cumulative Frequency Interval (Number of Observations up to the class intervals)@ (fc_i)*
2000–3000	2	2
3000–4000	5	7
4000–5000	6	13
5000–6000	4	17
6000–7000	3	20

*The frequency is represented by small letter f with a subscript '*i*' indicating the frequency of the i^{th} interval. Thus: $f_1 = 2$, $f_2 = 5$, $f_3 = 6$, $f_4 = 4$, and $f_5 = 3$.

@ fc_i represents the cumulative frequency up to the i^{th} interval. Thus $fc_1 = 2$, $fc_2 = 7$, $fc_3 = 13$, $fc_4 = 17$, and $fc_5 = 20$. It may be noted that the cumulative frequency up to the last class interval is the total frequency, i.e. the number of observations.

8.3 DIAGRAMMATIC/GRAPHICAL PRESENTATION

There are several diagrams/graphs used for presentation of data. However, we shall discuss only the following diagrams/graphs that we consider as most appropriate, in consonance with the objective of this book:

- Bar Chart
- Pareto Chart

- Pie Chart
- Histogram
- Ogive
- Line Graph
- Lorenz Curve

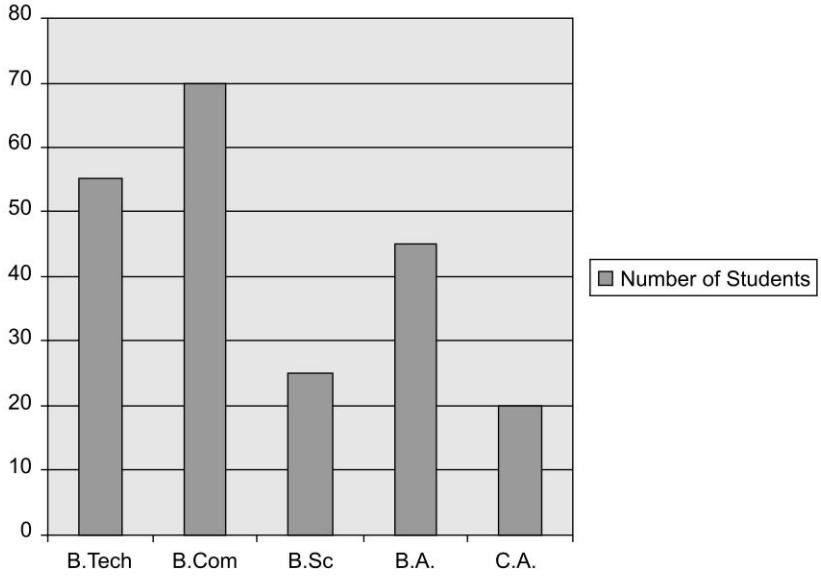
(i) Bar Chart

It comprises a series of bars of equal width—the base of the bars being equal to the width of the class interval of a grouped data. The bars 'stand' on a common base line, the heights of the bars being proportional to the frequency of the interval.

The following data gives the distribution of 215 MBA students at a management institute, according to educational qualifications

Educational Qualification	*Number of Students*
B.Tech	55
B.Com	70
B.Sc.	25
B.A.	45
C.A.	20

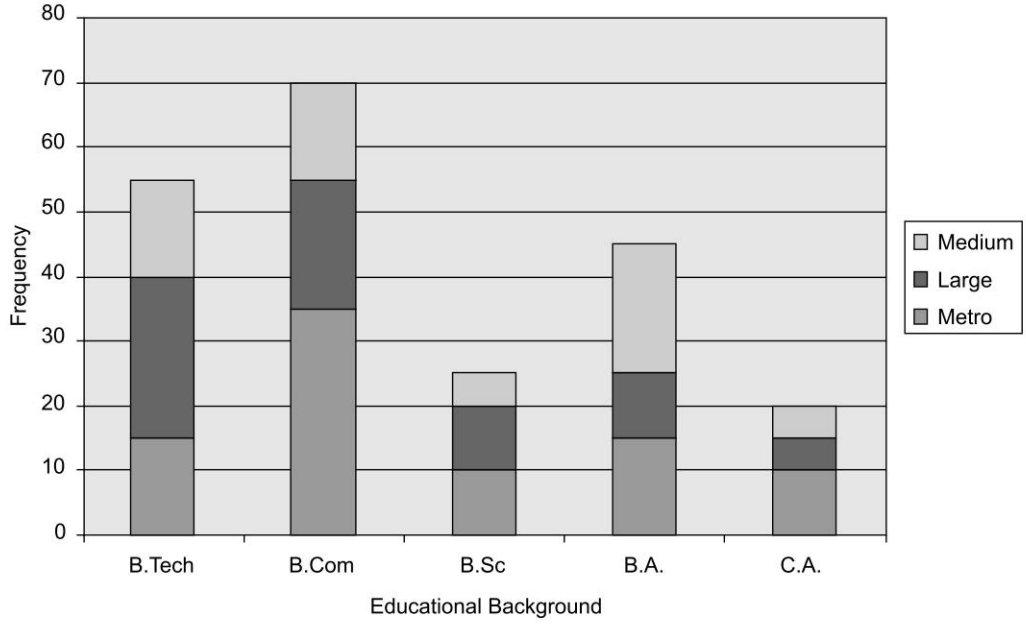
The bar chart for the above data is given below:



Subdivided Bar Chart A Subdivided Bar Chart is a bar chart wherein each bar is divided into further components. In the above example, if the information about the cities from where the students have graduated, is also available as given below.

Educational Qualification	*Metro*	*Large*	*Medium*	*Total*
B.Tech	15	25	15	55
B.Com	35	20	15	70
B.Sc.	10	10	5	25
B.A.	15	10	20	45
C.A.	10	5	5	20

The entire information is presented with the help of subdivided bar chart depicted as follows.



Multiple Bar Chart Multiple bar chart is one in which two or more bars are placed together for each entity. The bars are placed together to give comparative assessment of values of some parameter over two periods of time or at two different locations, etc. This chart is normally used when we wish to present visual comparison of two years' data for several entities, brands, etc.

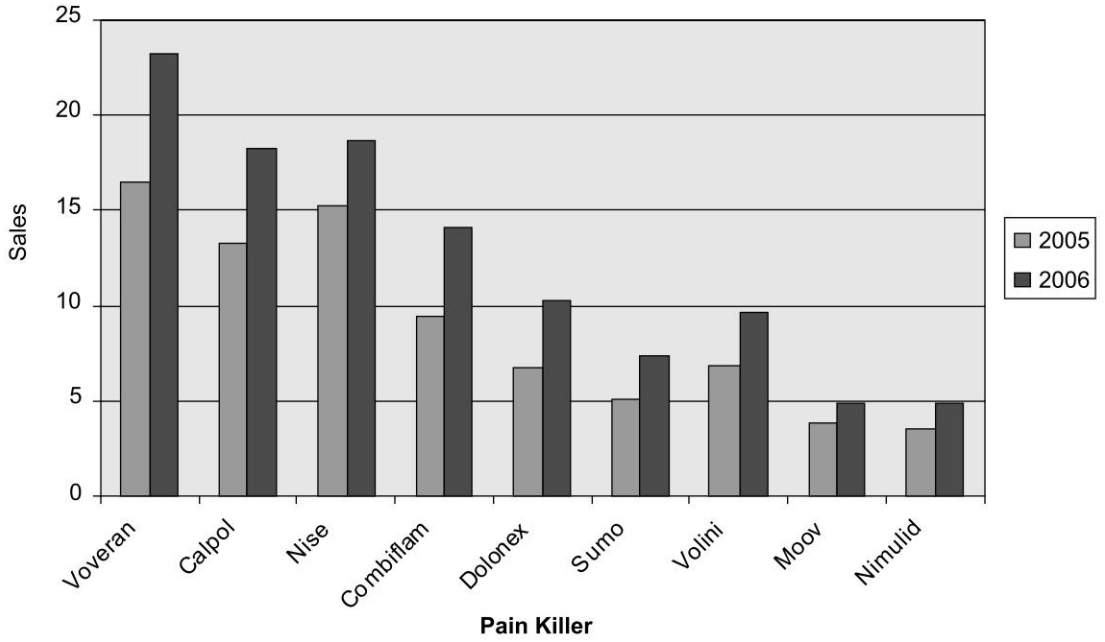
As an illustration of the Multiple Bar Chart, we consider the following data giving sales of top market brands among pain killers in India.

(Rs. in Crores)

Pain Killer	*2005*	*2006*
Voveran	16.5	23.2
Calpol	13.2	18.2
Nise	15.2	18.6
Combiflam	9.4	14.1
Dolonex	6.8	10.3
Sumo	5.1	7.4
Volini	6.9	9.6
Moov	3.8	4.9
Nimulid	3.5	4.9

Source: Economic Times dt. 9th October 2006.

The multiple bar chart for presentation of the above data is as follows:

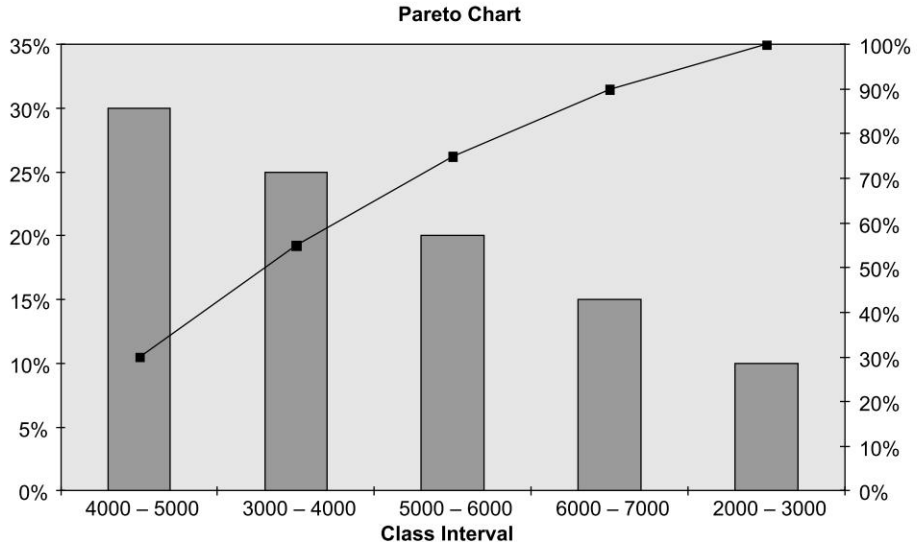


(ii) Pareto Chart

This specialised bar chart, named after the famous Italian economist, is used to classify a variable into groups or intervals from largest to smallest frequency.

It facilitates identification of the most frequent occurrence or causes of an event or phenomenon. It is used for sorting data by using any criteria like geographical regions, organisations like management institutes, banks, countries, cities, etc., and time zones like day/week/month, etc.

As an illustration, the Pareto's chart for the Billionaire's data is as follows:



(iii) Pie Chart

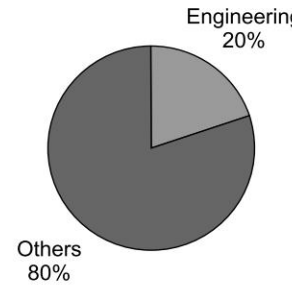
It is one of the most popular charts for presenting the 'whole' into parts. It is a circular chart divided into sectors representing relative magnitude of various components.

Pie Chart

When a question is asked as to why the pie chart is called by that term, a majority of persons who have some idea about the chart respond that it is something to do with π, because they feel that the chart is circular in shape, and the circumference of a circle is 2π r. They forget that the symbol π has the spelling "pi" while the chart is pie chart. The widely held belief is that since pie is a dish like cake, with circular shape, and it is cut into pieces which resemble the chart, it is called a pie chart. However, there is one more belief. Pie was the name of a cook in a royal palace in France. Instead of arranging different dishes in different plates, as we do at home or in parties, he used to place all the dishes in each and every plate so that one could pick up all the items from one plate instead of moving from plate to plate. He used to arrange the dishes on the plates just like the pie chart—voluminous items like chips getting more space than heavier items like biscuits. Because of this, it is believed that the chart is named after him.

The methodology of preparing the chart is explained below:

A pie chart is obtained by dividing a circle into sectors such that these sectors have areas or centre angles proportional to different components given in the data. The total angle at the center is 360. For example, suppose 20% of the MBA students in a management institute have engineering background, the angle at the cent$\frac{20}{100}$e will be $\times 360° = 72°$, as shown in the pie chart as follows.



This chart is generally used when the emphasis is on visual presentation of data for easy comprehension.

For example, the following data, giving the sources of funds in Government of India's budget for the year 2007-08, can be presented in the form of a pie chart, indicating the sources of funds.

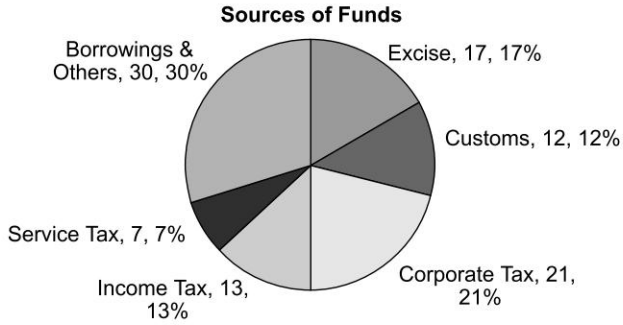
Government of India's Budget 2007-08

Sources of Funds	*Percentage of Total*	*Uses of Funds*	*Percentage of Total*
Excise	17	Central Plan	20
Customs	12	Non-plan Assistance and Expenditure	23
Corporate Tax	21	Defense	12
Income Tax	13	Interest Payments	20
Service Tax	7	States' Share	18
Borrowings & Others	30	Subsidies	7
Total	100		100

The above information relating to sources of funds is presented in a tabular form as follows:

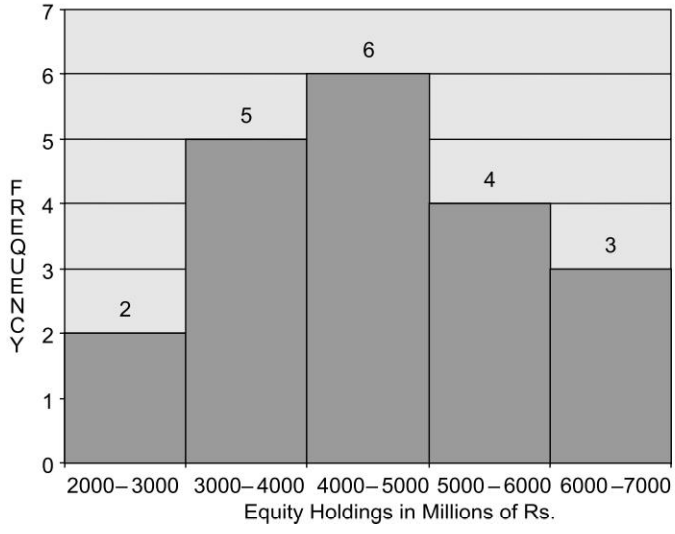
Sources of Funds	*Percentage of Total*	** Size of Segment (Degrees)*
Excise	17	61.2
Customs	12	43.2
Corporate Tax	21	75.6
Income Tax	13	46.8
Service Tax	7	25.2
Borrowings & Others	30	108
Total	100	360

*Size of segment is derived by dividing the percentage value by 100 and then multiplying by 360°. For example, for Income Tax, the segment = (13/100) × 360° = 46.8°. The pie chart for the above data is presented below:

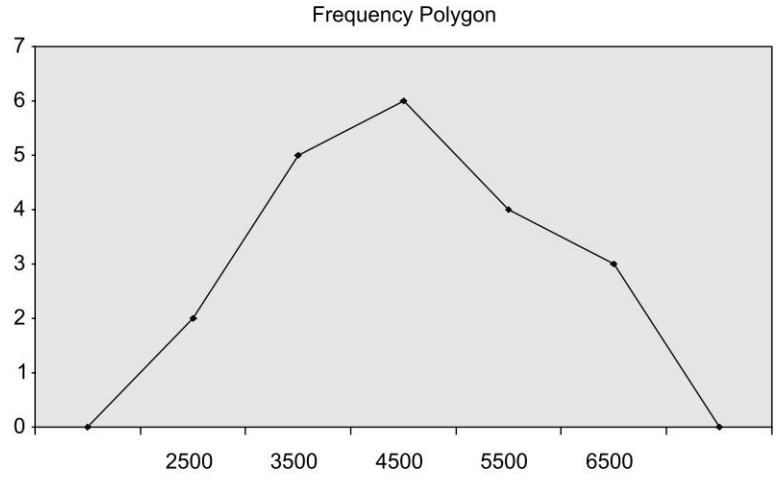


(iv) Histogram/Frequency Polygon

A histogram comprises of vertical rectangles whose base is proportional to the class interval and height is proportional to the frequency of an interval. For the billionaire's data, it is depicted below:

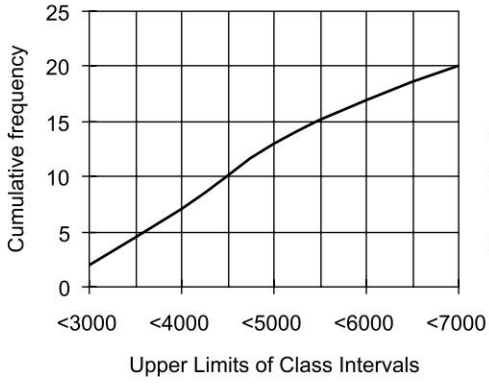


The polygon formed by joining the middle points of the above rectangles is known as the frequency polygon. It gives an idea of the shape of the distribution of frequencies.

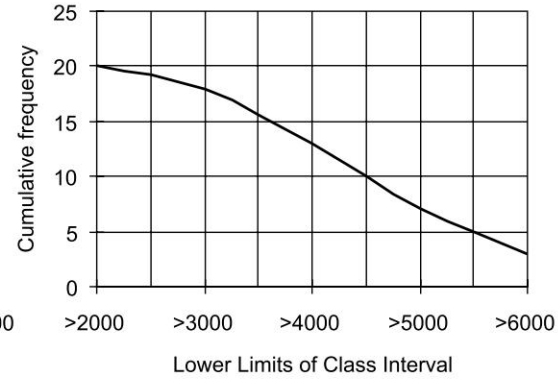


(v) Ogive

An ogive graph gives an idea about the number of observations less or greater than the values in the range of the variable. Accordingly, there are two types of ogives, viz. **'Less Than'** and **'Greater Than'**.

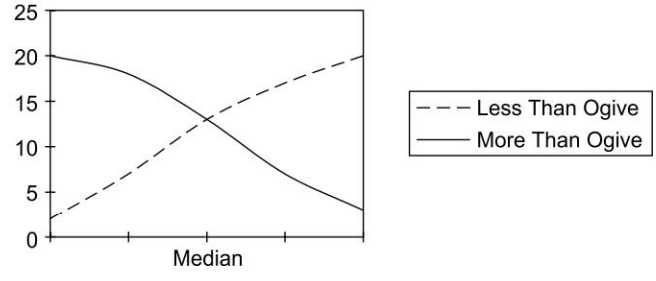


Less than Ogive



More than Ogive

If we combine both the ogives in the same diagram, the two together would like an arch. This shape is very popular in old royal structures from where the name ogive is derived.



The point of intersection of the two ogives is the median, as shown below. It is described, in detail, in Chapter 9. The

most important characteristic of median is that it divides the data into two equal parts such that 50% of the observations have value less than this, and the other 50% have values more than this.

(vi) Line Graph

A line graph is a visual presentation of a set of data values joined by straight lines. The data values could be over a period of time or over a set of entities like banks, individuals, etc.

The following graph relates to the productivity of staff i.e. 'Business per Employee' for some of the public sector banks.

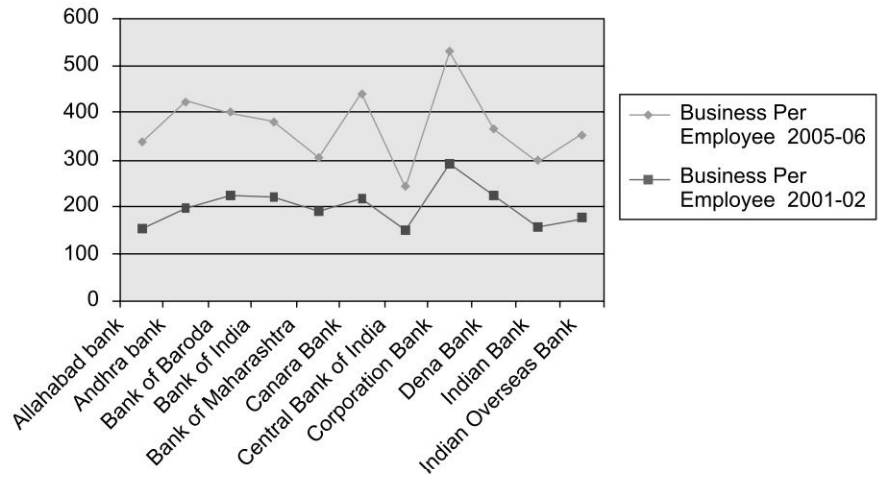
Business per Employee in Some Public Sector Banks

(Rs in Lakhs)

Bank	*Business Per Employee 2005-06*	*Business Per Employee 2001-02*
Allahabad Bank	336	153
Andhra Bank	426.75	195.96
Bank of Baroda	396	222.76
Bank of India	381	218.74
Bank of Maharashtra	306.18	191.44
Canara Bank	441.57	214.88
Central Bank of India	240.46	148.77
Corporation Bank	527	290.44
Dena Bank	364	221
Indian Bank	295	156
Indian Overseas Bank	354.73	175.41

Source: Economic Times dt. 23rd October 2006.

Business per Employee in Banks



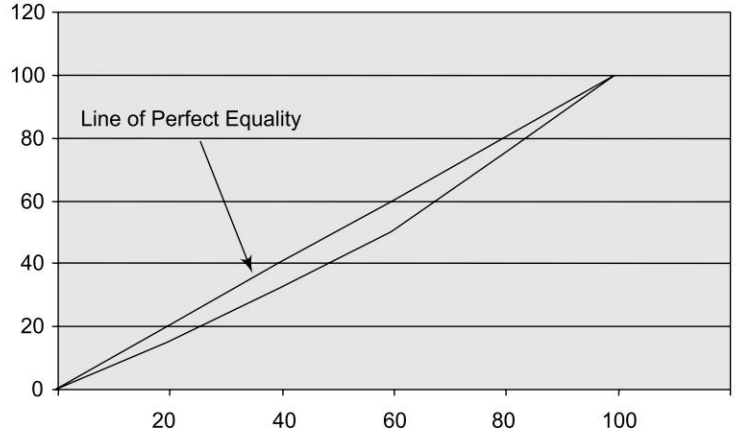
(vii) Lorenz Curve

Inequality is a hallmark of any financial system. The Lorenz Curve is a graph plotting the cumulative distribution of the amount of the variable concerned against the cumulative frequency distribution of the individuals possessing the amount.

The following Table has been derived from the data about 20 Billionaires given earlier in Section 8.2.

Cumulative Percentage of Billionaires	*Cumulative Equity Holding*	*% of Total Equity Holding*
20	16420	18
40	33423	37
60	52135	57
80	71330	78
100	91308	100

In the above data if the equity holding was equally distributed, 20% of equity would have been held by 20% of billionaires, 40% of equity would have been held by 40% of billionaires, and so on. However, it is not so in the above data because of inequality in distribution of equity holdings. The extent of this inequality is shown in the Lorenz curve depicted in the graph given below. The gap between the line of perfect equality and the line of actual equity holding can be shaded to indicate departure from equality. Thus shaded area would indicate the extent of departure of equality in the data from perfect equality; more the area more is the inequality.



8.3 USE OF GRAPHS AS A MANAGEMENT TOOL: A CASE STUDY

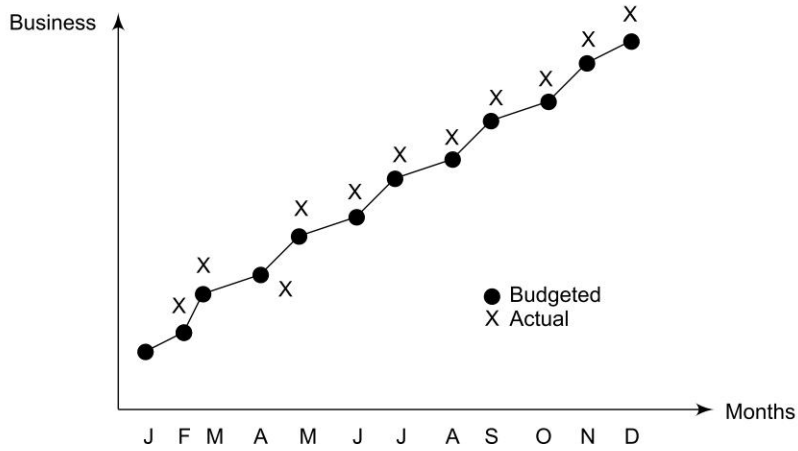
Late Dr. Y.B., Damle, ex-executive Director, Reserve bank of India, was once a senior manager in a public sector bank. Being in the Management Information System Department, he used to provide the requisite information to the Chairman. Dr. Damle developed the system of providing information on small cards which could be easily kept in a purse—just like credit cards. The Chairman found the 'information purse' highly useful while attending top level meetings and duly appreciated the services of Dr. Damle.

After sometime, Dr. Damle was posted as a branch manager. He looked forward to demonstrate the use of information systems even at the branch level.

There were 10 officers at the branch. He made each officer responsible for the different types of business, comprising deposits, advances and other miscellaneous business. He asked each one of them to prepare a graph depicting cumulative business at the end of every month for their slice of business during the next year. Of course, he provided requisite guidance to them in facilitating finalisation of their individual goals. Dr. Damle, then finalised the budget with his regional manager in the presence of all the officers.

Thereafter, he asked each of the officers to make a big size graph showing growth of business each month, and pin it on the board in his cabin. Thus, there were 10 graphs in his cabin within the first week of the first month of the new financial year.

Dr. Damle instructed that after the end of the month, each officer should plot the actual business value vis-à-vis the budgeted value as shown in the concerned graph. Dr. Damle felt that no officer would like to plot actual value less than the budgeted value—specially since the budget was decided by the officer himself. And, it really happened that way. As a result, the branch showed a phenomenal growth in its business. When the news reached the chairman, he was surprised, and visited the branch to appreciate the approach and efforts of Dr. Damle. He also gave a letter of appreciation to Dr. Damle, and promoted him as regional manager. Dr. Damle was happy that he had proved his conviction that the combination of psychology and graphs could do wonders. He continued to adopt the same strategy with the branch managers of his region. This approach which *inter-alia* included use of graph as a management tool did wonders, and soon he was picked up as a top level executive in another organisation.



Important developments having impact on business were recorded under the date / week / month when the development took place. Thus while looking at the graph, one could make out the impact of the development, and other factors including seasonality, if any.

Normally, graphs are not used as an integral part of a Management Information System. However, these can be used very effectively as planning and monitoring tools for effective management of a system.

Growth and fluctuations in the volume of any business activity like sales, profit, etc. or any other parameter of environment like prices of commodities and stocks, etc. do get reflected in the graphs.

As mentioned earlier, it is said that instead of our looking at them, they 'look' at us. They speak the truth. They reflect performance of the management or the system and provide motivation to improve or take corrective action, if necessary.

8.4 USING EXCEL

A data could be presented through charts and graphs easily with the help of MS Excel. The methodology of ungrouped (raw) data with frequency distribution and various charts and graphs, are described in enclosed CD.

Basic Analysis of Data

9

Contents

LEARNING OBJECTIVES

This chapter provides an understanding of relevance and need for calculation of:

- Various Measures of Location such as average or mean, median, mode, etc. as also their relative advantages and limitations.
- Various Measures of Variation or Dispersion such as range, standard deviation, mean deviation, etc., as also their relative advantages and limitations.
- Various measures of uniformity, consistency, disparity, volatility, etc. For example, an investor, in addition to expected return from an investment, may also like to assess the volatility or risk in returns.
- Measure of symmetry/skewness in data. In real life not all data are symmetrical with reference to their central value, and it is worthwhile to have a measure of the deviation from symmetry for comparing two or more sets of data.

Relevance

The 'Reliable' company had two plants to manufacture a type of bearing. Because of the competition and the criticality of the item, the company decided to pay utmost attention to ensure that the quality was as per the specification. However, due to staff problems, the company started noticing high rate of rejection of the item from the users. On analysing the quality of bearings, the company noted that the bearings from the two plants were getting rejected because of two

different reasons. While, the bearings from one plant were getting rejected because they were not measuring to the specified dimension – in fact it was lesser than specified, the bearings from other factory were getting rejected because of variation in the dimension – some were less than the specification and some were more than the specification. When the company referred the problem to a consultant, he pointed out that it was not only necessary to ensure quality in terms of specified average of dimension but also in terms of variation in the dimension from bearing to bearing. Further investigations revealed that in the first plant, the problems were mostly on account of the technical staff while in the second plant, it was mostly due to unrest among the operators. The problems were sorted out accordingly, and the company was able to retrieve the loss of image it had suffered.

9.1 SIGNIFICANCE AND INTRODUCTION

Whenever, we talk or discuss about any data or characteristic of a group of people, companies, nations, etc., we tend to summarise the data with one number. For example, while pointing to the decrease in age of powerful persons in India, the following statement was published by *India Today* in it's issue of 20th March 2006.

"Average age of top 50 powerful persons of 2006 in India decreased from 58 years in 2003 to 54 years in 2006."

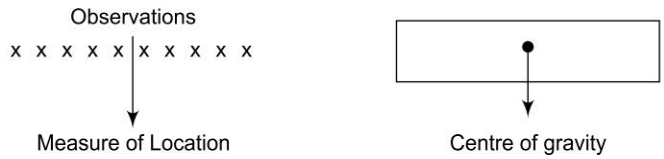
In this statement, the ages of powerful persons are summarised by one number i.e. 58 years in 2003, and 54 years in 2006. The simplest type of statistical analysis of a data containing a set of observations such as the equity holdings data given in Illustration 9.1, is the calculation of a single value which could be taken as representative of the entire data. For example, what is the value which could represent the equity holdings of all the 20 billionaires? There are several measures for arriving at this value, and are known as measures of central tendency or location.

Another type of statistical analysis involves measurement of dispersion or variation which indicates the extent to which the observations differ from each other. In the context of equity holdings data, this measure will give an idea of the variation or disparity in the equity holdings of the billionaires. Given this measure for two sets of data, one could compare their variability.

These measures as also some other related measures are described in this chapter.

9.2 MEASURES OF LOCATION OR CENTRAL TENDENCY

These measures indicate a value, which all the observations tend to have, or a value where all the observations can be assumed to be **located** or concentrated. The concept is similar to the centre of gravity where the entire mass of an item is assumed to be located or concentrated at a point, as shown in the diagram given below:



There are three such measures:

(i) Mean
(ii) Median, Quartiles, Percentiles
(iii) Mode

Before discussing these measures in the subsequent sections, we may mention some of the desirable properties of such a measure. This would help us to appreciate the usefulness and make a comparative evaluation of these measures.

- It should be easy to understand and calculate
- It should be based on all observations
- It should not be much affected by a few extreme observations
- It should be amenable to mathematical treatment. For example, we should be able to calculate the combined measure for two sets of observations given the measure for each of the two sets

9.2.1 Mean

There are three types of means viz.,

- Arithmetic Mean
- Harmonic Mean
- Geometric Mean

These are described below along with their relative advantages and limitations:

Arithmetic Mean

The arithmetic mean is defined for ungrouped data as well as for grouped data. We shall discuss both, separately.

(a) Ungrouped (Raw) Data

For ungrouped data, the arithmetic mean is defined as follows:

Arithmetic Mean (A.M.) of a set of n values, say, $x_1, x_2, x_3, \ldots. x_i, \ldots. x_n$, is defined a

$$\bar{x} = \frac{\text{Sum of Observations}}{\text{Number of Observations}} \tag{9.1}$$

$$= \frac{x_1 + x_2 + \cdots x_i + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Illustration 9.1

The following data, given in Section 8.2.1, gives value of equity holdings of 20 of the India's billionaires, in ascending order.

Name	*Equity Holdings (Millions of Rs)*
Kiran Mazumdar-Shaw	2717
The Nilekani family	2796
The Punj family	3098
Karsanbhai K. Patel & family	3144

(*Contd*)

(*Contd*)

Shashi Ruia	3527
K. K. Birla	3534
B. Ramalinga Raju, Rama Raju & Family	3862
Habil F. Khorakiwala	4187
The Murthy family	4310
Keshub Mahindra	4506
The Kirloskar family	4745
M.V. Subbiah & family	4784
Ajay G. Piramal	4923
Uday Kotak	5034
S.P. Hinduja	5071
Subhash Chandra	5424
Adi Godrej	5561
Vijay Mallya	6505
V.N. Dhoot	6707
Naresh Goyal	6874

The above equity holdings are tabulated below:

Table 9.1 Equity Holdings of 20 Indian Billionaires

(Rs in Millions)

2717	2796	3098	3144	3527
3534	3862	4187	4310	4506
4745	4784	4923	5034	5071
5424	5561	6505	6707	6874

For the above data, the A.M. is

$$\bar{x} = \frac{2717 + 2796 + \cdots + 3534 + \cdots + 4506 + 4745 + \cdots + 5424 + \cdots + 6874}{20}$$

= Rs **4565.4** Millions

The interpretation of this value is that the equity holdings of all the 20 billionaires could be considered to be concentrated or located at this value. In fact, if all the twenty billionaires are to be represented by only one value, then that value is Rs 4565.4 Millions. It may be noted that there is no billionaire with the equity holdings as Rs 4565.4 Millions but nevertheless this value is the representative of all the values. In fact, if this value is multiplied by the number of billionaires, viz. 20, we get the product as 91308 Millions which is the total equity holdings of the 20 billionaires.

(b) Grouped Data

When the data is grouped, and the following type of frequency table is prepared

Class Interval	*Mid-point of Class Interval* (x_i)	*Frequency* (f_i)
–	x_1	f_1
–	–	–
–	x_k	f_k

then

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \qquad \textbf{(9.2)}$$

where x_i is the middle point of the ith class interval, f_i is the frequency of the ith class interval, $f_i x_i$ is the product of f_i and x_i, and k is the number of class intervals.

The above formula is justified as follows:

The basic definition of the mean remains the same i.e.

$$\bar{x} = \frac{\text{Sum of Observations}}{\text{Number of Observation}}$$

To find the sum of all the observations, all we know is that f_i observations are in the i^{th} class interval. Since individual observations are not available, we assume that all the f_i observations are equal to the middle point (x_i) of the i^{th} class interval. Thus the total of f_i observations in the i^{th} class interval is equal to $f_i x_i$. It is well appreciated that all the f_i observations cannot be equal to x_i, but it is expected that some observations will be less than x_i and some observations will be more than x_i . Thus, on the average, positive and negative errors will cancel out each other. Further, this is the best that can be assumed when the individual observations are not known. Thus, the sum of all the observations in the data is to take the summation of all products $f_i x_i$ i.e. $\Sigma f_i x_i$. To get the A.M., this sum is divided by the total of the frequencies of all class intervals viz. Σf_i.

The following illustration explains calculation of arithmetic mean from a grouped data.

Illustration 9.2

The calculation is illustrated with the data relating to equity holdings of the group of 20 billionaires given in Illustration 9.1.

*Class Interval** *(1)*	*Frequency* (f_i) *(2)*	*Mid Value of Class Interval* (x_i) *(3)*	$f_i x_i$ *Col.(4) = Col.(2) × Col.(3)*
2000 – 3000	2	2500	5000
3000 – 4000	5	3500	17500
4000 – 5000	6	4500	27000
5000 – 6000	4	5500	22000
6000 – 7000	3	6500	19500
Sum	$\Sigma f_i = 20$		$\Sigma f_i x_i = 91000$

*(Intervals include the upper class value but not the lower)

Substituting values of Σf_i and $\Sigma f_i x_i$, in formula $\frac{\sum_{i=1}^{5} f_i x_i}{\sum_{i=1}^{5} f_i}$, we get

$$\bar{x} = 91{,}000 \div 20$$
$$= \mathbf{4550}$$

It may be noted that the A.M. of the ungrouped data worked out as 4565.4 is not the same as the value obtained in the case of grouped data. In fact, it need not be so because while calculating A.M. from the grouped data, it is assumed that all the observations in a class interval have the same value viz. the middle point of the interval. **The value of A.M. obtained from the grouped data is only an approximation of the value obtained from ungrouped data.**

Combined A.M. of Two Sets of Data

Let there be two sets of data with

Number of observations = n_1 and n_2

A.Ms. = $\bar{x}_1$ and $\bar{x}_2$

If these two data are combined, the combined mean $\bar{x}$ is given by

$$\bar{x} = \frac{\text{Sum of observation in the two data}}{\text{Number of observations in the two data}}$$

$$= \frac{\text{Sum of observation in first data} + \text{Sum of observations in second data}}{n_1 + n_2}$$

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \qquad \textbf{(9.3)}$$

This formula for combined mean has five quantities viz. $\bar{x}, \bar{x}_1, \bar{x}_2$, n_1 and n_2. If four of these quantities are known, the fifth one can be found out.

Weighted Arithmetic Mean

In calculating the arithmetic mean, we assume that all the observations have equal importance or 'weight'. However, sometimes, it may not be so. For example, while calculating average price, the price of sugar may be more important to a consumer than the price of salt. Similarly, for admission to Science stream in Class XI, marks in Mathematics might be given higher weightage than marks in language, say English. As further explanation, suppose, the price of sugar has increased by 20% during a period and the price of rice has increased by 10% during the same period. Can we say that the average price rise during the period is 15% i.e. average of 10 and 20? The concept of weightage is explained below.

Illustration 9.3

Let a family's average monthly consumption of sugar and rice be 5 kg. and 20 kg., respectively. Further, assume that the price of sugar increased from Rs 15 to 18 per kg., and price of rice increased

from Rs 10 to 11 per kg. It may be intuitively noted that the impact of price rise in rice is more than the impact of price rise in sugar as the consumption of rice is more. Mathematically, it can be shown below. It may be noted that the price rise in sugar has been given a weight of 5, and price rise in rice has been given a weight of 20 based on their monthly consumption.

Item	*Monthly Consumption*	*Weight* (w_i)	*Rise in Price (Percentage)* (p_i)	$w_i p_i$
Sugar	5	5	20	100
Rice	20	20	10	200

Therefore, the average price rise could be evaluated as

$$\bar{p} = \frac{\sum w_i p_i}{\sum w_i} = \frac{100 + 200}{5 + 20} = \frac{300}{25} = 12.$$

Thus the average price rise is 12 %.

In general, if the values $x_1, x_2\ x_3, \ldots.\ x_i, \ldots.x_n$ have weights $w_1, w_2, w_3, \ldots., w_i, \ldots., w_n$ then the weighted mean of x is given as

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \qquad \textbf{(9.4)}$$

Advantages and Disadvantages of Arithmetic Mean

Advantages	*Disadvantages*
• Easy to understand and calculate.	• Unduly influenced by extreme values.
• Makes use of full data.	• Cannot be calculated if values of all observations are not known. For example, in an ungrouped data 5 , 8 , 10 , 15 , > 20 wherein, all we know about the fifth observation is that it is more than 20, but we do not know its actual value. Thus, it is not possible to find out sum of all the observations, and hence A.M. cannot be calculated.
• Only sum of values and their number need be known for determination.	• It cannot be calculated for a grouped data, if even one class interval is open-ended.
• The sum of observations can be calculated with its knowledge as also the number of observations. For example, if average of 5 observations is 10, then the sum of all the 5 observations is $5 \times 10 = 50$.	• It cannot be located or comprehended graphically.
• It is amenable to mathematical treatment. For example, given the means of two sets of data, we can find the combined mean for the data formed by combining the two sets of data. This property of A.M. forms the basis for most of statistical analysis.	• It is suitable mostly for data which are symmetrical or near symmetrical around some central value.
• It is least susceptible to change from sample to sample	

Geometric Mean

While calculating arithmetic mean, equal weight is attached to all the observations except while calculating weighted mean. This may, at times, lead to arithmetic mean not being a true representative of the observations. For example, the arithmetic mean of 4 observations 1, 2, 3 and 10 is 16/4 = 4 which is obviously not a true representative of the data. One of the measures of location that could be used in such a situation is the **Geometric Mean**.

The Geometric Mean (**G.M.**) of a series of observations with $x_1, x_2, x_3, \ldots, x_n$ is defined as the n^{th} root of the product of these values. Mathematically

$$\textbf{G.M.} = \{(x_1)\ (x_2)\ (x_3)\ \ldots\ (x_n)\}^{(1/n)} \qquad \textbf{(9.5)}$$

It may be noted that the G.M. cannot be defined if any value of *x* is *zero* as the whole product of various values becomes zero.

For example, the G.M of 0, 2, 4, 5, 9, and 10 is 0.

Further, G.M. is also not defined for negative values of observations.

If, it is found difficult to calculate the value of G.M. by the above formula, one may take logarithms on both sides to get

$$\log \text{G.M.} = \frac{1}{n}\{\log x_1 + \log x_2 + \log x_3 + \ldots + \log x_n)$$

$$= \frac{1}{n} \Sigma \log x_i$$

or

$$\textbf{G.M.} = \textbf{Antilog}\left\{\frac{\mathbf{1}}{\boldsymbol{n}} \boldsymbol{\Sigma} \textbf{log}\ \boldsymbol{x_i}\right\} \qquad \textbf{(9.6)}$$

This formula is, generally, found to be more convenient.

Illustration 9.4

For the data with values, 2, 4, and 8,

$$\text{G.M.} = (2 \times 4 \times 8)^{(1/3)}$$
$$= (64)^{1/3}$$
$$= \mathbf{4}$$

It can also be calculated by taking logarithm on both sides as illustrated below:

$$\log \text{G.M.} = (1/3) \log 64 = (1/3) \times (1.8062\ ) = 0.6021$$

Therefore, $\text{G.M.} = \text{antilog}\ (0.6021) = \mathbf{4}$

Average Rate of Growth of Production/Business or Increase in Prices:

If P_1 is the production in the first year and P_n is the production in the nth year, then the average rate of growth is given by $(G - 100)\%$ where,

$$G = 100\ (P_n/P_1)^{1/(n-1)} \qquad \textbf{(9.7)}$$

or,
$$\log G = \log 100 + \{1/(n-1)\}\ (\log P_n - \log P_1) \qquad \textbf{(9.8)}$$

The annual growth rate calculated with the help of G.M. is also called Compound Annual Growth Rate or Average Rate of Return on Investment.

Advantages and Disadvantages of Geometric Mean

Advantages	*Disadvantages*
• Makes use of full data. • Can be used to indicate rate of change. • Extreme values have lesser impact on the value of this mean as compared to arithmetic mean. • Specially useful for studying rate of growth.	• Cannot be determined if any value is zero. • More difficult to calculate and less easily understood • If any one value is negative, it cannot be calculated.

Harmonic Mean

The harmonic mean (H.M.) is defined as the reciprocal of the arithmetic mean of the reciprocals of the observations.

For example, if x_1 and x_2 are two observations, then the arithmetic means of their reciprocals viz. $1/x_1$ and $1/x_2$ is

$$\{(1/x_1) + (1/x_2)\}/2 = (x_2 + x_1)/2\ x_1x_2$$

The reciprocal of this arithmetic mean is $2\ x_1x_2/(x_2 + x_1)$. This is called the harmonic mean.

Thus the harmonic mean of two observations x_1 and x_2 is

$$\frac{2\,x_1\,x_2}{x_1 + x_2} \qquad \textbf{(9.9)}$$

This mean is used in averaging rates when the time factor is variable and the act being performed is the same.

The situation in example relating to average speed of car, discussed earlier in this Section, is of this type.

It may be noted that the distance traveled from *A* to *B* and *B* to *A* is the same but time taken to travel is different because of different speeds of 40 and 60 kms/hr. Therefore, the harmonic mean giving the average speed can be calculated as follows:

$$\text{Harmonic mean or Average Speed} = \frac{2}{1/40 + 1/60} = \frac{2 \times 40 \times 60}{40 + 60}$$

$$= \frac{4800}{100}$$

$$= 48 \text{ km/hr.}$$

9.2.2 Median

While the mean is an appropriate measure of location in most of the applications, there are situations that have extreme values either on lower side or on the higher side. For example, if the data comprises of values 2, 8, 9, 11, the mean works out to be 30/4 = 7.5 which may not be considered as a representative of the data as three out of four value are more than this value. Similarly, if the data comprises of values 8, 9, 11, 22, the mean works out to be 12.5 which again may not be considered as a good representative of the data.

Further, some times, exact values may not be available at either end of the range of values. For example, the land holding of 5 farmers could be:

less than (<) 5 acres, 10, 12,15, 20 acres

or, the annual income of five persons could be:

1.2 lakhs, 2.5, 4.0, 4.5, more than (>) 5 lakhs

In both the cases mentioned above, calculation of arithmetic mean is not possible because it is defined as equal to

$$\frac{\text{Sum of all the observations}}{\text{Number of observations}}$$

Thus, calculations of A.M. requires sum of all the observations which cannot be calculated because specific values of observations are not available.

Similarly, if there is a frequency distribution with open ended interval like

Class Interval	*Frequency*
<5	4
5 – 10	8
10 – 15	10
15 – 20	6
20 – 25	2

Or

Class Interval	*Frequency*
0 – 5	2
5 – 10	5
10 – 15	10
15 – 20	8
>20	3

then also A.M. cannot be calculated because we cannot find out the middle point of the 1st interval, in the first case, and middle point of the last interval, in the second case.

Thus, whenever there are some extreme values in the data, calculation of A.M. is not desirable. Further, whenever, exact values of some observations are not available, A.M. cannot be calculated. In both the situations, another measure of location called Median is used.

Median of a set of values is defined as the middle most value of a series of values arranged in ascending/descending order. In general, if there are n observations arranged in ascending or descending order, median is defined as the value corresponding to the $(n + 1)/2^{th}$ observation.

If the number of observations is odd, the value corresponding to the middle observation is the median. For example, if the observations are 2, 3, 5, 8 and 10, the median is the value corresponding to the 3rd observation i.e. 5.

If the number of observations is even then the average of the two middle most values is the median. For example, if the observations are 2, 3, 6, 8, 10 and 12, then the median is the value corresponding to the 3.5^{th} observation or the average of the 3rd and 4th observations viz. 6 and 8 i.e. 7.

The methodology of calculating the median for a given data is described below.

(a) Ungrouped Data

First the data is arranged in ascending/descending order.

In the earlier example relating to equity holdings data of 20 billionaires given in Table 9.1, the data is arranged as per ascending order as follows

2717 2796 3098 3144 3527 3534 3862 4187 4310 4506 4745 4784 4923
5034 5071 5424 5561 6505 6707 6874

Here, the number of observations is 20, and therefore there is no middle observation. However, the two middle most observations are 10^{th} and 11^{th}. The values are 12112 and 12388. Therefore, the median is their average.

$$\text{Median} = \frac{4506 + 4745}{2} = \frac{9251}{2}$$

$$= \mathbf{4625.5}$$

Thus, the median equity holdings of the 20 billionaires is Rs 4625.5 millions. It may be noted that 10 billionaires have equity holdings less than this value and 10 billionaires have equity holdings more than this value.

It may be recalled that the mean equity holdings calculated from the above data was Rs **4565.4** Millions

(b) Grouped Data

The median for the grouped data is also defined as the value corresponding to the $((n+1)/2)^{th}$ observation, and is calculated from the following formula:

$$\textbf{Median} = L_m + \frac{((n/2) - f_c)}{f_m} \times w_m \qquad \textbf{(9.10)}$$

where,

L_m is the lower limit of the median class internal i.e. the interval which contains $n/2^{th}$ observation

f_m is the frequency of the median class interval

i.e. the class interval which contains the $((n)/2)^{th}$ observation

f_c is the cumulative frequency up to the interval just before the median class-interval

w_m is the width of the median class-interval, and n is the number of total observations.

The justification for using the ratio $n/2$ in the numerator is explained with a numerical example below:

Consider the following data:

Class Interval	*Frequency*	*Cumulative Frequency*
0-10	4	4
10-20	5	9
20-30	3	12

We have to find out the value corresponding to the $(13/2)^{th}$ observation. From the above table, we conclude that the median – the value corresponding to the 6.5^{th} observation lies in the second interval. Since 4 observations are in the earlier viz. first interval, the median is the value corresponding to the $2{\cdot}5^{th}$ observation in the interval 10 to 20. Since, there are 5 observations in that interval, we can presume that, all the five observations are equally spaced at 11,13,15,17 and 19, as shown below

10 * 11 * 12 * 13 * **14** * 15 * 16 * 17 * 18 * 19 * 20

This implies that the 5^{th} observation is at 11, the sixth observation is at 13, and the 7^{th} observation is at 15, and so on.

Thus, the median corresponding to the 6.5^{th} observation is midway between 13 and 15 i.e. 14. This also amounts to saying that only two parts of the five parts (frequency of median interval being 5) are added to the lower point of the class interval. Thus, in grouped data, the median is really the value corresponding to the sixth (i.e.12/2) and not 6.5^{th} observation. This justifies the value (n/2) in the numerator of the formula for calculating the median from the grouped data.

The calculation of median for the above grouped data is illustrated below.

$$\text{Median} = L_m + \frac{((n/2) - f_c)}{f_m} \times w_m$$

$$= 10 + \frac{((12/2) - 4)}{5} \times 10$$

$$= 10 + \frac{(6-4)}{5} \times 10 = 10 + \frac{2}{5} \times 10 = \mathbf{14.}$$

We shall now calculate the median for the grouped equity holdings data given in Illustration 9.2.

The data is presented in tabular form as follows:

Class Interval	*Frequency*	*Cumulative frequency*
2000-3000	2	2
3000-4000	5	7
4000-5000	6	13
5000-6000	4	17
6000-70000	3	20

Here, $n = 20$, the median class interval is from 4000 to 5000 as the 10^{th} observation lies in this interval.

Further,

$$L_m = 4000$$
$$f_m = 6$$
$$f_c = 7$$
$$w_m = 1000$$

Therefore,

$$\text{Median} = 4000 + \frac{20/2 - 7 \times 1000}{6}$$

$$= 4000 + 3/6 \times 1000 = 4000 + 500 = \mathbf{4500}$$

It may be recalled that the A.M. calculated from the same grouped data is **4550.**

It may be added that the median divides the data into two parts such that the number of observations less than the median are equal to the number of observations more than it. This property makes median a very useful measure when the data is skewed like income distribution among persons/households, marks obtained in competitive examinations like that for admission to Engineering/Medical Colleges, etc.

Frequencies as Percentages:

Sometimes, in a grouped data, instead of the frequencies in class intervals, we are given percentage of total frequency in each class interval.

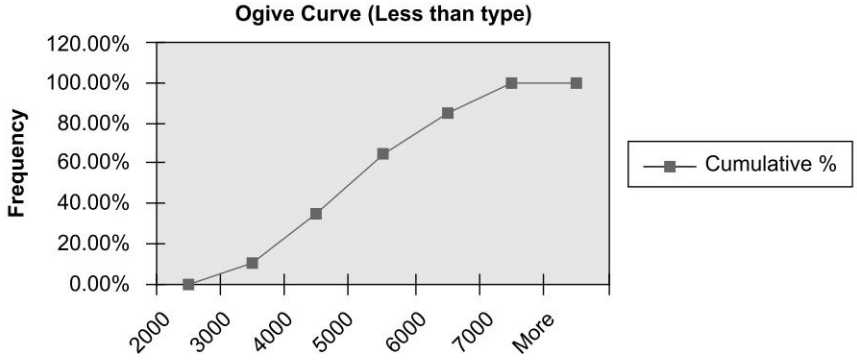
Advantages and Disadvantages of Median

The following table gives the advantages and disadvantages of using median as the measure of location.

Advantages	*Disadvantages*
• **Simple to understand—divides the sample/ population into two parts such that 50% are less than this value and 50% are more than this value.**	• **Arrangement of data in ascending or descending order may be tedious if the number of values is large**
• **Simple to calculate especially from ungrouped data**	• **Cannot be used to find out the total of the values Does not lend itself to mathematical operations**
• **Extreme values do not affect its value.**	
• **Can be determined even when all the values are not known.**	• **May not be representative of the data, if the number of observations are small. For example, if there are 3 observations, say 1, 2 and 10, the median is 2 which is not representative.**
• **Specially useful where data is skewed like income/asset distribution.**	• **It is susceptible to more variation from sample to sample**

9.2.3 Quartiles

Median divides the data into two parts such that 50% of the observations are less than it and 50% are more than it. Similarly, there are "Quartiles". There are three Quartiles viz. Q_1, Q_2 and Q_3. These are referred to as first, second and third quartiles. The first quartile, Q_1, divides the data into two parts such that 25% (Quarter) of the observations are less than it and 75% more than it. **The second quartile, Q_2, is the same as median**. The third quartile divides the data into two parts such that 75% observations are less than it and 25% are more than it. All these can be determined, graphically, with the help of the Ogive curve as shown below (equity holdings data).



As regards determining the quartiles mathematically, just like Median is defined as the value corresponding to the $\{(n + 1)/2\}$th observation for ungrouped data, and the value corresponding to the $(n/2)$th observation in the grouped data, Q_1 and Q_3 are defined as values corresponding to an observation given below:

	Ungrouped Data (arranged in ascending order)	**Grouped Data**
Lower Quartile Q_1	$\{(n + 1)/4\}^{th}$	$(n/4)^{th}$
Median Q_2	$\{(n + 1)/2\}^{th}$	$(n/2)^{th}$
Upper Quartile Q_3	$\{3\ (n + 1)/4\}^{th}$	$(3n/4)^{th}$

Calculations of Q_1 and Q_3 for grouped data are similar to median, and are illustrated below (for equity holdings data)

$$Q_1 = L_{Q_1} + \frac{(n/4) - f_c}{f_{Q_1}} \times w_{Q_1}$$

$$Q_3 = L_{Q_3} + \frac{(3n/4) - f_c}{f_{Q_3}} \times w_{Q_3}$$

For equity holdings data, the first and third quartiles are calculated as follows:

$$Q_1 = 3000 + \frac{((20/4) - 2)}{5} \times 1000$$

$$= 3000 + \frac{(5 - 2)}{5} \times 1000$$

$$= 3000 + \frac{3000}{5} = 3000 + 600$$

$$= \mathbf{3600}$$

The interpretation of this value of Q_1 is that 25% billionaires have equity holdings less than Rs 3,600 millions.

$$\mathbf{Q_3} = 5000 + \frac{(15 - 13)}{4} \times 1000 = 5000 + \frac{2}{4} \times 1000$$
$$= \mathbf{5500}$$

The interpretation of this value of Q_3 is that 75% billionaires have equity holdings less than Rs 5500 millions.

Incidentally, the average of first and third quartiles could also be considered as a measure of location.

Thus $(Q_1 + Q_3)/2 = (3600 + 5500)/2 = 9100/2 =$ **4550**

can also be considered as a measure of location for the equity holdings data.

9.2.4 Percentiles

Just like quartiles divide the data into 4 quarters, there are percentiles which split the data into several parts, expressed in percentages. A percentile also known as centile, divides the data in such a way that "given percent of the observations are less than it". For example, 95% of the observations are less than the 95th percentile. **It may be noted that the 50th percentile denoted as P_{50} is the same as the median.** The percentiles can be calculated just as median and quartiles.

As an illustration,

$$\mathbf{P_{95}} = L_{P95} + \frac{(95/100) \times n - f_c}{f_{P\,95}} \times w_{P95}$$

where, L_{P95} is the lower point of the class interval containing 95th percent of total frequency, f_c is the cumulative frequency up to the 95th percentile interval, f_{P95} is the frequency of the 95th percentile interval and w_{P95} is the width of the 95th percentile interval. Referring to the Table in Illustration 4.2, and substituting the respective values, we get

$$\mathbf{P_{95}} = 6000 + \frac{(19/20)\times 20 - 17}{3} \times 1000$$

$$= 6000 + \frac{2}{3} \times 1000$$

$$= 6000 + 667$$

$$= \mathbf{6667}$$

The interpretation of this value is that 95% of the billionaires have equity holdings less than Rs 6667 Millions.

9.2.5 Mode

In addition to mean and median, **Mode** is yet another measure of location or central tendency. Mode is a French word meaning fashion. Accordingly, it is defined in such a way that it represents the 'fashion' of the observations in a data. Mode is defined as the 'most fashionable' value, which maximum number of observations have or tend to have as compared to any other value. For example, if, in a group of 10 boys, 3 are wearing white shirts, 4 are wearing **blue** shirts, 1 is wearing a red shirt and 2 are wearing yellow shirts, the fashion or mode could be taken as 'blue' as it is the colour of shirts of maximum number of boys. Similarly, if the observations are 2, 4, 4, 5, 8, 8, 8, and 9, the mode is the number 8 because 3 observations have this value. It is possible to have more than one mode in a data. For instance, in the data comprising of the observations, 3, 5, 5, 9, 9 and 10, there are 2 modes viz. 5 and 9.

In a grouped data, the mode is calculated by the following formula:

$$\textbf{Mode} = \boldsymbol{L_m} + \frac{2}{3} \times \boldsymbol{w_m} \qquad \textbf{(9.11)}$$

where,

L_m is the lower point of the modal class interval

f_m is the frequency of the modal class interval

f_0 is the frequency of the interval just before the modal interval

f_2 is the frequency of the interval just after the modal interval

w_m is the width of the modal class interval

The method of calculating mode from a grouped data is illustrated with the help of the equity holdings data given in Illustration 9.2, and reproduced below:

Class Interval *(1)*	*Frequency* (f_i) *(2)*
2000 – 3000	2
3000 – 4000	5
4000 – 5000	**6**
5000 – 6000	4
6000 – 7000	3
Summation	$\Sigma f_i = 20$

It may be noted that the modal interval i.e., the class interval with the maximum frequency (6) is 4000 to 5000. Further,

$L_m = 4000$
$w_m = 1000$
$f_m = 6$
$f_0 = 5$
$f_2 = 4$

Therefore

$$\text{Mode} = 4000 + \frac{(6-5)}{2\times 6-5-4} \times 1000$$

$$= 4000 + \frac{1}{3} \times 1000 = 4000 + 333.3$$

$$= \mathbf{4333.3}$$

Thus the modal equity holdings of the billionaires is Rs 4333.3 millions.
It may be recalled that the mean and median from the same data are 4550 and 4500.
The calculation of mode is further explained through an illustration below.

4.2.6 Features of a Good Statistical Average

Having discussed all the measures of location or 'average', it may be worthwhile to list the distinguishing features of a 'good' statistical average. These are as follows:

- Readily computable, comprehensible and easily understood
- It should be based on all the observations
- It should be reliable enough to be taken as true representative of the population
- It should not be much affected by the extreme values in the data
- It should be amenable to further mathematical treatment. This properly helps in assessing the reliability of conclusions drawn about the population value with the help of sample value
- Should not vary much from sample to sample taken from the same population

Somehow, it is observed that whenever there are several statistical measures of a parameter, the measure that is more difficult to calculate is also *more reliable*. It is in consonance with general observation in life that *better* things are *harder* to get! Further, in general, a measure that uses full information contained in a data is more reliable than the measure that uses only partial information.

9.2.7 Comparison of Measures of Location

The comparative features of some of the measures of location are presented in a tabular form as follows:

Advantages	*Disadvantages*
Arithmetic Mean	
• Easy to understand and calculate • Makes use of full data • Only number and sum of the observations need be known for its calculation	• Unduly influenced by extreme values • Cannot be calculated from the data with open-end class- intervals in grouped data or when values of all observations are not available—all that is known is that some observations are either less than or greater than some value, in ungrouped data

(*Contd*)

(*Contd*)

Geometric Mean	
• Makes use of full data	• Cannot be calculated if any observation has the value zero or is negative
• Extreme large values have lesser impact	• Difficult to calculate and interpret
• Useful for data relating to ratios and percentages	
• Useful for rate of change/growth	
Median	
• Simple to understand	• Arranging values in ascending /descending order may sometime be tedious
• Extreme values do not have any impact	• Sum of the observations cannot be found out, if only Median is known
• Can be calculated even if values of all observations are not known or data has open-end class intervals	• Not amenable for mathematical operations
• Used for measuring qualities and factors which are not quantifiable	
• Can be approximately determined with the help of a graph (ogives)	

9.2.8 Five Number Summary

Sometimes, a data is summarised with five numbers, and is known as five-number summary. It comprises:

1. The minimum or smallest observation
2. The lower quartile or first quartile
3. The median
4. The upper quartile or third quartile
5. The maximum or largest observation

The five–number summary is sometimes represented as in the following table:

Median
1st Quartile 3rd Quartile
Minimum Maximum

Thus, for the equity holding data of 20 billionaires, the five numbers summary is

4500
3600 5500
2717 6874

9.3 MEASURES OF VARIATION/DISPERSION

The measures of location provide a value on which all the observations could be assumed to be located or concentrated. However, these measures do not give an idea of the variation or dispersion that is present among the observations i.e. how much they are scattered or differ from each other.

For example, consider the following four sets of data

	50	49	40	10
	50	50	50	50
	50	51	60	90
Mean	**50**	**50**	**50**	**50**

The mean in all the sets of data is the same, and thus if we were to draw any conclusion from these sets of data, only on the basis of mean, then all the data would be considered the same. But we do note the distinction among the four sets of data. The observations in the first set are all identical having the same value as 50, the observations in the second set are quite close to each other varying from each other by just ± 1, the observations in the third set are varying by as much as ± 10 where as in the fourth set the variation among observations is maximum by ± 40.

It is, therefore, desirable to have some measures that could provide an idea of the extent of variation present among the observations. These are:

(i) Range
(ii) Semi Inter-Quartile Range
(iii) Standard Deviation (Variance)
(iv) Coefficient of Variation

These have been described in subsequent Sections.

9.3.1 Range

It is the simplest measure of variation, and is defined as the difference between the maximum and the minimum values of the observations:

Range = Maximum Value – Minimum Value

The range for the above four sets of data are 0, 2, 20 and 80, respectively. This does appeal to the common sense about the comparative presence of variation in the four sets of data.

However, since the range depends only on the two viz. the minimum and the maximum values, and does not utilize the full information in the given data, it is not considered very reliable or efficient as brought out in Chapter 11 on Statistical Inference.

However because of simplicity or ease in its calculation, it is widely used in control charts used for controlling the quality of manufactured items.

Coefficient of Scatter is another measure based on the range of a data. It is defined as the ratio,

$$\frac{\text{Range}}{\text{Maximum} + \text{Minimum}} = \frac{\text{Maximum} - \text{Minimum}}{\text{Maximum} + \text{Minimum}}$$

and is called **"Relative Range", "Ratio of the Range"** or **"Coefficient of Scatter"**, and gives an indication about variability in the data. For the above four sets of data, the coefficients of scatter are 0, 0.02, 0.2 and 0.8, respectively. It may be noted that, being the ratio, this measure is a pure number and has no unit of measurement.

9.3.2 Semi Inter-Quartile Range or Quartile Deviation

The difference between the third Quartile (Q_3) and first Quartile (Q_1) i.e $Q_3 - Q_1$ is referred to as **Inter-Quartile Range (IQR)**. It gives an idea of the range in the sense that it gives the range

within which the middle 50 % observations lie. With reference to the equity holdings data, it works out to

$$Q_3 - Q_1 = 5500 - 3600 = 1{,}900$$

This implies that 50% of the billionaires have equity holdings between 3,600 to 5,500, the range of their holdings being 1,900.

However, a much more popular measure of variation is **Semi Inter-Quartile Range** or **Quartile Deviation,** and is defined as

$$\frac{Q_3 - Q_1}{2} \tag{9.12}$$

This is especially useful if sample mean cannot be calculated because of the open ended class interval(s).

For the equity holdings data in Table 4.1, its value, vide Section 4.2.3, is (5500 – 3600) ÷ 2 = 1900/2 = **950**.

9.3.3 Variance and Standard Deviation

While calculating mean deviation, the absolute values of observations from the mean were taken because without doing so, the total deviation was zero for the data comprising values 1, 2 and 3 even though there was variation present among these observations. However, another way of getting over this problem of total deviation being zero is to take the squares of deviations of the observations from the mean as shown below:

	Observation x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	1	–1	1
	2	0	0
	3	+1	1
Sum	6	0	2
Mean	2	0	**2/3 (= 0.67)**

Thus, the sum of squares of the deviations from the mean is 2. This forms the basis for defining another measure of variation called Variance or Standard deviation. The sum of squares divided by number of observations is known as the **variance** and its square root is known as the **standard deviation.** Karl Pearson introduced these terms.

In the above table, variance is 0.67, and its square root 0.82 is the standard deviation. Their calculation is illustrated below for ungrouped as well as grouped data.

(a) For Ungrouped Data

For the data with n observations $x_1, x_2, \ldots, x_n$, the variance is defined as

$$\textbf{Variance} = \frac{1}{n}\sum(x_i - \bar{x})^2 \tag{9.13}$$

It is also written as

$$= \frac{1}{n}\sum x_i^2 - \bar{x}^2 \tag{9.14}$$

In real life situations, whenever a data is collected or obtained as a sample from a population, the variance of the sample values of the observations is defined as

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

for certain theoretical reasons, discussed in details in Chapter 11 on Statistical Inference. In fact many of the calculating devices provide variances for both population as well as sample. Even the template titled 'Elementary Statistics Chapter 9' for this chapter gives values of both the variances. However, in this chapter for all the manual calculations, we have assumed the given data as the population itself, unless mentioned as sample s.d.

For calculation of variance we can use either of the above two formulas. If it is easy to comput** (x_i –), the first formula can be used; otherwise the second formula is to be used.

The standard notation for variance is σ^2 where σ is a Greek symbol anologous to small *s* in English language. It may be recalled that Σ is a Greek symbol anologous to capital *S* in English.

The square root of σ^2 i,e, σ is known as the standard deviation. For the above example,

$$\text{Variance } (\sigma^2) = 1/3 \sum_{i=1}^{3} (x_i - \bar{x})^2$$
$$= 1/3\ (2) = 2/3$$
$$= 0.67$$
$$\text{Standard Deviation } (\sigma) = \sqrt{0.67}$$
$$= 0.82$$

There is another measure of deviation known as Root Mean Square Deviation abbreviated as RMSD, and is defined as follows:

Root Mean Square Deviation

$$\text{RMSD} = (1/n)\ \Sigma\ (x_i - A)^2$$

where, *A* is some arbitrary value. It can be proved that

$$\text{RMSD} = \text{Variance} + d^2 \text{ where } d = A - \bar{x}$$

Standard Deviation is *minimum* when the deviations are measured from the arithmetic mean.

It may be noted that, as mentioned earlier, while Mean Deviation is minimum when the deviations are measured from the median, the standard deviation is minimum when the deviations are measured from the mean.

Impact of Change in Origin and Scale

If a constant *c* is added to each of observations $x_1, x_2, \ldots, x_n$, the standard deviation of new observations $(x_1 + c), \ldots (x_n + c)$ is unchanged.

For example, as shown above, standard deviation of the observations 1, 2 and 3 is 0.82. Suppose 10 is added to all the three observations so that the new observations become 11, 12, and 13. It can be verified that the standard deviation of 11, 12 and 13 remain unchanged as 0.82.

Thus, in general, if a constant is added or subtracted from original observation, the s.d. of new observations remains the same as of original observations.

For example, if s.d. of the variable x is 7, then the s.d. of the variable $x + 5$ is also 7.

However, if all the original observations are multiplied by a constant, say *c*, the s.d. of new observations is *c* times the s.d. of original observations.

As another example, if the s.d. of variable x is 7, then the s.d. of variable $5x$ is $5 \times 7 = 35$ and the s.d. of variable $x/5$ is $7/5 = 1.4$.

As yet another example, since s.d. of 1, 2 and 3, as shown above, is 0.82, s.d. of these observations each multiplied by 100 viz. 100, 200 and 300 also gets multiplied by 100, and is $= 100 \times 0.82 = 82$. Now, if these observations are each divided by 10 so that the new observations are 10, 20 and 30, the s.d. of these observations also gets divided by 10, and is = 8.2.

(b) Calculation of Standard Deviation from Grouped Data

The formulae for calculation of variance and s.d. from a grouped data are as follows

$$\text{Variance} = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i} \tag{9.15}$$

$$= \frac{\sum f_i x_i^2 - \sum f_i (\bar{x})^2}{\sum f_i} \tag{9.16}$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

Calculation of Variance and Standard Deviation for the Billionaires Data Given in Illustration 9.2

Class Interval	*Mid Point of Class Interval* (x_i)	*Frequency* (f_i)	$f_i x_i$	$f_i x_i^2$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$
2000–3000	2500	2	5000	12500000	-2050	4202500	8405000
3000–4000	3500	5	17500	61250000	-1050	1102500	5512500
4000–5000	4500	6	27000	121500000	-50	2500	15000
5000–6000	5500	4	22000	121000000	950	902500	3610000
6000–7000	6500	3	19500	126750000	1950	3802500	11407500
Sum		20	91000	443000000		10012500	28950000
**verage ()			4550			Variance =	1447500

Thus the variance for the given data by using the formula (9.15) is 1447500 and Standard Deviation is 1203.1.

The variance can also be calculated by the formula (9.16) by substituting the relevant values.

It may be verified that, the variance of the equity holdings for ungrouped data is 1217.8.

Combining Variances of Two Populations Suppose a set of data has n_1 values whose variance is σ_1^2, and another set of data has n_2 values whose variance is σ_2^2, and both the sets of data have the same mean, then if these two sets of data are combined to have $n_1 + n_2$ values, then the variance σ^2 of this combined data is

$$\frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2}$$

Interpretation of S.D.

It may be noted that while s.d. is most useful in describing or analysing a data, there is no physical interpretation of s.d. like that of mean or median. It is useful only for comparative purposes. For example, if s.d. of population '*A*' is 4 and the s.d. of population '*B*' is 6, we conclude that dispersion among data values in population '*B*' is more than dispersion among data values in population '*A*'.

9.3.4 Chebychef's Lemma

As per Chebychef's Lemma, Standard Deviation (σ) helps to find out the percentage of population lying within certain limits from the mean of the variable. As per the Lemma:

"For any set of data (population or sample), and for any constant $k > 1$, at least $(1 - 1/k^2)$ of the observations must lie within k s.ds. on either side of the mean."

For example, at least 75 % of the observations lie between mean ± 2σ, and at least 81% of the observations lie between mean ± 3σ. Thus, if mean and s.d. of a set of observations are 100 and 10, respectively, at least 81% of the observations lie between $100 \pm 3 \times 10$ i.e. 100 ± 30.

Illustration 9.5

A cold drink bottling plant fills bottles of 500 ml. capacity with mean of 500 ml. and s.d. of 5 ml. At least what percentage of bottles would contain cold drink between 490 and 510 ml.?

Solution:

In the illustration, $k = 2$ and s.d. $\sigma = 5$. Therefore, minimum percentage of bottles with drink equal to within ± 2σ from the mean would be greater than $1 - 1/4 = 3/4 = 75\%$.

9.3.5 Coefficient of Variation or Dispersion

If there are two or more sets of data which have means either quite different from each other, or in different units like one is in Rs. and the other in Days, then comparision of their variances is not valid and could lead to deceptive conclusions.

For illustration, let us consider the following three sets of observations

9	49	99
10	50	100
11	51	101

If we calculate the standard deviation for each of the above three sets of data, we get the same value viz. $\sqrt{\left(\frac{2}{3}\right)} = 0.82$ as shown below.

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	9	-1	1	49	–1	1	99	–1	1
	10	0	0	50	0	0	100	0	0
	11	+1	1	51	+1	1	101	+1	1
Sum	30		2	150		2	300		
Variance	2/3= 0.67			2/3= 0.67			2/3= 0.67		
Standard Deviation	**0.82**			**0.82**			**0.82**		

Thus, if standard deviation alone were to be taken as an index of variation, one would conclude that all the three sets have the same extent of variation. But, it may be noted that the variation in the first set is around 1 in the average value of 10, in the second set it is around 1 in the average value of 50, and in the third set, it is around 1 in the average value of 100. Thus, it appears intuitively that the variation is least in the third case and maximum in the first set. It was, therefore, felt necessary to evolve a measure of variation, which could take this aspect into account, and also which is a pure number without any units. Such measures are described in the next paragraph.

Coefficient of Variation (C.V.) or **Coefficient of Dispersion** (**C.D**.) or **Relative Dispersion** is a statistical measure introduced by Karl Pearson. It helps in studying the relative dispersion of two or more sets of data. It is defined as the ratio of standard deviation to the mean, and is, usually, expressed in % form. If m and σ are mean and standard deviation of a data, then its

$$\text{C.V.} = \frac{\sigma}{m} \times 100 \tag{9.17}$$

For example, if $m = 10$, and $\sigma = 2$, then

$$\text{C.V.} = \frac{2}{10} \times 100$$
$$= 20\%$$

For the above three sets of data with three observations each, Coefficients of Variation are:

$= (0.82/10) \times 100 = \mathbf{0.0820} \times 100 = \mathbf{82.0\%}$ (for values 9, 10 and 11)
$= (0.82/50) \times 100 = \mathbf{0.0164} \times 100 = \mathbf{16.4\%}$ (for values 49, 50 and 51)
$= (0.82/100) \times 100 = \mathbf{0.0082} \times 100 = \mathbf{8.2\%}$ (for values 99, 100 and 101)

Thus, the maximum value of C.V. is for the first set, and the minimum value is for the third set, and it gives true extent of variation in the three sets of data.

For the equity holdings data, given in Illustration 9.1, C.V. is worked out below.

For ungrouped data,

$$\text{C.V.} = (1217.8/4565.4) \times 100 = \mathbf{26.7\%}$$

For grouped data,

$$\text{C. V.} = (1203.1/4550) \times 100 = \mathbf{26.44\%}.$$

Incidentally, because of the above relationship among C.V., m and σ, given any two of these quantities, the third quantity can be found easily.

For example, if mean is 10, and C.V. is 5, then σ can be found as follows:

$$5 = \frac{\sigma}{10} \times 100$$

Therefore, $50 = 100\ \sigma$

or $$\sigma = \frac{50}{100} = 0.5$$

Interpretations of C.V. The coefficient of variation can have different interpretations in different applications.

While studying the performance of individuals (like cricketers) and teams, coefficient of variation can be defined as a measure of ***consistency*** in performance. Incidentally, it may be noted that, lesser the coefficient of variation, more the consistency. It may be interesting to have a look at the

runs scored by some of the Indian batsmen in the Cricket World Cup-2003 as also the measures of their consistency in the box given below:

WC 2003	*AUS (FINAL)*	*KEN (SEMI)*	*NEWZ*	*SRILA*	*KEN*	*PAK*	*ENG*	*NAMI*	*ZIM*	*AUS*	*HOLL*	*TOTAL*	*NO. MATCHES*	*AVERAGE*	*SD*	*CV*
PLAYER																
SACHIN	4	83	15	97	5	98	50	152	81	36	52	673	11	61.18	46	75.19
SEHWAG	82	33	1	66	3	21	23	24	36	4	6	299	11	27.18	26.35	96.94
GANGULY	24	111	3	48	107	0	19	112	24	9	8	465	11	42.27	45.42	107.44
DRAVID	47	1	53	18	32	44	62	—	43	1	17	318	10	31.80	21.53	**67.7**
KAIF	0	15	68	19	5	35	5	—	25	1	9	182	10	18.20	20.76	114.07
YUVARAJ	24	16	—	5	58	50	42	—	1	0	37	233	9	25.89	21.86	84.44

It may be noted that **Dravid** was the most consistent Indian batsman in the **World Cup 2003**.

While studying the per capita income of different states in India or different countries in the world, it can be defined as a measure of **disparity**. Disparity indices can be used for various developmental parameters like literacy, life expectancy, income, bank credit as proportion of deposit, etc. This can be attempted for among districts, states, among countries, etc.

While studying the return on equity capital invested in some shares, it could be defined as a measure of **volatility** or **risk** (lesser C.V. lesser the **risk**).

While studying the workload on different counters in banks/booking offices, or for studying wages in different orgnisations, or compensation offered to MBA students in different Institutes, etc., it may be defined as a measure of **uniformity.**

9.3.6 Desirable Characteristics of Measures of Dispersion

Before discussing the comparative features of various measures of dispersion, we list their desirable characteristics in a tabular form given below:

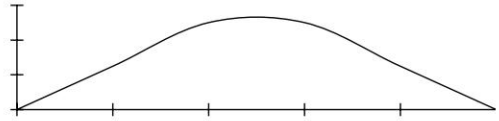
Features	*Desirable Characteristics*
• Understanding	• It should be easily understood
• Utilising available information	• It should utilise all the values of the observations in the data
• Ease of calculation	• It should be easy to calculate
• Sampling variation	• If samples are taken from a population, the values calculated from various samples should not vary much from each other. If it is so, the measure is considered reliable.
• Mathematical manipulation	• It should be amenable to mathematical manipulation so that further meaningful conclusions about the data can be made.

9.4 MEASURE OF SKEWNESS

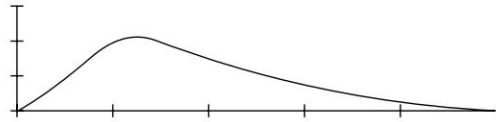
Skewness is an indicator of lack of symmetry in a data.

The histogram or the curve approximating the histogram of a data could be one of the following three types.

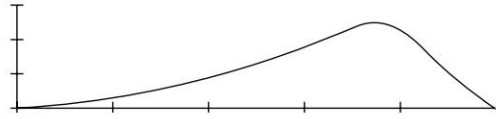
(i) Symmetric



(ii) Positively Skewed



(iii) Negatively Skewed



If the distribution is skewed, the extent of skewness could be measured by the following measure, known as **"Bowley's Coefficient of Skewness"**.

$$S_{kb} = \frac{(Q_3 - \text{Med}) - (\text{Med} - Q_1)}{(Q_3 - Q_1)} \tag{9.18}$$

It varies from –1 to +1.
If this measure is greater than zero, it is called **positively** skewed.
If this measure is lesser than zero, it is called **negatively** skewed.

Pearson's Measure of Skewness
Karl Pearson has also defined a measure of skewness as follows:

$$= \frac{\text{Mean} - \text{Mode}}{\sigma}$$

If calculation of mode is difficult, it can be calculated as

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

9.5 STANDARDISED VARIABLE

Standardised variable is a variable whose origin is shifted to its arithmetic mean, and which is then scaled by its standard deviation. Thus if x is a variable with mean m and s.d. σ, then $(x - m)/\sigma$ is a standardised variable, and is usually denoted by the letter z. It is also known as standardised score.

Such standardised variable, usually denoted by the letters 'u' or 'z', has mean as 0 and standard deviation as 1, as shown below:

x_i	$x_i - m$	$(x_i - m)^2$	$z_i = \frac{x_i - m}{\sigma}$	z_i^2
1	–1	1	$\frac{-1}{\sqrt{(2/3)}} = -\sqrt{\frac{3}{2}}$	$\frac{3}{2}$
2	0	0	0	0
3	1	1	$\frac{1}{\sqrt{(2/3)}} = \sqrt{\frac{3}{2}}$	$\frac{3}{2}$
Sum = 6	0	2	0	3
Average = 3	0	2/3	0	3/3 = 1

9.6 STANDARDISED MARKS—AN ILLUSTRATION

An all-India organisation wanted to award scholarships to 100 meritorious children of its employees who opted to study for graduation. The merit was to be decided on the basis of marks obtained in Class XII examination. However, a problem arose as to how to prepare the list of top 100 students which took into account the varying marks obtained by students of different disciplines like arts, science and commerce and different boards of examinations at all-India as well as state levels. Accordingly, an attempt was made to standardize the marks. These marks were derived to compare performance of students of different disciplines like Arts, Science and Commerce, or of different Educational Boards in different States in India or of different Universities.

The basic premise was as follows:

In an examination conducted by an Educational Board, a student 'A' from Arts stream gets, 65% marks, a student 'B' from Commerce stream gets 80% marks and a student 'C' from Science stream gets 85% marks. From this data, could we conclude that the performance of Student C is the best ? This might not be true, as it is well known that it is easier to get higher marks in Science as compared to Arts or Commerce streams. One way of getting over this problem was to formulate relative marks of these students by comparing their score with scores of toppers in the respective streams. Suppose, the topper in Arts stream got 80% marks, topper in Commerce got 85% marks and topper in Science got 95% marks. Then the standardised marks of the three students were taken as follows:

$$A = \frac{65}{80} \times 100 = 81.25\%$$

$$B = \frac{80}{85} \times 100 = 94.12\%$$

$$C = \frac{85}{95} \times 100 = 89.47\%$$

The merit list was prepared on the basis of these standardised marks derived for all applicants.

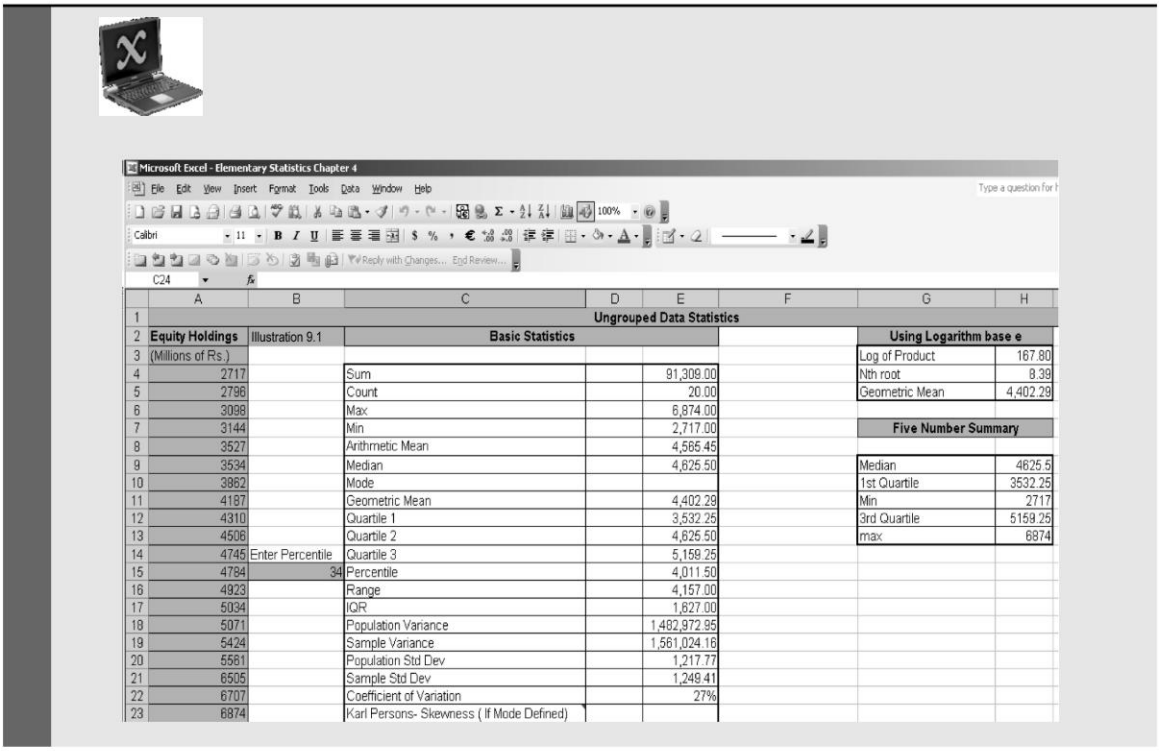
It may be noted that standardised marks for the student 'B' is 94.12% as compared to 89.47% for student 'C'. Thus, the performance of student 'B' may be considered better than that of 'C'. However,

such comparison of students is valid for the limited purpose of comparing relative performance in the examination, and does not reflect the relative intelligence of students.

9.7 USING EXCEL

For the convenience of the readers, we have provided some templates in the CD available with this book. We would now explain the use of the template for 'Elementary Statistics Chapter 4'. This template has two worksheets

- Elementary Statistics Ungrouped
- Elementary Statistics Grouped



Microsoft Excel - Elementary Statistics Chapter 4

	A	B	C	D	E	F	G	H
1	Ungrouped Data Statistics							
2	**Equity Holdings**	Illustration 9.1	**Basic Statistics**				**Using Logarithm base e**	
3	(Millions of Rs.)						Log of Product	167.80
4	2717		Sum		91,309.00		Nth root	8.39
5	2796		Count		20.00		Geometric Mean	4,402.29
6	3098		Max		6,874.00			
7	3144		Min		2,717.00		**Five Number Summary**	
8	3527		Arithmetic Mean		4,565.45			
9	3534		Median		4,625.50		Median	4625.5
10	3862		Mode				1st Quartile	3532.25
11	4187		Geometric Mean		4,402.29		Min	2717
12	4310		Quartile 1		3,532.25		3rd Quartile	5159.25
13	4506		Quartile 2		4,625.50		max	6874
14	4745	Enter Percentile	Quartile 3		5,159.25			
15	4784	34	Percentile		4,011.50			
16	4923		Range		4,157.00			
17	5034		IQR		1,827.00			
18	5071		Population Variance		1,482,972.95			
19	5424		Sample Variance		1,561,024.16			
20	5561		Population Std Dev		1,217.77			
21	6505		Sample Std Dev		1,249.41			
22	6707		Coefficient of Variation		27%			
23	6874		Karl Persons- Skewness (If Mode Defined)					

In the above snapshot of the template for ungrouped data, one could enter the data in the cell A4 downwards. Once the data is entered, the template automatically calculates the values for different measures given in the boxes above.

The various measures given in the template are:

Measure	*Description*
Sum	Sum of all observations
Count	Number of observations
Max	Maximum of the observations

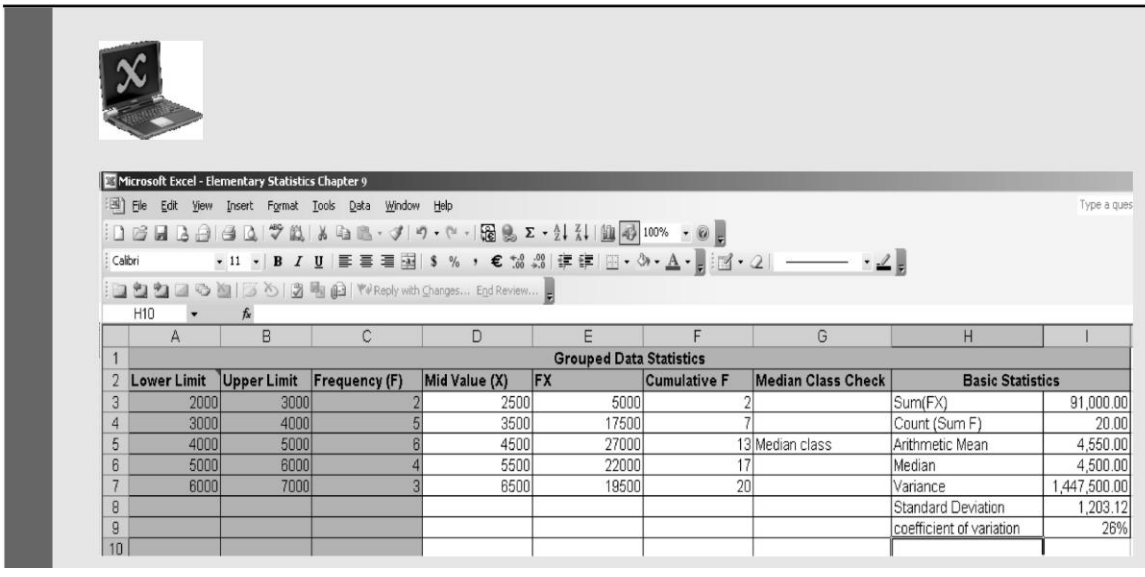
(*Contd*)

(*Contd*)

Min	Minimum of the observations
Arithmetic Mean	Mean of the observations
Median	Median of the observations
Mode	Mode of the observations
Geometric Mean	Geometric mean of the observations
Quartile 1	First quartile of the observations
Quartile 2	Second quartile or Median of the observations
Quartile 3	Third quartile of the observations
Percentile	Percentile of the observations for given percentage
Range	Range of the observations
IQR	Inter Quartile Range of the observations
Population variance	Population Variance of the observations
Sample Variance	Sample Variance of the observations
Population Std Deviation	Population Std Dev of the observations
Sample Std Deviation	Sample Std Dev of the observations
Coefficient of Variation	Coefficient of variation of the observations
Karl Pearson's- Skewness (If Mode Defined)	Karl Pearson's Skewness of the observations*
Geometric Mean	Geometric mean of the observations
Five Number Summary	Five number summary of the observations

**Note:* Karl Pearson's Skewness can be calculated only if mode is defined.

We may recall the Illustration 9.2 relating to Billionaire's data. We have solved this illustration with the help of the above template.



	A	B	C	D	E	F	G	H	I
1	Grouped Data Statistics								
2	Lower Limit	Upper Limit	Frequency (F)	Mid Value (X)	FX	Cumulative F	Median Class Check	Basic Statistics	
3	2000	3000	2	2500	5000	2		Sum(FX)	91,000.00
4	3000	4000	5	3500	17500	7		Count (Sum F)	20.00
5	4000	5000	6	4500	27000	13	Median class	Arithmetic Mean	4,550.00
6	5000	6000	4	5500	22000	17		Median	4,500.00
7	6000	7000	3	6500	19500	20		Variance	1,447,500.00
8								Standard Deviation	1,203.12
9								coefficient of variation	26%
10									

The above snapshot gives the worksheet containing the template for grouped data. In this template, if one enters the lower limit, the upper limit and the frequency of the class interval, the template would automatically calculate the rest of the columns like mid value, f.x, cumulative frequency, and the basic statistics.

This template can be used to solve all the problems given in this chapter.

OBJECTIVE TYPE QUESTIONS

1. Which of the following is not a measure of location?
   (a) Arithmetic Mean (b) Median (c) Range (d) Harmonic Mean
2. The arithmetic mean of a set of 10 values is 100. If a value 55 is added, the arithmetic mean of these 11 values is
   (a) 102 (b) 101 (c) 105 (d) 110
3. If a car covers a distance of 120 kms. from A to B in 2.5 hrs and from B to A in 1.5 hrs, the average speed of the car is
   (a) 50 km/hr (b) 55 km/hr (c) 60 km/hr (d) 65 km/hr
4. If the mean of a variable is 50, then the mean of the new variable (x – 30)/20 is:
   (a) 0.5 (b) 1.0 (c) 1.5 (d) 2.0
5. Which of the following statements is not true about arithmetic mean?
   (a) Easy to understand
   (b) Makes use of full data
   (c) Can be located graphically
   (d) Least susceptible to change from sample to sample
6. Which of the following is not a disadvantage of Geometric mean?
   (a) Cannot be determined if any value is 0
   (b) Extreme values have more impact on the value of arithmetic mean as compared to Geometric mean.
   (c) More difficult to calculate
   (d) If one value is negative, Geometric mean cannot be calculated
7. Which one of the following is not a desirable feature of a good statistical average?
   (a) Based on all observations
   (b) Sensitive to extreme values in the data
   (c) Readily computable
   (d) Amenable to mathematical treatment
8. Which of the following measures is not included in Five number summary?
   (a) Mean (b) Median (c) First Quartile (d) Third Quartile
9. Which of the following statements is not true?
   Standard deviation is unchanged if
   (a) a constant is added to all the observations
   (b) a constant is subtracted from all observations
   (c) a constant is either added or subtracted from all observations
   (d) all the observations are multiplied by a constant
10. If a variable x has standard deviation as 20, then standard deviation of $5(x - 10)$ is
    (a) 25 (b) 10 (c) 30 (d) 100
11. If the s.d. of a data is 10 and C.V. is 50, then the mean of the data is
    (a) 15 (b) 20 (c) 15 (d) 2

12. C.V. is useful for studying
    (a) consistency (b) disparity (c) risk (d) all of above

EXERCISES

1. Following is the list of countries having GDP more than one Trillion Dollar for the year 2006-07.

Trillionaires Club GDP in $ Trillion

Country	*GDP in $ Trillion*
US	13.46
Japan	4.46
Germany	2.89
China	2.55
UK	2.36
France	2.23
Italy	1.84
Spain	1.22
Canada	1.27
Brazil	1.07
India*	1.01

* For 2006–07, at 26th April's exchange rate. All others from Credit Suisse estimate for Dec. 2006.
Source: Times of India, dt. 27th April 2007.

Calculate the coefficient of variation among the above countries with respect to GDP. Interpret the coefficient of variation in the given context.

If the GDP of each country increases by 10% next year, what would be the value of this coefficient of variation?

2. The following data relates to sales of top market brands among pain killers in India.

Painkiller	*Sales (Rs. Crores) July-August*		*Growth (%)*
	2005	*2006*	
Voveran	16.5	23.2	40.6
Calpol	13.2	18.2	37.8
Nise	15.2	18.6	22.3
Combiflam	9.4	14.1	50.0
Dolonex	6.8	10.3	51.4
Sumo	5.1	7.4	45.0
Volini	6.9	9.6	39.1
Moov	3.8	4.9	28.0
Nimulid	3.5	4.9	40.0

Source: Economic Times, dt. 9th October 2006.

Calculate the average growth for all the brands as a whole.

3. The following table gives the 'real' income that senior managers actually take home in certain countries, including India. These have been arrived at, after adjusting for the cost of living, rental expenses and purchasing power parity.

Rank	*Country*	*Amount (in Euros)*
1	Turkey	79,021
2	India	77,665
3	Russia	77,355
4	Switzerland	76,913
5	Brazil	76,449
8	Germany	75,701
9	Japan	69,634
13	USA	61,960
23	UK	46,809
26	China	42,288

Source: Hay Group's (USA) World Pay Report published in 10th September 2006 issue of Business Today.

Calculate appropriate measures of location and dispersion.

4. The following data gives certain socio-economic parameters for some countries.

	Life Expectancy (Years)	*School Enrolment (%)*	*GDP per Capita (PPP US$)*	*Population (2004)*	
				(Million)	*% Urban*
Norway	79.6	100	38,454	4.6	77.3
Iceland	80.9	96	33,051	0.3	92.7
US	77.5	93	39,676	295.4	80.5
Thailand	70.3	74	8,090	63.7	32.0
China	71.9	70	5,896	1,308	22.0
Sri Lanka	74.3	63	4,390	20.6	15.2
India	63.6	62	3,139	1,087.1	28.5
Pakistan	63.4	38	2,225	154.8	34.5
Bangladesh	63.3	57	1,870	139.2	24.7
Nigeria	44.6	21	779	13.5	16.7

Source: Hindustan Times dt. 10th November 2006.

Calculate the following indicators for all the ten countries as a whole:

(i) Average life expectancy
(ii) Average school enrollment
(iii) Average GDP per capita
(iv) Average percentage of urban population

5. Jyoti and Anuj are the two eligible candidates for the position of Chief Manager (HRD). Their overall rating (expressed in %) for the last five years are summarised below.

Year	Jyoti	Anuj
2001	70%	81%
2002	81%	82%
2003	90%	79%
2004	74%	78%
2005	85%	75%

Analyse the above data to have an assessment of their past performance.

6. If the price of a commodity doubles in a period of 4 years, what is the average percent increase per year?
7. Given below is a frequency distribution of the sundry debtor status for a sample of 50 accounts of a supplier of goods, before and after introduction of an incentive scheme for prompt payment.

Number of Days Outstanding	Number of Accounts	
	Before	After
31–40	8	10
41–50	26	22
51–60	20	15
61–70	10	8
71–80	6	5

Find out the extent to which the incentive scheme has been successful. Also, has the variability in 'Number of Days Outstanding' reduced?

8. The following Table gives the return on investment (ROI) of 60 companies.

ROI (%) Intervals	Number of Companies
0 –5.0	10
5.0–10.0	25
10.0–15.0	12
15.0–20.0	8
20.0–25.0	5

(Intervals include the upper class mark but not the lower.)

Find the mean, median, mode, and coefficient of variation for the ROI. Give the rating, on a scale from 0 to 100, to a company whose ROI is 12%.

9. Given below is the pattern of deposits of a bank and the corresponding interest rates. Calculate the appropriate average interest rate cost of deposits.

	Rate (%)	Amount Outstanding (in Crores of Rs.)
1. Current	0	27
2. Savings	3.5	39
3. Fixed	—	—

(Contd)

(Contd)

Less than 6 months	5.0	61
6 months – 1 year	6.0	46
1 year – 3 years	7.0	32
3 years – 5 years	8.0	10
Over 5 years	9.5	32

Simple Correlation and Regression

10

Contents

LEARNING OBJECTIVES

This chapter provides the requisite knowledge and expertise to:

- Understand the relevance and applications of relationship between two variables. For example, one could explore whether advertising expenses of a company are related to overall sales of the company?
- Determine the nature of mathematical relationship if it exists; for example whether it is linear i.e. a straight line?
- Derive the exact equation, if the relationship is found to be linear or otherwise. The equation gives an idea of the extent to which advertising can influence the sales.
- Do similar analysis and derive another equation to measure the extent of the influence of another variable like R&D expenses on the sales.
- Compare the influences of two variables on one variable like advertising and R&D expenses on sales.
- To forecast one variable with the help of the other variable like forecasting the sales given the budget for advertising or the budget for R&D.
- Measure the extent of association between two variables, which are available in categorical form. Sometimes, the data is available in categorical form as follows. For example, to assess the taste of new toothpaste among users, the users may be divided in three categories viz. “young”, “middle aged” and “old”, and the response may be categorised as “liked”, “indifferent” and “did not like”.
- Measure ‘rank correlation’ between two variables whose ranks on some criteria are available rather than their numerical values. For instance, ranks of 10 companies as per turnover and profit in a year could be analysed to calculate correlation between ‘sales turnover’ and ‘profit’ of these companies. Similarly, overall ranks of ten companies could be studied for the years 2005 and 2006 to assess correlation in the performance of these companies in the years 2005 and 2006.

Relevance

The new CEO of a "Healthcare" pharmaceutical company called a meeting of the Heads of various departments to discuss the strategy for future. While he expressed satisfaction over the growing sales of the company, he also emphasised the need for giving a further boost to the sales and image of the company. The Head of the R&D unit suggested more funds for innovating new products and improving the existing ones. He pointed that out the R&D had most significant contribution to the sales of the company. The Head of the Marketing Department emphasized the importance of marketing strategies for boosting the sales of the company. He, therefore, wanted more funds to be made available for the purpose. The Head of HRD Department suggested need for more staff as also new training programmes for improving the sales skills and motivation for sales force. This, he said would improve the sales of the company, very significantly. The CEO agreed, in principle, with them but wanted some analysis of quantitative facts and figures to evaluate the claims of the Heads of Departments, and commit funds for the new strategies. The job was entrusted to a consultant who analysed the data, using statistical techniques, in general, and correlation and regression analysis, in particular to assess the impact of R&D, Marketing and HRD initiatives in boosting the sales of the company, and thus facilitated the CEO in taking appropriate decisions based on analytical approach.

10.1 SIGNIFICANCE AND INTRODUCTION

People have always taken interest in talking about and studying the relationship between various phenomena. Astrologers relate the welfare and various aspects of a human being to the position of the stars. Palmists do the same thing for an individual with the help of the lines on his/her palm.

The first reported scientific study of relationship, based on collected data, was carried out by Francis Galton in 1887 when he wanted to study the relationship between heights of sons and their fathers. It set the trend for quantifying the type and extent of relationship and using it to forecast a variable with the help of another variable or forecast a variable with the help of a number of other variables. Such situations could be of two types. One situation is where it is easier to forecast a variable or even a number of variables, which have an impact on a variable, which is relatively difficult to forecast. For example, for leaves of a particular plant, the weight and surface area are related – more the surface area more the weight. It is easier to measure the weight of a leaf than its surface area. This feature could be used to establish a relationship between the surface area and the weight, and use the relationship to estimate the surface area of a leaf if its weight is given. Another situation could be where a variable is dependent on another variable which can be planned or projected with relative ease. By establishing a relationship between these two variables, the dependent variable could be estimated or forecasted, given the planned/projected value of the other variable. For example, in an organisation, the manpower depends on the volume of business/sales. After establishing a relationship between the manpower and the volume of business/sales and given the planned/projected volume of business/sales, one could estimate the manpower requirement in the organisation.

10.1.1 Simple Correlation and Regression Analysis–Illustrative Applications

The following is a list of a few applications of Simple Correlation and Regression Analysis.

- Relationship between Sales of a Company and Expenditure on Advertisement or Investment on Research & Development(R&D)
- Relationship of Sales of a Company and Earning per Share or Price-Earning Ratio of its stock
- Relationship of Price and Demand of a Product
- Relationship between Rate of Inflation and Gold Price
- Relationship between I.Q. and Performance in Entrance Examination to MBA
- Relationship between Speed of Conveyor Belt in a factory and Percentage of Defectives in the output.

The above list can be expanded by including studies in an organisation over a period of time and among organisations at the same point of time.

10.2 CORRELATION ANALYSIS

Correlation is defined as the degree of relationship between two or more variables. It is also referred to as covariation (variation in one variable affecting the variation in the other variable). The degree of correlation between two variables is called **simple** correlation. The degree of correlation between one variable and several other variables is called **multiple** correlation. This chapter is confined to discussing simple correlation and simple regression analysis between two variables.

The relationship between two variables, say 'x' and 'y', can be studied by collecting data on pairs of these variables. A typical sample of size n, i.e. n pairs of values of x and y, are given in the following table:

Table 10.1 Typical Data for Correlation and Regression Analysis

x	y
x_1	y_1
x_2	y_2
—	—
—	—
x_i	y_i
—	—
x_n	y_n

These values of x and y can be depicted with the help of the rectangular co-ordinate system by plotting the observed **pairs of values** of x and y, as shown below. This diagram is known as the **scatter** diagram.

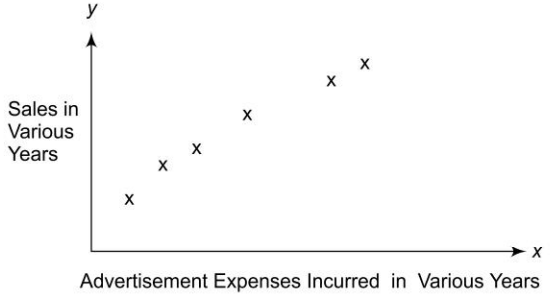


Fig. 10.1 *Scatter Diagram of Sales and Advertisement Expenses in Various Years*

The diagram shows the joint variation among the pairs of values (x_1, y_1), (x_2, y_2), ……, (x_n, y_n), and gives an idea of the relationship between x and y.

If the points are scattered around a straight line, the correlation is **linear** (as shown above), and if the points are scattered around a curve, the correlation is **non-linear** as shown below. If the points are scattered all over without any pattern, there may not be any correlation between the variables x and y.

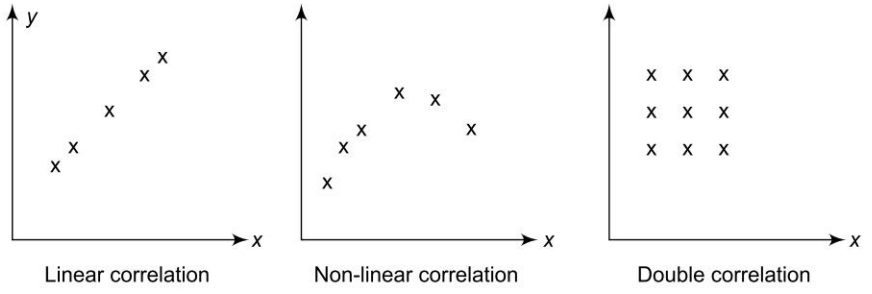


Fig. 10.2 *Scatter Diagrams Indicating Different Types of Correlation*

Two variables may have a positive correlation or negative correlation or they may be uncorrelated. The variables are said to be positively correlated, if they tend to change in the same direction i.e. if they tend to increase or decrease together, as shown below:

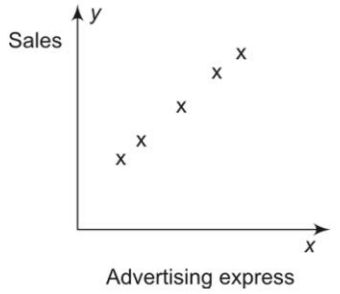


Fig. 10.3 *Scatter Diagram Showing Positive Correlation*

On the other hand, two variables are said to be negatively correlated if they tend to change in the opposite direction i.e. when one increases, the other decreases or if one decreases the other increases. For example the demand of a non-essential product goes down if its price is increased, as depicted below:

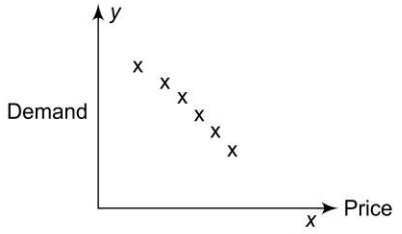


Fig. 10.4 *Scatter Diagram Showing Negative Correlation*

One interesting feature to be noted in day-to-day life is that a **mother's joy** increases as the **level of milk** in the feeding bottle decreases! As yet another example of negative correlation could be the correlation between **contentment** and **urge** to progress in a person!

However, two variables are uncorrelated when they tend to change with no connection to each other as shown below in the case of Intelligent Quotient (I.Q.) and Height of persons.

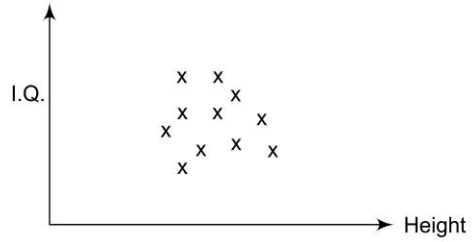


Fig. 10.5 *Scatter Dagram Showing No Correlation*

Following scatter diagrams indicating positive, negative and zero correlation are placed together for easy comprehension.

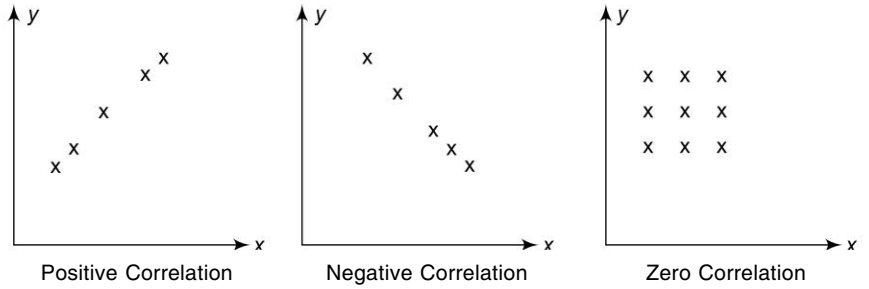


Fig. 10.6 *Scatter Diagrams Depicting Positive, Negative and Zero Correlations*

It may be noted that all the points in the scatter diagram tend to lie near a line or a curve with a positive slope when two variables are positively correlated. Similarly, the points in the scatter diagram tend to lie near a line or curve with a negative slope when two variables are negatively correlated. However, in case of zero correlation, the points do not tend to lie on a line or a curve.

In the above Fig. 10.6, both the scatter diagrams with positive and negative correlation indicate linear relationship. However, there are situations when the scatter diagram suggests curvilinear correlation. A couple of situations are shown below:

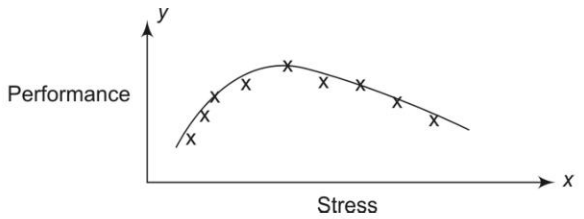


Fig. 10.7 *Scatter Diagram Indicating Relationship between Levels of Stress and Performance*

It may be noted that as the stress increases, the performance also increases, but beyond a point when stress increases further, the performance drops. Similar curvilinear pattern is observed while studying relationship between efforts and results, as indicated below.

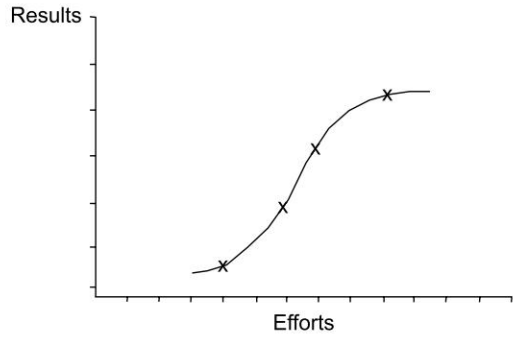


Fig. 10.8 *Scatter Diagram Indicating Relationship between Efforts and Results*

It may be noted that initially when the efforts increase, the results also improve; but after a certain level, when efforts are increased further, the results start declining. For example, we may improve our results by studying for more and more hours; but can we improve results by increasing the study time to, say 20 or more hours?

In this chapter, however, we confine to only linear correlation.

10.3 MEASURE OF LINEAR CORRELATION

The scatter diagram gives a rough indication of the nature and extent/strength of the relationship between the two variables. The quantitative measurement of the degree of correlation between two variables, say 'x' and 'y', is given by a parameter called **correlation coefficient**. It was developed by **Karl Pearson.** That is how, sometimes, it is referred to as "Pearsonian Correlation Coefficient". It is denoted by the Greek letter ρ **(rho)**, if it is calculated from the population values, and '**r',** if it is calculated from a sample. The use of letter 'r' for correlation could be traced to Galton's use of the word regression, and subsequent developing correlation from regression. In the context of correlation study, the population data comprises of pairs of values of x and y. Generally, we deal with only samples in various studies. A sample contains pairs of values of x and y as given earlier, in Table 10.1.

The correlation coefficient is defined by the formula

$$r = \frac{\text{Covariance of } x \text{ and } y}{(\text{Standard Deviation of } x)\,(\text{Standard Deviation of } y)}$$

where **covariance** of x and y, abbreviated as **cov(x, y)**, is a measure of joint variation in x and y. Just like $(1/n)\Sigma(x_i - \bar{x})^2$ is the variance of x, and $(1/n)\Sigma(y_i - \bar{y})^2$ is the variance of y, similarly, $(1/n)\Sigma(x_i - \bar{x})(y_i - \bar{y})$ **is the covariance of x and y. Further, as variance gives an idea of the variation in a single variable, like x or y, covariance gives an idea of joint variation in both the variables taken together.**

The correlation coefficient is also expressed as

$$\rho(\text{Greek letter pronounced as rho}) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \textbf{(for population)} \tag{10.1}$$

where, σ_{xy} is the covariance between x and y, σ_x is the s.d. of x, and σ_y is the s.d. of y.

However, we are concerned only with the sample, and, therefore,

$$r = \frac{s_{xy}}{s_x s_y} \qquad \textbf{(for sample)} \tag{10.2}$$

where s_{xy} is the covariance between x and y, s_x is the s.d. of x, and s_y is the s.d. of y, all calculated from a sample.

$$= \frac{(1/n)\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{((1/n)\Sigma(x_i - \bar{x})^2)}\sqrt{(1/n)\Sigma(y_i - \bar{y})^2}}$$

$$= \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}} \tag{10.3}$$

However, if the variables x_i and y_i are measured from their means i.e. $\bar{x}$ and $\bar{y}$; respectively, the formula for r can be written as

$$= \frac{(1/n)\Sigma X_i Y_i}{\sqrt{(1/n)\Sigma X_i^2}\sqrt{(1/n)\Sigma Y_i^2}}$$

$$= \frac{\Sigma X_i Y_i}{\sqrt{(\Sigma X_i^2)}\sqrt{(\Sigma Y_i^2)}} \tag{10.4}$$

where $\boldsymbol{X_i = (x_i - \bar{x})}$ and $\boldsymbol{Y_i = (y_i - \bar{y})}$ are the variables **measured from their means**.

The correlation coefficient can also be calculated by the following formula derived from formula (10.3)

$$= \frac{(\Sigma x_i y_i - n\bar{x}\bar{y})}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)}\sqrt{(\Sigma y_i^2 - n\bar{y}^2)}} \tag{10.5}$$

using the results:

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - n\bar{x}$$

and $$\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - n\bar{x}^2$$

and $$\Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - n\bar{y}^2$$

It may be noted that the correlation coefficient is just a ratio, and has no dimension like time, money, etc. It is independent of the dimensions of variables *x* and *y*.

10.3.1 Calculation of Correlation Coefficient

The calculation of correlation coefficient from a sample of pairs of observations is illustrated below.

Illustration 10. 1

The following data gives Sales and Net Profit for some of the top Auto-makers during the quarter July-September 2006, Find out the correlation coefficient.

Sales and Net Profits: Some Top Auto-makers

Company	*Average Sales Estimates (Rs Crores)*	*Year to Year Growth (%)*	*Average Net Profit Estimates (Rs Crores)*	*Year to Year Growth (%)*
Tata Motors	6,484.8	36	466.0	38
Hero Honda	2,196.5	1.4	224.2	6
Bajaj Auto	2,444.7	31	345.4	19
TVS Motor	1,032.9	31	35.1	10
Bharat Forge	461.6	23	63.4	22
Ashok Leyland	1,635.8	31	94.7	26
M&M	2,365.5	24	200.6	28
Maruti Udyog	3,426.5	13	315.7	20

Source: Economic Times, 11th October 2006.

Note: The sales and profit figures are rounded for ease of calculations, and placed in tabular format as follows.

Company	*Average Sales Estimates (Rs Hundred Crores)*	*Average Net Profit Estimates (Rs Ten Crores)*
Tata Motors	65.00	47
Hero Honda	22.00	22
Bajaj Auto	24.00	34.5
TVS Motor	10.00	3.5
Bharat Forge	5	6
Ashok Leyland	16.00	9
M& M	24.00	20
Maruti Udyog	34.00	32

If one wants to calculate the correlation coefficient between the average sales and net profit of these automobile companies, one may take the average sales as variable x and average net profit as y. The scatter diagram is given below:

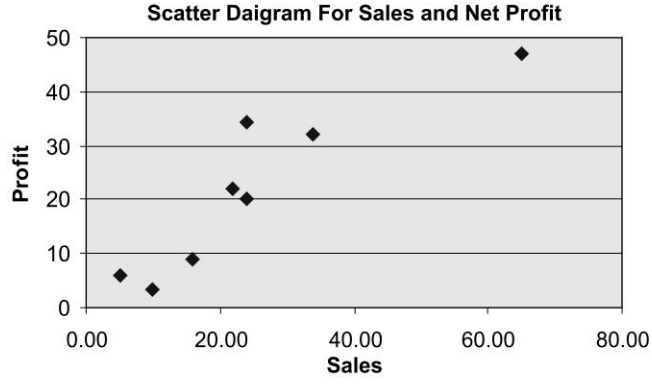


Fig. 10.9 *Scatter Diagram Showing Automobile Sales and Profit*

It may be noted that all the points representing the pairs of values lie around a straight line with a positive slope. This indicates that the correlation is positive and is linear.

For calculating the correlation coefficient by the formula (10.1), one requires the following expressions to be calculated:

Σx_i(for calculating $\bar{x}$), Σy_i(for calculating $\bar{y}$)

Σx_i^2, Σy_i^2 and $\Sigma x_i y_i$.

The calculations are detailed below:

Automobile Company	*Average Sales* x_i	*Average Profit* y_i	x_i^2	y_i^2	$x_i y_i$
Tata Motors	65	47	4,225	2209	3055
Hero Honda	22	22	484	484	484
Bajaj Auto	24	34.5	576	1190.25	828
TVS Motor	10	3.5	100	12.25	35
Bharat Forge	5	6	25	36	30
Ashok Leyland	16	9	256	81	144
M&M	24	20	576	400	480
Maruti Udyog	34	32	1,156	1024	1088
Sum (Σ)	**200**	**174**	**7,398**	**5,436.50**	**6,144**
Average	**25($\bar{x}$)**	**21.75($\bar{y}$)**			

Substituting the values of $\bar{x}$, $\bar{y}$, Σx_i^2, $\Sigma y^2{}_i$, $\Sigma x_i y_i$, in formula (10.5), we get

$$r = \frac{6144 - 8 \times 25 \times 21.75}{\sqrt{(7398 - 8 \times 25^2)}\sqrt{(5436.5 - 8 \times (21.75)^2)}}$$

$$r = \frac{6144 - 4350}{\sqrt{(7398 - 5000)}\sqrt{(5436.5 - 3784.5)}}$$

$$r = \frac{1794}{\sqrt{2398}\sqrt{1652}}$$

$$r = \frac{1794}{1990.35}$$

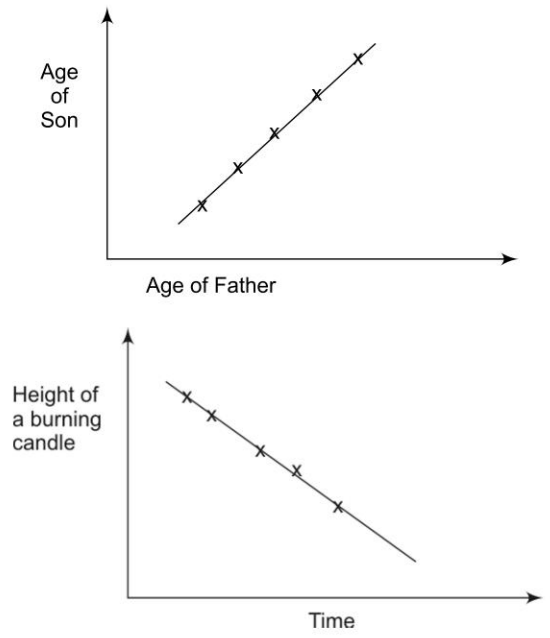
$$= \mathbf{0.90135}$$

10.3.2 Numerical Value of Correlation Coefficient and Its Interpretation

While, in the above illustration, the value of correlation coefficient is positive, as mentioned earlier, it could be negative as well.

The value of correlation coefficient r lies between –1 and +1. The value +1 implies that there is a perfect positive correlation, and the pairs of values of x and y lie on a straight line with **positive** slope, as shown in the diagram given on next page.

The value –1 implies that there is a perfect negative correlation. One example of such a situation is the relationship between time and the height of a burning candle as shown in the diagram given below. The pairs of values of x and y lie on a straight line with **negative** slope.



However, in real life situations, when the variables are random variables, the correlation coefficient does not equal the value 1 (+1 or –1). Its value nearing 1, indicates strong linear relationship, and its value nearing 0 indicates absence of linear relationship.

It is to be emphasized that the value of correlation coefficient *r* indicates the extent or intensity of *only* linear relationship. The low or near zero value of ***r*** merely means that the relationship is not linear but there could be other type of relationships.

10.3.3 Spurious Correlation

Sometimes, the value of correlation coefficient could be misleading, as illustrated by the following study.

Correlation Between Yield of Potatoes and Number of Marriages!

One research worker conducted a study of the number of marriages and the yield of potatoes over a period of time in a certain geographical area. He observed that both were increasing over the period, and the value of correlation coefficient was quite high. He, therefore, concluded that the yield of potatoes affects the number of marriages, and more the yield of potatoes more the number of marriages!

The real reason was that over the period, the population was increasing leading to increased number of marriages. Also, the yield of potatoes was increasing due to better methods of cultivation and better quality of seeds and fertilisers. Thus both the variables were increasing due to different factors and not due to interdependence on each other. Such correlation is termed as ***spurious*** or ***nonsense*** correlation.

Yet another example could be the correlation between the number of students getting graduate degrees every year and the number of auto accidents in the country!

It is, therefore, necessary that before using correlation analysis in any situation, and using it for studying cause – effect relationship, one must ascertain that there is some prima-facie reason to believe that the two chosen variables are interrelated. However, if two variables are observed to be increasing/decreasing together or if one is increasing, the other is decreasing or vice-versa, time and again, then even if there is no apparent known relationship, it may be worth investigating the reasons for the observed pattern.

10.3.4 Coefficient of Determination

The square of the correlation coefficient ***r*,** expressed as r^2, is known as the **coefficient of determination**. It indicates the extent to which variation in one variable is explained by the variation in the other. For example, let two variables, say x and y, be inter-dependent, and variation in x causes variation in y. Further, let the correlation coefficient work out to be, say, 0.9. The coefficient of determination, in this case, is square of 0.9 i.e. 0.81. It implies that 0.81 or 81% of the variation in y is due to variation in x or explained by the variation in x. The remaining 19% (= 100% – 81%) is due to or explained by some other factors. Incidentally, $(1 - r^2)$ is referred to as **coefficient of alienation**, and gives the percentage of variation in the dependent variable, not explained by the independent variable.

Consider the following illustration for calculation and interpretation of Coefficient of Determination.

Illustration 10.2

Five students of a Management Programme at a certain Institute were selected at random. Their Intelligent Quotient (I.Q.) and the marks obtained by them in the paper on Decision Science (including Statistics) were as follows:

I.Q.	*Marks in Decision Sciences (Out of 100)*
120	85
110	80
130	90
115	88
125	92
120	87

In this example, we may take **I.Q.** as the independent variable as ***x*,** and **Marks** in Decision Science as dependent variable ***y***. This is so, because the marks obtained, would generally depend on the I.Q. of a student.

10.4 SPEARMAN'S RANK CORRELATION

So far, we have discussed correlation between two variables, which can be measured and quantified in appropriate units of money, time, etc. However, sometimes, the data on two variables

is given in the form of the ranks of two variables based on some criterion. In this Section, we discuss the correlation between ranks of the two variables rather than their absolute values.

For example, instead of the final grades and salary offered in campus placement of the top 10 students, one may have the data about their final grade ranks from 1 to 10, and rank in terms of salary offered, from 1 to 10. Such data, for a Management Institute for the Batch (Marketing) of 2006, is given below. Incidentally, all these students are those who had no work experience:

	Rank as per Final Grade	*Rank as per Salary Offered*
Simran	1	1
Sajay	2	3
Saluni	3	2
Sumil	4	4
Raunak	5	6
Anuj	6	5
Shreya	7	9
Abhishek	8	8
Karan	9	10
Gyanesh	10	7

With this type of data, we can have an idea of the correlation between 'Grade' and 'Salary' of the students through a measure called '**Spearman Rank Correlation**', introduced by Charles Spearman in 1904, and described in this section.

Spearman defined the rank correlation as

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} \tag{10.6}$$

where, d_i is the difference in the ranks of i^{th} individual or unit, and n is the number of individuals or units. In the above example, $n = 10$. The value of d_1 for Simran is 0, the value of d_2 for Sajay is 3 – 2 = 1, the value of d_3 for Saluni is 3 –2 = 1, and the value for Sumil is 4 – 4 = 0. It may be noted that while taking the difference between the ranks, we take only the absolute value of the difference without bothering for the sign.

For calculating the value of r_s, the following table is prepared:

	Rank as per Final Grade	*Rank as per Salary Offered*	d_i	d_i^2
Simran	1	1	0	0
Sajay	2	3	–1	1
Saluni	3	2	1	1
Sumil	4	4	0	0
Raunak	5	6	–1	1
Anuj	6	5	1	1
Shreya	7	9	–2	4
Abhishek	8	8	0	0
Karan	9	10	–1	1
Gyanesh	10	7	3	9
Sum			**0**	**18**

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6\times 18}{10(100-1)} = 1 - \frac{108}{990} = 1 - 0.11$$

$$= 0.89$$

Thus, the value of Spearman's rank correlation between Final Grade and Salary offered is 0.89.

The value of r_s, like correlation coefficient, lies between –1 and +1. The value + 1 implies that there is perfect correlation in the ranks i.e. the ranks are the same for all the individuals/units. The value –1 implies that the two sets of ranking are just reverse of each other. Thus, if there are three students ranked 1, 2 and 3 on one parameter, they would be ranked 3, 2 and 1 on the second parameter to get a rank correlation of –1. In the above example, if the data were as follows:

Name of Student	*Rank as per Final Grade*	*Rank as per Salary*
Simran	1	10
Sajay	2	9
Saluni	3	8
Sumil	4	7
Raunak	5	6
Anuj	6	5
Shreya	7	4
Abhishek	8	3
Karan	9	2
Gyanesh	10	1

i.e. Simran who was ranked first as per final grade was ranked the last #10 as per salary offered, Sajay who was ranked second as per final grade was ranked #9, and so on, the Spearman's rank correlation would have been –1. This can be verified by calculating the rank correlation for the above Table.

Sometimes, it may happen that two individuals or entities may have the same rank. For instance, in the above example, suppose Saluni and Sumil had the same CGPA; in that case both of them have to be given the same rank. This is done by splitting the total of their ranks equally between them. Thus, both Saluni and Sumil would get rank as (3 + 4)/2 = 3.5. Rest of the methodology remains the same. The formula does undergo a change in such cases but if the number of observations having same rank is not large, the above formula is a good approximation. It may be verified that in the above case, the rank correlation would work out to be 0.881. As a further possibility, suppose three students, say Simran, Sajay, and Saluni had the same CGPA, the rank allocated to them would have been total of their positions viz. 1 + 2 + 3 divided by 3. Thus each one would have been assigned the rank 2. It may be verified that in such a case, the rank correlation would work out to be 0.8807.

The above methodology assures justice to all the individuals whose ranks are tied without changing the sum of ranks for all the individuals or entities as the case might be.

The Spearman's correlation coefficient is, in general, easier to calculate than Karl Pearson's correlation. However, it is less reliable. Besides, it has the following limitations:

(i) It is quite cumbersome to calculate, if number of observations are large.

(ii) It cannot be calculated from a grouped data. Incidentally, Pearson's correlation coefficient can be calculated from a grouped data, but it has not been discussed in this book.

However, it does not require any assumption about the distributions of x and y, while Pearson's correlation coefficient requires the assumption of normal distributions for both x and y. That is how, rank correlation is said to be a distribution free or non-parametric method of assessing correlation. This is a positive point, if one is not sure of the distributions of the variables x and y.

It is interesting to note that the value of Spearman's rank correlation can be found by taking the two ranks as two variables x and y, and calculating Pearson's correlation coefficient between x and y. However, Spearman's formula is simpler to calculate as compared to Pearson's formula.

We have discussed several examples of rank correlation to indicate its application in variety of situations, and because one could have a quick assessment of the correlation between two variables without the use of any calculating aid.

Example 10.1

As per a study, the following are the ranks of priorities for ten factors taken as 'Job Commitment Drivers' among the executives in Asia Pacific (AP) and India. Calculate the rank correlation between priorities of 'Job Commitment Drivers' among executives from India and Asia Pacific.

Favourable Rank			*Percent*	
India	*Asia Pacific*	*Job Commitment Drivers*	*India*	*AP*
1	1	Job Satisfaction	71%	63%
2	2	Work Environment	70.8%	60%
3	4	Teamwork	67%	54%
4	3	Communication	64%	56%
5	5	Performance Management	54%	50%
6	6	Innovation	53%	49%
7	9	Leadership	52%	40%
8	7	Training and Development	51%	45%
9	8	Supervision	50%	42%
10	10	Compensation/Benefits	39%	30%

Source: Watson Wyatt India, *Hindustan Times*, dt. 25th April 2006.

Solution:

The Spearman's rank correlation is defined as

$$r_{\text{rank}} = r = 1 - \frac{6\Sigma di^2}{n(n^2 - 1)}$$

where, d_i is the difference in ranks of the values x and y, and n is the number of pairs of observations. In this Example, $n = 10$.

It may be verified that the rank correlation is **0.9515**.

Thus there is high degree of similarity in preferences for job commited drivers in India and Asia Pacific.

Example 10.2

The churn in the market capitalisation of the top companies during the years 2005 and 2006 was reported in the *Times of India*, dt. 12th October 2006, as follows.

Market Capitalisations of Top 17 Companies

2006			*2005*		
Rank	*Company*	*Market Capitalisation (Rs Crores)*	*Rank*	*Company*	*Market Capitalisation (Rs Crores)*
1	RIL	1,62,971	1	ONGC	1,46,835
2	ONGC	1,61,536	2	RIL	1,08,011
3	Infosys	1,12,180	3	NTPC	85,094
4	NTPC	1,07,521	4	Infosys	71,550
5	TCS	1,05,973	5	TCS	69,022
7	Bharti	90,159	6	Bharti	64,717
8	Wipro	78,164	7	Wipro	54,923
9	ITC	69,886	8	Indian Oil	53,431
10	Indian Oil	66,309	9	SBI	48,606
11	ICICI Bank	61,628	10	ITC	48,202
12	BHEL	56,969	11	ICICI Bank	38,932
13	SBI	54,177	12	HLL	38,686
6	HLL	51,222	13	BHEL	28,776
14	HDFC	37,432	14	HDFC	24,516
15	L&T	35,613	15	SAIL	23,688
16	Tata Motors	34,708	16	Tata Motors	20,332
17	SAIL	34,551	17	L&T	19,479

Note: Reliance Communications, Suelos Energy and Bajaj Auto (Ranked 8th, 16th and 20th) are not included in 2006 and GAIL, HDFC Bank and Tata steel (Ranked 16th, 17th and 18th) are not included in 2005 as these Companies were not listed within 20 ranks in both the years.

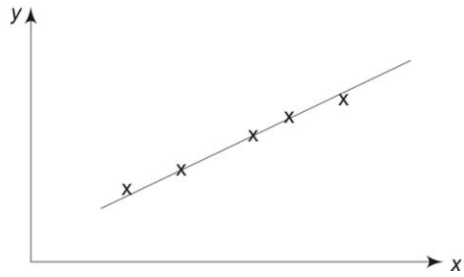
Use the above data to calculate Spearman's rank correlation as well as Pearson's correlation coefficient and comment on the results.

It may be verified that while the Pearson's correlation coefficient is 0.9897, the Spearman's correlation coefficient is 0.9314.

10.5 REGRESSION ANALYSIS

Correlation analysis deals with exploring the correlation that might exist between two or more variables (so far, only two variables have been discussed).

The regression analysis, on the other hand, studies the relationship among two or more variables. The relationship can be described and measured in a functional form. If the relationship between two variables, one dependent and the other independent or explanatory variable, is a linear function



or a straight line, like shown below, then the linear function is called simple regression equation, and the topic under study is referred to as simple regression analysis.

Actually when we talk of relationship between two variables, the scenario could be as follows – in the order of ignorance to knowledge:

- **Ignorance:** We just do not know whether there exists any relationship at all
- We know that the two are related but do not know anything beyond this
- We know that the two are positively or negatively related. Positive relation implies that if one increases, the other also increases, or if one decreases the other also decreases. Negative relation implies that if one increases the other decreases or vice versa. But, beyond this aspect of positive and negative correlation, we are ignorant
- We know only the nature of relationship; whether it is linear or curvilinear.
- **Knowledge:** We know exactly the mathematical equation of this relationship, so that if one of the variables is known, the other can be derived from the equation.

In real life, the 'knowledge' types of situations are very rare, and we have to be contended with the next best i.e. statistical relationships. These can be estimated by the Principle of Least Squares as indicated later in this Chapter.

10.5.1 Types of Regression Analysis

There are two types of regression analysis viz. **Simple** and **Multiple**. While Simple Regression Analysis deals with only two variables, the Multiple Regression Analysis deals with more than two variables.

10.5.2 Estimation of Relationship i.e. Regression Equation

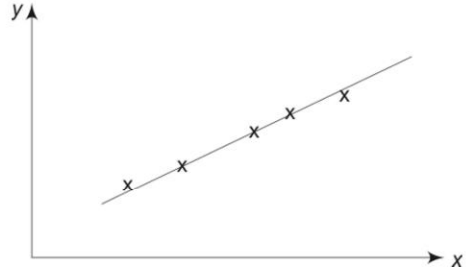
Given a sample of n pairs of values of the variables x and y, one may like to estimate the relationship between the two variables.

Let the data collected on two variables be in the form of n pairs of observations, as follows:

x	y
x_i	y_i
x_2	y_2
:	:
x_n	y_n

The first step is to plot the points to get the scatter diagram.

The simplest type of linear relationship between two random variables x and y can be expressed graphically as a straight line, shown below:



and mathematically as

$$y = a + bx$$

where 'b' is the inclination or slope of the straight line and 'a' is the intercept of the line on the y-axis. The value of 'b' is the amount of increase in y when x increases by 1. The graph is as follows:

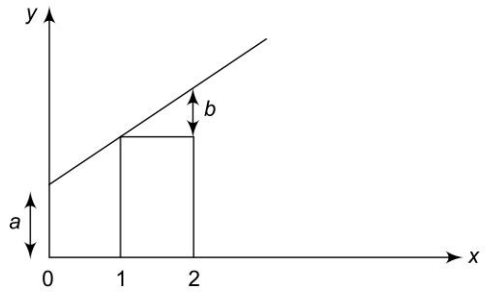


Fig. 10.10 *Graphical Presentation of Values of 'a' and 'b'*

It may be mentioned that the equation

$$y = a + bx$$

is the equation for the sample. The equation for the population from which sample is drawn is

$$y = \alpha + \beta x$$

The values of the parameters α and β are estimated by a and b, respectively. The exercise of estimating the relationship amounts to determining the values of 'a' and 'b'. This is done through the **Principle of Least Squares** discussed below.

Principle of Least Squares For a given value of x as x_i, the observed value of y, as per the sample, is y_i. For the same value of x_i, the estimated value of y, say $\hat{y}_i$, as per the equation $y = a + bx$ is $\hat{y}_i = a + bx_i$.

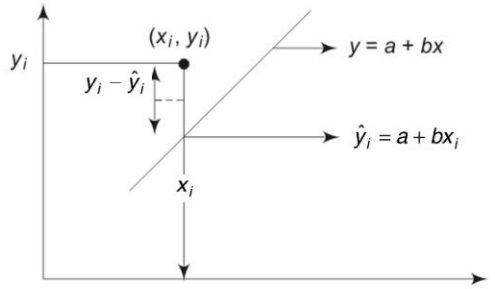


Fig. 10.11 *Difference Between Observed and Estimated Values as per Equation*

The Principle of Least Squares provides the criterion for selecting that line for which the sum of squares of differences between the observed values and the estimated values is *minimum.*
The values of 'a' and 'b' are obtained as

$$b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\{\Sigma(x_i - \bar{x})^2\}}} \tag{10.7}$$

or,
$$= \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{\{\Sigma x_i^2 - n\bar{x}^2\}}} \tag{10.8}$$

and,
$$a = \bar{y} - b\bar{x} \tag{10.9}$$

The above equation $y = a + bx$ is called the **regression equation of *y* on *x***. It indicates the sensitivity of y to changes in x i.e. how y changes with respect to changes in x.

The slope 'b' is called the **regression coefficient of *y* on *x*.** It gives the amount of change in y when x changes by one unit. Thus, when x changes by one unit, y changes by 'b' units. For example, if the relationship is

y (yield of rice in kgs.) = 200 + 5x (rainfall in cms.)

it implies that if rainfall increases by 1 unit i.e. 1 cm., yield of rice increases by 5 units i.e. by 5 Kgs. Thus, one has to be careful while interpreting the estimated or forecasted value of y for a given value of x.

The intercept 'a' has only mathematical interpretation in the sense that it is the value of y for x = 0, but it need not have any physical interpretation. In fact, sometimes its physical interpretation can lead to nonsensical conclusion.

Illustration 10.3

We refer to the data given in Illustration 10.2 of studying correlation between I.Q. and Marks, one can estimate the regression equation of Marks on I.Q. of the type

y (Marks) = $a + bx$ (I.Q.)

by representing Marks by the variable y and I.Q. by the variable x.

This is illustrated below.

For estimating the value of b, by the Formula (10.8), we require the values of, $\bar{x}$, $\bar{y}$, $\Sigma x_i y_i$ and Σx_i^2. These quantities are already calculated above as:

$$\bar{x} = 20$$
$$\bar{y} = 7$$
$$\Sigma x_i y_i = 960$$
$$\Sigma x_i^2 = 2650$$

Substituting these values, we get

$$\boldsymbol{b = 0.48}$$

(*Note:* when we calculated correlation coefficient in this example, we changed origin for x by 100 and for y by 80 by subtracting respective numbers. The regression coefficient will remain same by change of origin but "a" will be different; hence we will find "a" by substituting actual means)

and
$$a = \bar{y} - b\bar{x} = 87 - 0.48 \times 120 = 29.4$$

The physical interpretation of the regression coefficient 'b' is that an increase of I.Q. by one will increase the marks by 0.48.

The intercept a = 29.4 which gives the value of y when x = 0, has no physical interpretation as it means that a student with even I.Q. as 0, will score **29.4** marks.

However, instead of the variables x and y, if we consider the variables $X_i(= x_i - \bar{x})$ and $Y_i(= y_i - \bar{y})$ i.e. the variables measured from their means, then the regression equation of Y on X is

$$Y = bX \tag{10.10}$$

and the regression coefficient 'b' is

$$b = \frac{\Sigma X_i Y_i}{\Sigma X_i^2} \tag{10.11}$$

It may be noted that there is no intercept in the equation (10.10), as its value is 0. The value of 'b', however, remains unchanged.

It may be added that if the variables X and Y are measured from $\bar{x}$ and $\bar{y}$, respectively or even some arbitrary values like 5 and 25, respectively, then value of 'b' remains unchanged. This property can be used to simplify calculations if the values of x and y are large like in Example 10.2, which will be solved using this property.

The formula (10.11) can also be written as

$$b = \frac{s_{yx}}{s_x^2} \tag{10.12}$$

where, s_{yx} is the sample covariance of x and y, and s_x^2 is the sample variances of x. For the population values, the value of b is wriiten as

$$b = \frac{\sigma_{yx}}{\sigma_x^2} \tag{10.13}$$

where, σ_{yx} is the population covariance of x and y, and σ_x^2 is the population variances of x.

10.5.3 Correlation Coefficient between Two Standardised Variables

If the variables x and y are standardized as X and Y, i.e.

$$X = \frac{x - \bar{x}}{s_x}$$

$$Y = \frac{y - \bar{y}}{s_y}$$

then the correlation coefficient between X and Y is the same as between x and y. However, the regression equation of Y on X is

$$\boldsymbol{Y = BX} \tag{10.14}$$

where the value of B is different from that of 'b'. In fact,

$$B = b\frac{s_y}{s_x} \tag{10.15}$$

In computrtised output such as SPSS, Bs i.e. regression coefficients between standardized variables are referred to as **'beta'. It may be clarified that this 'beta' is different from 'beta' of a stock described in Section 10.7.**

It may be noted that X and Y, being standardised variables, have mean as 0 and s.d. as 1. Further, they have no units of measurement like time, weight, money, etc. Thus, for an increase in X by 1 causes a change in Y by B.

In fact, correlation and regression between standardised variables solves the problem of dealing with different units of measurements of *x* and *y*. The values of '*B*s' in different regression equations can be compared to assess the impact of the change in *Y* due to change in *X*.

10.6 STANDARD ERROR OF ESTIMATOR

Once the regression equation is estimated as

$$y = a + bx$$

from the sample of n pairs of values (x_1, y_1), (x_2, y_2), …, (x_i, y_i) …, (x_n, y_n), the value of y can be estimated for each observed value of x. Thus, for example, for $x = x_i$, while the observed value of y is y_i, the estimated value $\hat{y}_i$ is equal to $(a + bx_i)$ vide Fig. 10.11. The difference between these two viz. $y_i - a - bx_i$ is known as the error committed while estimating the value of y by the regression equation. The sum of such squares of errors for all the observed values in the sample divided by the sample size n, is known as the **Error Variance**, and is given as

$$\sigma_e^2 = \frac{\Sigma\{y_i - \hat{y}_i\}^2}{n} = \frac{\Sigma(y_i - a - bx_i)^2}{n}$$

This is also called the **residual variance**. The square root of this variance is called the standard error of the estimator. The variance of the deviations of estimated values viz. $\hat{y}_i$ from the sample mean of y i.e. $\bar{y}$ is given as

$$\sigma_r^2 = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{n}$$

and is termed the variance in the values of y as **explained by the regression equation** of y on x. Incidentally, the total variation of y is equal to the sum of the **residual or error variation** due to fitted equation, and the **explained variation** by the regression equation.

Mathematically speaking

$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2 \qquad \textbf{(10.16)}$$

Total Variation = Residual /Error/Unexplained + Explained Variation

The above equation is also written in the form:

Total Sum of Squares = Error Sum of Squares + Regression Sum of Squares

Incidentally, the **Coefficient of Determination** measures the effectiveness or reliability of the regression analysis. It is defined as the **proportion of regression variation or explained variation to the total variation.** Being a square, its value is always positive, and being proportion, it is always less than or equal to 1. **Thus, the coefficient of determination lies between 0 and 1. Within this range, more the value better is the effectiveness or reliability of the regression equation.** Mathematically,

$$r^2 = \frac{\sigma_r^2}{\sigma_y^2} = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$= 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

$$= 1 - \frac{\sigma_e^2}{\sigma_y^2}$$

The standard error can also be written in terms of correlation coefficient as follows:

$$\sigma_e^2 = \sigma_y^2(1 - r^2) \qquad \textbf{(10.17)}$$

10.7 CROSS-SECTIONAL AND TIME SERIES CORRELATION ANALYSIS

Time series data is the data about a variable over a period of time; for example, the data about sales of a company over a period of, say, 10 years. Another example could be the data about average salary offered to MBA graduates of an Institute for the last, say 5 years.

Further, the cross-sectional data is the data about a variable taken at the same period of time over a number of units under study e.g. companies, banks, countries, individuals, etc. For example, data about sales of 10 companies, for the year 2006-07.

Another example could be the data about the salary offered to toppers of various specialisations in a Management Institute in the year 2006.

Whenever, we consider correlation between two variables, say between Sales Revenue and Advertising Expenses, it can be studied either for:

(i) Time series data i.e. data about sales revenues and advertising expenses for the same company over a period of time, or for

(ii) Cross-sectional data i.e. data about sales revenues and advertising expenses during the year 2006-07 for a number of companies.

While the results and conclusions for time series data are valid only for one company, the results and conclusions for cross-sectional data are valid for a group of companies at industry level. The results could be more meaningful if the group of companies belong to a specific sector like automobile, banking, telecom, etc.

For example, one could work out regression equation indicating relationship between, say, Expenditure on R&D and sales revenue for different sectors of industries, say, Pharmaceutical, IT, Telecom, etc.

Yet, another example could be the study of correlation in the prices of stocks on BSE and NSE. This can be done by studying the data about prices of a stock, say Reliance Industries on certain number of days, as given below:

Closing Prices of RIL on BSE and NSE

Date	*BSE*	*NSE*
5-10-2006	1155.05	1154.90
6-10-2006	1163.05	1162.85
9-10-2006	1154.10	1153.45
10-10-2006	1150.50	1150.90

(*Contd*)

(*Contd*)

11-10-2006	1143.20	1143.30
12-10-2006	1169.50	1168.75
13-10-2006	1190.15	1190.70
16-10-2006	1213.40	1215.20
17-10-2006	1216.05	121340
18-10-2006	1208.00	1208.50

The relationship between BSE and NSE stock prices could also be studied by taking into consideration the stock prices of each of a number of stocks, say 10 stocks, on a particular day, say 5th October 2006, and recording the data as follows:

Closing Prices of Some Stocks on BSE and NSE on 5th October 2006

Stock	*BSE*	*NSE*
Reliance	1155.05	1154.90
ONGC	1143.60	1141.75
Infosys Technologies	1865.70	1865.20
ICICI Bank	705.45	705.85
Bharti Airtel	459.10	459.95
Reliance Communication	348.30	349.60
HDFC	1452.95	1448.10
HLL	249.95	249.95
ITC	187.40	187.20
Tata Motors	894.75	893.70

This correlation between BSE and NSE prices gives an idea of correlation for a wider cross-section of industry.

10.8 BETA (β) OF A STOCK/SHARE

Beta is a statistical measure which reflects the sensitiveness of a stock to movement in the stock market index like BSE – SENSEX or NSE – NIFTY, as a whole. The beta value for the market index is taken as one. The 'beta' of a particular stock could be less than one or greater than one or even equal to one. A stock with beta more than one, say 2, would rise twice as much as the market index or would fall twice as compared to this index. Similarly, a stock with beta equal to 0.8 would rise by 80% of the rise in the market index or would fall by 80% of the fall in this index.

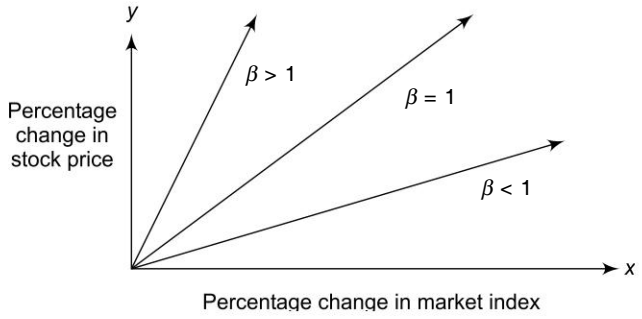
The volatility of a stock, in relation to the volatility of the market as a whole, is measured by its ***beta*** value (represented by the Greek letter b**)**. This statistical measure is the most popular indicator of risk associated with the stocks. An aggressive investor might invest in a stock with beta more than one with a view to beat the market when it goes up, but he does not mind losing more than the market index. A conservative investor, who wants to be protected against high downward swing in the market, might better opt for a stock with beta less than one.

It is measured by fitting a simple linear regression equation. The percentage daily/weekly/monthly change in the stock is taken as the dependent variable and percentage daily/weekly/monthly change

in the 'market' index such as BSE or NSE is taken as the independent variable. Thus the regression equation fitted to such data is of the form $y = a + b\,x$, as follows:

% daily change in the stock price = a + b (% daily change in the market index) (10.18)

A stock's ***beta*** measures the relationship between the stock's rate of return (variable y) and the average rate of return for the market as a whole (variable x). If ***beta*** is > 1, the stock is said to be 'aggressive'.



From the above regression equation (10.18), it may be noted that 'beta' of a stock is the covariance between the returns on a stock and index (like BSE or NSE), divided by the variance of index returns i.e.

$$\beta = \frac{\text{Covariance}\,(x, y)}{\text{Var}\,(x)}$$

where, x represents the index returns and y represents the stock returns.

The coefficient of determination i.e. r^2 derived from the data on percentage daily changes in a stock and percentage daily changes in market index provides a **measure of volatility explained in a stock's price by the market.**

Incidentally 'beta' values of stocks are available on the website of Stock Exchange Mumbai (BSE).

Example 10.3

The following data relates to the closing BSE Sensex and the stock price of RIL, for 10 trading days during the period from 5th October to 18st October 2006. Calculate the beta measure of the stock of RIL.

Date	*BSE*	*Stock Price of RIL*
5.10.06	12389	1155.05
6.10.06	12373	1163.05
9.10.06	12366	1154.1
10.10.06	12364	1150.5
11.10.06	12353	1143.2
12.10.06	12538	1169.5

(*Contd*)

(*Contd*)

13.10.06	12736	1190.15
16.10.06	12928	1213.4
17.10.06	12884	1216.05
18.10.06	12858	1208

The following table gives the percentage changes in BSE (x_i) as well as RIL (y_i) and the calculation requires for calculating covariance x_i, y_i and variance of x. It may be noted that the percentage change in BSE index on 6/10/2006 is worked out as:

{(the value of BSE index on 6/10/2006 – the index on 5/10/2006) × 100 ÷ value of BSE index on 5/10/2006}

The other values of x_i and y_i have been derived, accordingly

Date	*BSE* x_i	*RIL* y_i	x_i^2	$x_i y_i$
6.10.06	–0.1291	0.69261	0.016679	–0.08945
9.10.06	–0.0566	–0.7695	0.003201	0.043536
10.10.06	–0.0162	–0.3119	0.000262	0.005045
11.10.06	–0.089	–0.6345	0.007915	0.056451
12.10.06	1.49761	2.30056	2.242841	3.445346
13.10.06	1.5792	1.76571	2.49387	2.788411
16.10.06	1.50754	1.95354	2.27267	2.945028
17.10.06	–0.3403	0.21839	0.115836	–0.07433
18.10.06	–0.2018	–0.662	0.040724	0.133588
Sum	3.751339	4.552866	7.193997	9.253626
Average	0.416815	0.505874		

$$\beta = \frac{\text{Covariance}\,(x, y)}{\text{Var}\,(x)}$$

$$= \frac{\sum y_i x_i - n\bar{y}\bar{x}}{\{\sum x_i^2 - n\bar{x}^2\}}$$

Substituting the values of $\bar{y}$, $\bar{x}$, $\Sigma y_i x_i$ and Σx_i^2, in the formula, we get

$$\beta = \frac{9.2536 - 9 \times 0.4168 \times 0.5059}{7.194 - 9 \times (0.4168)(0.4168)}$$

$$= \frac{9.2536 - 1.8977}{7.194 - 1.559}$$

$$= 1.306$$

This implies that the RIL stock was 30.6% more aggressive than BSE SENSEX.

The objective of the above analysis is only to explain the calculations involved in determining the beta of the stock. The objective is not to draw any inference for RIL stock. For drawing any valid conclusions about a stock, the price of the stock is to be studied over a much longer period.

Incidentally, as per a study published in *Economic Times* dated 25th February 2007, there is a trend in the increase in correlation of Indian stock markets with the other stock markets in the world. It is mentioned therein that the link-up with equity indices is helping people take a call on how the Indian market will perform that day. The data containing the correlation of Indian markets with some of the other markets over the period 2000 to 2006 is given below:

Sticking Together

Correlation of Indian Markets with Other Markets

Year	*FTSE –UK*	*Nikkei –Japan*	*Dow Jones –US*	*NYSE COM –US*	*NASDAQ –US*	*Strait Times –Singapore*	*CAC40 –France*
2000	0.15	0.22	-0.07	0.01	0.07	0.28	0.1
2001	0.21	0.31	0.18	0.15	0.11	0.4	0.17
2002	0.14	0.2	0.03	0.01	0.03	0.22	0.11
2003	0.24	0.25	0.11	0.14	0.11	0.35	0.25
2004	0.24	0.3	0.14	0.18	0.16	0.42	0.27
2005	0.23	0.29	0.07	0.08	0.06	0.3	0.3
2006	0.37	0.39	0.1	0.23	0.11	0.53	0.38

It may be noted that the correlations in 2006 are more than the correlations in the year 2000.

10.9 ASSUMPTIONS IN USING REGRESSION EQUATION

The regression equation

$$y = a + bx$$

fitted to a data given in Table 10.1, could be written as

$$y_i = a + b\,x_i + e_i$$

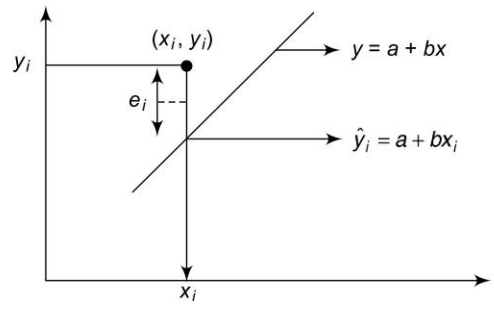
where, 'e_i' is the error term of the ith pair of observations (x_i, y_i). It is equal to the difference between observed and estimated value of y_i for the value of x_i. It indicates the error committed in estimating y with the help of the regression equation

$$y = a + bx$$

These error terms are positive or negative. They can even be zero for those points that lie on the regression equation, as illustrated in the following figure:

There are certain assumptions about these error terms that ought to hold good for the regression equation to be useful for drawing conclusions from it or using it for prediction purposes. These are:

(i) The distribution of $e_i s$ is normal
(ii) $E(e_i) = 0$
(iii) Var $(e_i) = \sigma^2$ for all values of i
(iv) $r(e_i, e_j) = 0$



The implication of the first assumption is that the errors are symmetrical with both positive and negative values. The second assumption implies that the sum of positive and negative errors is zero, and thus they cancel out each other. The third assumption means that the variance or fluctuations in all error terms are of the same magnitude. The fourth assumption implies that the error terms are uncorrelated with each other, i.e. one error term does not influence the other error term.

These are discussed in detail in Chapter 14 on Multivariate Statistical Techniques.

10.10 USING EXCEL

In this section, we explain the procedure for calculations with the help of MS Excel software.

Fitting a Regression Line

We have provided the templates in the CD with this book. These templates make the calculations easy. We would now explain the use of the template for 'Simple Correlation & Regression Analysis'.

In the snapshot shown on next page, the data is entered in column *A* for independent variable (x) and in column *B* for dependent variable (y). Once the data is entered, the calculations are carried out automatically by the template.

We may recall the Illustration 10.1 relating to Automakers. The same is solved above with the help of template. This template gives the correlation coefficient r, the coefficient of determination R^2, the scatter diagram, the values of $\hat{y}$ (estimated value of y) and the error (difference between the actual and the estimated value from the line of best fit.), the values of a and b in the regression equation line $y = a + b\,x$. It also gives the value of the dependent variable (in this case net profit), for a given value of independent variable (in this case Average Sales).

This template can be used for solving any problem on regression and correlation.

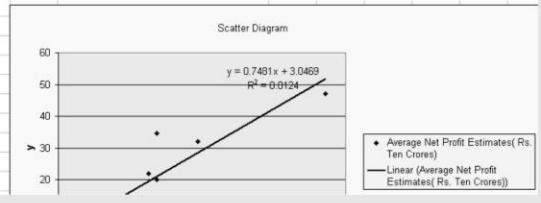
Beta of a Stock

We have also provided another template for finding 'Beta of Stocks/Shares'.

In this template, one has to enter the stock value of a share and the market index, say x and y, and the template would automatically calculate the Beta of stocks, correlation of two variables and also give the interpretation of Beta.

Microsoft Excel - Simple Correlation and Regression

	A	B	C	D	J	K	L	M	N	O	P	Q
1	*x*	*y*										
2	Average Sales Estimates (Rs. Millions)	Average Net Profit Estimates(Rs. Ten Crores)	$\hat{y}$	Error		Illustration 10.1 Automakers						
3	65	47	51.67493745	-4.674937448								
4	22	22	19.50562969	2.494370309								
5	24	34.5	21.00187656	13.49812344								
6	10	3.5	10.52814846	-7.028148457		Correlation Coefficient	r	0.901349				
7	5	6	6.787531276	-0.787531276								
8	16	9	15.01688907	-6.016889074		Coefficient of Determination	R^2	0.812429				
9	24	20	21.00187656	-1.001876564								
10	34	32	28.48311093	3.516889074		Regression Coefficients For Line y = a + b x						
11												
12						(y is dependent Variable and x is	a	3.046914				
13						independent Variable	b	0.748123				
14												
15						Estimated value of y for given value of x	x_i	45				
16							$\hat{y}_i$	36.71247				
17												
18												
19												
20												
21												
22												
23												
24												
25												
26												
27												



GLOSSARY

Scatter Diagram	Values of *x* and *y* depicted with the help of the rectangular co-ordinate system, by plotting the observed **pairs of values** of *x* and *y*
Correlation Coefficient	The quantitative measurement of the degree of linear relationship between two variables, say '*x*' and '*y*'
Coefficient of Determination.	The square of the correlation coefficient**.**
Spurious Correlation	When two variables change due to different factors and not due to interdependence on each other
Rank Correlation	Association between two variables where data is given in the form of the ranks of two variables based on some criterion
Regression Analysis	Study of the relationship among two or more variables
Simple Regression Analysis	Study of the relationship between two variables

Multiple Regression Analysis	Study of the relationship among more than two variables
Principle of Least Squares	Sum of squares of differences between the observed values and the estimated values is *minimum*
Beta of Stock	A statistical measure which reflects the sensitiveness of a stock to movement in the stock market index like BSE – SENSEX or NSE – NIFTY, as a whole
Beta	Regression coefficients of standardised independent variables
Concurrent Deviation	Measure of correlation that depends only on the sign (and not magnitude) of the deviations of the two variables, say x and y, recorded at the same point of time from their values at the preceding point of time

OBJECTIVE TYPE QUESTIONS

1. Pearson's correlation coefficient is a measure of:
   (a) Relationship between two variables
   (b) Linear relationship between two variables
   (c) Independence between two variables
   (d) Dependence between two variables
2. The correlation coefficient between two independent variables is
   (a) 0 (b) +1 (c) 0.5 (d) –1
3. If correlation between two variables is uncertain, it implies that
   (a) $r = +1$ (b) $r = -1$ (c) $r = 0$ (d) $r = 0.5$
4. Scatter diagram depicts:
   (a) variations in x and y
   (b) joint variation in x and y
   (c) variation in y for given values of x
   (d) variation in x for given values of y
5. If the correlation between x and y is positive:
   (a) y increases as x increases
   (b) y decreases as x increases
   (c) x decreases as y increases
   (d) x increases as y decreases
6. Which of the following values of correlation coefficient is likely to be correct, if the correlation is quite high?
   (a) 0.6 (b) –0.8 (c) +0.7 (d) +0.5
7. Which of the following values of correlation coefficient is likely to be correct, if the correlation is quite low?
   (a) 0.5 (b) 0.8 (c) –0.6 (d) –0.7
8. Which type of data is not suited for calculation of Pearson's correlation coefficient?
   (a) Nominal (b) Ordinal (c) Interval (d) Ratio

9. We should expect negative correlation while studying relationship between
   (a) Income and expenditure of a family
   (b) Price and demand of a product
   (c) Sales and profit of a company
   (d) None of above
10. Which value of r indicates highest degree of correlation?
   (a) + 0.5 (b) +0.8 (c) –0.2 (d) –0.9
11. Which values of r indicates lowest degree of correlation?
   (a) +0.1 (b) +0.9 (c) –0.2 (d) –0.7
12. If $r = 0.8$, the coefficient of determination is:
   (a) 0.81 (b) 0.64 (c) –0.64 (d) 0.49
13. If coefficient of determination is 0.81, the correlation coefficient is:
   (a) +0.9 (b) –0.9 (c) ±0.9 (d) None of the above
14. If the correlation coefficient between the sales of a company and promotional expenses are reported to be high, which of the following values of r is more likely to be true?
   (a) –0.9 (b) 0.6 (c) 0.85 (d) 0.8
15. In a study involving two variables, the product of the two regression coefficients is found to be equal to 1. This implies that:
   (a) $r = 0$ (b) $r = +1$ (c) $r = –1$ (d) $r = \pm 1$
16. In a study involving two variables, the two regression equations are found to be perpendicular to each other. It implies that:
   (a) $r = 0$ (b) $r = +1$ (c) $r = –1$ (d) $r = \pm 1$
17. If the regression equation of y on x is $y = 1 – 2x$, the correlation coefficient between y and x is:
   (a) Positive (b) Negative
   (c) Zero (d) Any one of the above
18. Which of the following assumptions is not valid in using a regression equation?
   (a) Distribution of error terms (e_i) is normal
   (b) Expected value of e_i is 0
   (c) Variance of e_i is σ_i^2 for each value of i
   (d) Correlation coefficient between two error terms is 0
19. If beta (β) of a stock is 1.2, it implies that:
   (a) The stock price rises by 120% when the market index rises by 100%
   (b) The stock price rises by 20% more than the market index
   (c) The stock price declines by 20% less than the market index
   (d) The stock price rises by 20% when market index rises by 20%
20. Which of the following could be termed as spurious correlation? Correlation between:
   (a) Sales revenue and expenditure on R&D
   (b) Net profit and share price
   (c) Net profit and number of products manufactured
   (d) Sales revenue and manpower/staff strength

EXERCISES

1. The following data was published by *Economic Times* in the publication ET 500 containing salient details of top 500 companies:

Sr. No.	*Name of the Company*	*Net Sales (Sept, 2005) (Rs Cr)*	*Net Profit (Sept, 2005) (Rs Cr)*	*P/E Ratios on 31st Oct., 2005*
1.	Infosys	7836	2170.9	32
2.	TCS	8051	1831.4	30
3.	WIPRO	8211	1655.8	31
4.	Bharti	9771	1753.5	128
5.	Hero Honda	8086	868.4	16
6.	ITC	8422	2351.3	20
7.	Satyam computers	3996	844.8	23
8.	HDFC	3758	1130.1	21
9.	Tata Motors	18363	1314.9	14
10.	Siemens	2753	254.7	38

Fit regression equations of (i) Net Profit on Net Sales, and (ii) P/E Ratio on Net Sales for group of these companies.

2. In the list of top 500 companies published as ET 500 in February 2006, the following are the ranks of the top ten companies according to their (i) Overall rating (ii) Market capitalisation, and (iii) Net Profit.

Name of the Company (Abbreviated)	*Over all Rank in February 2006*	*Rank as per Market Capitalisation in Oct., 2005 (Rank within these 10 companies)*	*Rank as per Net profit (Rank within these 10 companies)*
Infosys	1	1	2
TCS	2	2	3
WIPRO	3	4	5
Bharti	4	3	4
Hero Honda	5	5	1
ITC	6	8	8
Satyam Computers	7	9	9
HDFC	8	6	7
Tata Motors	9	7	6
Siemens	10	10	10

Calculate

(i) the rank correlation coefficients between the overall rank and market capitalisation rank
(ii) the overall rank and rank as per Net Profit, and
(iii) market capitalisation rank and rank as per Net Profit.

3. The following data gives the closing prices of BSE Sensex, and the stock prices of three individual companies viz. ICICI Bank, L&T and Reliance Industries Ltd. during the 10 trading days during the period from. 6th to 21st March 2006.

Date	*BSE*	*ICICI Bank*	*Reliance Industries*	*L&T*
6-3-2006	10735	613.20	731.90	2413.40
7-3-2006	10725	600.65	731.85	2493.80
8-3-2006	10509	590.55	719.45	2439.40
9-3-2006	10574	601.75	726.70	2442.35
10-3-2006	10765	612.90	732.15	2512.65
13-3-2006	10804	603.10	732.35	2536.25
16-3-2006	10878	607.50	768.50	2507.35
17-3-2006	10860	605.25	774.85	2466.75
20-3-2006	10941	605.40	776.20	2461.45
21-3-2006	10905	597.80	780.30	2466.05

(a) Find the following correlation coefficients between the stock prices of the companies and comment.
  (i) ICICI Bank and Reliance Industries
  (ii) ICICI Bank and L&T, and
  (iii) L&T and Reliance Industries

(b) Calculate the 'Beta' measures of all the three stocks, and comment.

4. A company's past records contain the following data relating to sales revenue and expenditure on advertisements for six years, as follows:

Year	*Sales Revenue (Rs Crores)*	*Advertising Expenditure (Rs Crores)*
2001	125	15
2002	132	16
2003	145	20
2004	150	21
2005	160	23
2006	170	25

Calculate the appropriate regression equation, and estimate the sales in the next year when the advertisement expenses are budgeted as Rs 30 Crores.

5. A company wanted to assess the consistency between two HRD executives who were to recruit MBA students for summer placements. They were asked to assess the 12 trainee executives recruited from the last batch, and give their rankings. The rankings given by the two executives are as follows:

Trainee Executive	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Executive 1	1	11	8	2	12	10	3	4	7	5	6	9
Executive 2	4	12	11	2	5	10	1	3	9	8	6	7

Find the correlation coefficient and comment on the result.

6. The following data gives expenditure on R&D and profit of a company

(in Crores of Rs)

Profit	*Expen diture on R&D*
50	0.40
60	0.40
40	0.30
70	0.50
85	0.60
100	0.80

(a) Find the regression equation of profit on R&D Expenditure.
(b) Estimate the profit when expenditure on R&D is budgeted at Rs 1 Crore.
(c) Find the correlation coefficient.
(d) What proportion of variability in profit is explained by variability in expenditure on R&D.

7. Following are the ranks of ten different banks on the basis of customer satisfaction and their market share of deposits.

Customer satisfaction	3	2	1	5	6	4	7	8	9	10
Market Share	1	3	4	5	6	2	10	7	8	9

Find the correlation coefficient and comment.

8. The following sample data shows the demand for a product in thousands of units and its price (in Rs) charged in six different market areas:

Price	:	10	18	14	11	16	13
Demand	:	125	58	90	100	72	85

(i) Fit a least square line from which we can predict the demand for the product in terms of its price
(ii) Estimate the demand for the product in a market where it is priced at Rs 15
(iii) Determine the percentage of the variation in the demand for the product that is due to differences in price.

9. The following data gives the I.Q. and marks in Statistics of 6 students

I.Q.	:	100	130	110	125	115	120
Marks	:	70	90	70	83	76	75

(a) Find the regression equation of marks on I.Q.
(b) Estimate the marks of a student whose I.Q. is 120
(c) Find correlation co-efficient between I.Q. and Marks.
(d) What proportion of variability in Marks in explained by the variability in I.Q.

10. A personnel manager is interested in assessing whether the performance on the job and the academic score in a management programme are correlated. The following data gives the academic scores and the scores on the job performance of a random sample of 12 executives.

Sample No.	*Academic Score*	*Score on Job*	*Sample No.*	*Academic Score*	*Score on Job*
1	65	37	7	91	46
2	72	54	8	72	47
3	78	42	9	56	30
4	84	58	10	92	52
5	89	44	11	68	32
6	52	40	12	77	50

Calculate the correlation coefficient between job performance and academic score in management programme. Also, find the coefficient of determination, and offer comments.

Regression Analysis: Using Weather to Predict Sales*

Tesco, the U.K.'s largest grocery chain, has set up its own weather team to forecast temperatures and its relationship with consumer demand. The team has created its own software based on the weather and shopping patterns of 12 British regions over the last three years. It calculates how shopping patterns change "for every degree of temperature and every hour of sunshine." Thus, Tesco uses the forecasts to help it reduce costs and avoid wasting food.

*Based on a reprot published in N Y Times dt. 1st Sept. 2009.

Statistical Inference

11

Contents

LEARNING OBJECTIVES

This chapter provides the requisite knowledge and expertise to

- Understand the two aspects of statistical inference, viz. 'Estimation' and 'Testing of Hypothesis', based on a sample of observations from a population. While estimation involves estimating some parameter of a population like 'mean life of a brand of car battery', testing of hypothesis implies testing of some assumption about the population like whether the mean life of a brand of car battery is 3 years?
- Understand the properties of good estimators, and scope of estimation with respect to accuracy and confidence in an estimate. It is to be appreciated that an estimate based on a sample from population cannot be 100% accurate, and one cannot swear by it with 100% confidence. A compromise is stuck by defining confidence intervals with certain extent of accuracy, say ±5%,

and also on the confidence, say 99%. in the statement that the true value lies within ±5% of the sample estimate. Thus a typical conclusion could read like "I am 99% confident that the true value of population mean lies within ±5% of the sample mean."

- Understand the types of errors committed while drawing conclusions with the help of a sample from a population about some assumption made about the population parameter. While Type-I error is "to reject an hypothesis when it is true", the Type-II error is to "accept a hypothesis when it is false". These errors can only be minimised but not eliminated.
- Conduct various tests of significance. A test of significance implies the procedure for testing whether the assumption made about the population parameter is true or not. This is based on the analysis of observations in the sample. There are many tests of significance, like (i) Mean life of a brand of car battery is 3 years, (ii) Proportion of defectives in a manufactured lot of electric bulbs is 2%, (iii) the mean lives of car batteries manufactured in two different plants are equal (iv) Whether the expenses on R&D have any impact on the sales of a company? (v) Whether the daily sales of a retail outlet follow any particular pattern or distribution, etc.?

Relevance

The new General Manager of Everbright Light Company manufacturing tube lights is concerned about the dwindling profits of the company. The main reason is that the company provides a guarantee of 1 year of life, and undertakes to replace a tube light if it fails within one year. Since a good number of tube lights are failing in less than a year and are being replaced free of cost, they are lowering the company's profitability and also causing loss of reputation. The General Manager intuitively feels that the guaranteed life must be such that the percentage of tube lights failing within that period is quite small; say 5% or 10%, so as to keep the cost of replacement low. Since it may not be appropriate to reduce the guaranteed life, the only alternative is to increase the life of the tube light. After careful consideration, he outlines the following steps:

- Estimate the average life of tube lights, as well as the variation in their lives.
- Take action to increase the life of tube light with the help of improved technology and better management of the production process.
- Test whether the actions taken have increased the life, and by how much?
- Fix the price and guarantee period in such a way as to ensure adequate increase in profits.

The subject of statistical inference, as described in this chapter, could play a useful role in these steps.

11.1 SIGNIFICANCE AND INTRODUCTION

As mentioned earlier, one of the objectives of Statistics is to use the information contained in a small sample of observations for drawing a conclusion or making an inference about the larger population. This area of Statistics is known as **'Statistical Inference'**. It involves inductive approach that is making inferences about population parameters like mean and standard deviation in the Normal distribution or the proportion/percentage in the Binomial distribution. Such inference may be in the form of **Estimation** or **Testing** of certain **Assumption** or **Hypothesis.** For example, either one could estimate the average life of light bulbs, produced by a company, or one could test the claim of the

company that the average life of bulbs is, say 2,500 hours. Similarly, either one could estimate the percentage of defective bulbs produced in the factory, or one could test the claim of the factory that the percentage of defective bulbs is less than or equal to 5%. As another example, suppose an investor is looking for avenues to invest his recently acquired wealth into mutual funds and stocks of reputed companies. After discussions with friends, he decides to invest into those stocks and funds which:

(a) Provide higher yield/returns as compared to other stocks, funds and the overall market.
(b) Are less volatile as compared to other stocks and funds. This criteria is meant to minimise risk.

Statistical inference provides adequate tools to facilitate decision-making in the above situation.

In this chapter, we shall describe various methods of estimation, as also various methods of testing hypotheses. We shall also discuss the criteria evolved to measure their 'goodness'. This helps in differentiating among various estimators or tests of hypotheses, and selecting the most appropriate ones. However, before discussing the theoretical aspects, we indicate below, the relevance of estimation and testing of hypothesis in business environment.

The Reliable Company is doing extremely well in terms of growth in sales and profits. However, the Head of Finance feels that the company can generate more profits if the ever-growing expenditure on advertisements is reduced. The Head of Marketing feels that one of the main reasons for the growth of the company is the aggressive advertising campaigns. To resolve this contentious issue, Statistical Inference can be used to create an analytical approach that assesses the impact of advertisements on sales, and provides various scenarios of sales with various levels of expenditure on advertisement.

The new Chairman of Evershine Detergent Company realised that the sales force in the company was not performing up to the desired standards. He asked the Human Resources Department to organise a two-week comprehensive training programme to increase the marketing skills as well as motivation of the sales force. The company obtained two proposals from two Management Institutes with about the same financial implications. The Chairman felt that before giving the entire assignment to one particular Institute, the HR Department could organise two training programmes, one by each of the two Institutes. Based on the comparative effectiveness, evaluated with the help of statistical inference, of the two programmes, the contract could then be awarded to one of the Management Institutes.

A pharmaceutical company has developed a medicine for curing insomnia. However, before introducing it in the market, the company would like to test the effectiveness of the medicine on a certain number of patients. Such tests could be designed, analysed and evaluated with the help of statistical inference.

11.2 ESTIMATION

The topic of estimation in Statistics deals with estimation of population parameters like mean of a statistical distribution. It is assumed, that the concerned variable of the population follows a certain distribution with some parameter(s). For instance, it may be assumed that the life of the electric bulbs follows a normal distribution which has two parameters viz. mean (m) and standard deviation (σ). While one of the parameters, say, standard deviation is known to be equal to 200 hours from

past experience, the other parameter, viz. the mean life of the bulbs, is not known, and which we wish to estimate.

Given a sample of observations x_1, x_2, x_3, …, x_n, one is required to determine with the aid of these observations, an estimate in the form of a specific number like 2500 hrs., in the above case. This number can be taken to be the best value of the unknown mean. Such single value estimate is called **'Point'** estimate. The estimation could also be in the form of an interval, say 2,300 to 2,700 hrs. This can be taken to include the value of the unknown mean. This is called **'Interval Estimation'**. An example of point and interval estimation could be provided from our day-to-day conversation when we talk about commuting time to office. We do make statements like "It takes about 45 minutes ranging from 40 to 50 minutes depending on the traffic conditions." The statistical details of these two types of estimation are described below.

11.3 INTERVAL ESTIMATION

So far, we have seen that the estimate is a single value. This is called point estimation, and is the best estimate of the parameter from the collected data. However, if another data is collected, the point estimate may change. In fact each sample could lead to a different estimate. Now the dilemma is as to which estimate should be taken as the real estimate of the population parameter which is unknown.

For example, suppose there is a population of 30 units as follows, and we wish to estimate its mean:

8	10	14	9	12	7	6	14	13	8	9	11	12	14	13
11	7	11	14	10	9	10	8	10	7	10	9	11	13	12

The population mean, say m, is

$$m = \frac{\textbf{Sum of all 30 observation}}{\textbf{Number of observations}} = \frac{\textbf{312}}{\textbf{30}}$$

$$= \textbf{10.4}$$

Let three samples of 10 observations be taken from this population as follows:

Sample I:	8	9	6	8	12	11	14	10	7	11
Sample II:	10	12	14	9	13	7	10	8	10	13
Sample III:	14	7	13	11	14	11	9	10	9	12

The estimate of the population mean 10.4 is provided by the three sample means which can be worked out as 9.6, 10.6 and 11.0. Thus, while in real life we have no alternative but to accept the estimate obtained from the sample as an estimate of the population mean, we should realise that the population mean may not be equal to the sample mean, and could be around this value. Thus it may be more logical to assume that the population value lies in an interval containing the sample value. This is where the relevance of interval estimation comes into picture. **In the interval estimation, instead of trying to pinpoint a single value for the unknown parameter, we derive an interval that is expected to include the true value of the parameter with the desired level of confidence; hence, the name 'Confidence Interval'.**

Ideally, one would like the interval to be as narrow as compared to a wide interval. For instance, in the above example, if one gives the interval as 5 to 15, it may not serve much purpose. Obviously,

the width of the interval has to be within reasonable limits. This is where the subject of Statistics plays a role in finding out these limits. The intervals or limits, so arrived, are referred to as **confidence intervals** or **confidence limits**. The word confidence implies the degree of confidence one has that the population value would lie in the interval or lie within the limits. This aspect is discussed, in detail in the next Section.

11.3.1 Factors Affecting the Width of a Confidence Interval

The width of the confidence interval depends on several factors described below:

(i) Sample Size

The width of the interval is inversely proportional to the sample size. More the sample size, the narrower would be the interval. In the extreme case, when sample size becomes equal to population size, the sample parameter like mean is the same as the population mean and, therefore, the width of the interval is zero.

(ii) Variability in the Population

The width of the interval is directly proportional to the variation in the data. Thus, more the variability in the data more would be the width of the interval.

(iii) Confidence Desired

So far a sample is used to estimate population parameters, one can never make any estimation with 100% confidence, and one has to settle for a confidence less than 100%. Usually, 95% confidence level is considered adequate allowing 5% of margin. For the justification or relevance of 5% margin, one may recall that in day-to-day life while comparing if there is a little difference, one makes a statement "Oh! There is a difference of only 19-20" implying thereby that a difference of 5% is acceptable. However, if the situation so desires, one may like to have even a confidence level of 99%. But, as we shall see later, more the level of confidence, wider will be the confidence interval. Since a wider interval, like "life of a car battery is between 2 and 6 years", does not serve much purpose, a compromise has to be made in deciding the level of confidence.

11.3.2 Confidence Intervals for Various Parameters

The parameters for which the confidence intervals have been discussed in this chapter are:

(i) Mean (Normal Distribution)
(ii) Proportion (Binomial Distribution)

Confidence Interval for Mean: (population standard deviation i.e. σ known) Suppose, one wants a confidence interval estimation for the mean of a variable which follows the normal distribution with mean m and standard deviation σ.

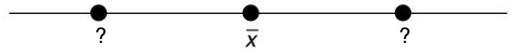
The value of the estimator, the sample mean, is given by

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

The distribution of $\bar{x}$ is normal with mean 'm' and s.d. $\frac{\sigma}{\sqrt{n}}$.

Now we want the confidence interval such that we are 95% confident that the actual mean would lie within that range. This interval has to be around the estimated value $\bar{x}$, and is derived below.

We assume that the interval is symmetric around $\bar{x}$ as follows:



We want that the absolute difference between $\bar{x}$ and the population mean m i.e. $|\bar{x} - m|$ is to be less than a value, say d, with probability 0.95 i.e.

$$P(|\bar{x} - m| \leq d) = 0.95$$

It may be derived that the 95% confidence interval for mean is

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \text{ to } \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

The interpretation of this confidence interval is that if there is a variable x which is normally distributed with mean m and s.d σ, and a sample mean of n observations of the variable is calculated as $\bar{x}$, then we are 95% confident that the mean m would lie in the above interval.

Similarly, the 99% confidence interval can be derived by finding out the values of z in the standard normal curve, such that the area between those points is 0.99. From Table T1, relating to area under standard normal curve, we see that the values are –2.575, and +2.575.

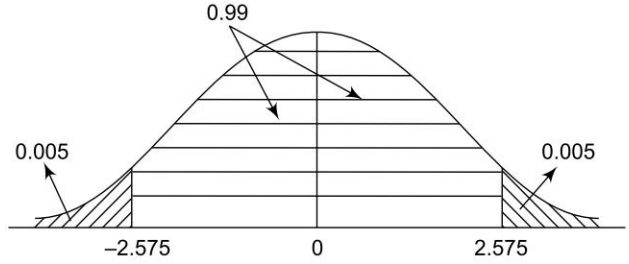


Fig 11.1 *Area under Standard Normal Curve*

Thus, the 99% confidence interval is

$$\bar{x} - 2.575\frac{\sigma}{\sqrt{n}} \text{ to } \bar{x} + 2.575\frac{\sigma}{\sqrt{n}}$$

The other two commonly used confidence intervals are:
The 90% Confidence Interval:

$$\bar{x} - 1.645\frac{\sigma}{\sqrt{n}} < m < \bar{x} + 1.645\frac{\sigma}{\sqrt{n}}$$

and the 50% confidence interval is

$$\bar{x} - 0.6745\frac{\sigma}{\sqrt{n}} \text{ to } \bar{x} + 0.6745\frac{\sigma}{\sqrt{n}}$$

In general, (100 – α)% confidence interval is derived by finding those values of z, under standard normal curve, such that they contain (100 – α)% of the area. Such values of z are written as $z_{\alpha/2}$. The sum of area up to $-z_{\alpha/2}$ and area beyond $z_{\alpha/2}$ is equal to α.

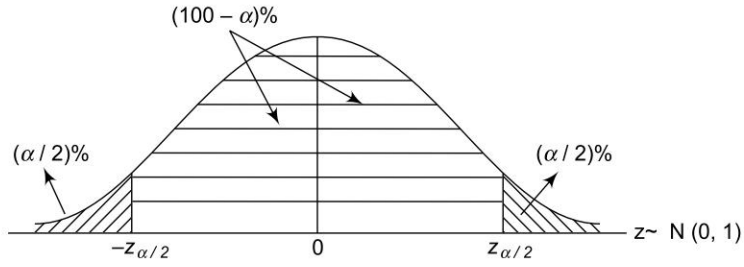


Fig. 11.2 *Area under Standard Normal Curve*

In general, (100 – α)% confidence interval for population mean is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \textbf{(11.1)}$$

where,

n is the sample size
σ is the s.d of the population
$z_{\alpha/2}$ is the point on the standard normal curve area beyond which is α/2%
$(1 - \alpha)$ is the level of confidence.

It may be noted that most samples taken in practice are without replacement. In such cases, if population is finite, i.e. if sample size is greater than 5% of population size, then the standard error of mean is $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$. The term $\sqrt{\frac{N-n}{N-1}}$ is called as **correction factor**. For infinite population, this term approximates to 1, leading standard error to be $\frac{\sigma}{\sqrt{n}}$. This is explained in Chapter 4, Section 4.9.6.

The calculation of confidence intervals is illustrated through examples given below:

Confidence Interval for Mean (population standard deviation σ known):

Example 11.1

In order to introduce some incentive for higher balance in savings accounts, a random sample of size 64 savings accounts at a bank's branch was studied to estimate the average monthly balance in savings bank accounts. The mean and standard deviation were found to be Rs. 8,500 and Rs. 2000, respectively. Find (i) 90%, (ii) 95%, (iii) 99% confidence intervals for the population mean.

Solution:

Confidence limits with confidence level (100 – α)% (or (100 – α)% confidence limits) for average monthly balance in savings accounts are given as:

(i) 90% confidence limits:

$$\bar{x} \pm 1.645\frac{\sigma}{\sqrt{n}}$$

$$8500 \pm 1.645\frac{2000}{\sqrt{64}}$$

$$8500 \pm \frac{3290}{8}$$

Rs. **8500 ± 411.25**

(ii) 95% confidence limits:

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

$$8500 \pm 1.96\frac{2000}{\sqrt{64}}$$

$$8500 \pm \frac{3920}{8}$$

Rs. **8500 ± 490**

(iii) 99% confidence limits,

$$\bar{x} \pm 2.575\frac{\sigma}{\sqrt{n}}$$

$$8500 \pm 2.575\frac{2000}{\sqrt{64}}$$

$$8500 \pm \frac{5150}{8}$$

Rs. **8500 ± 644**

It may be noted that the interval or limits get wider as the desired level of confidence is increased.

Confidence Interval for Mean (population standard deviation, σ, unknown) In the above examples, the value of population standard deviation was given. If not given, it has to be estimated from the sample itself. In that case, however, the distribution of $\bar{x}$ is not normal but student's '*t*' distribution. So, instead of referring to the Table for normal distribution, we refer to the Table T2 for '*t*' distribution.

Thus the $(100 - \alpha)\%$ confidence interval for the population mean is:

$$\bar{x} - t_{\alpha/2,\,(n-1)}\frac{s}{\sqrt{n}} \text{ to } \bar{x} + t_{\alpha/2,\,(n-1)}\frac{s}{\sqrt{n}} \qquad \textbf{(11.2)}$$

where, $\bar{}$ and s are the mean and the standard deviation of the observations in a sample of size n, and $t_{\alpha/2,\,(n-1)}$ is the value of student's '*t*' with $(n - 1)$ d.f., area beyond which, on either side is $\alpha/2$.

Example 11.2

For assessing the number of monthly transactions in credit cards issued by a bank, transactions in 25 cards were analysed. The analysis revealed an average of 7.4 transactions and sample standard deviation of 2.25 transactions. Find confidence limits for the monthly number of transactions by all the credit card holders of the bank?

Solution:

When population standard deviation is not known, and values of sample mean and sample standard deviation are given, the 95% confidence interval for the population mean is given by

$$\bar{x} - t_{\alpha/2,\,(n-1)}\frac{s}{\sqrt{n}} \text{ to } \bar{x} + t_{\alpha/2,\,(n-1)}\frac{s}{\sqrt{n}}$$

where $\bar{x}$ and s are the sample mean and sample standard deviation, respectively, and $t_{\alpha/2,\,n-1}$ is the point in the t distribution with $(n-1)$ d.f. such that the total area, on both the sides of the curve, beyond which is equal to α. For example if $\alpha = 0.05$ and $n = 25$ (d.f. = 25 – 1 = 24), the value of 't' is 2.06 vide Table T2 relating to 't' distribution

Since, level of confidence is not given, we assume it to be 95%, and thus $\alpha = .05$. It is given that $\bar{x} = 7.4$, $s = 2.25$, $n = 25$ and $t_{\alpha/2,\,n-1} = t_{.025,\,24} = 2.06$.

Thus the confidence limits for the average number of transactions per credit card are

$$\bar{x} \pm 2.06\frac{2.25}{\sqrt{25}}$$

or, $$7.4 \pm \frac{4.635}{5}$$

or, $$7.4 \pm 0.927$$

or, **6.473 to 8.327**

For large values of *n*, say (≥ 30), '*t*' distribution may be approximated by a normal distribution. Thus, the limits are given by the formula 11.1, using *s* in place of σ.

Confidence Interval for Proportion The confidence intervals for the proportion are worked out on the assumption that a proportion follows a Binomial distribution. If the proportion in the population is p_o, the sample proportion $\hat{p}$ follows the Binomial distribution with mean p_o and s.d. as $\frac{p_o q_o}{\sqrt{n}}$

where $q_o = 1 - p_o$, and n is the size of the sample or the number of observations in the sample is n. Since the tables showing the probabilities for various values of p and n are not feasible (would require many tables for different values of n), we use the property that a Binomial distribution approaches normal distribution as the value of n become large and p is not near to 0 or 1. Thus, if $\hat{p}$ is the proportion in a sample of size n, then

$$\frac{\hat{p} - p_o}{\sqrt{p_o q_o / n}} \text{ is distributed as N (0, 1)}$$

where, p_0 is the proportion in the population.

In general, the confidence interval for the population proportion at α level of significance is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad \textbf{(11.3)}$$

Thus the 95% confidence interval for the population proportion is

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} \text{ to } \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Since p_o is not known its estimate $\hat{p}$ is used in the expression, and thus the limits are

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} \text{ to } \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

The calculation of limits is illustrated through the following numerical example:
An insurance company policy sells foreign travel policy to those going abroad. The company in reputed to settlle the claims within a period of 2 months. However, the new CEO of the company came to know about the delay in settling claims. He, therefore, ordered the concerned official to take a sample of 100 claims, and report the proportion of cases which were settled within 2 months. The CEO received the proportion as 0.6. What are the 95% confidence limits for such proportion?
Here, $n = 100$, and $\hat{p} = 0.6$.
Therefore, 95% confidence limits are

$$0.6 - 1.96\sqrt{\frac{0.6 \times 0.4}{100}} \text{ to } 0.6 + 1.96\sqrt{\frac{0.6 \times 0.4}{100}}$$

or, $0.6 - 0.096$ to $0.6 + 0.096$
or, **0.504 to 0.696**
Thus, the CEO could be 95% confident that about 50 to 70% of claims are settled within 2 months.

It may be noted that most samples taken in practice are without replacement. In such cases if population is finite i.e. if sample size is greater than 5% of population size, then the standard error of proportion is $\sqrt{\frac{\hat{p}\hat{q}}{n}} \times \sqrt{\frac{N-n}{N-1}}$. The term $\sqrt{\frac{N-n}{N-1}}$ is called as finite population correction factor. For infinite population this term approximates to 1, leading standard error to $\sqrt{\frac{\hat{p}\hat{q}}{n}}$. This is explained in Chapter 4, Section 4.9.6.

11.4 SAMPLE SIZE REQUIRED TO ESTIMATE A PARAMETER WITH DESIRED CONFIDENCE AND ACCURACY

As we have discussed above, the confidence interval for a parameter like mean and proportion, depends on stipulated level of confidence like 95%, margin of error or tolerance like 5%, variation present in the population size as measured by s.d., and sample size. In this section, we discuss the methodology for deciding the sample size which is required to meet the criteria for level of confidence and margin of error or tolerance. **The sample size, so determined is the minimum size as any sample size more than this value would automatically satisfy all the criteria.** We discuss sample size required only for two parameters viz. mean and proportion.

11.4.1 Sample Size Required for Estimating Mean

In this case, the problem involves finding the sample size required to be confident to the extent of (100 – α%) that the sample mean will not deviate from the population mean by more than specified margin, say 4.

Illustration 11.1

A company wants to determine the average time to complete a certain job. The past records show that the s.d. of the completion times for all the workers in the company has been 10 days, and there is no reason to believe that this would have changed. However, the company feels that because of the procedural changes, the mean would have changed. Determine the sample size so that the company may be 95% confident that the sample average remains within ±2 days of the population mean.

The company wants to be 95% confident that the difference between the sample mean ($\bar{x}$) and actual population mean (m) should be ≤ 2. This can be mathematically expressed as

$$P\{|\bar{x} - m| \leq 2\} = 0.95$$

(dividing both sides of the inequality within brackets by $\sigma/\sqrt{n}$, we get)

$$P\left\{\left|\frac{\bar{x} - m}{\sigma/\sqrt{n}}\right| \leq \frac{2}{\sigma/\sqrt{n}}\right\} = 0.95$$

Since $\dfrac{\bar{x} - m}{\sigma/\sqrt{n}}$ is distributed normally with mean 0 and s.d. 1, usually represented by z, we have

$$P\left(|z| \leq \frac{2}{\sigma/\sqrt{n}}\right) = 0.95$$

From, standard normal distribution Table T1, giving area under the curve, we see that the value of z area beyond which is 2.5% or 0.025 is 1.96 i.e.

$$P(|z| \leq 1.96) = 0.95$$

Therefore,

$$\frac{2}{\sigma/\sqrt{n}} = 1.96$$

or,

$$2\sqrt{n} = 1.96\sigma$$

Squaring both sides and substituting value of σ as 10, we get,

$$4n = (1.96)^2 \times (100)$$

or

$$\boldsymbol{n = 96.04}$$

Thus, a sample size of at least 97 is required to be 95% confident that the difference between the sample and population means will be less than 2.

11.4.2 Sample Size Required for Estimating Proportion

In this case, the problem involves finding the sample size required to be confident to the extent of (100 – α%) that the sample proportion will not deviate from the population proportion by more than specified margin, say 2%.

Illustration 11.2

The production manager of a manufacturing company wants to asses the percentage of items which do not meet the specifications, and are thus labeled as defective. He wants to be 95% confident that the percentage has been estimated to be within 1% from the true value. What is the most conservative sample size needed for the situation?

The given criteria can be expressed in a equation form as follows:

$$P\{|\hat{p} - p_0| \leq \text{margin}| = 1 - \alpha$$

where p_0 the true value of proportion defective, margin is the amount of accepted or tolerated difference between the true and estimated proportion, and $1 - \alpha$ is the level of confidence required.

$$P\{\hat{p} \leq 0.01\} = 0.95$$

$$P\left\{\frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0 q_0}{n}}} \leq \frac{0.01}{\sqrt{\frac{p_0 q_0}{n}}}\right\} = 0.95$$

$$P\left\{|z| \leq \frac{0.01 \times \sqrt{n}}{\sqrt{p_0 q_0}}\right\} = 0.95$$

(As mentioned earlier in Section 11.3, $(\hat{p} - p_0)/\sqrt{(p_0 q_0)/n}$ is distribution as $N(0, 1)$)

From the property of normal distribution, we know that the value of z is 1.96

Therefore,
$$1.96 = \frac{.01}{\sqrt{\frac{p_0 q_0}{n}}}$$

or,
$$1.96\sqrt{\frac{p_0 q_0}{n}} = .01$$

Squaring on both sides, we have,

$$(1.96)^2 \frac{p_0 q_0}{n} = 0.0001$$

or,
$$n = (1.96)^2 \frac{p_0 q_0}{0.0001}$$

For most conservative estimate of n i.e. the least value of n for which the criteria is satisfied, the value of the product $p_0 q_0$ has to be maximum. It can be shown mathematically that the value of the product of two fractions is maximum when both the fractions are equal to ½. The maximum value of $p_0 q_0$ is thus equal to ½ × ½ = ¼.

It is also shown in the table given below.

p_0	q_0	$p_0 \times q_0$
0.1	0.9	0.09
0.2	0.8	0.16
0.3	0.7	0.21
0.4	0.6	0.24
0.5	**0.5**	**0.25**
0.6	0.4	0.24
0.7	0.3	0.21
0.8	0.2	0.16
0.9	0.1	0.09

Thus

$$n = \frac{3.8416 \times (1/4)}{0.0001}$$

$$= 9604$$

Thus the minimum or the most conservative sample size for ascertaining the percentage of defectives to be within ±1% of the true value with 95% confidence is 9604.

It may be verified that if the margin is increased from 1% to 3%, the required sample size comes down to 1067.

It may also be verified that if the margin remains the same i.e. 1%, and if the level of confidence is increased from 95% to 99%, the requisite sample size increases to 16,641.

Formulas for Calculation of Sample Size In fact, the value of the sample size can be readily found from the following formulas:

(i) For Mean:

$$n \quad \left(\frac{z_{\alpha/2} \times \sigma}{\text{Margin of Error}} \right)^2 \qquad \textbf{(11.4)}$$

where, $z_{\alpha/2}$ is the value of z, total area beyond which on both sides is α.

Illustration 11.3

For the given data as follows:

Level of Confidence	99%
Margin of Error	40
σ	200

the sample size

$$n \quad \left(\frac{2.575 \times 200}{40} \right)^2$$

166

Thus, if we want to estimate mean of a population whose s.d. is 200, within a margin of 40 of the population value, the minimum sample size required is 166.

(ii) For Proportion:

$$\boldsymbol{n} \quad \left(\frac{z_{\alpha/2} \times \sqrt{p(1-p)}}{\text{Margin of Error}} \right)^2 \qquad \textbf{(11.5)}$$

where, $z_{\alpha/2}$ is the value of z, total area beyond which on both sides is α

p is the value of unknown proportion in the population

Illustration 11.4

For the given data as follows:

Level of Confidence	95%
Margin of Error	2%

the sample size

$$n \left(\frac{1.96 \times \sqrt{0.5 \times 0.5}}{0.02} \right)^2$$

2401.

Thus, if we want to estimate the proportion in a population, within a margin of 2% of the population value, then the minimum sample size required is 2401.

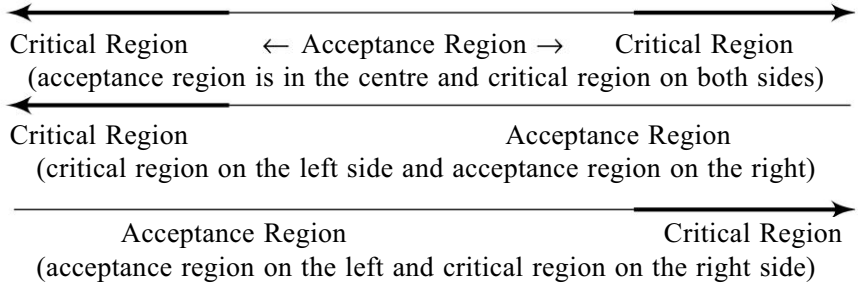
11.5 TESTING OF HYPOTHESIS OR TESTS OF SIGNIFICANCE

In the earlier sections, we discussed the issues relating to estimation of parameters like mean and proportion of a population. In this section, we shall discuss the other aspect of statistical inference viz. testing of hypothesis or assumption about a parameter like mean and proportion of a population.

A statistical hypothesis is an assumption about any aspect of a population. It could be the parameters of a distribution like mean of a Normal distribution, describing the population, the parameters of two or more populations, correlation or association between two or more characteristics of a population like age and height, etc.

In fact, hypothesis is an integral part of any research or investigation. Many a time, initially experiments or investigations are carried out to test an hypothesis, and the ultimate decisions are taken on the basis of the collected information and the result of the test.

To test any statistical hypothesis on the basis of a random sample of n observations, we divide the n-dimensional sample space into two regions. If the observed sample point $X(x_1, x_2, x_3, \ldots, x_n)$, falls into a region called **critical or rejection region**, the hypothesis is rejected, but if the sample falls into the complementary region (total sample space minus critical region), the hypothesis is accepted. The complementary region is also called the **acceptance region.** However, for a better understanding of the concept, the entire region is mapped on to a straight line which is then divided into two parts viz. acceptance and rejection regions as follows:



It may be noted that there is no lower limit or upper limit for critical or rejection regions.

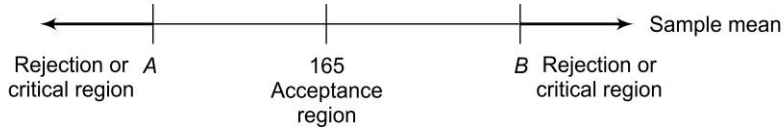
As an illustration, suppose, one makes a statement that a particular High School has tall students, and the average height of its Class X students is 165 cms. One alternative, to test this statement, is to measure the height of all the High school students and accept or reject the statement. However, if the time and other resources to measure the height of all students are not available, only a sample could be drawn from the 'population' of all High school students, and decision taken to accept or reject the hypothesis based on this sample. Now, let a sample of size, say 25, be taken from the total population of, say of 250 students. Obviously the common sense approach would be

to calculate the average height of the students in the sample and compare with the stated average of 165.

Now it is too much to accept that even if the statement is true, the average height of the sample would also be 165 cms.

Even without the knowledge of Statistics, the common sense approach dictates that if the sample mean is close to 165 cms. we can accept the hypothesis that the whole school mean is 165, but if it is not so close, we can reject the hypothesis.

The following diagram can illustrate this. If the sample means lies between A and B – we can call it the '**acceptance region',** we accept the hypothesis. However, if the sample mean lies beyond this region i.e. either less than A or greater than B—this region can be called the **rejection or critical region**—we reject the hypothesis.



Now the problem is to find out the values of A and B. Here the commonsense fails to specify the exact values of A and B. This can be done with the help of the theory of Statistical Inference.

Before, we proceed to find out the values of A and B, a few points have to be considered. First of all, whenever we take a decision about a population (number of students in Class X, in this case) based on a sample, the decision cannot be 100% foolproof or reliable. The following possibilities exist:

(i) The hypothesis might be true i.e. the average height of all Class X students is 165 cms but the 25 students in the sample could be relatively shorter, and, therefore, their average may work out to less than A or they could be relatively taller, and their average may work out to more than B. In such a situation, we would reject the hypothesis even if it is true. This type of judgmental error is called **Type–I error**.

(ii) The hypothesis might be false i.e. the average height of the Class X students is not 165 cms. but the 25 students in the sample could be such that their average height works out to be in the region from A to B, and, therefore, we would accept the hypothesis that the average height of Class X students is 165 cms. This type of judgmental error is called **Type–II error**.

Incidentally, in military applications such as radar or sonar, Type-I error is called a 'miss' (e.g. an enemy missile has been missed by the detection system), Type-II error is called a 'false alarm' (e.g. detection system is sensing an enemy missile even though it is not there).

We can represent various possibilities in decision making about a population from a sample with the help of the following diagram.

		Hypothesis	
		True	False
Decision	Accept	Right Decision	**Type–II Error**
	Reject	**Type–I Error**	Right Decision

Whenever, a decision about a population is based on a sample, these two errors cannot be eliminated. These can only be minimised. However, it is not possible to reduce both the errors simultaneously, for the same sample size; the moment we want to reduce one error, the other gets increased. This is further explained in Section 11.5.5.

It is, therefore, customary, while evolving various tests, to fix the Type-I error and minimise the Type-II error. All the tests of significance discussed in this book are based on this premise.

As a convention in Statistics, Type-I error is always denoted by α, and Type-II error by β. The quantity 1- β is called the 'power' of a test, signifying its ability to reject a hypothesis when it is false, α is also referred to as level of significance, and 1 – α is called confidence coefficient.

Type-I Error as 5%

It is customary to fix Type - I error, in most cases, as 5%. Sometimes, it is taken as 1%, or some other %. However, in Statistics, unless otherwise stated, it is taken as 5%. This convention started in UK/USA. It is interesting to note that, in India, it is quite common in Hindi/Urdu speaking regions to make a statement about similarity of quality, while comparing qualities of two items, "*Oh!, bus unnis-bees ka fark hai*". In English, the statement means that it is just a difference between 19 and 20, signifying that a difference of 5% is considered negligible or tolerable! Incidentally, it may be noted that on the computer key board, 5 and % are on the same key!

11.5.1 Examples of Type-I and Type-II Errors from Indian Epics

There are two excellent and most appropriate situations from our epics, viz. "Abhigyan Shakuntalam"and "Mahabharata", where Type-I and Type-II errors were committed. These are described below:

Type-I and Type-II Errors from Indian Epics

It may be recalled that in "Abhigyan Shakuntalam", King Dushyanta had married Shakuntala when he met her in her village, while wandering in a jungle. He gave her his royal ring as a gift which could also serve as her identity when she would come to meet him, in future. However, while going to meet him, she lost the ring in the river. When she reached Dushyant's palace and met him, he failed to recognise her especially since she did not have the ring. Thus Dushyanta committed Type-I error as he rejected Shakuntala as his wife when, in fact, she was his true wife.

In Mahabharata epic, Dronacharya—the 'guru' of both Pandavas and Kauravas–was fighting from the Kaurva's side. However, he had taken a vow that he would stop fighting if and when his son Ashwathama was killed in the war. It so happened that during the war, one elephant named Ashwathama was killed. Lord Krishna—the mentor for Pandavas –thought of a strategy to make Dronacharya lay down his arms. Yudhishthir, on the advice of Lord Krishna, went to Dronacharya and pronounced *"Ashwathama*

hato–naro va kunjaro?" **i.e. Ashwathama was dead–but was it a human or an elephant? Dronacharya, on listening the first part of Yudhishthir's sentence, presumed that his son was dead, and he left for his heavenly abode without waiting to listen to the second part of Yudhishthir's sentence. Thus, Dronacharya could be said to have committed Type-II error i.e. accepting a statement when it was not true.**

11.5.2 Level of Significance for a Test

It is used as a criterion for rejecting the null hypothesis. It is expressed as a percentage like 5% or 1%, or sometimes as 0.05 or 0.01. If the probability of finding the difference between the observed value of the statistic and the value of statistic as per the null hypothesis, defined in the next section, exceeds the specified level of significance, we conclude that the difference is significant; otherwise we conclude that the said difference is insignificant. Level of significance is also the probability of committing Type-I error. **Thus, specifying level of significance as 5% implies fixing Type – I error as 5%.**

Incidentally, the terminology Level of Significance is said to have been introduced by R.A. Fisher who is said to have recommended its value to be taken as 5%.

11.5.3 Methodology of Carrying Out Tests of Significance

First of all, the population and its characteristic have to be defined. For instance, in the above example relating to height of students, the population is the total of 250 students of Class X in the school, and the characteristic is the height of students. The testing procedure has to be evolved allowing the Type-I error to be fixed, say 0.05 or 5%.

Secondly, we have to set up the hypothesis. Suppose, we want to test the statement that the average height of Class X students is 165 cms. The given statement or the statement to be tested is designated as **Null Hypothesis**. The word Null is used because the nature of testing is such that we try our best to nullify this hypothesis on the basis of the sample collected, and if we do not find sufficient evidence from the sample to do so, we have no alternative but to accept it. It is represented by $\boldsymbol{H_0}$; H with zero as subscript.

This is analogous to a situation when the police present an 'accused' before the judge for prosecution. The judge starts with the presumption that the accused is innocent. The police collect and present relevant facts and evidences to 'nullify' the assumption of the judge. But if the police is not able to do so, they have no alternative but to accept the presumption of the judge that the accused is innocent.

Incidentally, both the judicial system and Statistics concentrate on disproving the null hypothesis rather than proving the alternative hypotheses. While an attempt is made to reject the null hypothesis on the basis of evidence, if the evidence is not sufficient to reject the null hypothesis, the null hypothesis is accepted. If the null hypothesis does not get rejected, the alternative hypothesis is accepted.

Historically, the term 'Null' hypothesis is said to have been introduced by R.A. Fisher and the notation 'H_0' for specifying null hypothesis said to have been introduced by E.S. Pearson and J. Neyman.

After setting up the null hypothesis, one has to set up an **alternative** hypothesis i.e. a statement which is intended to be accepted if the null hypothesis is rejected. It is denoted by $\mathbf{H_1}$. In the above case relating to height of students, the alternative statement could be that the average height is not 165 cms. These hypotheses could be written as:

Null Hypothesis : H_0 : $m = 165$ cms.
Alternative Hypothesis : H_1 : $m \neq 165$ cms.

Obviously, one of the two hypotheses is to be accepted based on the calculations from the values obtained through the sample.

In this case, let the sample values (heights of 25 students selected in the sample) be:

161	163	168	162	163
165	164	170	168	167
167	166	166	169	166
169	160	166	169	164
165	168	172	166	166

Now, an estimate of the population mean is obtained from this sample. Thus

$$\bar{x} = \frac{\sum x_i}{25} = 166$$

where x be the variable representing the characteristic i.e. height of students in this case, and x_i the height of the i^{th} student in the sample.

Further, let us assume that the height of the students follows the normal distribution with mean m (which is unknown) and standard deviation as 3. That is,

$$x \sim N(m, 3^2)$$

As mentioned in Section 10.3 of this chapter, the sample mean $\bar{x}$, based on sample size n, is distributed as Normal with mean m and s.d. as $3/\sqrt{n}$ i.e. $3/\sqrt{25}$ = **3/5**, in the above case.

Now the test statistic is formed as follows:

$$z = \frac{\bar{x} - 165}{3/5}$$

It may be noted that the test statistic z is formed by subtracting assumed population mean from sample mean i.e. $(\bar{x} - 165)$, and then dividing the same by the standard deviation of $\bar{x}$. Thus, z is a standardised variable with mean 0 and s.d. as 1. This is abbreviated as $z \sim N(0, 1)$. Further, let the Type-I error be fixed as 0.05 i.e. 5%

The **acceptance** and **critical** regions are derived as follows. It may be recalled that if a variable x has mean m and s.d. σ, then the sample mean $\bar{x}$ of n values of this variable is distributed normally with mean m and s.d. $\sigma/\sqrt{n}$. Thus, the variable

$$\frac{\bar{x} - m}{\sigma/\sqrt{n}}$$

is distributed normally with mean 0 and s.d. 1. Similarly, the above variable z is also distributed normally with mean 0 and s.d. as 1.

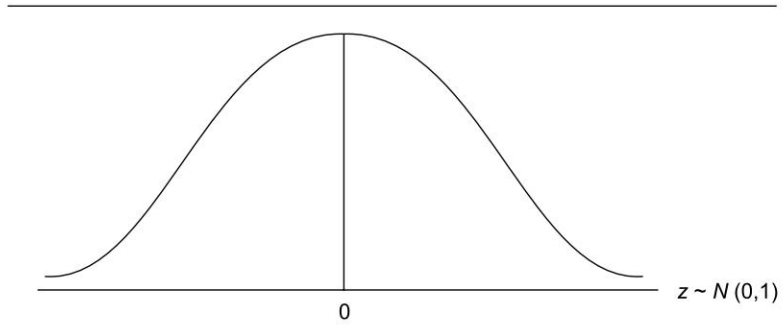


Fig 11.2 *Standard Normal Curve*

Now the calculated value of z is

$$\frac{166-165}{3/5} = \frac{5}{3} = 1.67$$

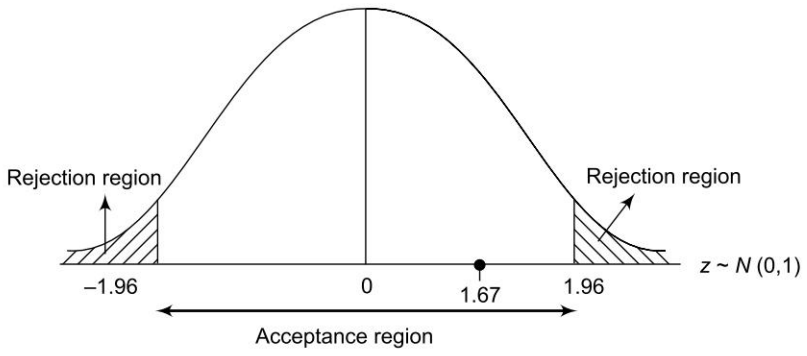


Fig 11.3 *Acceptance and Critical/Rejection Regions*

This value of z i.e, 1.67 lies in the acceptance region.

Thus the null hypothesis cannot be rejected. This implies that the statement that the average height of the students of Class X is 165 cms. cannot be rejected. This also implies that we have not collected sufficient evidence to nullify the hypothesis that the average height of the students of class X is 165 cms, and therefore, cannot reject it. The interpretation of the difference between $\bar{x}$ and 165 is that it is due to the random or chance factors arising due to sampling.

Illustration 11.5

A new Production Manager of a manufacturing unit making high intensity light bulbs, wants to improve the quality of bulbs as measured by their lives. He, therefore, decides to estimate the life of bulbs, in the existing system, and states that he would like to bring about changes in the manufacturing process unless the average life of bulbs, as evidenced by continuous burning of bulbs, is more than 650 hours. He is informed that the standard deviation of the life of bulbs is 25 hours.

A sample of 100 bulbs is selected from a week's production, and the average life of these bulbs works out to be 670 hrs. Does this indicate that the quality of bulbs meets the criteria stipulated by the Production Manager?

Let us assume that the life of the bulbs represented by the random variable x follows the normal distribution with mean 650 hrs. and standard deviation 25, and the sample of 100 bulbs is from the same population. It reduces to the statement that the mean value of the life, in the population (from which the sample is selected) is 650 and its s.d. is 25.

Now, to set up the null hypothesis and the alternative hypothesis, we note that the null hypothesis is that the average life is 650 hrs. and the alternative hypothesis, desired by the Production Manager, is that the average life is more than 650 hrs. We may add that even though H_0 is that $m = 650$, actually because of H_1 being $m > 650$, what it implies is that $m \leq 650$. Thus, if H_0 is rejected, it means that the average life of bulbs is > 650 hrs. Common sense dictates that for H_0 to be rejected and H_1 accepted, the average life of sample bulbs has to be more than 650 hrs. If sample average life is less than 650, the test need not be even conducted, and one could accept H_0 or reject H_1.

Thus we have, in this case

$$H_0: \quad m = 650 \text{ hrs.}$$
$$H_1: \quad m > 650 \text{ hrs.}$$

The alternative hypothesis is set up depending on the requirement of the situation. Here, the manager wants the assurance that the average life of all the bulbs manufactured by the unit i.e. population mean is more than 650 hrs, and, therefore, the alternative hypothesis H_1 is that m exceeds 650 hrs.

The statistic to be used in this case is

$$z = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}}$$

Its sampling distribution is distributed normally with mean 0 and s.d. as 1.
From the sample values, it is given that

$$\bar{x} = 670, \quad m_o = 650, \quad n = 100, \quad \sigma = 25$$

Thus, we have

$$z = \frac{670 - 650}{25/\sqrt{100}} = \frac{20}{25/10} = \frac{20}{2.5}$$

$$= 8$$

Since z is a standard normal variable with mean 0 and s.d. 1, the point for critical region at 5% level of significance is the point the area to the right of which is 5%. Referring to the Table T1, giving area under the standard normal curve, we note that this point, as shown below, is 1.645.

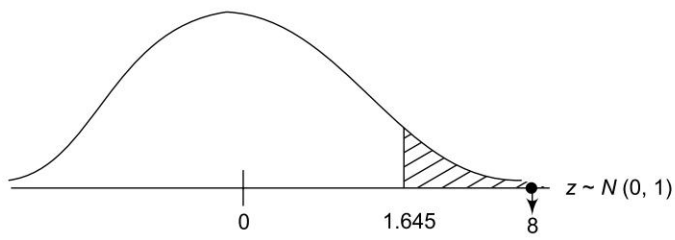


Fig 11.4 *Area under Standard Normal Curve*

Now since the calculated value of z is more than the tabulated value, it falls in the critical region. Therefore, the null hypothesis is to be rejected i.e. the mean of the bulbs is not 650 hrs, and the alternative hypothesis i.e. the mean life is greater than 650 hrs. is to be accepted. Thus the Production Manager should be convinced that the average life of the bulbs being made by the existing production process is up to his level of expectation, i.e. exceeding 650 hrs.

11.5.4 Steps for Conducting Tests of Significance for Mean

The methodology explained in the previous section is summarised into a number of steps to be followed for carrying out a test of significance for mean. In fact, the same steps are followed for conducting all tests of significance.

(i) Set up the Null Hypothesis It is in the form

$$H_o : \quad m = m_o$$

m_0 is the value which is assumed or claimed for the population characteristic. It is the reference point against which the alternative hypothesis is set up, as explained in the next step. However, sometimes, H_1 is set up first, and the form of H_0 is decided accordingly. This is explained below.

(ii) Set up the Alternative Hypothesis It is in one of the following forms

$$H_1 : m \neq m_o$$

or, $$H_1 : m > m_o$$

or, $$H_1 : m < m_o$$

One has to choose from the above three forms depending on the situation posed, as explained below.

In the example relating to the heights of students discussed above in Section 11.5.2, the situation involved only testing the statement made about the average height of Class X students. Therefore, the alternative was of the form $\boldsymbol{m \neq m_o}$.

In the example about the life of bulbs discussed above, the Production Manager wanted to test whether the average life was more than 650 hrs. Therefore the alternative hypothesis was of the form $\boldsymbol{m > m_o}$.

In yet another example, the manufacturer of a reputed brand of cigarettes ordered that the nicotine content in a cigarette should not exceed the stipulated level of 30 mgs. In order to check this, he selected a sample of 200 cigarettes from the lot to be packed, and found that the average nicotine content was 28.5 mgs. Could he be reasonably sure that the stipulations laid down by him were being met? In this case, the manufacturer wanted the nicotine content to be less than a particular value, and therefore, the alternative hypothesis is of the form $m < \boldsymbol{m_o}$.

(iii) Decide the Level of Significance Usually, it is fixed as 5%, or sometimes 1%; if one wants to decrease the chance of rejecting when it is true. However, other values of the level of significance like 2%, 3% etc, are also possible.

(iv) Decide on the appropriate Statistic like *z*, above.

(v) Indicate the Critical Region The critical region is formed based on following factors:

(a) Distribution of the Statistic i.e, whether the statistic follows the normal, 't', χ^2 or 'F' distribution.

(b) Form of alternate hypothesis. If the form has $\neq$ sign, (e.g. $m \neq m_0$), the critical region is divided equally in the left and right tails/sides of the distribution.

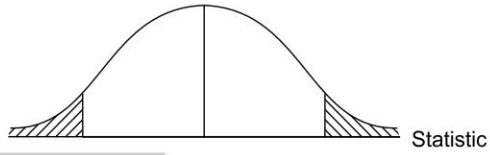


Fig 11.5 *Two Sided/Tailed Critical Region*

If the form of alternative hypothesis has < sign (e.g. $m < m_o$), the entire critical region is taken in the left tail of the distribution.

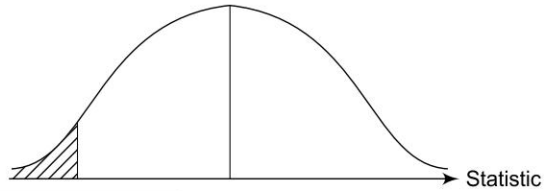


Fig 11.6 *One Sided (left) Critical Region*

If the form of alternative hypothesis has > sign (e.g. $m > m_o$), the entire critical region is taken on the right side of the distribution.

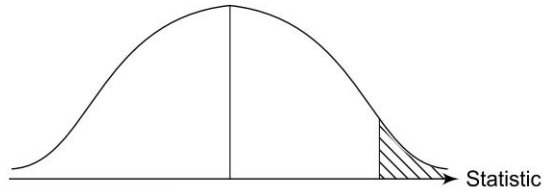


Fig 11.7 *One Sided (right) Critical Region*

(vi) Ascertain Tabulated Values Find out the tabulated values of the statistic based on the value of the level of significance and indicate the critical region—it can be one-sided, i.e, the entire region is on one side or it can be both sided as shown below.

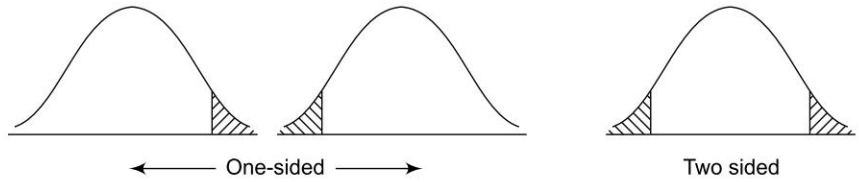


Fig 11.8 *Critical Regions*

(vii) Calculate the value of the statistic from the given data It is to be emphasised that the data should be collected only after setting up the hypotheses. **Further, a statistic is always calculated on the assumption that the null hypothesis is true.**

(viii) Accept or Reject the Null Hypothesis If the calculated value of the statistic falls in the critical region, reject the null hypothesis; otherwise accept the null hypothesis.

All the above steps are explained with the help of several examples given below. However before proceeding further, we would like to explain as to why the two types of errors viz. Type I and Type II cannot be reduced simultaneously through an Illustration in Section 11.5.5.

Incidentally, it may be noted that the null hypothesis is in the form of an equation like $m = 10$, or inequality like $m \leq 10$ or $m \geq 10$. The alternative hypothesis, can, however, be in the form of either

not equal to ($\neq$), less than ($<$), or greater than ($>$).

11.5.5 Simultaneous Reduction of Two Types of Errors

It was mentioned in Section 11.5, that both Type I and Type II errors cannot be reduced simultaneously, for the same sample size. This is illustrated below:

Suppose there is a population whose s.d. (σ) is known to be equal to 10, and we want to test the hypothesis that its mean is 30 against the alternative that it is > 30. The hypothesis set up is:

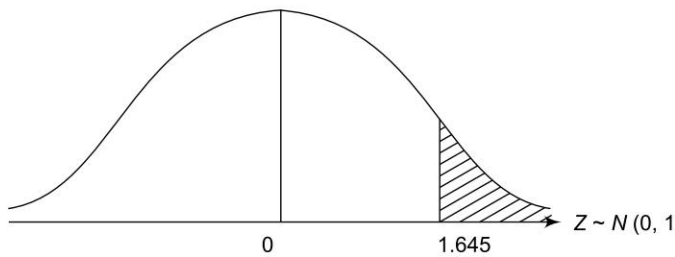
$$H_o: \quad m = 30$$
$$H_1: \quad m > 30$$

The test statistic is

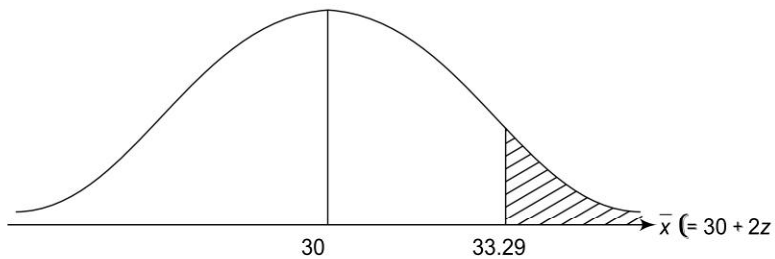
$$z = \frac{\bar{x} - 30}{\sigma/\sqrt{n}}$$

$$= \frac{\bar{x} - 30}{10/5} \quad \text{(assuming } n = 25\text{)}$$

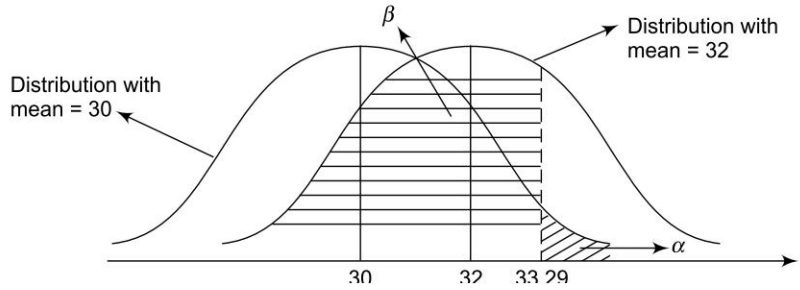
If the Type I error is fixed as $\alpha = 0.05$, the critical region (shaded) is as shown below:



It can also be shown as below in the diagram for distribution of $\bar{x}$.



Now, suppose the mean is actually 32 so that the distribution is as shown below



Now, even if the alternative hypothesis is true, i.e. $m = 32$, if the sample mean is less than 33.29, we will accept the null hypothesis i.e. we will commit Type II error. Thus the shaded region marked by horizontal lines gives the Type II error i.e. β. The two diagrams given above indicate that if we reduce α, i.e. shaded area marked by slanting lines, the β i.e. shaded area marked by horizontal lines increases.

11.6 VARIOUS TESTS OF SIGNIFICANCE

Various tests of significance have been developed to meet various types of requirements in social, economic, financial, operational and business environments, etc. We have selected the following tests for discussion in this chapter.

- Z-test based on Normal distribution
- 't'-test based on student's 't' distribution
- Chi-square test based on the χ^2 distribution
- 'F'-test based on Fisher's 'F' distribution

However, we have classified the tests, described below, according to parameters. All these tests are based on the assumption that the observations are drawn from a normal distribution. The tests of significance developed for situations when this condition is not satisfied are given in the next Chapter i.e. 13 on Non-Parametric or Distribution Free Tests.

(i) Mean (population s.d. σ known)

We have illustrated three tests for three different types of alternative hypotheses, viz.:

(a) Not equal to m_o
(b) Greater than m_o
(c) Less than m_o

(a)

$$\boldsymbol{H_0}: \quad m = m_0$$
$$\boldsymbol{H_1}: \quad m \neq m_0$$

$$z = \frac{\bar{x} - m_o}{\sigma\sqrt{n}} \qquad \text{which is distributed as } N(0, 1) \qquad \textbf{(11.6)}$$

If $|z| \leq 1.96$ Accept H_0
> 1.96 Reject H_0

Example 11.3

A random sample of 100 students from the current year's batch gives the mean CGPA as 3.55 and variance 0.04. Can we say that this is same as the mean CGPA of the previous batch which was 3.5?

Solution:

Since we have to test the hypothesis that the population mean is equal to 3.5, the null and alternative hypotheses are set up as

$$H_0: \quad m = 3.5$$
$$H_1: \quad m \neq 3.5$$

Even though population variance is not given, but the sample size is so large that the sample variance can be taken to be equal to population variance i.e. 0.04. Thus, the test statistic used is

$$z = \frac{\bar{x} - m_o}{\sigma\sqrt{n}} \qquad \text{which is distributed as } N(0,1)$$

We are given,

$$\bar{x} = 3.55,\ m_0 = 3.5,\ n = 100 \quad \text{and} \quad \sigma = \sqrt{0.04} = 0.2,$$

Therefore,

$$z = \frac{3.55 - 3.5}{0.2/\sqrt{100}} = \frac{0.05}{0.2/10} = \frac{0.05}{0.02}$$

$$= 2.5$$

Since, the calculated value of $z(2.5)$ is more than the tabulated value of z, at 5% level of significance, 1.96, the null hypothesis is rejected, as shown below:

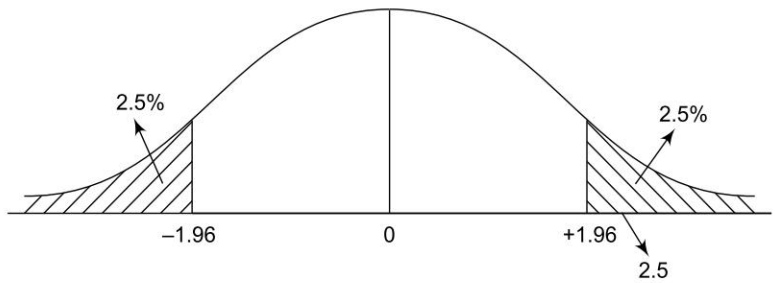


Fig 11.9

Thus the current year's mean CGPA is not the same as the last year's mean CGPA.

(b)

$$H_0: \quad m = m_o$$
$$H_1: \quad m < m_o$$

Example 11.4

It has been found from experience that the mean breaking strength of a brand of thread is 500 gms. with a s.d. of 40 gms. From the supplies, received during the last month, a sample of 16 pieces of thread was tested which showed a mean strength of 450 gms. Can we conclude that the thread supplied is inferior?

Solution:

In this case

$$H_0: \quad m = 500$$
$$H_1: \quad m < 500$$

The test statistics is $z = \dfrac{\bar{x} - m_0}{\sigma/\sqrt{n}}$ which is distributed as $N(0,1)$

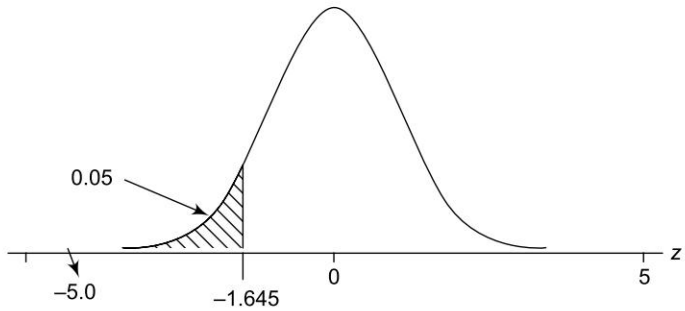
We are given that

$$\bar{x} = 450,\ m_0 = 500,\ n = 16 \quad \text{and} \quad \sigma = 40,$$

Therefore,

$$z = \frac{450 - 500}{40/\sqrt{16}} = \frac{-50}{40/4} = \frac{-50}{10}$$

$$= -5.0$$



Here the test statistic falls in the rejection region, and hence we reject the null hypothesis i.e. the sample indicates that the thread is inferior.

(c) $\boldsymbol{H_0}: m = m_o$

$\boldsymbol{H_1}: m > m_o$

Example 11.5

A telephone company's records indicate that individual customers pay on an average Rs 155 per month for long-distance telephone calls with standard deviation Rs 45. A random sample of 40 customers' bills during a given month produced a sample mean of 160 for long-distance calls. At 5% significance, can we say that the company's records indicate lesser mean than the actual i.e. actual mean is more than 155 mts.?

Solution:

Here the two hypotheses are set up as follows.

$$\boldsymbol{H_0}: \quad m = 155$$
$$\boldsymbol{H_1}: \quad m > 155$$

The test statistic is

$$z = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}}$$ which is distributed as $N(0,1)$

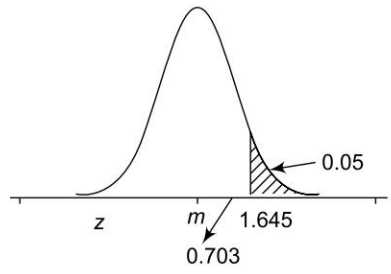
We are given that

$$\bar{x} = 160,\ m_0 = 155,\ n = 40 \quad \text{and} \quad \sigma = 45,$$

Therefore,

$$z = \frac{160 - 155}{45/\sqrt{40}} = \frac{5}{45/6.325} = \frac{5}{7.115}$$

$$= 0.703$$



Here, the test statistic falls in the acceptance region. Hence we do not reject the null hypothesis i.e. there is no evidence to infer that records indicate lesser mean than the actual.

(ii) Mean (σ unknown)

Here again, we have illustrated three tests for three different types of alternative hypotheses viz.

(a) Not equal to m_o
(b) Greater than m_o
(c) Less than m_o

(a) $\quad \boldsymbol{H_0}: \ m = m_0$

$\quad \boldsymbol{H_1}: \ m \neq m_0$

The test statistic formed is:

$$\text{'}t\text{'} = \frac{\bar{x} - m_o}{s/\sqrt{n}} \sim t_{\alpha/2,(n-1)} \qquad \textbf{(11.7)}$$

where, s is the sample s.d, i.e,

$$s = \sqrt{\left\{\frac{\Sigma(x_i - \bar{x})^2}{n-1}\right\}}$$

If the calculated value of t lies in the critical (shaded) region then the null hypothesis is rejected, otherwise it is accepted.

Example 11.6

A sample size of 10 drawn from a normal population has mean as 31, and variance as 2.25. Is it reasonable to assume that the mean of the population is 30? Assume $\alpha = 0.01$.

Solution:

This is the situation corresponding to testing the hypothesis that the population mean has a specified value m_0 when s.d. (σ) of the population is not known. The appropriate test is the students 't' test with:

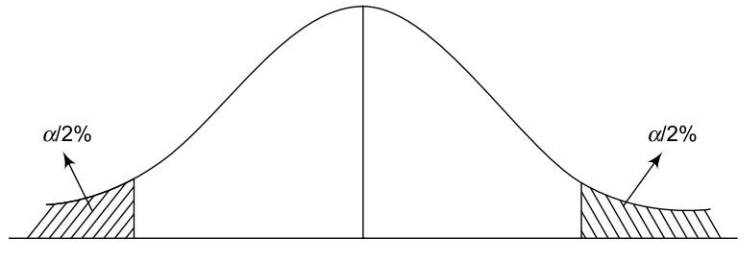


Fig 11.10 *Area Under Student's 't' Distribution*

$$H_0: \quad m = 30$$
$$H_1: \quad m \neq 30$$

The test statistic is

$$t = \frac{\bar{x} - m_0}{s/\sqrt{n}} \qquad \sim \text{students '}t\text{' distribution with } (n-1) \text{ d.f.}$$

where,

$\bar{x}$ is the sample mean, and

s is the sample standard deviation

In the given case

$$n = 10$$
$$\bar{x} = 31$$
$$s = \sqrt{2.25} = 1.5$$

Thus,

$$\text{'}t\text{'} = \frac{31-30}{1.5/\sqrt{10}} = \frac{1}{0.474} = 2.11$$

The tabulated value of 't' with $(n-1) = (10-1) = 9$ d.f. at 1% level of significance vide Table T2 is 3.25.

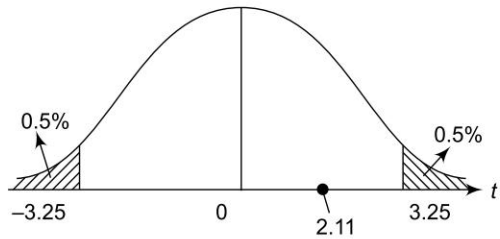


Fig 11.11 *Area Under 't' Distribution*

Since the calculated value 2.11 is less than 3.25 and falls in the acceptance region, we do not reject the null hypothesis H_0 that the population mean age of the whole batch is 30. It is therefore reasonable to assume that the population mean is 30.

If the sample size is large say ≥ 30, '*t*' distribution may be approximated by the normal distribution and instead of *t* statistic, one may use *z* statistic = $\dfrac{\bar{x} - m_0}{s/\sqrt{n}} \sim N(0, 1)$

(iii) Proportions:

There are three tests for three different types of alternative hypotheses, viz.:

(a) Not equal to p_o
(b) Greater than p_o
(c) Less than p_o

(a) $H_0 : \; p = p_o$
$H_1 : \; p \neq p_o$

Calculate $\hat{p} = \dfrac{r}{n}$

$$z = \frac{|\hat{p} - p_0|}{\sqrt{\dfrac{p_0 q_0}{n}}} \qquad z \sim N(0, 1) \qquad \textbf{(11.8)}$$

If $z \leq 1.96$ Accept H_0
If $z > 1.96$ Reject H_0

(b) $H_0 : \; p = p_o$
$H_1 : \; p > p_o$

(c) $H_0 : \; p = p_o$
$H_1 : \; p < p_0$

Example 11.7

A manufacturer of LCD TV claims that it is becoming quite popular, and that about 5% homes are having LCD TV. However, a dealer of conventional TVs claims that the percentage of homes with LCD TV is less than 5%. A sample of 400 household is surveyed, and it is found that only 18 households have LCD TV. Test at 1% level of significance whether the claim of the company is tenable.

Solution

Let p be the proportion of defective parts.

In this case, we take the null hypothesis as the claim of the company is 5%, that is to be tested against the alternative that the percentage of LCD TV is less than 5%. Thus, the null and alternative hypotheses are set up as follows:

$H_0 : \; p = 0.05$
$H_1 : \; p < 0.05$

The appropriate test statistic to be used is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}} \sim \text{N}(0,1)$$

In this given situation

$$p_o = 0.05$$
$$n = 400$$
$$q_o = 1 - p_o = 0.95$$

$$\hat{p} = \frac{18}{400}$$

$$= -0.045$$

Therefore,

$$z = \frac{0.045 - 0.05}{\sqrt{\frac{0.05 \times 0.95}{400}}}$$

$$= \frac{-0.005}{0.011}$$

$$= -0.4545$$

Now, since z is a standard normal variable, we find the value of z such that the area under the standard normal curve beyond that point is 1%.

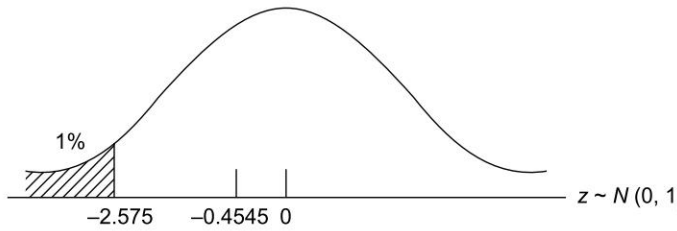


Fig 11.12 *Critical Region Under Normal Curve*

Since the calculated value of z (–0.4545) is more than the tabulated value (–2.575) and falls in the acceptance region, we conclude that the null hypothesis is not to be rejected. Thus, the company's claim is tenable.

11.6.1 Tests of Significance Involving Two Populations

So far, we have discussed tests of significance relating to mean and proportion from one population. Now, we discuss the tests relating to means and proportions from two populations.

Equality of Two Means For testing the equality of means of two populations, it is assumed that their variances are equal. We discuss below two tests, one for the situation when the two variances are known, and other for the situation when the two variances are unknown.

Let,

n_1 : be the sample size from first population.
n_2 : be the sample size from second population.
$\bar{x}_1$: be the mean of first population
$\bar{x}_2$: be the mean of other population

σ : be the pooled s.d. of both populations, and defined as

$$\sigma : \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2}}$$

σ_1^2 : being the variance of first population

σ_1^2 : being the variance of second population

(a) $\boldsymbol{H_0}$: $m_1 = m_2$

$\boldsymbol{H_1}$: $m_1 \neq m_2$

($\sigma_1 = \sigma_2 = \sigma$ known)

The test statistic is

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad z \sim N(0, 1) \qquad \textbf{(11.9)}$$

Assuming level of significance to be 5%, accept H_0, if $|z| \leq 1.96$, otherwise, reject H_0, and accept H_1 .

(b) $\boldsymbol{H_0}$: $m_1 = m_2$

$\boldsymbol{H_1}$: $m_1 \neq m_2$

($\sigma_1 = \sigma_2$ but unknown)

Since σ_1 and σ_2 are unknown, these have to be estimated from the samples. However, it is assumed that the variances in the two populations are equal, say σ. The estimate of this is obtained by 's' through pooling the variances of the two samples, as shown below:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

where, $$s_1^2 = \frac{1}{(n_1 - 1)}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 \qquad \text{(s.d. of first sample)}$$

$$s_2^2 = \frac{1}{(n_2 - 1)}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2 \qquad \text{(s.d. of second sample)}$$

x_{1i} : is the ith observation in first sample

x_{2i} : is the ith observation in second sample

$\bar{x}_1$: is the mean of first sample

$\bar{x}_2$: is the mean of second population

The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \textbf{(11.10)}$$

which is distributed as student's 't' with $n_1 + n_2 - 2$ d.f.

If the absolute value of the calculated value of 't' is more than the tabulated value of 't' at (n_1 + n_2 – 2) d.f. H_0 is rejected.

Example 11.8

A car manufacturer is procuring car batteries from two companies. For testing whether the two brands of batteries, say 'A' and 'B', had the same life, the manufacturer collected data about the lives of both brands of batteries from 20 car owners – 10 using 'A' brand and 10 using 'B' brand. The lives were reported as follows:

Lives in Months

Battery 'A' :	50	61	54	60	52	58	55	56	54	53
Battery 'B' :	65	57	60	55	58	59	62	67	56	61

Test whether both the brands of batteries have the same life?

Solution:

The hypotheses to be tested are:

$H_0 : m_1 = m_2$ (m_1 and m_2 being average lives of

$H_1 : m_1 \neq m_2$ batteries 'A' and 'B', respectively)

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}$$

where, $$s_1^2 = \frac{1}{(n_1-1)}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 \quad \text{(s.d. of first sample)}$$

$$s_2^2 = \frac{1}{(n_2-1)}\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2 \quad \text{(s.d. of second sample)}$$

x_{1i} : is the ith observation in first sample

x_{2i} : is the ith observation in second sample

$\bar{x}_1$: is the mean life of battery 'A'

$\bar{x}_2$: is the mean life of battery 'B'

The 't' statistic to be calculated is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

(with $n_1 + n_2 - 2 = 10 + 10 - 2 = 18$ d.f.)

where

n_1 and n_2 are number of observations for 'A' and 'B'

The requisite calculations are as follows:

Lives of Battery 'A'	*Lives of Battery 'B'*	$u = x_i - 50$	$v_i = y - 60$	u^2	v^2
50	65	0	5	0	25
61	57	11	–3	121	9
54	60	4	0	16	0
60	55	10	–5	100	25
52	58	2	–2	4	4
58	59	8	–1	64	1
55	62	5	2	25	4
56	67	6	7	36	49
54	56	4	–4	16	16
53	61	3	1	9	1
	Sum	53	0	391	134
	Average	5.3	0		

Note: Here we are using $u = x - 50$ and $v = y - 60$, for simplicity of calculations. We are using property of variance that it is unaffected by change of origin.

From the above Table, we have:

$\Sigma u_i = 53$ $\qquad$ $\Sigma v_i = 0$

$\bar{u} = 5.3$ $\qquad$ $\bar{v} = 0$

$\Sigma u^2 = 391$ $\qquad$ $\Sigma v^2 = 134$

$s_1^2 = (\Sigma u^2 - 10(\bar{u})^2)/9 = 12.23$ $\qquad$ $s_2^2 = (\Sigma v^2 - 10(\bar{v})^2)/9 = 14.89$

Thus,

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

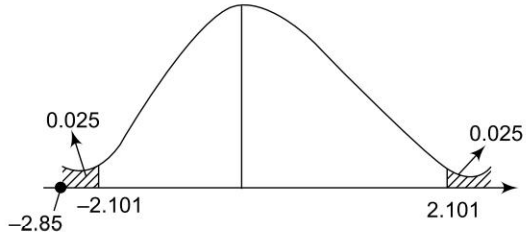
$= 13.56$

$s = 3.68$

$\bar{x}_1 = \bar{u} + 50 = 5.3 + 50 = 55.3$ $\qquad$ $\bar{x}_2 = \bar{v} + 60 = 0 + 60 = 60$

$$\text{'}t\text{'} = \frac{55.3 - 60}{\sqrt{13.54\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{-4.7}{\sqrt{13.54 \times 0.2}} = \frac{-4.7}{1.65}$$

$= -2.85$



This calculated value of 't' is less than tabulated value of t with 18 d.f. at 5% level of significance (2.101) which lies in the shaded region, which is the rejection region. Therefore, the null hypothesis that $m_1 = m_2$ is rejected. Thus, the lives of the two batteries are not the same.

(c) No Significant Change (Paired '*t*' Test)

On several occasions, we are required to find out the effectiveness of a 'treatment'. The treatment could be in the form of a medicine for curative purpose, training salesmen for improving sales, advertisement campaign for boosting sales, an exercise for improving, say, a swimmer's performance, etc. For testing whether a 'treatment' has been effective or not, certain number of persons are selected and their performance recorded before the 'treatment' is given. The performance is recorded again after the 'treatment'. Based on the analysis, as described below, a conclusion is reached as to whether the 'treatment' has been effective or not.

It is to be appreciated that the 'treatment' might cause some change but what is tested is whether the change is statistically significant or not.

$$\boldsymbol{H_0}: \quad m_1 = m_2 \qquad \text{(no significant change)}$$
$$\boldsymbol{H_1}: \quad m_1 \neq m_2 \qquad \text{(significant change i.e. increase or decrease)}$$

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad \sim t_{(n-1)} \text{ d.f.} \qquad \textbf{(11.11)}$$

where,

$$d_i = x_i(\text{after}) - x_i(\text{before}).$$

is the difference in 'before' and 'after' values, for ith person, measured by variable x, $\bar{d}$ is the mean of d_is, and s_d is the s.d. of d_is.

Illustration 11.6

Suppose a pharmaceutical company wants to test the effectiveness of a medicine to reduce the level of blood-sugar in diabetic patients. Obviously, the company would like to administer the medicine to some patients and record the level of blood-sugar in those patients before the start of taking medicine and certain days after taking the medicine.

For such situation, the data is collected in the form:

Value of variable (x) relating to various persons		*Difference between the values*	
x_i (Before)	x_i (After)	$d_i = (x_i \text{ (Before)} - x_i \text{ (After)})$	di^2
x_1 ( )	x_1 ( )	d_1	d_1^2
:	:	:	
x_i ( )	x_i ( )	d_i	d_i^2
:	:	:	
x_n ( )	x_n ( )	d_n	d_n^2
Sum		Σd_i	Σd_i^2
Average		$\bar{d}$	

The null and alternative hypotheses are:

$$\boldsymbol{H_0}: \quad m_1 = m_2 \qquad \text{(no significant change)}$$
$$\boldsymbol{H_1}; \quad m_1 \neq m_2 \qquad \text{(significant change either increase or decrease)}$$

and, the test statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad \text{with } (n-1) \text{ d.f.}$$

where $d_i = x_i$ (after) $- x_i$ (before), and s_d is the standard deviation of the differences d_is, and is, as usual, calculated as

$$s_d = \frac{\sum d_i^2 - n(\bar{d})^2}{n-1}$$

Example 11.9

As per ET-TNS Consumer Confidence Survey, published in *Economic Times* dt. 10th November 2006, the consumer confidence indices for some of the cities changed from December 2005 to September 2006, as follows. Is the difference significant?

City	*December 2005*	*September 2006*
Delhi	106	83
Jaipur	117	142
Mumbai	112	126
Ahmedabad	123	108
Kolkata	83	84
Bhubaneshwar	137	144
Bangalore	137	138
Kochi	113	134

Solution:

The following table gives the data collected and also the calculations required to test the hypothesis that the consumer confidence indices did not change from December 2005 to September 2006.

City	*Dec-05*	*Sep-06*	d_i	d_i^2
Delhi	106	83	23	529
Jaipur	117	142	–25	625
Mumbai	112	126	–14	196
Ahmedabad	123	108	15	225
Kolkata	83	84	–1	1
Bhubaneshwar	137	144	–7	49
Bangalore	137	138	–1	1
Kochi	113	134	–21	441
Sum (Σ)			–31	2067
Average ($\bar{d}$)			–3.875	

The statistic to be used is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad \text{is distributed as } t_{n-1} \text{ d.f.}$$

In the above example, $n = 8$, $\bar{d} = -3.875$, and $\Sigma d_i^2 = 2067$

Thus,

$$s_d^2 = \frac{\Sigma(di - \bar{d})^2}{7} = \frac{\Sigma di^2 - 8\bar{d}^2}{7}$$

$$= \frac{2067 - 8 \times (-3.875)^2}{7} = 16.68$$

Therefore,

$$s_d = 5.90$$

$$'t' = \frac{-3.875}{16.68/\sqrt{8}} = \frac{-3.875}{5.90}$$

$$= -0.657$$

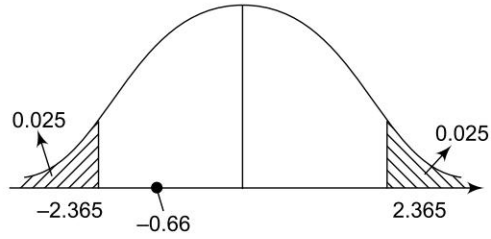


Fig. 11.14 *Critical Region Under 't' Distribution*

The tabulated value of '*t*' at 7 d.f. and 5% level of significance is 2.365.

Since the calculated value is less than the tabulated value, and it does not fall in the critical region, the null hypothesis is not rejected. This implies that there is no apparent difference in the consumer indices from December 2005 to September 2006.

Since the calculated value is more than the tabulated value, it falls in the critical region and the null hypothesis is rejected. Thus, we may conclude that the training has been effective.

(d) Equality of two Means—One Tailed

$$H_0 : \quad m_1 = m_2$$
$$H_1 : \quad m_1 > m_2$$

Here the test statistic is the same as for the alternative $H_0 : m_1 = m_2$, and $H_1 : m_1 \neq m_2$
The '*t*' statistic to be calculated is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

which is distributed as student's '*t*' with $n_1 + n_2 - 2$ d.f.
Here,

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

where, $s_1^2 = \frac{1}{(n_1 - 1)} \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2$ (s.d. of first sample)

$s_2^2 = \frac{1}{(n_2 - 1)} \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2$ (s.d. of second sample)

The statistic for the test is
where,

x_{1i} : is the ith observation in first sample
x_{2i} : is the ith observation in second sample
$\bar{x}_1$: is the mean of first sample
$\bar{x}_2$: is the mean of second population

However, the entire critical region is on one side i.e. on the right because the alternative is $m_1 > m_2$.

The choice of m_1 and m_2 is arbitrary. In any given case, the greater value of variable may be taken as m_1 and smaller variable be taken as m_2. Thus, having test for alternative hypothesis $H_1 : m_1 < m_2$ is not necessary.

(ii) Equality of Two Proportions

(a) Two Sided or Two Tailed:

$\boldsymbol{H_0}$: $p_1 = p_2$
$\boldsymbol{H_1}$: $p_1 \neq p_2$

The statistic for this test is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left\{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \qquad z \sim N(0, 1) \qquad \textbf{(11.12)}$$

where

$$\bar{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_2 + n_2}$$

and $\bar{q} = \bar{p}$

Example 11.10

A firm wanted to choose a popular actor to be the brand ambassador for the firm's product. However, before taking the final decision, the firm conducted a market survey to know the opinion of its customers in Mumbai and Delhi. The surveys conducted in the two cities revealed that while 290 out of 400 customers favoured the choice, in Mumbai, only 160 out of 300 customers favoured the choice in Delhi. Can the firm conclude that the proportions of customers who favoured the actor in Mumbai and Delhi are the same?

Solution:

Let $\hat{p}_1$ denote the proportion of customers favouring the choice of the actor in Mumbai.
Let $\hat{p}_2$ be proportion of customers favouring the choice of the actor in Delhi.

Then the alternative hypothesis that majority of customers are

$$\hat{p}_1 = 290/400 = 0.725 \quad (n_1 = 400)$$

$$\hat{p}_2 = 160/300 = 0.533 \quad (n_2 = 300)$$

The null and alternate hypothesis are as follows:

$$H_1 : \quad p_1 \neq p_2$$

$$H_0 : \quad p_1 = p_2$$

The test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left\{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}}$$

which is distributed as Normal with mean as 0 and s.d. as 1, and where,

$$\bar{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_2 + n_2} = \frac{290 + 160}{400 + 300} = \frac{450}{700} = 0.643$$

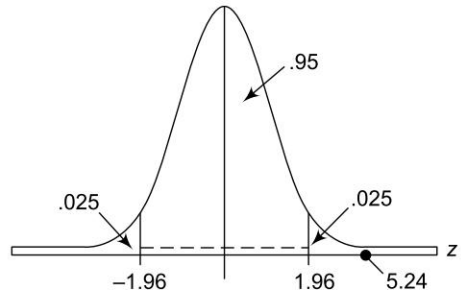
and $\bar{q} = 1 - \bar{p} = 0.357$

Therefore,

$$z = \frac{0.725 - 0.533}{\sqrt{0.23\left(\frac{1}{400} + \frac{1}{300}\right)}} = \frac{0.192}{0.037}$$

$$= 5.24$$

Since level of significance is not given, we presume it to be 0.05. From the Table *T*1 of area under normal curve, we note the two tail value for level of significance 0.05 as 1.96, as shown below:



Since the calculated value of *z* is in the rejection region, the null hypothesis may be rejected. It indicates that there is a significant difference in the customers' responses in the two cities for the choice of the actor.

(b) One Sided or One Tailed:

$$\boldsymbol{H_0} : \quad p_1 = p_2$$

$$\boldsymbol{H_1} : \quad p_1 > p_2$$

The statistic for this test is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left\{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \qquad z \sim N(0, 1)$$

where

$$\bar{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_2 + n_2}$$

and $\bar{q} = 1 - \bar{p}$

(c) One Sided or One Tailed:

$\boldsymbol{H_0}$: $p_1 = p_2$

$\boldsymbol{H_1}$: $p_1 < p_2$

Just like testing equality of means, choice of p_1 and p_2 is arbitrary. The bigger ratio may be taken as $p_1$and the smaller as p_2. Thus, having test for alternative $H_1 : p_1 < p_2$ is not necessary.

11.10.2 Test for Association/Dependence

In correlation/regression analysis, we deal with the existence of correlation or dependence between two variables which can be measured like sales, net profit, etc. But sometimes, a variable may not be measurable, like day of the week or it could be in the form of an attribute like pass/*fail, or good*/*average,* etc. In such cases, instead of using the word 'correlation', we use the word 'association'/dependence, and instead of the variables, we use the word "factors". The word dependence/independence could be used in both the cases. Some of the situations involving association between two attributes are as follows:

- Credit worthiness of borrowers for personal loans and their age groups.
- Association between training received and performance of salesmen.
- Returns on an individual stock and return on stocks of a sector of stocks like Banking, Pharmaceutical, Information Technology, etc.
- Salary Level of Employees and Level of job satisfaction.
- Attitude (Bearish, Neutral, Bullish) towards stock market and age of investors.
- Impact of a TV campaign and Category of viewers viz. Urban/Metropolitan, Semi Urban and Rural

In such cases, the null and alternative hypotheses to be tested are:

H_0 : There **is no** association or dependence of one factor on the other

H_1 : There **is** association or dependence of one factor on the other

Such hypotheses are tested with the help of χ^2 test.

Uses of χ^2 Test:

There are two broad uses of χ^2 test. These are:

(i) χ^2 test as a test of Significance for Association/Dependence

(ii) χ^2 test as a test of Goodness of Fit

(The phrase 'goodness of fit' is used because the hypothesis tested is as to how 'good' the observed frequencies fit a specified assumption, pattern or distribution).

We now illustrate the Chi-Square tests with simple examples.

Illustration 11.7

From a hospital record, the following data was obtained about the births of new born babies on various days of the week during the past year:

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Total
184	148	145	153	150	154	116	1050

Can we conclude that the birth of a child is independent of the day in a week?

Solution:

In this case, the null and alternative hypotheses are

H_0 : The birth of child is **not** associated with any day of the week or it is independent of the day of a week.

H_1 : The birth of a child is associated with days of the week.

In such cases, the test statistics to be used is χ^2, which is evaluated as follows:

$$\chi^2_{(k-1)} = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \quad \textbf{(11.13a)}$$

$$= \sum_{i=1}^{k} \frac{O_i^2}{e_i} - n \quad \textbf{(11.13b)}$$

where, O_i is the observed frequency, e_i, is the expected frequency for each of the seven days of the week, and n is the total number of observations (1050 in this case). The expected values e_i are derived from the equation

$$e_i = n\, p_i$$

where p_i is the probability of an observation lying in the ith interval (i^{th} day of the week in this case).

One can compute χ^2 with either of the above two formulas depending on the values of $(O_i - e_i)$ and the ease of finding squares of O_is. If e_is are full numbers and their difference from O_i s can be found easily, then the first expression may be more convenient as we have to square smaller quantities. However, if we can square O_is, conveniently, then second expression can be used.

In the above data relating to birth of children, under the null hypothesis that the birth of a child is independent of the day of the week, the probability that a child is born on any one particular day of the week is 1/7. Thus all p_is are equal to 1/7. Thus the expected frequency, e_i for each day of the week is $1050 \times 1/7 = 150$. If we want to use the first expression for calculating the value of χ^2, then the table is prepared as follows:

Table 11.1 Observed and Accepted Frequencies

O_i	e_i	$O_i - e_i$	$(O_i - e_i)^2$	$(O_i - e_i)^2/e_i$
184	150	34	1156	7.71
148	150	–2	4	0.03

(*Contd*)

(*Contd*)

145	150	–5	25	0.17
153	150	3	9	0.06
150	150	0	0	0
154	150	4	16	0.11
116	150	–34	1156	7.71
			χ^2 =	**15.77**

Now, the test statistic χ^2 is calculated by substituting values of $(O_i - e_i)^2/e_i$s in formula (11.14a) above, we get

$$\chi^2 = 15.77$$

The number of d.f. for this χ^2 are $(k - 1)$ where k is the number of classes (days in this case). Since number of days are 7, the d.f. are 7–1 = 6. Since the level of significance is not specified, it is presumed as 5%.

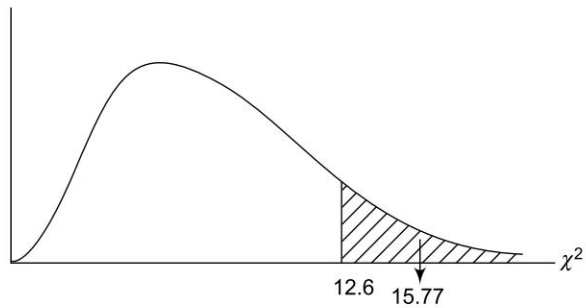


Fig 11.15 Critical Region Under χ^2 Distribution

The tabulated value of χ^2 at 6 d.f. and 5% level of significance is 12.6.

Since the calculated value is more than the tabulated value, and falls in the critical region, the null hypothesis is rejected i.e. the birth of a child in the hospital is associated with days of the week. Incidentally, this may be so because the births by operation are avoided on Sundays, and postponed to Mondays.

Example 11.11

A behavioural scientist is conducting a survey to determine if the financial benefits, in terms of salary, influence the level of satisfaction of employees, or whether there are other factors such as work environment which are more important than salary in influencing employee satisfaction. A random sample of 300 employees is given a test to determine their level of satisfaction. Their salary levels are also recorded. The information is tabulated below:

Level of Satisfaction	*Annual salary (Rs. Lakhs)*			
	Up to 5	*5 – 10*	*More than 10*	*Total*
High	10	10	10	30
Medium	50	45	15	110
Low	40	15	5	60
Total	100	70	30	200

At 5% level of significance, determine whether the level of employee satisfaction is influenced by salary level?

Solution:

We are required to test the hypothesis that the Annual Salary and the Level of Satisfaction are independent of each other. This hypothesis can be tested with the help of χ^2 test, relating to test of independence.

The null and alternative hypotheses are as follows:

H_0 : Annual salary and the level of satisfaction are independent of each other

H_1 : Annual salary and the level of satisfaction are not independent of each other

The statistic χ^2 is defined as

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(O_{ij} - e_{ii})^2}{e_{ij}} \qquad \textbf{(11.14)}$$

where,

O_{ij} is the observed frequency of ith row (i = 1 or 'High' and i = 2 for 'Medium' and i = 3 as 'Low') and jth column (j = 1 for ' Upto 5' and j = 2 for '5 to 10' and for j = 3 'More than 10')

e_{ij} is the expected frequency of ith row and jth column, and is derived by the following expression

$$e_{ij} = \frac{(\text{Total observations in } i\text{th row}) \times (\text{Total observations in } j\text{th column})}{\text{Total number of Observations}}$$

n = Total number of observations

Applying this formula for expected frequencies, we get the various expected frequencies as under:

$$e_{11} = \frac{\text{Row 1} \times \text{Column 1}}{\text{Total number of Observations}} = \frac{100 \times 30}{200} = 15$$

$$e_{12} = \frac{\text{Row 1} \times \text{Column 2}}{\text{Total number of Observations}} = \frac{70 \times 30}{200} = 10.5$$

$$e_{13} = \frac{\text{Row 1} \times \text{Column 3}}{\text{Total number of Observations}} = \frac{30 \times 30}{200} = 4.5$$

$$e_{21} = \frac{\text{Row 2} \times \text{Column 1}}{\text{Total number of Observations}} = \frac{100 \times 110}{200} = 55$$

$$e_{22} = \frac{\text{Row 2} \times \text{Column 2}}{\text{Total number of Observations}} = \frac{70 \times 110}{200} = 38.5$$

$$e_{23} = \frac{\text{Row 2} \times \text{Column 3}}{\text{Total number of Observations}} = \frac{30 \times 110}{200} = 16.5$$

$$e_{31} = \frac{\text{Row 3} \times \text{Column 1}}{\text{Total number of Observations}} = \frac{100 \times 60}{200} = 30$$

$$e_{32} = \frac{\text{Row 3} \times \text{Column 2}}{\text{Total number of Observations}} = \frac{70 \times 60}{200} = 21$$

$$e_{33} = \frac{\text{Row 3} \times \text{Column 3}}{\text{Total number of Observations}} = \frac{30 \times 60}{200} = 9$$

The following table is prepared with the help of the above expected frequencies.

O_i	e_i	$O_i - e_i$	$(O_i - e_i)^2$	$(O_i - e_i)^2/e_i$
10	15	–5	25	1.667
10	10.5	–0.5	0.25	0.024
10	4.5	5.5	30.25	6.722
50	55	– 5	25	0.455
45	38.5	6.5	42.25	1.097
15	16.5	–1.5	2.25	0.136
40	30	10	100	3.333
15	21	– 6	36	1.714
5	9	– 4	16	1.778
			χ^2 =	**16.93**

Now the value of statistic χ^2 is calculated as:

= 16.93

The tabulated value of χ^2 for 5% level of significance at (number of rows –1)(number of columns – 1) = (3 –1) × (3 –1) = 4 d.f. is 9.49

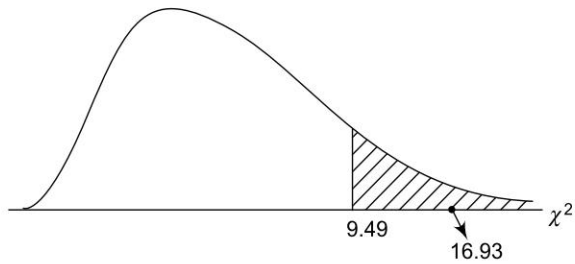


Fig 11.16 *Critical Region Under χ^2 Distribution*

Since the calculated value of (16.93) is more than the tabulated value (9.49), it falls in the critical region, the null hypothesis is to be rejected.

Thus, it is to be concluded that the Annual salary and the level of satisfaction are not independent of each other.

Example 11.12: Uniform Distribution: Market Shares of Different Types of Cars

A marketing manager wants to test his belief that four different categories of cars share the auto market in a particular segment, equally in Delhi city. These four categories of cars are brand *A*, brand B, brand C, and imported cars. He stood at a busy intersection, and collected following data on 1,000 such cars:

Brand	*Number of Cars (Observed Frequency)* O_i
A	235
B	255
C	240
Imported	270
Total	1,000

With the support of this data, we can help him in solving the problem. We first set up the hypotheses

H_0: Same market shares (uniform distribution)

H_1: Different market shares (non-uniform distribution)

The chi-square test statistic defined in Formula 11.13 can be used to perform the hypothesis test. When the null hypothesis is true, there should be 250 cars of each brand. The requisite calculation are shown in the following table. This implies that the expected frequency should be 250 for each car.

Brand	*Number of Cars (Observed Frequency)* O_i	*Expected Frequency* e_i	$O_i - e_i$	$(O_i - e_i)^2$	$(O_i - e_i)^2/e_i$
A	235	250	–15	225	0.9
B	255	250	5	25	0.1
C	240	250	–10	100	0.4
Imported	270	250	20	400	1.6
Total	1,000	1,000			$\chi^2 = 3$

It may be noted that the calculated value of χ^2 is 3.

The tabulated value of χ^2 with 3 d.f. vide Table *T*3 is 7.81. Since calculated value (3) is less than this, we cannot reject the null hypothesis that the data has uniform distribution.

Conditions for applying χ^2 Test:

Before proceeding further with χ^2 test, it may be mentioned that certain conditions are required to be fulfilled before applying this test.

In the first place, the total sample size '*n*' must be reasonably large to ensure similarity between the theoretical χ^2 distribution and sampling distribution of χ^2 statistic. It is difficult to say what constitutes largeness, but as a general rule, χ^2 test should not be used when *n* is less than 50, even when the number of cells are low.

Secondly, expected frequency in any cell should not be less than 5. When the expected frequencies are too small, the value of χ^2 is overestimated and results in too many rejections of the null hypothesis.

11.6.3 Test for Goodness of Fit

As stated earlier, the χ^2 test is also useful for testing goodness of fit which implies testing whether the observed data fits some given pattern or distribution. For example, we may like to fit a distribution like Binomial, Poisson and Normal to a given set of data. If the fit is found to be good, the knowledge of the distribution could be used for detailed analysis.

11.7 'F'-TEST AND ITS APPLICATIONS

The '*F*' distribution is the distribution of ratio of variances. Accordingly, this test is used to test the equality of variances of two populations. It is also used for testing significance of a regression equation and testing equality of several population means, simultaneously. These two types of 'F' test are described in details in Chapter 12 on Analysis of Variance.

If σ_1^2 and σ_2^2 are the variances of two populations, the null and alternative hypotheses are:

$$H_o : \quad \sigma_1^2 = \sigma_2^2$$

$$H_1 : \quad \sigma_1^2 \neq \sigma_2^2$$

Some of the applications of this test are in the situations described below. All these situations involve comparing variances in two populations.

- Quality of items, as measured by some indicators like diameter, length, etc. manufactured by two machines. The machine with lower variance, implying more consistent quality, is preferred.
- Price or Earning Per Share (EPS) of two shares on day-to-day basis, say for a week or month, etc., The share with lower variance implying lesser volatility could be considered less risky and consequently preferred by some investors.
- Service time taken by two systems or agencies—the one with lower variance is preferred.

Illustration 11.8

Let A and B be two agricultural regions. The data below presents the yields, in quintals, of 10 plots (of equal area) from each of the two regions.

Region *A* :	12	7	15	10	13	8	7	10	10	8
Region *B* :	10	9	6	7	8	7	10	15	12	9

Let us now test whether the above random samples taken from the two regions have the same variance at 5% level of significance.

Solution:

F-Test is the appropriate test for testing equality of variances of the two populations. In the given case, regions are treated as populations, and the yields of plots as individual observations.

The null and alternative hypotheses are formulated as follows:

H_0 (null hypothesis) : $\sigma_1^2 = \sigma_2^2$ (two populations have the same variance)

H_1 (alternative hypothesis): $\sigma_1^2 \neq \sigma_2^2$ (two populations do not have equal variance)

	x_{1i}	x_{1i}^2	x_{2i}	x_{2i}^2
	12	144	10	100
	7	49	9	81
	15	225	6	36
	10	100	7	49
	13	169	8	64
	8	64	7	49
	7	49	10	100
	10	100	15	225
	10	100	12	144
	8	64	9	81
Sum	100	1064	93	929
Average	10	—	9.3	—

For carrying out the test of significance, we calculate the statistic '*F*' which is defined as the ratio of sample variances as follows:

$$F = \frac{s_1^2}{s_2^2} \qquad \textbf{(11.15)}$$

where,

$$s_1^2 = \frac{\Sigma(x_{1i} - \bar{x}_1)^2}{n_1 - 1} = \frac{\Sigma x_{1i}^2 - n_1 \bar{x}_1^2}{(n_1 - 1)}$$

$$s_2^2 = \frac{\Sigma(x_{2i} - \bar{x}_2)^2}{n_2 - 1} = \frac{\Sigma x_{2i}^2 - n_2 \bar{x}_2^2}{(n_2 - 1)}$$

It may be recalled that in the '*F*' ratio, the numerator is greater than the denominator i.e.

$$F = \frac{\text{Larger estimate of population variance}}{\text{Smaller estimate of population variance}}$$

Thus the larger of the two sample variances is to be taken in the numerator. Let us assume that $s_1^2 > s_2^2$. However, if $s_1^2 > s_2^2$, we would have defined the '*F*' ratio as s_2^2 / s_1^2.

The '*F*' statistic has two degrees of freedom: one for the numerator and one for the denominator. These are written as d_1, d_2 where

$d_1 = n_1 - 1$ = degrees of freedom for sample having larger variance.
$d_2 = n_2 - 1$ = degrees of freedom for sample having smaller variance.

Since, F test is based on the Ratio of two variances, it is also known as the Variance Ratio Test. Now we calculate both s_1^2 and s_2^2 to see which of the two is larger.

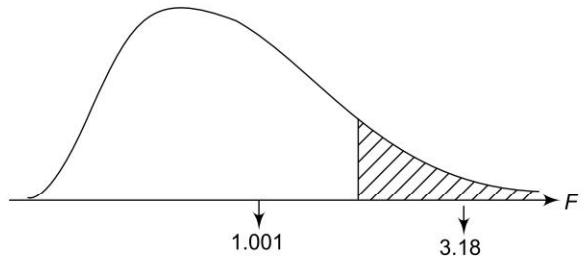
where, $s_1^2 = \frac{64}{9} = 7.11$

$$s_2^2 = \frac{64.1}{9} = 7.12$$

Since s_2^2 is larger we define *F* as:

$$F = \frac{s_2^2}{s_1^2}$$

Therefore, $F = \frac{7.12}{7.11} = 1.001$



The value of F for 9, 9 d.f. at 5% level of significance is 3.18. Thus the calculated value falls in the acceptance region.

Thus, we accept the Null hypothesis and conclude that the samples come from populations having equal variance.

Assumptions for *F*-test:

The following assumptions are required for the validity of the 'F' test for comparing variances of two populations.

- **Normality**: i.e the values in each population are normally distributed.
- **Homogeneity**: i.e the variance within each population are equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). The assumption is needed in order to combine or pool the variances within the populations into a single source of variation 'within populations'.
- **Independence of errors**: It stipulates that the error, i.e. variation of each value around its own population mean, should be independent of the value.

11.7.1 Relationship of '*F*' Statistic with χ^2 Statistic and '*t*' Statistic

It has been mentioned earlier that 'F' statistic is defined as the ratio of two independent χ^2 variables divided by their respective d.f.

The 'F' statistic and 't' statistic are related to each other by the formula

$$F_{1,n} = t_n^2$$

11.8 TEST FOR SIGNIFICANCE OF REGRESSION COEFFICIENT

As discussed in Chapter 10 on Correlation and Regression Analysis, the regression equation of a dependent variable, say 'y' on an independent variable, say 'x', is written as

$$y = a + bx$$

where, b is the regression coefficient (of y on x). The regression coefficient 'b' is said to be significant when it is different from 0 at a prescribed level of significance. It may be appreciated that even if its value is 0 in the population, the sample would result in some value different from 0, positive or negative, due to randomness of observations. Therefore, if the value of 'b' is within a small interval around 0, we tend to believe that its value in the population could be 0. But if its value is beyond this interval, we tend to believe that its value in the population is **significant** i.e. some positive or negative value different from 0.

The hypotheses in such case are written as

$$H_0: \quad b = 0$$
$$H_1: \quad b \neq 0$$

The procedure of conducting the test is explained through an illustration given below:

The test statistic is

$$t = \frac{\hat{b} - b_0}{\hat{s}_b} \quad \text{distributed as students '}t\text{' with } (n-2) \text{ d.f.} \qquad \textbf{(11.16)}$$

where, n is the number of pairs of observations on x and y, viz. (x_1, y_1), (x_2, y_2), … (x_n, y_n), $\hat{b}$ is the estimate of 'b', b_0 is the specified value of 'b' to be tested, which in this case is 0, and $\hat{s}_b$ is the standard error of $\hat{b}$, defined as

$$\hat{s}_b = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{(n-2)\Sigma(x_i - \bar{x})^2}}$$

and $\hat{y}_i$ is the estimated value of y_i for $x = x_i$ as per the fitted regression equation.

Illustration 11.9

Let the observations on a pair of variable x and y be as follows:

x	y
x_1	y_1
x_2	y_2
⋮	
x_i	y_i
⋮	⋮
x_n	y_n

Let the regression equation for the above data be:

$$y = a + bx$$

Then, testing the significance of the regression coefficient implies testing the null hypothesis

$$H_0 : \quad b = 0$$

against the alternative hypothesis

$$H_1 : \quad b \neq 0$$

The statistic for testing this hypothesis is students 't' with n –2 d.f., and is defined as:

$$t = \frac{\hat{b} - 0}{\hat{s}_b}$$

where,

$\hat{b}$ = estimate of 'b' derived from sample data.

$$= \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$$

n = Sample size i.e. number of pairs of observation on the two variables x and y

$\hat{s}_b$ = Standard error of $\hat{b}$

$$= \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{(n-2)\Sigma(x_i - \bar{x})^2}}$$

$\hat{y}_i = \hat{a} + \hat{b}x_i$ ($\hat{y}_i$ is the estimated value of y_i for $x = x_i$)

$\hat{a} = \bar{y} - \hat{b}\bar{x}$ (estimated value of a from given data)

The test is explained with a example relating to correlation between sales and advertising expenses of a company.

Table 11.2 Sales and Advertising Expenses

Sales (y) (Rs. in Crores)	*Advertising Expenses (x) (Rs. in Crores)*
60	1.0
62	1.5
65	2.0
68	2.5
72	3.5
75	4.5

	Advertising Expenses (x) (Rs. in Crores)	*Sales (y) (Rs. in Crores)*	x^2	xy	y^2
	1	60	1	60	3600
	1.5	62	2.25	93	3844
	2	65	4	130	4225
	2.5	68	6.25	170	4624
	3.5	72	12.25	252	5184
	4.5	75	20.25	337.5	5625
Sum	15	402	46	1042.5	27102
Average	2.5	67			

The values of a and b are

$a = 55.97$ and $b = 4.41$

The regression equation of sales (y) on advertising expenses (x) is

$y = 55.97 + 4.41x$

The observed (y_i) and estimated values of y_i, ($\hat{y}_i$) are tabulated below:

	Advertising Expenses (x) (Rs. in Crores)	*Sales (y) (Rs. in Crores)*	$\hat{y} = 55.97 + 4.41x$	$y - \hat{y}$	$(y - \hat{y})^2$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	1	60	60.38	–0.38	0.1444	–1.5	2.25
	1.5	62	62.585	–0.585	0.342225	–1	1
	2	65	64.79	0.21	0.0441	–0.5	0.25
	2.5	68	66.995	1.005	1.010025	0	0
	3.5	72	71.405	0.595	0.354025	1	1
	4.5	75	75.815	–0.815	0.664225	2	4
Average	2.5			Sum	2.559	Sum	8.5

From the above table, we get $\Sigma(y_i - \hat{y}_i)$ as 2.56

and $$\hat{s}_b = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{(n-2)\Sigma(x_i - \bar{x})^2}} = \sqrt{\frac{2.56}{4 \times 8.5}} = \sqrt{\frac{2.56}{34}} = 0.2744$$

Now,

$$t = \frac{\hat{b}}{\hat{s}_b}$$

$$= \frac{4.41}{0.2744}$$

$$= 16.07$$

The tabulated value of 't' with $(n - 2)$ i.e., $6 - 2 = 4$ d.f., at 5% level of significance, is 2.776.

Since the calculated value is more than the tabulated value, and falls in the critical region, we reject the null hypothesis H_0, and accept the alternative hypothesis H_1. This implies that the regression coefficient is significant.

It may be mentioned that in correlation and regression analysis, it is presumed that both *x* and *y* are normally distributed. This is necessary for testing the significance of the regression coefficient either by *t* or *F* test, and finding its confidence limits.

11.9 TESTS FOR SIGNIFICANCE OF CORRELATION COEFFICIENT

One of such tests is used for testing the significance of the Pearson's correlation coefficient between two variables, say x and y. The hypotheses are set up as follows:

$$\mathbf{H_0}: \quad \rho = 0$$
$$\mathbf{H_1}: \quad \rho \neq 0$$

where, ρ is the correlation coefficient between the two variables in the population.

The appropriate test in such cases is the student's 't' test, and the 't' statistic is defined as

$$t = \frac{r}{\sqrt{\left(\frac{1-r^2}{n-2}\right)}} \quad \text{is distributed as '}t\text{' with } (n-2) \text{ d.f.} \qquad \textbf{(11.17)}$$

when r is the correlation coefficient calculated from n pairs of values of the two variables, say x and y.

We provide the example of the test for the above example of relationship between Sales and Advertisement Expenses.

The value of r for the above example is

$$r = \frac{\sum y_i x_i - n\bar{x}\bar{y}}{\sqrt{\sum (y_i^2 - n\bar{y}^2)\sum (x_i^2 - n\bar{x}^2)}}$$

$$r = \frac{1042.5 - 6 \times 2.5 \times 67}{\sqrt{(27102 - 6 \times 4489)(46 - 6 \times 6.25)}}$$

$$= 0.9924$$

Once the value of r is known, we calculate the 't' statistic as

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$$

$$= \frac{0.9924 \times \sqrt{4}}{\sqrt{1-(0.9924)^2}} = \frac{1.9848}{0.123}$$

$$= 16.123$$

The tabulated value of 't' for 4 d.f. at 5% level of significance is 2.78.

Since the calculated value (16.123) is more than the tabulated value, and falls in the rejection region, we reject the hypothesis that $\rho = 0$.

Thus there is significant correlation between Sales and Advertisement Expenses.

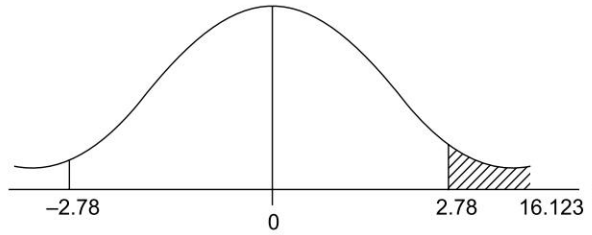


Fig. 11.17 *Critical Region for 't' Distribution*

Since z is distributed as normal with mean 0 and s.d. as 1, we can use the Table T1 of area under standard normal curve to take an appropriate decision. Since value of z(2.09) is more than 1.96 and lies in the critical region, we reject H_0 and conclude that the correlation is significant.

11.9.1 Minimum Value of Correlation Coefficient to be Significant

It may be noted that whether the value of r is significant or not greatly depends on the value of n, in addition to the computed value of r.

For example, for $n = 2$, the value of r will always be 1 even if there is no relationship between the values of x and y. For example, two points (x_1, y_1) and (x_2, y_2), will always lie on a straight line as shown below:

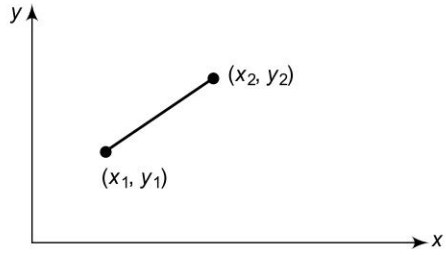


Fig 11.18 *Two Points Always Lie on a Line*

Thus, if $n = 2$, even the value of 1 may not be significant, and thus may not imply correlation between x and y.

In general, for lower values of n, even a high value of correlation coefficient may not be significant. On the other hand, for high values of n, even a low value of correlation coefficient may be significant, calculated with n pairs of observations, The value of r which can be considered significant at 5% level of significance is given by the following formula

$$r \geq \frac{t_{0.25,(n-2)}}{\sqrt{(n-2)+t^2_{0.25,(n-2)}}} \qquad \textbf{(11.18)}$$

For ready reference, the following table gives the minimum values of r which are considered significant at 5% level of significance for values various of n.

Table 11.3 Minimum Value of r Considered Significant for Various Values of n

n	r
3	0.997
4	0.950
5	0.878
…	…
…	…
10	0.633

The minimum values of r, for various values of n, at 5% and 1% levels of significance are given in Table T8. It has been derived by the formula (11.18).

11.9.2 Test for Significance of Spearman's Rank Correlation Coefficient

In Section 11.4 of Chapter 10 on Simple Correlation and Regression Analysis, we have discussed Spearman's rank correlation coefficient. It gives the correlation between ranks given by two individuals to a set of persons or items. Like testing the significance of ordinary Pearson's correlation coefficient, we can also test the significance of rank correlation. The procedure is described in Section 13.11 of Chapter 13 on Non-parametric Tests.

11.10 *p*-VALUE

Computer Based Approach to Acceptance or Rejection of Hypothesis

This refers to level of significance, and is used in computer outputs. Instead of calculating test statistic and critical values, as is customary in manual system of calculation, the computer, generally calculates the p-value, and compares it with the corresponding level of significance. **This value indicates the maximum level of significance at which the Null hypothesis would be accepted**. If the level of significance is specified as 5% and the p-value generated by computer is 0.02, i.e. less than .05, the null hypothesis is rejected.

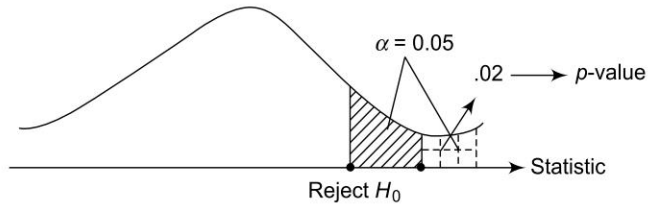


Fig 11.19 *Critical Region and p-value*

If the *p*-value is 0.10 i.e. ≥ 0.05, the null hypothesis is accepted.

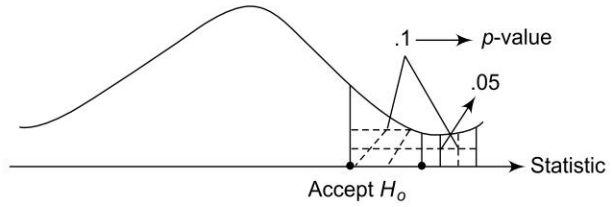


Fig 11.20 *Critical Region and p-value*

The *p*-value is the probability of getting the given sample when the null hypothesis is true. The smaller the p-value, the less likely it is that the observed sample would have come from the assumed population i.e. when the null hypothesis is true. The computer packages provide the *p*-value associated with any test of significance. Thus, one can compare the *p*-value to the level of significance, and conclude without referring to statistical tables.

The use of *p*-value in testing of hypotheses is explained with the help of the following example.

A random sample of 100 students from the current year's batch gives the mean CGPA as 3.55 and variance 0.04. Can we say that this is same as the CGPA of the previous batch which was 3.5?

Solution:

Since we have to test the hypothesis that the population mean is equal to 3.5, the null and alternative hypotheses are set up as

$$H_0: \quad m = 3.5$$
$$H_1: \quad m \neq 3.5$$

Even though population variance is not given, but the sample size is so large that the sample variance can be taken to be equal to population variance i.e. 0.04. Thus, the test statistic used is

$$z = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}} \qquad \text{which is distributed as } N(0,1)$$

We are given,

$$\bar{x} = 3.55,\ m_0 = 3.5,\ n = 100 \text{ and } \sigma = \sqrt{0.04} = 0.2$$

Therefore,

$$z = \frac{3.55 - 3.5}{0.2/\sqrt{100}}$$

$$= \frac{0.05}{0.2/10} = \frac{0.05}{0.02}$$

$$= 2.5$$

Since, the calculated value of z (2.5) is more than the tabulated value of z, at 5% level of significance, 1.96, the null hypothesis is rejected, as shown below:

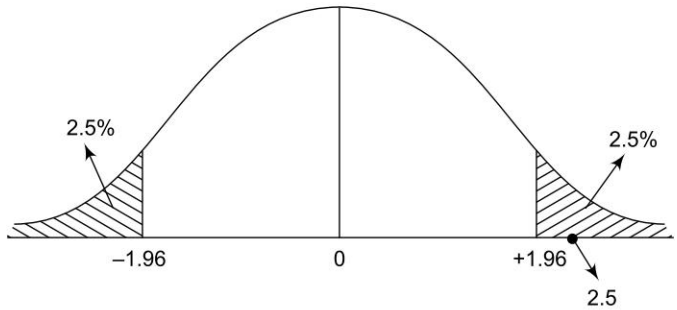


Fig. 10.21

Thus the current year's CGPA is not same as the last year's CGPA.

The *p*-value for this example works out to 0.0124 which is less than 0.05 (assumed level of significance). Hence the null hypothesis is rejected by this method as well.

11.11 USING EXCEL

The various tests described in the chapter could be done easily with the help of MS Excel. As discussed in the earlier chapters, one could enter the functions and select the functions from the 'Statistical Functions' list. The various functions that could be useful for this Chapter are, 'Confidence', 'Chidst', 'Chinv', 'Fdist', 'Finv' 'Normdist', 'Norminv', 'Tdist', 'Tinv', etc.

As a matter of convenience for ease in calculations, we have provided template for the calculations involved in this chapter. There are the following twelve worksheets:

1. Mean Estimation
2. Proportion Estimation
3. Sample Size Mean & Proportion
4. One Sample Test—Mean
5. One Sample Test—Proportion
6. Two Sample Mean Pop Sds Known
7. Two Sample Mean Pop Sds Unknown
8. Two Sample Mean Paired Diff
9. Two Sample Test– Proportion
10. Chi Square Goodness of Fit test

11. Chi Square Independence Test
12. *F* Test—Equality of Variances

Use of each of these worksheets is explained in this section, separately.

1. Mean Estimation

This worksheet is basically divided into two parts, viz.:

- Estimation with Raw Data
- Estimation with Sample Statistics

Estimation with Raw Data

This is at the left side of the worksheet. One has to enter the raw data in the column A from cell *A*4 downwards. The template automatically calculates the sample statistics like sample mean sample standard deviation and sample size.

For computation of confidence interval from raw data, one could use the upper left part of the worksheet if the population standard deviation is known, and lower left part if population standard deviation is not known.

If population standard deviation is known, if one enters the s.d. and the confidence level in the cells *D*8 and *D*9 respectively, the template automatically computes the confidence intervals.

If population standard deviation is not known, once the confidence level is entered in the cell *D*27, the template automatically computes the confidence intervals. We have given a drop box for entering the confidence level in which one could select from different values.

Estimation with Sample Statistics

This is at the right side of the worksheet. Here, one could directly enter the known sample statistics.

For computation of confidence interval from sample statistics, one could use the upper right part of the worksheet if the population standard deviation is known, and lower right part if population standard deviation is not known.

If population standard deviation is known, if one enters the sample mean, sample size population s.d. and the confidence level in the cells, *I*6, *I*7, *I*8 and *I*9 respectively, the template automatically computes the confidence intervals.

If population standard deviation is not known, if one enters the sample mean, sample size Sample s.d. and the confidence level in the cells, *I*24, *I*25, *I*26 and *I*27, respectively, the template automatically computes the confidence intervals.

If the finite population correction is applicable, one could enter the population size in the cell *D*11, if using raw data and in the cell *I*11, if using sample statistics.

2. Proportion Estimation

If one enters the sample size, sample proportion and the confidence level in the cells *B*5, *B*6 and *B*7, respectively, the template would automatically calculate the confidence intervals in the cells *B*10 and *C*10.

3. Sample Size Mean and Proportion

This gives the sample size for

- Mean, and
- Proportion

The upper part of the worksheet is for Mean and the lower part is for proportion.

For Mean, if one enters the width, population standard deviation and the confidence level in the cells *C*5, *C*6, and *C*7 respectively, the template automatically calculates the sample size.

We have solved Illustration 11.1, using this template.

For proportion, there are two possibilities, either the past proportion is known or it is not known. If the proportion is known, the left side part of the lower worksheet is to be used where one could enter the interval width, sample proportion and the confidence level in the cells *C*18, *C*19, and *C*20 respectively, the template would calculate the sample size. If there is no idea about the past proportion, one could use the right side part of the worksheet where one could enter the interval width and the confidence level in the cells *F*18, and *F*20 respectively, the template automatically calculates the most conservative sample size.

We have solved Illustrations 11.1 and 11.2, using this template.

4. ONE SAMPLE TEST—MEAN

This worksheet could be used for testing of population mean for a single sample.

Microsoft Excel - Statistical Inference

File Edit View Insert Format Tools Data Window Help

Arial 10 B I U

D12 H1: m ≠

	A	B	C	D	E	F	G	H	I	J	K	L
1				Hypothesis Testing - Population Mean						Hypothesis Testing - Population Mean		
2				Hight of Students						Illustration 10.4 High intensity light bulbs		
3		Values		Pop Stdev Known - Raw Data						Pop Stdev Known - Sample Statistics Given		
4	1	161										
5	2	165		Sample size	25					Sample size	100	
6	3	167		Sample Mean	166.00					Sample Mean	670.00	
7	4	169										
8	5	165		Population Stdev.	3					Population Stdev.	25	
9	6	163		Standard Error	0.6000					Standard Error	2.5000	
10	7	164		Test Statistic (Z)	1.6667		Alpha			Test Statistic (Z)	8.0000	Alpha
11	8	166		Null Hypothesis		p-value	5%			Null Hypothesis		p-value	5%
12	9	160		H1: m ≠	165	0.0956	cannot reject			H1: m ≠	650	0.0000	Reject
13	10	168		H1: m <	165	0.9522	cannot reject			H1: m <	650	1.0000	cannot reject
14	11	168		H1: m >	165	0.0478	Reject			H1: m >	650	0.0000	Reject
15	12	170										
16	13	166										
17	14	166		Correction for Finite Population						Correction for Finite Population		
18	15	172				Population size	500				Population size	500
19	16	162				Test Statistic (Z)	1.7083				Test Statistic (Z)	8.9353
20	17	168										
21	18	169										
22	19	169					Alpha					Alpha
23	20	166		Null Hypothesis		p-value	5%			Null Hypothesis		p-value	5%
24	21	163		H1: m ≠	165	0.0876	cannot reject			H1: m ≠	650	0.0000	Reject
25	22	167		H1: m <	165	0.9562	cannot reject			H1: m <	650	1.0000	cannot reject
26	23	166		H1: m >	165	0.0438	Reject			H1: m >	650	0.0000	Reject
27	24	164										
28	25	166										
29				Pop Stdev Unknown - Raw Data; Population Normal						Pop Stdev Unknown - Sample Statistics Given; Population Normal		
30				Sample size	25					Sample size	100	
31				Sample Mean	166.00					Sample Mean	670	
32				Sample Stdev.	2.872					Sample Stdev.	25	
33												
34				Test Statistic	1.7408					Test Statistic	8.0000	
35							Alpha					Alpha
36				Null Hypothesis		p-value	5%			Null Hypothesis		p-value	5%
37				H1: m ≠	165	0.0945	cannot reject			H1: m ≠	650	0.0000	Reject
38				H1: m <	165	0.0473	Reject			H1: m <	650	0.0000	Reject
39				H1: m >	165	0.9527	cannot reject			H1: m >	650	1.0000	cannot reject
40												

Above is the snapshot of the worksheet One Sample Test – Mean of the template, Statistical Inference. This worksheet is basically divided into two parts viz.

- Testing for population mean with Raw Data
- Testing for population mean with Sample Statistics

Testing for Population Mean with Raw Data

This is at the left side of the worksheet. One has to enter the raw data in the column A from cell *A*4 downwards. The template automatically calculates the sample statistics like sample mean, sample standard deviation and sample size.

For testing the hypothesis from raw data, one could use the upper left part of the worksheet if the population standard deviation is known, and lower left part if population standard deviation is not known.

If population standard deviation is known, one needs to enter the s.d., and the value of α in the cells *D*8 and *G*11 respectively. One has to also enter the 'Alternate Hypothesis'. The Null Hypothesis is assumed universally as H_0: $m = m_0$. We have given all three types of tests, two tailed, left tailed and right tailed. One can enter the hypothesised value of m in the cell *E*9. The template would automatically test all the three hypothesis (all the three tailed tests). One could consider the appropriate test for conclusion.

If population standard deviation is not known, once the value of α is entered in the cell *G*36, and the hypothesised value of m is entered in the cell *E*37, the template automatically tests the hypothesis. Rest of the explanation is same as for the above case.

Testing for Population Mean with Sample Statistics

This is at the right side of the worksheet. Here, one could directly enter the known sample statistics.

For testing of hypothesis from sample statistics, one could use the upper right part of the worksheet if the population standard deviation is known, and lower right part if population standard deviation is not known.

If population standard deviation is known, and if one enters the sample size, sample mean, population s.d., value of α and the hypothesised value of mean in the cells, *I*6, *I*7, *I*8 and *D*9 respectively, the template carries out all the three tests.

If population standard deviation is not known, if one enters the sample mean, sample size Sample s.d. and the confidence level in the cells *J*5, *J*6, *J*8, *L*11 and *J*12 respectively, the template automatically tests the hypotheses.

If the finite population correction is applicable, one could enter the population size in the cell *G*18 if using raw data and in the cell *L*18 if using sample statistics.

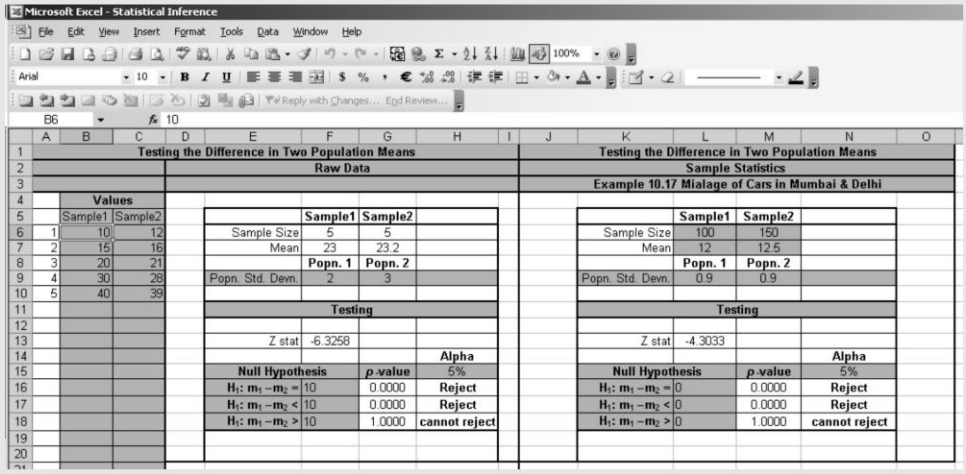
We have solved the example on heights of students, using raw data and Illustration 11.5, using sample statistics template.

5. One Sample Test-Proportion

If one enters the sample size, the number of successes, the value of α and the hypothesised value of proportion, in the cells *B*5, *B*6, *D*10, and *B*11, the template carries out all the three tests.

5. TWO SAMPLE MEAN POP SDS KNOWN

This worksheet could be used for testing of equality of means for two samples.



Microsoft Excel - Statistical Inference

File Edit View Insert Format Tools Data Window Help

B6 10

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Testing the Difference in Two Population Means									Testing the Difference in Two Population Means				
2		Raw Data									Sample Statistics				
3											Example 10.17 Mialage of Cars in Mumbai & Delhi				
4		Values													
5		Sample1	Sample2			Sample1	Sample2					Sample1	Sample2		
6	1	10	12		Sample Size	5	5				Sample Size	100	150		
7	2	15	16		Mean	23	23.2				Mean	12	12.5		
8	3	20	21			Popn. 1	Popn. 2					Popn. 1	Popn. 2		
9	4	30	28		Popn. Std. Devn.	2	3				Popn. Std. Devn.	0.9	0.9		
10	5	40	39												
11					Testing						Testing				
12															
13					Z stat	-6.3258					Z stat	-4.3033			
14								Alpha						Alpha	
15					Null Hypothesis		p-value	5%			Null Hypothesis		p-value	5%	
16					H_1: $m_1 - m_2 =$	10	0.0000	Reject			H_1: $m_1 - m_2 =$	0	0.0000	Reject	
17					H_1: $m_1 - m_2 <$	10	0.0000	Reject			H_1: $m_1 - m_2 <$	0	0.0000	Reject	
18					H_1: $m_1 - m_2 >$	10	1.0000	cannot reject			H_1: $m_1 - m_2 >$	0	1.0000	cannot reject	
19															
20															

Above is the snapshot of the worksheet Two Sample Mean Pop Sds Known of the template, Statistical Inference. This worksheet is basically divided into two parts viz.

- Testing for equality of means of two samples with Raw Data
- Testing for equality of means of two samples with Sample Statistics

Testing for Equality of Means of Two Samples with Raw Data

This is at the left side of the worksheet. One has to enter the raw data in the columns *B* and *C* from cell *B*6 and *C*6 downwards. The template automatically calculates the sample statistics like sample mean sample standard deviation and sample size, for the two samples.

One needs to enter the s.d.s, and the value of α in the cells *F*9, *G*9 and *H*15 respectively. One has to also enter the 'Alternate Hypothesis'. The Null Hypothesis is assumed universally as H_0: $m_1 - m_2 = d_0$ We have given all three types of tests, two tailed, left tailed and right tailed. One can enter the hypothesised value of the difference, d_0, in the cell F16. The template would automatically test all the three hypothesis (all the three tailed tests). One could consider the appropriate test for conclusion.

Testing for Equality of Means of Two Samples with Sample Statistic

This is at the right side of the worksheet. Here, one could directly enter the known sample statistics for the two samples.

One needs to enter the sample mean, sample size Sample s.d. and the confidence level in the cells *L*6, *M*6, *L*7, *M*7, *L*9, *M*9 *N*15 and *L*16 respectively, the template automatically tests the hypotheses.

6. TWO SAMPLE MEAN POP SDs UNKNOWN

This worksheet could be used for testing of equality of proportions for two samples.

Microsoft Excel - Statistical Inference

	A	B	C	D	E	F	G	H	I	J	K
1											
2				Equality of Means of Two Populations Using Raw Data				Equality of Means of Two Populations Using Sample Statistics			
3		Values		Example 11.8 Car Batteries							
4		Battery 'A'	Battery 'B'	Assuming Population Variances are Equal				Assuming Population Variances are Equal			
5	1	50	65		Sample1	Sample2			Sample1	Sample2	
6	2	61	57	Size	10	10		Size	10	10	
7	3	54	60	Mean	55.30	60.00		Mean	55.30	60.00	
8	4	60	55	Std. Deviation	3.50	3.86		Std. Deviation	3.50	3.86	
9	5	52	58								
10	6	58	59	Pooled Variance	13.56			Pooled Variance	13.57		
11	7	55	62	t Stat	-2.8539			t Stat	-2.8524		
12	8	56	67	DOF	18			DOF	18		
13	9	54	56								
14	10	53	61				Alpha				Alpha
15				Null Hypothesis		p-value	5%	Null Hypothesis		p-value	5%
16				$H_1: m_1 - m_2 =$	0	0.0105	Reject	$H_1: m_1 - m_2 =$	0	0.0106	Reject
17				$H_1: m_1 - m_2 <$	0	0.0053	Reject	$H_1: m_1 - m_2 <$	0	0.0053	Reject
18				$H_1: m_1 - m_2 >$	0	0.9947	cannot reject	$H_1: m_1 - m_2 >$	0	0.9947	cannot reject
19											
20											
21				Assuming Population Variances are Unequal				Assuming Population Variances are Unequal			
22				t Stat	-2.85388			t Stat	-2.85244		
23				DOF	17			DOF	17		
24							Alpha				Alpha
25				Null Hypothesis		p-value	5%	Null Hypothesis		p-value	5%
26				$H_1: m_1 - m_2 =$	0	0.0110	Reject	$H_1: m_1 - m_2 =$	0	0.0110	Reject
27				$H_1: m_1 - m_2 <$	0	0.0055	Reject	$H_1: m_1 - m_2 <$	0	0.0055	Reject
28				$H_1: m_1 - m_2 >$	0	0.9945	cannot reject	$H_1: m_1 - m_2 >$	0	0.9945	cannot reject
29											

Above is the snapshot of the worksheet Two Sample Mean Pop Sds unknown of the template, Statistical Inference. This worksheet is basically divided into two parts viz.

- Testing for equality of means of two samples with Raw Data
- Testing for equality of means of two samples with Sample Statistics

Testing for Equality of Means of Two Samples with Raw Data

This is at the left side of the worksheet. One has to enter the raw data in the columns *B* and *C* from cell *B*5 and *C*5 downwards. The template automatically calculates the sample statistics like sample mean sample standard deviation and sample size, for the two samples.

One needs to enter the value of α in the cell *G*15 respectively. One has to also enter the 'Alternate Hypothesis'. The Null Hypothesis is assumed universally as H_0: $m_1 - m_2 = d_0$. We have given all three types of tests, two tailed, left tailed and right tailed. One can enter the hypothesised value of the difference, d_0, in the cell *E*16. The template would automatically test all the three hypotheses (all the three tailed tests). One could consider the appropriate test for conclusion.

We have solved the Example 11.8 relating to car batteries using raw data template.

Testing for Equality of Means of Two Samples with Sample Statistics

This is at the right side of the worksheet. Here, one could directly enter the known sample statistics for the two samples.

One needs to enter the sample sizes, sample means, population s.ds. for both the samples, value of α and the hypothesised value of the difference in the cells, *I*6, *J*6, *I*7, *J*7, *K*15 and *I*16 respectively, the template automatically tests the corresponding hypotheses.

7. TWO SAMPLE MEAN PAIRED DIFF

This worksheet could be used for testing of equality of means for paired or dependent samples.

Microsoft Excel - Statistical Inference

File Edit View Insert Format Tools Data Window Help

Arial 10 100% Reply with Changes... End Review...

G17

	A	B	C	D	F	G	H	I
1		Hypothesis Testing - Equality of two Means - Paired Samples						
2		Example 11.9 ET-TNS Consumer Confidence Survey						
3		Dec-05	Sep-06	Difference	Pop Stdevs Unknown; Populations Normal			
4	1	106	83	23	Sample size	8		
5	2	117	142	-25	Mean of Difference	-3.88		
6	3	112	126	-14	Stdev. of Difference	16.677		
7	4	123	108	15				
8	5	83	84	-1	Test Statistic	-0.6572		
9	6	137	144	-7				Alpha
10	7	137	138	-1	Null Hypothesis		*p*-value	5%
11	8	113	134	-21	H_1: d ≠	0	0.5321	cannot reject
12					H_1: d <	0	0.2660	cannot reject
13					H_1: d >	0	0.7340	cannot reject
14								

Above is the snapshot of the worksheet Two Sample Mean Paired Diff of the template, Statistical Inference. One needs to enter the data in the columns *B* and *C* from the cells *B*4 and *C*4 downwards. the template automatically computes the two sample statistics like the difference, sample size, mean of difference, standard deviation of difference. One needs to also enter the value of α and the hypothesised value of difference in the cells *I*10 and *G*11 respectively. The template then computes the test statistic, *p*-value and tests the hypotheses.

We have solved Example 11.9 relating to consumer confidence survey, using the above template.

9. Two Sample Test

One could enter the sample sizes, number of successes for both the samples in the cells, *B*6, *C*6, *B*7, *C*7 respectively. If one enters the value of a and the hypothesised difference in the cells, *D*12 and *B*13 respectively, the template would carry out all the three tests for the data.

We have solved Example 11.10 using the above template.

8. CHI SQUARE TEST FOR GOODNESS OF FIT

This worksheet could be used for testing of goodness of fit for a distribution.

Microsoft Excel - Statistical Inference

File Edit View Insert Format Tools Data Window Help

Arial 10 B I U 100% Reply with Changes... End Review...

F12

	A	B	C	D	E	F	G	H
1	Goodness of Fit							
2	Illustration 11.7 Babies Born on Days of Week							
3	Observed Values	Expected Values	Deviation	Deviation^2	χ^2		Hypothesis Testing	
4	184	150	34	1156	7.71		DOF	6
5	148	150	-2	4	0.03		$\chi^{2\ stat}$	15.77
6	145	150	-5	25	0.17		*p*-Value	0.01502
7	153	150	3	9	0.06		Alpha	5%
8	150	150	0	0	0.00		Result	**Reject**
9	154	150	4	16	0.11			
10	116	150	-34	1156	7.71			

Above is the snapshot of the worksheet Chi Square Goodness of Fit test of the template, Statistical Inference. In this template one could give the observed and the expected frequencies in the columns *A* and *B* from the cells *A*4 and *B*4 respectively. The template automatically calculates the columns like deviation squared deviation, and χ^2 statistic. After entering the value of α in the cell, H7, the template carries out the test and gives the result in the cell H8.

We have solved Illustration 11.7 with the help of above template.

11. Chi Square Test for Independence

In this template, the data is in the form of contingency table form. The maximum dimension of the matrix is 10×10. In the above template, one could enter the data in the matrix form in the columns, *B*, *C*, … *K* and rows, 5, 6, ... 14.

Once the data is entered, the template calculates the expected frequencies, and the test statistics. After entering the value of a in the cell *O*7, it also tests the hypothesis and displays the result in the cell *O*8.

9. *F* TEST—EQUALITY OF VARIANCES

This worksheet could be used for testing of equality of variances for two samples.

Microsoft Excel - Statistical Inference

File Edit View Insert Format Tools Data Window Help

Arial 10 B I U 100%

Reply with Changes... End Review...

D17

	A	B	C	D	E	F	G	H	I	J	K
1		*F*-Test for Equality of Variances									
2					Illustration 11.8 Agricultural Regions				Example 10.33 IT and Pharmasutical Sectors		
3					Using Raw Data				Using Sample Statistics		
4		Values				Sample 1	Sample 2			Sample 1	Sample 2
5		Sample1	Sample2		Size	10	10		Size	9	8
6	1	12	10		Variance	7.11	7.12		Variance	20.00	13.00
7	2	7	9								
8	3	15	6		F Stat	**1.0016**			F Stat	**1.5385**	
9	4	10	7		DOF1	9			DOF1	8	
10	5	13	8		DOF2	9			DOF2	7	
11	6	8	7								
12	7	7	10				Alpha				Alpha
13	8	10	15		Null Hypothesis	*p*-value	5%		Null Hypothesis	*p*-value	5%
14	9	10	12		$H_1: \sigma^2_1 - \sigma^2_2 \neq 0$	0.9982	cannot Reject		$H_1: \sigma^2_1 - \sigma^2_2 \neq 0$	0.5837	cannot Reject
15	10	8	9		$H_1: \sigma^2_1 - \sigma^2_2 < 0$	0.5009	cannot Reject		$H_1: \sigma^2_1 - \sigma^2_2 < 0$	0.7082	cannot Reject
16					$H_1: \sigma^2_1 - \sigma^2_2 > 0$	0.4991	cannot Reject		$H_1: \sigma^2_1 - \sigma^2_2 > 0$	0.2918	cannot Reject

Above is the snapshot of the worksheet F Test–Equality of Variances of the template, Statistical Inference. This worksheet has two parts viz.

- Testing for population variance with Raw Data
- Testing for population variance with Sample Statistics

Testing for Population Variance with Raw Data

This is at the left side of the worksheet. One has to enter the raw data in the columns *B* and *C* from cells *B*6 and *C*6 downwards. The template automatically calculates the sample statistics like sample sizes and sample variances of the two samples.

If the value of α is entered in the cell *G*13, the template carries out the required test and gives the results.

Testing for Population Variance with Sample Statistics

This is at the right side of the worksheet. Here, one could directly enter the known sample statistics.

If one enters the sample sizes and the sample variances in the cells, *J*5, *K*5, *J*6, *K*6, respectively, and also the value of α in the cell *K*13, the template carries out the required test and gives the results.

We have solved Illustration 11.8 relating to Agricultural Regions, using raw data.

GLOSSARY

Term	*Definition or Explanation*
Alternative Hypothesis	Statement to be accepted if null hypothesis is rejected.
Chi Square Test	Test for association between two factors or Test for goodness of fit.
Confidence Coefficient	Complement of level of significance
Confidence Interval or limits	The interval or limits within which the true value of the parameter lies.
Confidence Level	It is expressed in percentage e.g. 95%, and indicates the degree of confidence that the true value of the parameter lies in the specified interval.
Consistent Estimator	A property which implies that the estimate tends to the true value of the parameter as the sample size increases.
Contingency Coefficient	Measure of association between two factors in a contingency table.
Efficient Estimator	A property which implies that the variance of the estimator is minimum as compared to any other estimator.
Estimator	A function of sample values to estimate a parameter of a population.
'*F*' Test	Test for equality of variances or Test for significance of regression equation.
Hypothesis	A statement or assumption or a claim about a parameter of a population.
Interval Estimation	Estimate in the form of an interval, say from 18 to 20.
Level of Significance	Same as Type – I error.
Null Hypothesis	An assumption or claim about a specific parameter
'*p*' Value	Maximum level of significance at which the null hypothesis would be accepted
Point Estimation	A single value estimate like 20.
Power of a Test $(1 - \beta)$	Refers to the probability that the test will lead to rejection of a statement when it is false. Equal to 1–Type II error.
Significant	Difference between the sample value and population value is said to be significant if the calculated statistic falls in the rejection/critical region.
Standard Error	Standard Deviation of an Estimate from a Sample.
Statistic	A function of sample values.
Student's '*t*' Test	Test for specified value of mean, Test for equality of two means, Test for significance of regression or correlation coefficient.
Sufficient Estimator	A property which implies that after the calculation of such estimator, the sample does not contain any worthwhile information about the estimated parameter.
Type–I Error (α)	Refers to a test of significance. Probability that the test will lead to rejection of a statement when it is true. Denoted by Greek letter α.
Type–II Error (β)	Refers to a test of significance. Probability that the test will lead to accepting a statement when it is false. Denoted by Greek letter β.
Unbiased Estimator	A property which implies that its expected value is equal to the population value.

OBJECTIVE TYPE QUESTIONS

1. Which of the following factor does not usually affect the width of a confidence interval?
   (a) Sample size
   (b) Confidence desired
   (c) Variability in the population
   (d) Population size
2. Given level of confidence as 95% and margin of error as 2%, the minimum sample size required to estimate the proportion is:
   (a) 1256 (b) 2009 (c) 2401 (d) 2815
3. Type-I error is defined as the probability to:
   (a) accept a hypothesis when it is true
   (b) accept a hypothesis when it is false
   (c) reject a hypothesis when it is true
   (d) reject a hypothesis when it is false
4. Which of the following is not an alternative hypothesis?
   (a) $H_1 : m \neq m_o$ (b) $H_1 : m > m_o$ (c) $H_1 : m < m_o$ (d) $H_1 : m = m_o$
5. If the alternative hypothesis is $m_1 > m_2$, the critical region will be on?
   (a) the left side (b) the right side
   (c) on both sides (d) any one of the above
6. Which of the following statement about confidence limits for population mean is not true?
   (a) 50% confidence limits are wider than 95%
   (b) 90% confidence limits are wider than 95%
   (c) 95% confidence limits are wider than 99%
   (d) 99% confidence limits are widest
7. If on the basis of a sample, a hypothesis is to be rejected, at 5% level of significance, then p – value will be:
   (a) = 0.05 (b) < 0.05 (c) > 0.05 (d) < 0.025
8. Which one of the following statements is false?
   (a) α is called Type-I error
   (b) 1– α is called power of the test
   (c) β is called Type-II error
   (d) 1– β is called power of the test.
9. Which one of the following is not a step in conducting a test of significance?
   (a) Set up the Null hypothesis
   (b) Decide the level of significance
   (c) Decide the power of the test
   (d) Decide on the appropriate statistic.
10. The p-value indicates the:
    (a) Minimum level of significance at which the Null hypothesis would be rejected
    (b) Maximum level of significance at which the Null hypothesis would be accepted
    (c) Maximum level of significance at which the Null hypothesis would be rejected
    (d) Minimum level of significance at which the Null hypothesis would be accepted

EXERCISES

1. A Department store wants to determine the % of shoppers who buy at least one item. A random sample of 500 shoppers leaving the shop showed that 150 did not buy any item. What is the 96% confidence interval for the true percentage of buyers?
2. A manager wants to determine the average time required to complete a job. As per the past data about completion of the job, the standard deviation is 5 days. How large should the sample be so that he may be 99% confident that the sample mean may lie within ±2 days of the actual mean?
3. An oil company has purchased a new machine which fills 1 litre tins with a type of oil. If the fill exceeds 1000 ml, there will be wastage of oil. If the fill is under 1000 ml, there will be complaints from the customers. To check the filling operation of the machine, 36 bottles are chosen at random and found to have a fill of 999.2 ml. The s.d. of the machine is known to be 1.2 ml. What is the hypothesis for such a test? Test the hypothesis using 1% level of significance.
4. A sample of 50 pieces of a certain type of string was tested. The mean breaking strength turned out to be 15 kgs. Test whether the sample is from a batch of strings having a mean breaking strength of 15.6 kgs. and standard deviation of 2.2 kgs. Use 1% level of significance.
5. A claim is made that a batch of bulbs has a mean life of 2000 hrs. From past experience, it is known that the s.d. of lives is 100 hrs. A buyer specifies that he wants to test the claim against the alternative hypothesis that the mean burning time is, in fact, below 2000 hrs at 2% significance level. A sample of size 25 is drawn, and the sample average is found to be 1950 hrs. What conclusion should the buyer make? Should the buyer accept the hypothesis that the mean life of all bulbs in the population is at least 2000 hrs?
6. A manufacturer of a patent medicine claims that it is effective in curing 90% of the people suffering from the disease. In a sample of 200 people using this medicine, 160 were relieved of suffering. Determine whether his claim is justified?
7. The owner of a workshop wants to know which of the 2 brands of hand gloves used in the workshop has longer lasting life than the other. He selected, at random, 40 workers who wear gloves of National firm, and their gloves lasted on an average for 80 days with s.d. 5.0 days; while another 40 randomly selected workers wear out the gloves of Liberty firm on an average in 84 days with s.d. of 4.0 days. Can he feel 95% confident that the difference between the two brands is significant?
8. The manager of a workshop wishes to determine if a new process would reduce the working time per unit manufactured on a given machine. He recorded the initial timings in minutes taken by 5 workers and the new timings by the same workers after introducing the process. He wishes to draw inference about the usefulness of the process from the observations given below:

Workers		1	2	3	4	5
Working time per unit	Before	8	4	9	8	6
	After	5	3	8	6	8

Use $\alpha = 0.05$ and test whether the new process has resulted in the reduction of mean working time.

9. A new petrol additive is being tested, in the hope that it will increase kms. per litre. A series of trials are carried out, with and without the additive. One hundred trials on a brand of car without the additive show an average petrol consumption of 15 km.p.litre. with the s.d. of 1.2. With the additive, the average of another 150 trials is 16.5 km.p.litre with a s.d. of 1.4. Do these figures establish, at 5% significance level, that the additive has increased the km.p.liter consumption?
10. A pharmaceutical company wants to estimate the mean life of a particular drug under typical weather conditions. The following results were obtained from a simple random sample of 25 bottles of the drug. The population s.d is given to be 3 months.

    Sample Mean = 30 months
    Population s. d. = 3 months

    Find interval estimates with confidence level of (i) 90% (ii) 95% and (iii) 99%.
11. The Value for Money, a consumer products firm, interested in promoting a new product, wishes to test the effectiveness of sponsoring a major TV movie. Of the 300 individuals surveyed, during the week preceding the movie, 45% were familiar with the new product. After the screening, a sample of 400 individuals were surveyed and the brand awareness found to be 51%. Can the firm conclude that the brand awareness was improved by sponsoring the movie, at $\alpha = 0.05$?
12. A new washing machine liquid detergent was introduced in the market, by using only cash discount incentive as a promotional drive. After about a month, a sample of 60 housewives were requested to rate the new detergent. After a month of intensive TV advertising, the same women were asked to rate the detergent once again. Using a scoring system, based on perceptions of product effectiveness, the difference in scores had mean 1.6 and s.d. 0.4. Is there evidence that perceptions of the products' effectiveness changed during the period of the advertising? Carry out the test at 0.05 level of significance.
13. In a management institute, the A+, A and B Grades allocated to students in their final examination, were as follows:

Specialisation	Grades		
	A+	A	B
Finance	20	25	10
Marketing	15	20	8
Operations	5	15	7

    Using 5% level of significance, determine whether the grading scale is independent of the specialisation?
14. The Progressive Bank allocates the loan applications in the order they are received to its four loan approval officers, one after the other in the same sequence, to avoid any bias of processing. The following data shows the loan approvals by the four officers:

Decision	Ms. Simran	Mr. Sajay	Ms. Saloni	Mr. Sumil
Approved	24	17	35	11
Rejected	16	18	15	20

Use an appropriate statistical test to determine if the loan approval decision is independent of the loan officer processing the loan application. Carry out the test at 5% level of significance.

15. The following results were obtained while studying the service time taken by two operators while serving the customers, selected at random:

Operator	Number of Customers	Mean Service time (in seconds)	Sum of Squares of Deviations from Mean
'A'	10	160	900
'B'	12	140	1080

Test the equality of variances of the service times of the two operators at 5% level of significance.

Analysis of Variance

12

Contents

LEARNING OBJECTIVES

The main objective of this chapter is to help in understanding the type of analysis that is required to test:

- Equality of means of more than two variables/factors representing life of picture tubes, returns on investments, effectiveness of training, impact of promotional strategies
- Validity or significance of multiple regression equation
- Equality of interactions between two factors like three advertising campaigns and three categories of populations like urban, semi-urban and rural.

Relevance

Mr. Pankaj, the CEO of a relatively new brand of television, was enjoying the bliss of continued phenomenal growth in sales during the last two years. However, to give further impetus to the sales, he planned to recruit about 200 field staff who could provide guidance to the staff of retail outlets as also the potential customers about the features of televisions, innovations taking place, etc., and thus bring out the cost-benefit aspect of the company's televisions. Mr. Pankaj intended to take fresh science graduates and provide them four weeks training at one of the three institutions which were equipped to provide such training. However, before awarding the contract, he thought of comparing the effectiveness of training imported by the three institutions. Out of the first batch of 30 officers, he deputed 10 officers to each of the three institutions where they were given training in three different modules, viz. technical, financial and behavioural. After the training, he hired the services of a reputed consulting organisation to conduct a quantitative assessment of the training imparted to the field officers. The consulting agency used ANOVA to evaluate the impact of training in the three modules as also the three institutions. Such an analysis helped Mr. Pankaj to award the training contract on objective basis, without any prejudice.

12.1 SIGNIFICANCE AND INTRODUCTION

The Analysis of Variance, popularly known as ANOVA, is a very useful technique for testing the equality of more than two means of populations. In Chapter 11 on Statistical Inference, we have discussed statistical tests, for testing the equality of means of two populations. In such cases, we can use either normal or 't' test described in Section 11.6.1. However, sometimes, we are required to test the equality of means of three or more populations For example, whether:

- The average life of light bulbs being produced in three different plants is the same.
- All the three varieties of fertilisers have the same impact on the yield of rice.
- The level of satisfaction (or any other parameter) among the participants in all IIMs is the same.
- The impact of training on salesmen trained in three institutes is the same.
- The service time of a transaction is the same on four different counters in a service unit.
- The average price of different commodities in four different retail outlets is the same.
- Performance of salesmen in four zones is the same.

The word 'analysis of variance' is used because the technique involves first finding out the total variation among the observations in the collected data, then assigning causes or components of variation to various factors and finally drawing conclusions about the equality of means.

ANOVA is also used to test the significance of a regression equation as a whole i.e. whether all the regression coefficients of a regression equation are all equal to 0.

The following example illustrates the concepts as well as the calculations involved in a typical ANOVA application.

Illustration 12.1

Three groups of five salesmen each, were imparted training related to marketing of consumer products by three Management Institutes. The amount of sales made by each of the salesmen, during the first month after training, were recorded and are given in Table 12.1

Table 12.1 Amount of Sales by Salesmen

		Salesmen				
Institutes		*1*	*2*	*3*	*4*	*5*
	1	67	70	65	71	72
	2	73	68	73	70	66
	3	61	64	64	67	69

The problem posed here is to ascertain whether the three Institutes' training programmes are equally effective in improving the performance of trainees. If m_1, m_2 and m_3 denote the mean effectiveness of the programmes of the three Institutes, statistically, the problem gets reduced to test the null hypothesis, i.e.

$$H_0 : \quad m_1 = m_2 = m_3$$

against the alternative hypothesis that it is not so, i.e.,

$$H_1 : \quad \textbf{All means are not equal}$$

It may be appreciated that the sales figures of all the 15 salesmen would have been varying from each other, even if they had attended the same training programme. This is due to inherent variations

that exist from person to person. Therefore, the variation in the 15 observations could be attributed to two factors – one, the training received at different institutes and the other, due to inherent factors present in the different salesmen and some other miscellaneous factors like their areas of operations, etc. The Analysis of Variance technique helps us to find out the variation due to both the factors, and also assess whether the variation among the three Institutes is significantly greater than the other factors. If if it so, the programmes of the three Institutes are not equally effective.

All the above 15 observations could be represented by a variable x_{ij} indicating score at i^{th} Institute for j^{th} salesman. It may be noted that while i varies from 1 to 3, j varies from 1 to 5. Further, let mean of salesmen at i^{th} Institute be

$$\frac{1}{5}\sum_{j=1}^{5} x_{ij} = \bar{x}_i$$

and mean of j^{th} Salesmen be $\quad \frac{1}{3}\sum_{i=1}^{3} x_{ij} = \bar{x}_j$

The methodology of carrying out the test is illustrated below.

The marginal row and column totals of values in Table 12.1, as also means of the three groups of salesmen trained at each Institute, are worked out and presented in Table 12.2.

Table 12.2 Mean Sales by Groups of Salesmen

		Salesmen (5 salesmen at each Institute)					*Total for Each Institute*	*Mean for Each Institute*
Institutes		*1*	*2*	*3*	*4*	*5*		$\bar{x}_i$
	1	67	70	65	71	72	345	$\bar{x}_1$ = **69**
	2	73	68	73	70	66	350	$\bar{x}_2$ = **70**
	3	61	64	64	67	69	325	$\bar{x}_3$ = **65**
Total for Three Salesmen		201	202	202	208	207	**1020**	← **Grand Total**
Mean for Three Salesmen		67	67.3	67.3	69.3	69	**68**	← **Grand Mean**

The grand mean sale of all the salesmen is worked out below:

$$\text{Grand Mean } (\bar{\bar{x}}) = \frac{\text{Grand Total}}{\text{Total Number of Observations}} = \frac{1020}{15}$$

$$= 68$$

Now, total variation among all the 15 observations = Sum of the squares of deviations of all the 15 observations from their grand mean viz. 68

$$= (66 - 68)^2 + \ldots + (61 - 68)^2 + \ldots + (69 - 68)^2$$

$$= 180$$

It may be noted that if all the salesmen were equally good with respect to their performance, and the training of all the three Institutes were equally effective, all the observations would have been identical, and this sum would have been zero.

Mathematically, this sum is expressed as:

$$\text{Total Variation or Total Sum of Squares} = \sum_I \sum_j (x_{ij} - \bar{\bar{x}})^2$$

where, x_{ij} is the sales by j^{th} salesman (j =1, 2, 3, 4, 5) trained by ith Institute (i = 1, 2, 3)
This can also be written as

$$= \sum_I \sum_j (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{\bar{x}})^2$$

where, $\bar{x}_i$ is the mean sale of the salesmen trained by i^{th} Institute.

$$= \sum_I \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_I \sum_j (\bar{x}_i - \bar{\bar{x}})^2 \quad \text{(proof not included)} \qquad \textbf{(12.1)}$$

It may be noted that the first term in the above expression indicates the sum of squares of the deviation of sales by jth salesman trained by ith Institute from the average sales of five salesmen trained by ith Institute. Further, the second term indicates the sum of squares of the deviations of the mean of all five salesmen trained by ith Institute and the grand mean for all 15 salesmen.

The first term, termed as **sum of squares within institutes**, is evaluated numerically as follows:

$$\begin{aligned}\sum_I \sum_j (x_{ij} - \bar{x}_i)^2 &= \{(x_{11} - \bar{x}_1)^2 + (x_{12} - \bar{x}_1)^2 + (x_{13} - \bar{x}_1)^2 + (x_{14} - \bar{x}_1)^2 + (x_{15} - \bar{x}_1)^2\} \\ &\quad + \{(x_{21} - \bar{x}_2)^2 + \ldots + (x_{25} - \bar{x}_2)^2\} \\ &\quad + \{(x_{31} - \bar{x}_3)^2 + \ldots + (x_{35} - \bar{x}_3)^2\} \\ &= (67 - 69)^2 + (70 - 69)^2 + \ldots + (72 - 69)^2 \\ &\quad + (73 - 70)^2 + \ldots (66 - 70)^2 \\ &\quad + (61 - 65)^2 + \ldots + (69 - 65)^2 \\ &= 110\end{aligned}$$

Further, the second term representing **sum of squares between institutes** is evaluated as follows:

$$\begin{aligned}\sum_I \sum_j (\bar{x}_i - \bar{\bar{x}})^2 &= 5 \sum_i (\bar{x}_i - \bar{\bar{x}})^2 \\ &= 5\ \{(\bar{x}_1 - \bar{\bar{x}})^2 + (\bar{x}_2 - \bar{\bar{x}})^2 + (\bar{x}_3 - \bar{\bar{x}})^2\} \\ &= 5\ \{(69 - 68)^2 + (70 - 68)^2 + (65 - 68)^2\} \\ &= 5(14) \\ &= 70\end{aligned}$$

Thus, we note that

Total Sum of Squares (180) = Sum of Squares within Institutes (110)
+ Sum of Squares between Institutes (70)

Now, the sum of squares $\sum_I \sum_j (x_{ij} - \bar{x}_i)^2$ in the expression (11.1) is called the **Sum of Squares (among salesmen) Within Institute**, and if all the salesmen in each Institute would have been same, this would have been zero as each x_{i1}, x_{i2}, x_{i3}, x_{i4} and x_{i5} in i^{th} Institute would have been therefore, equal to $\bar{x}_i$.

The sum of squares $\sum_I \sum_j (x_i - \bar{\bar{x}})^2$ in expression (12.1) is called **Sum of Squares Between Institutes**, and would have been zero if all the three Institutes would have been equally effective because then $\bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \bar{\bar{x}}$.

Thus, we can say that

Total Sum of Squares = Sum of Squares Within Institutes
+ Sum of Squares Between Institutes **(12.2)**

The above application is referred to as **one way** or **one factor** ANOVA as we have tested differences among only one factor i.e. Institute. The three Institutes are referred to as three levels or treatments.

Assuming all salesmen to be equally competent, the observations vary from each other due to one factor viz. training imparted by the institutes – the three Institutes being referred to as three levels of the factor. In general, the observations are collected in the following format:

		Observations for Each Treatment					
		1	*2*	*3*	...	*ni*	*Total for Each Treatment*
Treatments	1	x_{11}	x_{12}	x_{13}	...		$T_1 = \sum_{j=1}^{n_1} x_{1j}$
	2	x_{21}	x_{22}	x_{23}	...		T_2
	3	x_{31}	x_{32}				T_3
	4	x_{41}	x_{42}				T_4
	⋮	⋮	⋮				⋮
	i	x_{i1}	x_{i2}				$T_i = \sum_{j=1}^{n_i} x_{ij}$
	⋮						⋮
	k	x_{k1}	x_{k2}				$T_k = \sum_{j=1}^{n_k} x_{kj}$

It may be noted that total number of treatments are equal to k, and there are n_i observations for ith treatment.

The general result is of the form;

Total Sum of Squares = Sum of Squares Between Treatments or Due to Treatments
+ Sum of Squares Within Treatments or
'Sum of Squares Due to Error' or 'Error Sum of Squares'

There is a convenient mechanical way of performing various calculations and carrying out the requisite procedure for testing the equality of treatments, and is explained below:

Calculation of Grand Mean $\bar{\bar{x}} = \dfrac{\sum\sum x_{ij}}{n}$ **(12.3)**

$$\textbf{Correction Factor (Abbreviated as CF)} = \frac{\sum\sum x_{ij}^2}{n} \quad \textbf{(12.4)}$$

$$\textbf{Total Sum of Squares (TSS)} = \sum_i \sum_j x_{ij}^2 \; \textbf{– CF} \quad \textbf{(12.5)}$$

$$\textbf{Sum of Squares due to Between Treatments (SST)} = \sum_i \left(\frac{T_i^2}{n_i} \right) \textbf{– CF}$$

The expression (11.2) is generalised as follows

Total Sum of Squares = Sum of Squares Between Treatments + Sum of Squares within Treatments

The Sum of Squares within Treatments is referred to as Sum of Squares due to Error or Error Sum of Squares, and is abbreviated as **SSE.** This terminology is used to indicate that total variation is either due to differences among treatments or due to several other unknown and random factors. All these unknown and random factors are combined to cause 'Error', and the sum of squares due to these unknown and random factors is called 'Error Sum of Squarer' or 'Sum of Squares due to Error'.

Sum of Squares due to Error (SSE) = Total Sum of Squares – Sum of Squares Due to Treatments

SSE = TSS – SST

For testing the equality of means or equal effectiveness of all the k treatments, we compute

$$\textbf{Mean Sum of Squares Due to Treatments (MSST)} = \frac{\text{SST}}{k-1} \quad \textbf{(12.6)}$$

$$\textbf{Mean Sum of Squares Due to Error (MSSE)} = \frac{\text{SSE}}{n-k} \quad \textbf{(12.7)}$$

The Fisher's ratio 'F' **Statistic** is defined as

$$\boldsymbol{F} = \frac{\text{MSST}}{\text{MSSE}} \quad \textbf{(12.8)}$$

and is distributed as $\boldsymbol{F}$ **with** $(\boldsymbol{k-1}, \boldsymbol{n-k})$ **d.f.**

If the calculated value of F is more than the tabulated value of F at $(k-1, n-k)$ d.f., we reject the null hypothesis that all the means are equal. If, on the other hand, the calculated value is less than the tabulated value, we accept the null hypothesis.

It may be noted that more the variation among the Institutes, more will be SST and MSST, and accordingly, more the value of 'F' implying greater chances of the rejection of hypothesis about equality of means.

For the above example of the training programmes at three Institutes,

$$\text{Correction Factor (CF)} = \frac{\sum\sum x_{ij}^2}{15} = \frac{1020^2}{15}$$

$$= 69360$$

$$\text{Total Sum of Squares (TSS)} = \sum\sum x_{ij}^2 - \text{CF}$$

$$= 672 + 70^2 + \ldots 72^2 + 73^2 + 68^2 + 66^2 + \ldots + 69^2 - \text{CF}$$
$$= 69540 - 69360$$
$$= 180$$

Further,

$$\text{SST} = \sum\left(\frac{T_i^2}{n}\right) - \text{CF} = \frac{345^2}{5} + \frac{350^2}{5} + \frac{325^2}{5} - \text{CF}$$

$$= 69430 - 69360$$
$$= 70$$

Therefore,

$$\text{SSE} = \text{TSS} - \text{SST}$$
$$= 180 - 70$$
$$= 110$$

Thus,

$$\text{MSST} = \frac{\text{SST}}{k-1} = \frac{70}{3-1} = \mathbf{35}$$

$$\text{MSSE} = \frac{\text{SSE}}{n-k} = \frac{110}{15-3} = \mathbf{9.167}$$

$$F = \frac{\text{MSST}}{\text{MSSE}} \quad (k-1,\ n-k \text{ d.f.})$$

$$= \frac{35}{9.167}$$

$$= \mathbf{3.82}\ (2,\ 12 \text{ d.f.})$$

The tabulated value of F at 2, 12 d.f. vide Table *T*4 is 3.89. Since the calculated value is less than the tabulated value, and falls in the acceptance region, we do not reject the null hypothesis that all the training Institutes are equal with respect to the training programmes conducted by them.

All the above results are summarised in an ANOVA Table given below.

Table 12.3 ANOVA Table

Source of Variation	*Sum of Squares (SS)*	*Degrees of Freedom (d.f.)*	*Mean Sum of Squares (MSS) = SS ÷ d.f.*	*'F' = MSST ÷ MSSE*	*α = 5% F(Tabulated)*	*Decision*
Between Institutes	70	2	35	**3.82**	**3.89**	**Do not Reject H_0**
Within Institutes	110	12	9.167			
Total	180	14				

Thus, there is no significance difference among the means of sales of salesmen trained at three different institutes.

12.2 POST HOC TESTS–PAIRWISE COMPARISON OF TREATMENT MEANS

In ANOVA, when we reject the null hypothesis, all we conclude is that all the treatment means, say m_1, m_2 and m_3 are not equal. But the test does not indicate the comparative equality of pairs of means i.e., whether;

$$m_1 = m_2$$

or,

$$m_1 = m_3$$

or,

$$m_2 = m_3$$

Thus, for a complete analysis, one may like to test the following null hypotheses:

$H_o: \; m_1 = m_2 \qquad H_1: \; m_1 \neq m_2$

$H_o: \; m_1 = m_3 \qquad H_1: \; m_1 \neq m_3$

$H_o: \; m_2 = m_3 \qquad H_1: \; m_2 \neq m_3$

Such tests are called post hoc tests, as they are carried out based on the result of the earlier test.

We shall discuss two such tests viz:

(i) Tukey's HSD(Honestly Significant Difference) Test

(ii) Fisher's LSD (Least Significant Difference) Test

For the sake of simplicity, we have illustrated these tests with the following example wherein the number of observations for each treatment is equal.

12.2.1 Tukey's Honestly Significant Difference (HSD) Test for Multiple Comparison of Means

The test statistic 'w' for the Tukey's test for multiple comparisons of means is,

$$w = \frac{qs}{\sqrt{n}} \qquad \textbf{(12.9)}$$

where, s is the square root of total variance of the three groups (Institutes), n is the number of observations in each of the k groups, and q, called Studentised range is found from the Table 12.6 developed for the purpose for given level of significance and $k(n-1)$ d.f.

The value of s is calculated by the formula

$$s^2 = \frac{1}{k}\sum_{i=1}^{k}\{s_i^2\}$$

where s_i^2 is the sample variance of i^{th} group, and k = number of groups = 3, in the above illustration 12.1.

The difference between any two sample means can be compared with the value w. If any difference is more than this value of w, it implies than the means of those populations are not equal. For example, if $\bar{x}_1 - \bar{x}_3 > w$, it implies that $m_1 \neq m_3$.

The method is illustrated below through a numerical example.

Illustration 12.2

The following table gives the compensations for middle level executives belonging to three different sectors of banks. The compensation comprises salaries as well as bonuses. All the data is given in

Rs. per month. Further, the data represents independent random samples of total compensation for eight banking executives belonging to each of the 3 banking sectors.

Table 12.4 Executive Compensation

(Rs in '00 per month)

Corporate Banking	*Retail Banking*	*Personal Banking*
755	520	438
712	295	828
845	553	622
985	950	453
1300	930	562
1143	428	348
733	510	405
1189	864	938

We use the above data for illustrating the Tukey's test for comparing means of the three pairs of variables, viz.:

Corporate banking and Retail banking

Corporate banking and Personal banking

Retail banking and Personal banking

The data is also used to illustrate calculation of 95% confidence limits.

Before applying the Tukey's test, one has to ascertain that the differences among means are significant. Because, if the differences among means are insignificant, there is no use of using Tukey's test to test significant difference between any pairs of means. Thus, first we use ANOVA to test the following null hypothesis

$$H_0 : \quad m_1 = m_2 = m_3$$

where, m_1, m_2 and m_3 are the mean compensations for corporate, retail and personal banking.

The Tukey's test is to be used only if the above hypothesis is rejected i.e. the mean compensations of executives in the above sectors are not equal.

It may be verified that the ANOVA yields following results.

Table 12.5 ANOVA Table

Source of Variation	*Sum of Squares (SS)*	*Degrees of Freedom (d.f.)*	*Mean Sum of Squares (MSS) = SS ÷ d.f.*	*'F' = MSST ÷ MSSE*	*α = 5% F(Tabulated)@*	*Decision*
Between*	685129	2	342565 (MSST)	6.45	3.47	**Reject H_0**
Within*	1115232	21	53106 (MSSE)			
Total	1800362	23				

*Banking Sectors

$$*F = \frac{\text{MS Between Sector}}{\text{MS Within Sector}}$$

@Tabulated value of F at (2, 21) d.f. vide Table T4 is 3.47.

It may be noted that the null hypothesis about equality of means of the three sectors is rejected. We can, therefore, now proceed to illustrate Tukey's test. It may be appreciated that, if the null hypothesis for equality of means was accepted, there would have been no need to proceed further.

For calculating the value of the statistic

$$w = \frac{qs}{\sqrt{n}}, \quad \left(q = \frac{w}{s/\sqrt{n}} \text{ is called studentised range}\right)$$

we note from the following Table 12.6 that for $n = 8$, the value of q at (3, 21) d.f. at 5% level of significance = 3.56

Table 12.6 Values of q for some values of $k(n-1)$ d.f.

	k	
n	3	4
6	15 (3.67)	20 (3.96)
8	21 (3.56)	28 (3.86)
10	27 (3.51)	36 (3.81)

For calculating the values of s, we prepare the following Table to calculate sample variances of sectors (s_i's)

Table 12.7 Calculation of 's'

(Rs. in '00)

Corporate Banking x_1	$(x_{1i} - \bar{x}_1)^2$	*Retail Banking* x_2	$(x_{2i} - \bar{x}_2)^2$	*Personal Banking* x_3	$(x_{3i} - \bar{x}_3)^2$
755	41108	520	12377	438	18564
712	60393	295	113064	828	64389
845	12713	553	6123	622	2280
985	743	950	101602	453	14702
1300	117135	930	89252	562	150.06
1143	34318	428	41311	348	51189
733	50513	510	14702	405	28646
1189	53477	864	54173	938	132314
$\bar{x}_1 = 957.75$	Sum = 370398	$\bar{x}_2 = 631.25$	Sum = 432601	$\bar{x}_3 = 574.25$	Sum = 312233

From the above Table 12.6, we find,

Total Sample Variance $s^2 = \frac{\sum s_i^2}{3} = \frac{1}{3}((370398 + 432601 + 312233) \div 7) = 53106.32$

Therefore, $s = 230.45$

Therefore, the critical or honestly significant difference between the sample means $= w = \frac{qs}{\sqrt{n}}$

$= \frac{3.56 \times 230.45}{\sqrt{8}} =$ **290.05**

The value of s can also be obtained from ANOVA Table 12.5, as s^2 is equal to MSSE i.e. mean sum of squares within sectors. From the table, it is found to be equal to 53106 which is the same value as obtained from the above Table 12.7.

Now all the pairs of differences between the sample means are compared with this value 290.05. The values greater than this will imply significant difference between the group means.

From the above table the difference in group means are as follows:

Difference between Corporate Banking and Retail Banking

$$\bar{x}_1 - \bar{x}_2 = 957.75 - 631.25 = \mathbf{326.5}$$

Difference between Corporate Banking and Personal Banking

$$\bar{x}_1 - \bar{x}_3 = 957.75 - 574.25 = \mathbf{383.5}$$

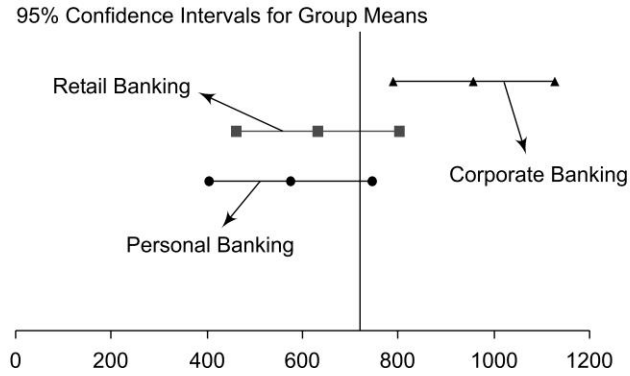
Difference between Retail Banking and Personal Banking

$$\bar{x}_2 - \bar{x}_3 = 631.25 - 574.25 = \mathbf{57}$$

Thus, we observe that the compensations between corporate banking and retail banking executives are significantly different and so is the compensation between corporate banking and personal banking executives. However, the difference in compensation between personal banking and retail banking executives is not significant.

It may be added that the 95% confidence limits for all the sector means have been derived and are given in tabular as well as pictorial form for the purpose of only better comprehension.

Corporate Banking	957.75 ± 169.44
Retail Banking	631.25 ± 169.44
Personal Banking	574.25 ± 169.44



12.2.2 Fisher's Least Significant Difference (LSD) Test

This test is analogous to student's 't' test for comparing means of two groups or sets of data where the data is available for more than two groups or sets. If there are three groups, the null hypotheses are:

$$H_o: \quad m_i = m_j \ (i, j = 1, 2, 3)$$

For the purpose of simplicity, it is assumed that all three groups have equal number of observations.

The test statistic is

$$t = |\bar{x}_i - \bar{x}_j| \qquad i, j = 1, 2, 3$$

where $|\bar{x}_i - \bar{x}_j|$ is the absolute difference between means of ith and jth groups.

To carry out the test, first we calculate LSD i.e. Least Significant Difference which is defined as,

$$\text{LSD} = t_{\alpha/2,\, k(n-1)} \sqrt{\text{MSSE}\left(\frac{1}{n} + \frac{1}{n}\right)} \qquad \textbf{(12.10)}$$

where n = number of observations in each of the three groups, k is the number of groups (assumed as 3), and MSSE is Within Groups Sum of Squares obtained from the ANOVA Table 12.5 prepared from the given data, $t_{\alpha/2,\, k(n-1)}$ is the value of Student's 't' at α% level of significance and $k(n-1)$ d.f. In our illustration, $n = 8$, $k = 3$, and within group sum of squares is the sum of squares within sectors calculated as 53106 in the ANOVA Table 12.5.

For the given example,

$$\sqrt{53106\left(\frac{1}{8} + \frac{1}{8}\right)}$$

$$= \sqrt{13276.5}$$

$$= 115.2$$

Value of 't' at 5% level of significance and $3(8-1) = 21$ d.f. $= 2.08$

Therefore,

$$\text{LSD} = 2.08 \times 115.2 = \textbf{239.6}$$

This is the Least Significant Difference between group means $\bar{x}_i$ and $\bar{x}_j$, for the null hypothesis $m_i = m_j$ to be rejected. It implies that the difference between two sample means to be significant must be at least 239.6. In the above example,

$$|\bar{x}_1 - \bar{x}_2| = 326.5$$

$$|\bar{x}_1 - \bar{x}_3| = 383.5$$

$$|\bar{x}_2 - \bar{x}_3| = 57$$

We note that the first two differences are greater than LSD (239.6) but the third one is less than LSD. Thus, while the hypotheses, H_o: $m_1 = m_2$ and H_o: $m_1 = m_3$ are rejected, the hypothesis,

H_o : $m_2 = m_3$ is not rejected.

It may be noted that the conclusions for comparison of group means are the same for Fisher's LSD test as for Tukey's HSD test.

12.3 TWO WAY OR TWO FACTOR ANOVA

In the above discussions, it may be noted that we have considered only one factor as a source of variation in the data. In Illustration 12.1, the factor was 'Training by Institute'. The different institutes contributed to the variation in performance of the salesmen. In Illustration 12.2, the factor responsible for the variation in the compensation was the banking sector – three different sectors contributed to variation in compensation.

Such type of analysis is called One Way or One Factor ANOVA. In this section, we discuss Two Way or Two Factor ANOVA. Here, the variation in the data is caused by two factors. This is illustrated with an example below, wherein the variation in the number of additional mobile phone subscribers is caused by different telecom companies as well as different periods of time, say months.

Example 12.1

The following Table gives the number of subscribers added by four major telecom players in India, in the months of August, September, October and November 2005. The data are given in 000's and are rounded off to the nearest 100, and are thus in lakhs.

Additions to Subscribers

(In Lakhs)

	Company			
Months	*Bharti*	*BSNL*	*Tata Indicom*	*Reliance*
August	6	6	2	5
September	7	6	2	3
October	7	6	6	4
November	7	8	7	4

(*Source:* Indiainfoline.com on India Mobile Industry)

Find out (at 5% level of significance):

(i) If the four companies significantly differ in their performance.

(ii) Is there significant difference between the months?

Solution:

First of all, the table is reconstructed by working out marginal totals for months and companies as follows:

Additions to Subscribers

(in Lakhs)

	Company				*Month*	
Months	*Bharti*	*BSNL*	*Tata Indicom*	*Reliance*	*Total*	*Average*
August	6	6	2	5	19	4.75
September	7	6	2	3	18	4.50
October	7	6	6	4	23	5.75
November	7	8	7	4	26	6.50
Company Total	27	26	17	16	Grand Total 86	
Average	6.75	6.5	4.25	4	Grand Mean 5.375	

From the above table, we work out the requisite sums of squares for preparing the ANOVA Table, as follows:

$$\text{Correction Factor (CF)} = \frac{(\sum\sum x_{ij})^2}{16} = \frac{86^2}{16}$$

$$= 462.25$$

$$\text{Total Sum of Squares} = \text{TSS} = \sum\sum x_{ij}^2 - \text{CF}$$

$$= 6^2 + 6^2 + \ldots + 7^2 + \ldots 4^2 - 462.25$$
$$= 514 - 462.25$$
$$= 51.75$$

$$\text{S.S. Between Months} = \sum\left(\frac{T_i^2}{n}\right) - \text{CF}$$

$$= \frac{19^2}{4} + \frac{18^2}{4} + \frac{23^2}{4} + \frac{26^2}{4} \; -]462.25$$

$$= 472.5 - 462.25$$
$$= 10.25$$

$$\text{S.S. Between Companies} = \frac{27^2 + 26^2 + 17^2 + 16^2}{4} - 462.25$$

$$= 487.5 - 462.25$$
$$= 25.25$$

Residual or Error Sum of Squares = 51.75 – 10.25 – 25.25 = 16.25

The above results are presented in the following ANOVA Table.

ANOVA Table

Source of Variation	*Sum of Squares (SS)*	*Degrees of Freedom (d.f.)*	*Mean Sum of Squares (MSS) = SS ÷ d.f.*	*'F' = MSST ÷ MSSE*	*α = 5% F(Tabulated)*	*Decision*
Between Months	10.25	3	3.42	3.42/1.81 = 1.89	3.86	
Between Companies	25.25	3	8.42	8.42/1.81 = 4.65	3.86	**Reject**
Error	16.25	9	1.81			
Total	51.75	15				

With calculated values of '*F*, given in the table, we can derive conclusions as follows:

(i) Since calculated value of *F* (4.65) is greater than 3.86, the tabulated value of F at 5% level of significance and 3, 9 d.f., it is concluded that there is a significant difference between the companies in terms of adding subscribers.

(ii) Since calculated value of *F* (1.89) is less than 3.86, the tabulated value of *F* at 5% level of significance and 3, 9 d.f., it is concluded that there is no significant difference between months in terms of total additional consumers.

12.4 TWO-WAY OR TWO-FACTOR ANOVA WITH INTERACTION

It has been observed that there are variations in the pay packages offered to the MBA students. These variations could be either due to specialisation in a field or due to the institute wherein they study.

The variation could also occur due to interaction between the institute and the field of specialisation. For example, it could happen that the marketing specialisation at one institute might fetch a better pay package rather than that at the other institute. These presumptions could be tested by collecting the following type of data for a number of students with different specialisations and different institutes. However, for the sake of simplicity of calculations and illustration, we have taken only two students each for each interaction between institute and field of specialisation.

Illustration 12.3

The data is presented below in a tabular format.

	Institute A	*Institute B*	*Institute C*
Marketing	8	10	8
	10	11	7
Finance	9	11	5
	11	12	6
HRD	9	9	8
	7	7	5

Here, the test of hypotheses will be

For Institute:

H_0 : Average pay packages for all the three institutes are equal

H_1 : Average pay packages for all the three institutes are **not** equal

For Specialisation:

H_0 : Average pay packages for all the specialisations are equal

H_1 : Average pay packages for all the three specialisations are **not** equal

For Interaction:

H_0 : Average pay packages for all the nine interactions are equal

H_1 : Average pay packages for all the nine interactions are **not** equal

Now we present the analysis of data and the resultant ANOVA Table.

Incidentally, this is an example of two-way or two-factor (Institute and Specialisation) with interaction.

First of all, a Table is reconstructed by working out marginal totals for Institutes and Specialisations as follows:

	Institute A	*Institute B*	*Institute C*	*Total*
Marketing	8	10	8	54
	10	11	7	
Finance	9	11	5	54
	11	12	6	
HRD	9	9	8	45
	7	7	5	
Total	54	60	39	153

From the above table, we work out the requisite sums of squares for preparing the ANOVA Table, as follows:

$$\text{Correction Factor (CF)} = \frac{(\text{Sum of All Observations})^2}{\text{Total number of Observations}}$$

$$= \frac{(153)^2}{18}$$

$$= 1300.5$$

$$\text{Total Sum of Squares (TSS)} = (\text{Sum of Squares of All Observations}) - \text{CF}$$
$$= 1375 - 1300.5$$
$$= 74.5$$

$$\text{Sum of Squares Between or due to Fields of Specialisation (Row: SSR)} = \frac{54^2}{6} + \frac{54^2}{6} + \frac{45^2}{6} - 1300.5$$
$$= 9$$

$$\text{Sum of squares due to Institutes Column: SSC} = \frac{54^2}{6} + \frac{60^2}{6} + \frac{39^2}{6} - 1300.5$$
$$= 39$$

Sum of squares due to Interaction between Institutes and fields of specialisations

$$\text{SSI} = n\sum\sum(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}..)^2 \qquad \textbf{(12.11)}$$

where, n is the number of observations for each interaction. In this example, it is equal to 2.

$\bar{x}_{ij}$ is the mean of the observations of i^{th} row and j^{th} column.

$\bar{x}_{i.}$ is the mean of the observations in the i^{th} row.

$\bar{x}_{.j}$ is the mean of the observations in the j^{th} column.

$\bar{x}..$ is the grand mean of all the observations.

These terms can be calculated by first calculating the means of all the interactions as also the means of corresponding rows and columns, as given below,

	Institute A	*Institute B*	*Institute C*	*Row Mean*
Marketing	9	10.5	7.5	**9**
Finance	10	11.5	5.5	**9**
HRD	8	8	6.5	**7.5**
Column Mean	**9**	**10**	**6.5**	**Grand Mean 8.5**

and then calculating the sum of squares for each interaction by the formula (12.5) as follows.

	Institute A	*Institute B*	*Institute C*	
Marketing	0.25	0	0.25	
Finance	0.25	1	2.25	
HRD	0	1	1	
			Total	**6**

Thus,

Interaction SS (SSI) = 2 × 6 = 12

Sum of Squares Due to Error = Total Sum of Squares – SS Due to Specialisation (Row: SSR) (column: SS – SS Due to Interaction – SS due to Institute

= TSS – SSR – SSC – SSI

= 74.5 – 9 – 39 – 12

= 14.5

Now, the ANOVA Table can be prepared as follows.

ANOVA Table

Source of Variation	*Sum of Squares (SS)*	*Degrees of Freedom (d.f.)*	*Mean Sum of Squares (MSS) = SS ÷ d.f.*	*'F' = MSST ÷ MSSE*	*α = 5% F(Tabulated)*	*Decision*
Specialisation	9	2	4.5	2.7931	4.2565	
Institute	39	2	19.5	12.103	4.2565	**Reject**
Interaction	12	4	3	1.8621	3.6331	
Error	14.5	9	1.61111			
Total	74.5	17				

From the above table, we conclude that the while the pay packages among the institutes are significantly different, there is no significant difference among the pay packages for fields of specialisations as also among interactions between the institutes and the fields of specialisations.

Example 12.2

RELIABLE tyre dealer wishes to assess the equality of lives of three different brands of tyres sold by it. It also wants to assess whether the lives of these tyres is the same for four brands of cars on which they are being used. Thus, each brand of tyre was tested on each of the four brands of cars. Further, the dealer wishes to ascertain the equality of lives of tyres for each combination of brand of tyre and car. The mileages obtained are given as follows:

(Mileage in '000 Kms)

		Car Brands			
		A	*B*	*C*	*D*
Tyre	I	32	30	34	36
		31	29	33	38

(*Contd*)

(*Contd*)

		33	28	36	39
Brands		31	30	35	40
	II	38	39	40	41
		37	40	41	39
		38	41	42	40
		39	39	43	42
	III	32	33	40	45
		30	32	42	43
		31	30	41	42
		33	31	40	46

This example is solved through template on ANOVA with Interaction, and is given in Section 12.7 titled USING EXCELL.

12.5 USE OF ANOVA FOR TESTING THE SIGNIFICANCE OF REGRESSION MODEL AND REGRESSION COEFFICIENTS

In Regression Analysis, we derive regression equations. However these equations cannot be used for further prediction of dependent variable or interpretation of regression coefficients unless the regression equation and regression coefficients are found to be statistically significant. While the significance of the regression equation is tested by '*F*' test, the significance of the regression coefficients can be tested with the help of '*t*' test.

The hypotheses to test the significance of a regression model,

$$y = b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

giving relationship of dependent variable y and independent variables $x_1, x_2, x_3, \ldots, x_k$, are,

$H_0 : b_1 = b_2 = b_3 = \ldots = b_k = 0$

H_1: at least one of the b's is not equal to zero

where, $b_1, b_2, \ldots, b_k$ are the regression coefficients for variables $x_1, x_2, \ldots x_k$, respectively.

We shall test the significance of the regression model with the help of the '*F*' test.
Here, the hypotheses are:

H_0: $b_1 = b_2 = 0$

H_1: At least one of the b's is not equal to zero

The '*F*' statistic is defined as

$$F = \frac{\text{Explained Variation}/k}{\text{Error SS}/(n-k-1)} = \frac{\text{Regression SS}/k}{\text{Error SS}/(n-k-1)} \qquad (12.12)$$

where, k is the number of independent variables (equal to 2 in the above equation), and n is the number of observations for each of the independent variables. Further details are given in Chapter 14 on Multivariate Statistical Techniques.

12.6 USING EXCEL

The major disadvantage of ANOVA is the extensive calculations associated with it. The calculations could be quite time consuming and complicated specially for the large data.

The various statistical computations described in this chapter can be performed with the help of MS Excel. One could use different formulae and functions for these computations.

As a matter of convenience, for ease in calculations, we have provided template 'ANOVA' for the calculations involved in this chapter. There are THREE different worksheets. These are:

- One Way
- Two Way
- Two Way with interaction

We now discuss the methodology to use each of these worksheets.

ONE WAY

This worksheet provides the calculations for one-way or one factor ANOVA.

Microsoft Excel - ANOVA

File Edit View Insert Format Tools Data Window Help

O19 =

	A	B	C	D	E	F	L	M	N	O	P	Q	R	S	T
1		One-Way ANOVA													
2		Illustration 12.1 Salesmen Training													
3								ANOVA Table							
4		Inst 1	Inst 2	Inst 3									Alpha		
5	1	67	73	61									5%		
6	2	70	68	64				Source	SS	df	MS	F	$F_{critical}$	*p*-value	Result
7	3	65	73	64				Between columns	70	2	35	3.8182	3.8853	0.0521	cannot reject
8	4	71	70	67				Within Columns	110	12	9.1667				
9	5	72	66	69				Total	180	14					
10															

The above is a snapshot of the worksheet 'One Way' of the template 'ANOVA'. In the above worksheet one has to enter the data in columns B. C, D, E, F, from 5th row downwards. The maximum number of variables possible is five. If one enters data in the above mentioned columns, the template automatically calculates the rest of the computations and gives ANOVA Table at the right side of the data.

The ANOVA Table has following columns:

- Source : Source of variation.
- SS : Sum of squares for respective source
- df : Degrees of freedom for the source
- MS : Mean Sum of Squares
- F : Calculated value of F statistic
- F critical : Table value of F statistic
- p-value : Probability value of the F statistic
- Result : The decision of accepting or rejecting the hypothesis

The source of variation contains three factors, columns, rows and total.

We have solved Illustration 12.1, relating to salesmen training from different institutes, using the above worksheet.

TWO WAY

This worksheet provides the calculations for two way or two factor ANOVA.

Microsoft Excel - ANOVA

File Edit View Insert Format Tools Data Window Help

E3 =

ANOVA

Example 12.1 Telecom Players

	C1	C2	C3	C4	C5
R1	6	6	2	5	
R2	7	6	2	3	
R3	7	6	6	4	
R4	7	8	7	4	

ANOVA Table

					Alpha 5%		
Source	SS	df	MS	F	$F_{critical}$	p-value	Result
Columns	25.25	3	8.42	4.66	3.86	0.0313	Reject
Rows	10.25	3	3.42	1.89	3.86	0.2014	cannot reject
Error	16.25	9	1.81				
Total	51.75	15					

The above is a snapshot of the worksheet 'Two Way' of the template 'ANOVA'. In the above worksheet, one has to enter the data in columns B, C, D, E, F, from 5th row downwards. The maximum number of columns possible is five. If one enters data in the above mentioned columns, the template automatically calculates the rest of the computations and gives ANOVA Table on the right side of the data.

The explanation of ANOVA Table is the same as in the previous worksheet, except that now the source contains four different factors, 'columns, rows, error and total'.

We have solved Example 12.1, relating to telecom players, using the above worksheet.

TWO WAY WITH INTERACTION

This worksheet provides the calculations for two-way or two factor ANOVA with interaction.

Microsoft Excel - ANOVA

File Edit View Insert Format Tools Data Window Help

F10

ANOVA

Example 12.2 (Milage car and tyre brands)

			A	B	C	D
I	1		32	30	34	36
			31	29	33	38
			33	28	36	39
			31	30	35	40
II	2		38	39	40	41
			37	40	41	39
			38	41	42	40
			39	39	43	42
III	3		32	33	40	45
			30	32	42	43
			31	30	41	42
			33	31	40	46

ANOVA Table

					Alpha 5%	
Source	**SS**	**df**	**MS**	***F***	$F_{critical}$	***p*-value**	Result
Columns	499.40	3	166.47	105.60	2.87	0.0000	**Reject**
Rows	338.67	2	169.33	107.42	3.26	0.0000	**Reject**
Interaction	199.67	6	33.28	21.11	2.36	0.0000	**Reject**
Error	56.75	36	1.58				
Total	1,094.48	47					

The above is a snapshot of the worksheet 'Two Way' of the template 'ANOVA'. In the above worksheet one has to enter the data in columns B. C, D, E, F, from 5th row downwards. The maximum number of columns possible is five. Here, each row has more than one observation. If one enters data in the above mentioned columns and rows, the template automatically calculates the rest of the computations and gives ANOVA Table at right side of the data.

The explanation of ANOVA Table is the same as in the previous worksheet, except that now the source contains five different factors, columns, rows, interaction, error and total.

We have solved Example 12.2, relating to mileage of tyres, using the above worksheet.

GLOSSARY

Analysis of Variance ANOVA	The process for splitting the variation of a group of observations into assignable causes and setting up various significance tests
One-Way or One-Factor ANOVA	When the source of variation in the observations is primarily due to one factor

Two-Way or Two-Factor ANOVA	When there are two factors as sources of variation in the observations
Interaction of Factors	Joint impact of two factors
Post hoc Test	Test carried out based on the result of the earlier test

EXERCISES

1. In order to promote use of credit cards by a bank, the users of all three types of card holders viz. 'Gold International', 'Gold' and 'Silver' were offered 5% discount on their bills. After the offer period, five cards were selected at random from each category. Percentage of increase in the bills of each type are given below:

	Increase in Bill Amounts in %		
Card No.	*'Gold International'*	*'Gold'*	*'Silver'*
1	10	15	40
2	20	20	35
3	15	30	30
4	15	20	40
5	25	25	45

   Are the increases among different cards bill amounts equal? What does it imply? Use Tukey' HSD and Fisher's LSD tests to carry out pair wise comparison of three types of cards.
2. Three groups of almost equally effective salesmen for consumer products were deputed to sales training programmes conducted by three different training institutes. The amounts of sales made by each of the 15 salesman during the first week, after completing the training, are as follows:

   Institute A : 65, 68, 64, 70, 71, 75
   Institute B : 73, 68, 73, 69, 64
   Institute C : 64, 64, 66, 69

   Can the difference among mean sales by the three groups be attributed to chance at the level of significance = 0.05?
3. The following data gives monthly rates of returns for the shares of three companies over the six month period from October 2006 to March 2007.

Month	*A*	*B*	*C*
October '06	3.5	5.2	4.0
November '06	–2.5	–4.0	3.6
December '06	–5.6	5.4	6.0
January '07	4.0	–4.6	–3.5
February '07	5.0	6.6	6.0
March '07	7.5	8.0	5.2

   Can we conclude that the average monthly rates of returns on all the shares have been the same?

4. A company dealing with office equipment has offices in 3 cities. The company wants to compare the volume of sales in the offices during the five day promotional period. The data about sales on each of the five days are given below:

Office	*Sales in Rs. 1000's*				
Mumbai	85	125	110	93	160
Delhi	124	75	82	135	105
Chennai	95	130	145	190	170

Test for differences between the sizes of sales among the 3 offices.

5. The following table gives the retail prices of a certain commodity in some selected shops in four cities.

City	*Prices*
A	62, 58, 60, 59
B	50, 48, 52
C	70, 65, 68, 64, 63
D	80, 85, 82, 78

Can we say that the prices of the commodity differ in the four cities?

6. It is required to assess the life of three different types of tyres. To eliminate, the effect of the brand of cars on which they are being used, each type of tyre was tested on each of the brand of cars. The mileages obtained are given as follows:

	Tyre Brand Mileage (in '000 Kms.)		
Car Brand	*I*	*II*	*III*
A	32	38	32
B	30	39	33
C	35	40	40
D	38	41	45

Use an appropriate test to assess whether there is any association between car brands and types of tyres.

7. A paint manufacturing company is marketing paint tins whose maximum retail price is Rs 600. The salesmen are free to negotiate the price with the retailers, subject to a minimum of Rs. 400. Three of its salesmen viz. '*A*', '*B*' and '*C*' have reported the following prices with the five shops each

	Sales Person		
Retail Shop	*A*	*B*	*C*
1	450	430	420
2	435	410	200
3	425	405	400
4	400	420	410
5	440	425	415

Test whether the average prices negotiated by the three salesmen are equal.

Non-Parametric Tests

13

Contents

1. Relevance– Advantages and Disadvantages
2. Tests for
   - Randomness of a Series of Observations – Run Test
   - Change in value or Preference—Sign Test
   - Specified Mean or Median of a Population – Signed Rank Test
   - Goodness of Fit of a Distribution – Kolmogorov-Smirnov Test
   - Comparing Two Populations – Kolmogorov-Smirnov Test
   - Equality of Two Means – Mann-Whitney ('U') Test
   - Equality of Several Means
     - — Wilcoxon-Wilcox Test
     - — Kruskel-Wallis Rank Sum ('H') Test – One Way ANOVA
     - — Friedman's (F') Test – Two Way ANOVA
   - Rank Correlation – Spearman's
   - Rank Correlation-Kendal's Tau

LEARNING OBJECTIVES

This chapter aims to

- Highlight the importance of non-parametric tests when the validity of assumptions in tests of significance, described in Chapters 11 and 12 on Statistical Inference and ANOVA, respectively, is doubtful.
- Describe certain non-parametric tests of significance relating to randomness, mean of a population, means of two or more than two populations, rank correlation, etc.

Relevance

The General Manager of **Evergrowing Corporation** was going through the report of a consultant who had suggested certain measures to accelerate the growth of the Corporation. He was bit surprised, that the report highlighted the importance of training to the staff and contained several options to impart the training. However, the selection of options was left to the Corporation. The General Manager requested the Chief of HRD Department to evaluate the effectiveness of the various options suggested and recommend appropriate training strategies, at the earliest. The HRD Chief was aware of the use of statistical tests to compare the effectiveness of training programmes, but was not too sure of the assumptions that were required for application of those tests. He, therefore, discussed the matter with his friend who was

teaching Statistics at a management institute. His friend was also not too sure of the validity of the assumptions in the training environment in the Corporation. He, however, advised, that since the time available was short and the effectiveness of training was not such a precise variable like, physical measurements of an item or monetary evaluation of an option, the HRD chief could use non-parametric tests to evaluate and compare the effectiveness of various training options. This advice helped the HRD Chief to evaluate the effectiveness of various training programmes and submit the report, well in time, to the General Manager.

13.1 SIGNIFICANCE AND INTRODUCTION

All the tests of significance, discussed in Chapters 10 and 11, are based on certain assumptions about the variables and their statistical distributions. The most common assumption is that the samples are drawn from a normally distributed population. This assumption is more critical when the sample size is small. When this assumption or other assumptions for various tests described in the above chapters are not valid or doubtful, or when the data available is 'ordinal' (rank) type, we take the help of non-parametric tests. For example, in the student's '*t*' test for testing the equality of means of two populations based on samples from the two populations, it is assumed that the samples are from normal distributions with equal variance. If we are not too sure of the validity of this assumption, it is better to apply the test given in this chapter.

While the parametric tests refer to some parameters like mean, standard deviation, correlation coefficient, etc., the non-parametric tests, also called as distribution-free tests, are used for testing other features also, like randomness, independence, association, rank correlation, etc.

In general, we resort to use of non-parametric tests where

- The assumption of normal distribution for the variable under consideration or some assumption like homogeniety of variances of two or more populations for a parametric test is not valid or is doubtful.
- The hypothesis to be tested does not merely relate to the parameter of a population but also aspects like randomness, correlation and independence.
- The numerical accuracy of collected data is not fully assured and therefore the data is converted into order or rank from.
- The data is available only in terms of order or rank.
- Results are required rather quickly through simple calculations.

However, the non-parametric tests have the following limitations or disadvantages:

- They ignore a certain amount of information.
- They are often not as efficient or reliable as parametric tests.

The above advantages and disadvantages are in consonance with the general premise in statistics that is, a method that is easier to calculate or does not utilise the full information contained in a sample, is less reliable.

The use of non-parametric tests, involves a trade-off. While the 'efficiency or reliability' is 'lost' to some extent, the 'ability' to use 'lesser' information and to calculate 'faster' is 'gained'.

There are a number of tests in statistical literature. However, we have discussed only the following tests:

Types and Names of Tests for

- Randomness of a Series of Observations–Run Test.
- Specified Mean or Median of a Population–Signed Rank Test.
- Goodness of Fit of a Distribution – Kolmogorov-Smirnov Test.
- Comparing Two Populations – Kolmogorov- Smirnov Test.
- Equality of Two Means – Mann - Whitney ('U') Test.
- Equality of Several Means:
  - — Wilcoxon-Wilcox Test
  - — Kruskel-Wallis Rank Sum ('H') Test
  - — Friedman's ('F')Test
- Rank Correlation – Spearman's

13.2 TEST FOR RANDOMNESS IN A SERIES OF OBSERVATIONS—THE RUN TEST

This test has been evolved for testing whether the observations in a sample occur in a certain order or they occur in a random order. The hypotheses are:

H_0: The sequence of observations is random

H_1: The sequence of observations is not random

The only condition for validity of the test is that the observations in the sample must be obtained under similar conditions.

The procedure for carrying out this test is as follows.

First, all the observations are arranged in the order they are collected. Then the median is calculated. All the observations in the sample larger than the median value are given a + sign and those below the median are given a – sign. If there are an odd number of observations, the median observation is ignored. This ensures that the number of + signs is equal to the number of – signs.

A succession of values with the same sign is called a **run** and the number of runs, say **R**, gives an idea of the randomness of the observations. This is the test statistic. If the value of R is low, it indicates a certain trend in the observations, and if the value of R is high, it indicates presence of some factor causing regular fluctuations in the observations.

The test procedure is numerically illustrated below.

Illustration 13.1

The Director of an institute wanted to peruse the answer books of 20 students out of 250 students in the paper on Statistical Methods. The Registrar selected 20 answer books one by one from the entire lot of 250. The books had the marks obtained by the students indicated on the first page. The Director wanted to ascertain whether the Registrar had selected the answer books in a random order. He used the above test for this purpose.

He found out that the marks, given below, have median as 76. The observations less than the median are under marked as the sign –, and observations more than the median are under marked as the sign +.

58	61	78	72	69	65	79	75	80	81	82	79	58	77	71	73	74	79	81	82
–	–	+	–	–	–	+	–	+	+	+	+	–	+	–	–	–	+	+	+
	-(1)-	(2)		-(3)-		(4)	(5)		-(6)-			(7)	(8)		-(9)-			-(10)-	

It may be noted that the number of (+) observations is equal to number of (–) observations; both being equal to 10. As defined above, succession of values with the same sign is called a run. Thus the first run comprises observations 58 and 61, with the – sign, second run comprises only one observation i.e. 78, with the + sign, the third observation comprises of three observations 72, 69 and 65, with the negative sign, and so on.

Total number of runs $R = 10$.

This value of R lies inside the acceptance interval found from Table T11 as from 7 to 15 at 5% level of significance. Hence the hypothesis that the sample is drawn in a random order is accepted.

Applications

(i) Testing Randomness of Stock Rates of Return The number-of-runs test can be applied to a series of stock's rates of return, for each of the trading day, to see whether the stock's rates of return are random or exhibit a pattern that could be exploited for earning profit.

(ii) Testing the Randomness of the Pattern Exhibited by Quality Control Data Over Time If a production process is in control, the distribution of sample values should be randomly distributed above and below the center line of a control chart. We can use this test for testing whether the pattern of, say, 10 sample observations, taken over time, is random,

13.3 SIGN TEST

The sign test is a non-parametric statistical procedure for fulfilling the following objective:

(i) To identify preference for one of the two brands of a product like tea, soft drink, mobile phone, TV, and/or service like cellular company, internet service provider.

(ii) To determine whether a change being contemplated or introduced is found favourable.

The data collected in such situations is of the type, + (preference for one) or '–' (preference for the other) signs. Since the data is collected in terms of plus and minus signs, the test is called 'Sign Test'. The data having 10 observations is of the type:

$$+ , +, –, +, –, –, +, +, +, –$$

Illustration 13.2

The Director of a management institute wanted to have an idea of the opinion of the students about the new time schedule for the classes

8.00 a.m. to 2.00 p.m.

He randomly selected a representative sample of 20 students, and recorded their preferences. The data was collected in the following form:

Student No.	*In Favour of the Proposed Option: 1, Opposed to the Option: 2*	*Sign (+ for 1, – for 2)*
1	1	+
2	1	+
3	2	–
4	1	+
5	2	–
6	2	–

(*Contd*)]

(*Contd*)]

7	2	–
8	2	–
9	1	+
10	1	+
11	2	–
12	1	+
13	2	–
14	2	–
15	1	+
16	1	+
17	1	+
18	1	+
19	1	+
20	1	+

Here, the hypothesis to be tested is:

$$H_0 : p = 0.50$$
$$H_1 : p \neq 0.50$$

Under the assumption that

H_0 is true i.e. $p = 0.50$

the number of + signs follows a binomial distribution with $p = 0.50$.

Let x denote the number of '+' signs.

It may be noted that if the value of x is either low or high, then the null hypothesis will be rejected in favour of the alternative

$$H_1 : p \neq 0.50$$

For $\alpha = 0.05$, the low value is the value of 'x' for which the area is less than 0.025, and the high value is the value of 'x' for which the area is more than 0.025.

If the alternative is

$$H_1: p < 0.50$$

then the value of 'x' has to be low enough to reject the null hypothesis in favour of

$$H_1: p < 0.50.$$

If the alternative is

$$H_1: p > 0.50.$$

then the value of 'x' has to be high enough to reject the null hypothesis in favour of

$$H_1: p > 0.50.$$

With a sample size of n =20, one can refer to the following table showing the probabilities for all the possible values of the binomial probability distribution with $p = 0.5$:

Number of '+' Signs	*Probability*
0	0.0000
1	0.0000
2	0.0002
3	0.0011

(*Contd*)]

(*Contd*)]

4	0.0046
5	0.0148
6	0.0370
7	0.0739
8	0.1201
9	0.1602
10	0.1762
11	0.1602
12	0.1201
13	0.0739
14	0.0370
15	0.0148
16	0.0046
17	0.0011
18	0.0002
19	0.0000
20	0.0000

The binomial probability distribution as shown above can be used to provide the decision rule for any **sign** test up to a sample size of $n = 20$. With the null hypothesis $p = 0.5$ and the sample size n, the decision rule can be established for any level of significance. In addition, by considering the probabilities in only the lower or upper tail of the binomial probability distribution, we can develop rejection rules for one-tailed tests.

The previous table gives the probability of the number of plus signs under the assumption that H_0 is true, and is, therefore, the appropriate sampling distribution for the hypothesis test. This sampling distribution is used to determine a criterion for rejecting H_0. This approach is similar to the method used for developing rejection criteria for hypothesis testing given in the Chapter 11 on Statistical Inference.

For example, let $\alpha = 0.05$, and the test be two sided. In this case, the alternative hypothesis will be

$$H_1 : p \neq 0.50$$

and we would have a critical or rejection region area of 0.025 in each tail of the distribution. Starting at the lower end of the distribution, we see that the probability of obtaining zero, one, two, three or four plus signs is 0.0000 + 0.0000 + 0.0002 + 0.0011 + 0.0046 + 0.0148 = 0.0207. Note that we stop at 5 + signs because adding the probability of six + signs would make the area in the lower tail equal to .0 207 + .0370 = .0577, which substantially exceeds the desired area of **0.025**. At the upper end of the distribution, we find the same probability of 0.0207 corresponding to 15, 16, or 17, 18, 19 or 20 + signs. Thus, the closest we can come to $\alpha = 0.05$ without exceeding it is 0.0207 + 0.0207 = 0.0414. We therefore adopt the following rejection criterion:

Reject H_0 if the number of + signs is less than 6 or greater than 14.

Since the number of + signs in the given illustration are 12, we cannot reject the null hypothesis, and thus the data reveals that the students are not against the option.

In case, sample size is greater than 20, we can use the large-sample normal approximation of binomial probabilities to determine the appropriate rejection rule for the sign test.

13.3.1 Sign Test for Paired Data

The sign test can also be used to test the hypothesis that there is "no difference" between two distributions of continuous variables x and y.

Let $p = P(x > y)$, and then we can test the null hypothesis

$$H_0 : p = 0.50$$

This hypothesis implies that given a random pair of measurements (x_i, y_i), then both x_i and y_i are equally likely to be larger than the other.

To perform the test, we first collect independent pairs of sample data from the two populations as

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

We omit pairs for which there is no difference so that we may have a reduced sample of pairs. Now, let 'r' be the number of pairs for which $x_i - y_i > 0$. Assuming that H_0 is true, then 'r' follows a binomial distribution with $p = 0.5$. It may be noted that if the value of r is either low or high, then the null hypothesis will be rejected in favour of the alternative

$$H_1 : p \neq 0.50$$

For $\alpha = 0.05$, the low value is the value of 'r' for which the area is less than 0.025, and the high value is the value of 'r' for which the area is more than 0.025.

If the alternative hypothesis is $H_1 : p > 0.50$ means that 'x' measurements tend to be higher than 'y' measurements.

The right-tail value is computed by the total of probabilities that are greater than 'r', and is the p value for the alternative H_1: $p > 0.50$. If one alternative hypothesis is H_1: $p < 0.50$.

It means that 'y' measurements tend to be higher that 'x'.

Illustration 13.3

The HRD Chief of an organisation wants to assess whether there is any significant difference in the marks obtained by 12 trainee officers in the papers on Indian and Global Perspectives conducted after the induction training.

(72, 70)	(82, 79)	(78, 69)	(80, 74)	(64, 66)	(78, 75)
(85, 86)	(83, 77)	(83, 88)	(84, 90)	(78, 72)	(84, 82)

We convert the above data in the following tabular form where '+' indicates values of $x_i > y_i$.

Trainee Officer	$x_i > y_i$
1	+
2	+
3	+
4	+
5	
6	+
7	
8	+
9	
10	
11	+
12	+

It may be noted that the number of + signs is 8.

Here, we test the null hypothesis that there is "no significant difference" in the scores in the two papers, with the two-sided alternative that there is a "significant difference". In symbolic notation:

$$H_0: \ m_1 = m_2$$
$$H_1: m_1 \neq m_2$$

Number of '+' Signs	*Probability*
0	0.0002
1	0.0029
2	0.0161
3	0.0537
4	0.1208
5	0.1934
6	0.2256
7	0.1934
8	0.1208
9	0.0537
10	0.0161
11	0.0029
12	0.0002

Let the level of significance be $\alpha = 0.05$.

In a two-sided test, we would have a rejection region or an area of approximately 0.025 in each tail of the distribution.

From the above table, we note that the probability of obtaining zero, one, or two plus signs is 0.0002 + 0.0029 + 0.0161 = 0.0192. Thus, the probability of getting number of plus signs up to 3 is 0.0192. Since it is less than $\frac{1}{2}\alpha$ i.e. 0.025,

H_0 will be rejected.

Similarly, the probability of getting 10, 11 and 12 + signs is 0.0002 + 0.0029 + 0.0161 = 0.0192. Thus, , the probability of getting number of plus signs beyond 9 is 0.0192. Since it is less than $\frac{1}{2}\alpha$ i.e. 0.025,

H_0: will be rejected.

Thus, H_0 will be rejected, if the number of + signs is up to 3 or beyond 9, at 5% level of significance. We, therefore, follow the following rejection criterion:

Reject H_0 if the number of + signs is less than 3 or greater than 9

Since, in our case the number of plus signs is only 8, we cannot reject the null hypothesis. Thus, we conclude that there is no significant difference between the marks obtained in the papers on Indian and Global perspectives.

If we want to test that the average marks obtained in the paper on Indian Perspective is more than the average marks in the paper on Global Perspectives, the null hypothesis is

$$H_0: \ m_1 = m_2$$

and the alternative hypothesis is

$$H_1: \ m_1 > m_2$$

Thus, the test is one sided. If the level of significance is 0.05, the entire critical region will be on the right side. The probability of getting number of plus signs 8 or more than 8 is 0.0792, which is

more than 0.05, and, therefore, the null hypothesis cannot be rejected. It may be noted that if the number of observations for which ($x_i > y_i$) would have been 9 or more than null hypothesis would have been rejected.

13.4 TEST FOR SPECIFIED MEAN OR MEDIAN OF A POPULATION—THE SIGNED RANK TEST

This test has been evolved to investigate the significance of the difference between a population mean or median and a specified value of the mean or median, say m_0.

The hypotheses are as follows:

Null Hypothesis H_o: $m = m_0$
Alternative Hypothesis H_1: $m \neq m_0$

The test procedure is numerically illustrated below.

Illustration 13.4

In the Illustration 13.1 relating to answer books of MBA students, the Director further desired to have an idea of the average marks of students. When he enquired from the concerned Professor, he was informed that the Professor had not calculated the average but felt that the mean would be 70, and the median would be 65. The Director wanted to test this, and asked for a sample of 10 randomly selected answer books. The marks on those books are tabulated below as x_i s.

Null Hypothesis: H_o: $m = 70$
Alternative Hypothesis: H_1: $m \neq 70$

Sample values are as follows:

x_i (Marks)	55	58	63	78	72	69	66	79	75	80
$x_i - m_0$	–15	–12	–7	+8	+2	–1	–4	+9	+5	+10
$\|x_i - m_0\|$*	15	12	7	8	2	1	4	9	5	10
Ranks of $\|x_i - m_0\|$										
+ or – signs	1	2	6	5	9	10	8	4	7	3
Ranks with signs of respective $(x_i - m_0)$	–1	–2	–6	+5	+9	–10	–8	+4	+7	+3

Any sample values equal to m_0 are to be discarded from the sample.
*Absolute value or modulus of $(x_i - m_0)$.

Now, Sum of (+) ranks = 28 Sum of (–) ranks = 27

Here, the statistic 'T' is defined as the minimum of the sum of positive ranks and sum of negative ranks.

Thus, T = Minimum of 28 and 27 = 27

The critical value of T at 5% level of significance is 8 (for n = number of values ranked = 10) vide Table T12.

Since the calculated value 27 is more than the critical value, the null hypothesis is not rejected. Thus the Director does not have sufficient evidence to contradict the Professor's guess about mean marks as 70.

It may be noted that the criteria using rank methods is reverse of the parametric tests wherein the null hypothesis is rejected if the calculated value exceeds the tabulated value.

In the above example, we have tested a specified value of the mean. If the specified value of the median is to be tested, the test is exactly similar only the word 'mean' is substituted by 'median'.

For example, if the Director wanted to test whether the median of the marks was 70, the test would have resulted in same values and the same conclusions. It may be verified that the Director does not have sufficient evidence to contradict the Professor's guess about median marks as 68.

Example 13.1

The following table gives the 'real' annual income that senior managers actually take home in certain countries, including India. These have been arrived at, by US based Hay Group, in 2006, after adjusting for the cost of living, rental expenses and purchasing power parity.

Overall Rank	*Country*	*Amount (in Euros)*
5	Brazil	76,449
26	China	42,288
8	Germany	75,701
2	India	77,665
9	Japan	69,634
3	Russia	77,355
4	Switzerland	76,913
1	Turkey	79,021
23	UK	46,809
13	USA	61,960

Test whether (1) the mean income is equal to 70,000 (2) The median value is 70,000.

It may be verified that both the hypotheses are not rejected.

13.5 TEST FOR GOODNESS OF FIT OF A DISTRIBUTION (ONE SAMPLE)—KOLMOGOROV-SMIRNOV

In Chapter 11, we have discussed χ^2 test as the test for judging goodness of fit of a distribution. However, it is assumed that the observations come from a normal distribution. When this condition is doubtful to be valid, we may use Kolmogorove-Smirnov test.

The test is used to investigate the significance of the difference between observed and expected cumulative distribution function for a variable with a specified theoretical distribution which could be Binomial, Poisson, Normal or an Exponential. It tests whether the observations could reasonably have come from the specified distribution. Here,

Null Hypothesis H_o: The sample comes from a specified population

Alternative Hypothesis H_1: The sample does not come from a specified population

The testing procedure envisages calculations of observed and expected cumulative distribution functions denoted by $F_o(x)$ and $F_e(x)$, respectively, derived from the sample. The comparison of the two distributions for various values of the variable is measured by the test statistic

$$D = |F_o(x) - F_e(x)| \quad \textbf{(13.1)}$$

If the value of the difference of D is less, the null hypothesis is likely to be accepted. But if the difference is more, it is likely to be rejected. The procedure is explained below for testing the fitting of an uniform distribution.

Illustration 13.5

Suppose we want to test whether availing of educational loan by the students of 5 Management Institutes is independent of the Institute in which they study.

The following data gives the number of students from each of the five institutes viz. *A*, *B*, *C*, *D* and *E*. These students were out of 50 students selected randomly from each institute.

The relevant data and calculations for the test are given in the following table.

Institutes

	A	*B*	*C*	*D*	*E*
Out of Groups of 50 students	5	9	11	16	19
Observed Cumulative Distribution Function *Fo* (*x*)	5/60	14/60	25/60	41/60	60/60
Expected Cumulative Distribution Function *Fe* (*x*)	12/60	24/60	36/60	48/60	60/60
|*Fo* (*x*) – *Fe* (*x*)|	7/60	10/60	**11/60**	7/60	0

From the last row of the above table, it is observed that

Maximum Value of *D* i.e. Max *D* = 11/60 = **0.183**

The tabulated value of *D* at 5% level of significance and 60 d.f. = 0.176 vide Table T13.

Since calculated value is more than the tabulated value, the Null Hypothesis is rejected. Thus, availing of loan by the students of 5 management institutes is not independent of the institutes in which they study.

Comparing χ^2 and K-S Test

The Chi-square test is the most popular test of goodness of fit. On comparing the two tests, we note that the K-S test is easier to apply. While χ^2- test is specially meant for categorical data, the K-S test is also applicable for random samples from continuous populations. Further, the K-S statistic utilises each of the *n* observations. Hence, the K-S test makes better use of available information than Chi-square statistic.

13.6 COMPARING TWO POPULATIONS—KOLMOGOROV-SMIRNOV TEST

This test is used for testing whether two samples come from two identical population distributions. The hypotheses are:

$$\mathbf{H_0}: F_1(x) = F_2(x)$$

i.e. the two populations of random variables *x* and *y* are almost the same.

$$\mathbf{H_1}: F_1(x) \neq F_2(x)$$

i.e. the two populations are not same that is claimed.

There are no assumptions to be made for the populations. However, for reliable results, the samples should be sufficiently large, say 15 or more.

The procedure for carrying out this test is as follows.

Given samples of sizes n_1 and n_2 from the two populations, the cumulative distribution functions $F_1(x)$ can be determined and plotted. Hence the maximum value of the difference between the plotted values can thus be found and compared with a critical value obtained from the concerned Table. If the observed value exceeds the critical value, the null hypothesis that the two population distributions are identical is rejected.

The test is explained through the illustration given below.

Illustration 13.6

At one of the Management Institutes, a sample of 30 (15 from commerce background and 12 from Engineering background), Second Year MBA students was selected, and the data was collected on their background and CGPA scores at the end of the First year.

The data is given as follows.

Sr No	*CGPA Commerce*	*CGPA Engineering*
1	3.24	2.97
2	3.14	2.92
3	3.72	3.03
4	3.06	2.79
5	3.14	2.77
6	3.14	3.11
7	3.06	3.33
8	3.17	2.65
9	2.97	3.14
10	3.14	2.97
11	3.69	3.39
12	2.85	3.08
13	2.92	3.3
14	2.79	3.25
15	3.22	3.14

Here, we wish to test that the CGPAs for students with Commerce and Engineering backgrounds follow the same distribution.

The value of the statistic '*D*' is calculated by preparing the table given below.

The calculation of values in different columns is explained below.

Col. (iii) The first cumulative score is same as the score in Col. (ii). The second cumulative score is obtained by adding the first score to second score, and so on. The last, i.e. 15th score, is obtained by adding fourteen scores to the fifteenth score.

Col. (iv) The cumulative distribution function for any observation in second column is obtained by dividing the cumulative score for that observation by cumulative score of the last i.e. 15th observation.

Sr No (i)	*CGPA Commerce (ii)*	*Cumulative Score of CGPAs (iii)*	*Cumulative Distribution Function of CGPAs Fi(C) = Col.(iii)/47.25 (iv)*	*CGPA Engineering (v)*	*Cumulative Score of CGPAs (vi)*	*Cumulative Distribution Function of Fi(E)= Col.(v)/45.84 (vii)*	*Difference Fi(C)–Fi(E): Di (viii)*
1	3.24	3.24	0.06857	2.97	2.97	0.06479	0.003781
2	3.14	6.38	0.13503	2.92	5.89	0.12849	0.006536
3	3.72	10.1	0.21376	3.03	8.92	0.19459	0.019167
4	3.06	13.16	0.27852	2.79	11.71	0.25545	0.023065
5	3.14	16.3	0.34497	2.77	14.48	0.31588	0.029092

(*Contd*)

(*Contd*)

6	3.14	19.44	0.41143	3.11	17.59	0.38373	0.027703
7	3.06	22.5	0.47619	3.33	20.92	0.45637	0.01982
8	3.17	25.67	0.54328	2.65	23.57	0.51418	0.029101
9	2.97	28.64	0.60614	3.14	26.71	0.58268	0.023459
10	3.14	31.78	0.67259	2.97	29.68	0.64747	0.025123
11	3.69	35.47	0.75069	3.39	33.07	0.72142	**0.029265**
12	2.85	38.32	0.81101	3.08	36.15	0.78861	0.022393
13	2.92	41.24	0.8728	3.3	39.45	0.8606	0.012202
14	2.79	44.03	0.93185	3.25	42.7	0.9315	0.000351
15	3.22	47.25	1	3.14	45.84	1	0

Values in Col. (vi) and Col. (vii) are obtained just like the values in Col. (iii) and Col. (iv).

The test statistic D is calculated from Col. (viii) of the above Table as

Maximum of D_i = 'D' = Maximum of $\{F_i(C) - F_i(E)\}$ = **0.0293**

Since the calculated value of the statistic viz. 'D' is **0.0293** is less than its tabulated value 0.338

at 5% level (for n =15), the null hypothesis that both samples come from the same population is not rejected.

13.7 EQUALITY OF TWO MEANS—MANN-WHITNEY 'U' TEST

This test is used with two independent samples. It is an alternative to the '*t*' test without the latter's limiting assumptions of samples coming from normal distributions with equal variance.

For using the U test, all observations are combined and ranked as one group of data, from smallest to largest. The largest negative score receives the lowest rank. In case of ties, the average rank is assigned. After the ranking, the rank values for each sample are totaled. The U statistic is calculated as follows:

$$U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1 \qquad \textbf{(13.2)}$$

or,

$$U = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2 \qquad \textbf{(13.3)}$$

where,

n_1 = Number of observations in sample 1; n_2 = Number of observations in sample 2

R_1 = *Sum* of ranks in sample 1; R_2 = *Sum* of ranks in sample 2.

For testing purposes, the smaller of the above two *U* values is used.

The test is explained through a numerical example given below.

Example 13.2

Two equally competent groups of 10 salespersons were imparted training by two different methods *A* and *B*. The following data gives sales of a brand of paint, in 5 kg. tins, per week per salesperson after one month of receiving the training. Test whether both the methods of imparting training are equally effective.

Salesman Sr. No.	Training Method A Sales	Salesman Sr. No.	Training Method B Sales
1	1,500	1	1,340
2	1,540	2	1,300
3	1,860	3	1,620
4	1,230	4	1,070
5	1,370	5	1,210
6	1,550	6	1,170
7	1,840	7	950
8	1,250	8	1,380
9	1,300	9	1,460
10	1,710	10	1,030

Solution:

Here, the hypothesis to be tested is that both the training methods are equally effective, i.e.

$$H_0: m_1 = m_2$$
$$H_1: m_1 \neq m_2$$

where, m_1 is the mean sales of salespersons trained by method A, and m_2 is the mean sales of salespersons trained by method B.

The following table giving the sales values for both the groups as also the combined rank of sales for each of the salesman is prepared to carry out the test.

Training Method A			Training Method B		
Salesman Sr. No.	Sales 5 kg. Tins	Combined Rank	Salesman Sr. No	Sales 5 kg. Tins	Combined Rank
1	1,500	14	1	1,340	10
2	1,540	15	2	1,300	8.5
3	1,860	20	3	1,620	17
4	1,230	6	4	1,070	3
5	1,370	11	5	1,210	5
6	1,550	16	6	1,170	4
7	1,840	19	7	950	1
8	1,250	7	8	1,380	12
9	1,300	8.5	9	1,460	13
10	1,710	18	10	1,030	2
Average Sales	1,515	—	—	1,253	—
Sum of Ranks	—	**R_1 = 134.5**	—	—	**R_2 = 75.5**
'*U*' values	—	**20.5**	—	—	**79.5**

The '*U*' statistic for both the training methods are calculated as follows:

$$U = 10 \times 10 + \frac{10(10+1)}{2} - 134.5 \quad \textbf{= 20.5}$$

$$U = 10 \times 10 + \frac{10(10+1)}{2} - 75.5 \quad \textbf{= 79.5}$$

The tabulated or critical value of U with $n_1 = n_2 = 10$, for $\alpha = 0.5$, is 23, for a two-tailed test vide Table T14.

It may be noted that in this test too, like the signed rank test in Section 13.4, the calculated value must be smaller than the critical value to reject the null hypothesis.

Since the calculated value 20.5 is smaller than the critical value 23, the null hypothesis is rejected that the two training methods are equally effective. It implies that Method *A* is superior to Method *B*.

13.8 EQUALITY OF SEVERAL MEANS—THE WILCOXON-WILCOX TEST FOR COMPARISON OF MULTIPLE TREATMENTS

This test is analogous to ANOVA, and is used to test the significance of the differences among means of several groups recorded only in terms of ranks of observations in a group. However, if the original data is recorded in absolute values, it could be converted into ranks. The hypotheses, like in ANOVA are:

H_0: $m_1 = m_2 = m_3$

H_1: All means are not equal

However, there is one important difference between ANOVA and this test. While ANOVA tests only the equality of all means, this test goes beyond that to compare even equality of all pairs of means.

The procedure for carrying out this test is illustrated by an example, given below.

Example 13.3

A prominent company is a dealer of a brand of LCD TV, and has showrooms in five different cities viz. Delhi (D), Mumbai (M), Kolkata (K), Ahmedabad (A) and Bangalore (B). Based on the data about sales in the showroom for 5 consecutive weeks, the ranks of LCD TVs in five successive weeks are recorded as follows:

	Delhi	*Mumbai*	*Kolkata*	*Ahmedabad*	*Bangalore*
Week	*D*	*M*	*K*	*A*	*B*
1	1	5	3	4	2
2	2	4	3	5	1
3	1	3	4	5	2
4	1	4	3	5	2
5	2	3	5	4	1
Rank Sum*	7	19	18	23	8

* Rank Sum is calculated by adding the ranks for all the five weeks for each of the city.

From the values of 'Rank Sum', we calculate the net difference in 'Rank Sum' for every pair of cities, and tabulate as follows:

Difference in Ranks	*D*	*M*	*K*	*A*	*B*
D	0	12	11	**16***	1
M	12	0	1	4	11

(*Contd*)

(*Contd*)

K	11	1	0	5	10
A	**16***	4	5	0	**15***
B	1	11	10	**15***	0

The critical value for the difference in 'Rank Sums' for number of cities = 5, number of observations for each city = 5, and 5% level of significance is 13.6, vide Table T15.

Comparing the calculated difference in rank sums with this 13.6, we note that difference in rank sums in A (Ahmedabad) and D (Delhi) as also, difference in rank sums between A (Ahmedabad) and B (Bangalore) are significant.

Note: **In the above case, if the data would have been available in terms of actual values rather than ranks alone, ANOVA would just lead to conclusion that means of D, M, K, A and B are not equal, but would have not gone beyond that. However, the above test concludes that mean of Ahmedabad is not equal to mean of Delhi as also mean of Bangalore. Thus, it gives comparison of all pairs of means.**

13.9 KRUSKALL-WALLIS RANK SUM TEST FOR EQUALITY OF MEANS (OF SEVERAL POPULATIONS) (H-TEST) (ONE-WAY ANOVA OR COMPLETELY RANDOMISED DESIGN)

This test is used for testing equality of means of a number of populations, and the null hypothesis is of the type

$$H_0: m_1 = m_2 = m_3 \ \ldots \text{ (can be even more than 3)}$$

It may be recalled that H_0 is the same as in ANOVA. However, here the ranks of observations are used and not actual observations.

The Kruskall-Wallis test is a One-Factor or One-Way ANOVA with values of variable as ranks. It is a generalisation of the two samples Mann-Whitney, "*U*" rank sum test for situations involving more than two populations.

The procedure for carrying out the test involves, assigning combined ranks to the observations in all the samples from smallest to largest. The rank sum of each sample is then calculated. The test statistic H is calculated as follows, and is distributed as χ^2 with (k–1) d.f.

$$H = \frac{12}{n(n-1)} \sum_{j=1}^{k} \frac{T_j^2}{n_j} - 3(n+1) \qquad \textbf{(13.4)}$$

where

T_j = Sum of ranks for treatment j
n_j = Number of observations for treatment j
$n = \Sigma n_j$ = Total number of observations
k = Number of treatments

The test is illustrated through an example given below.

Example 13.4

A chain of departmental stores opened three stores in Mumbai. The management wants to compare the sales of the three stores over a six day long promotional period. The relevant data is given below.

(Sales in Rs. Lakhs)

Store 'A' Sales	*Store 'B' Sales*	*Store 'C' Sales*
16	20	23
17	20	24
21	21	26
18	22	27
19	25	29
29	28	30

Use the Kruskal-Wallis test to compare the equality of mean sales in all the three stores.

Solution:

The combined ranks to the sales of all the stores on all the six days are calculated and presented in the following Table.

Store 'A'		*Store 'B'*		*Store 'C'*	
Sales	*Combined Rank*	*Sales*	*Combined Rank*	*Sales*	*Combined Rank*
16	1	20	5.5	23	10
17	2	20	5.5	24	11
21	7.5	21	7.5	26	13
18	3	22	9	27	14
19	4	25	12	29	16.5
29	16.5	28	15	30	18
	$T_1 = 34.0$		$T_2 = 54.5$		$T_3 = 82.5$

It may be noted that the rank given for sales as 21 in Store '*A*' and for sales as 21 in Store '*B*' are given equal ranks as 7.5. Since there are six ranks below, the rank for 21 would have been 7, but since the value 21 is repeated, both the values get rank as average of 7 and 8 as 7.5. The next value 22 has been assigned the rank 9. If there were three values as 21, the rank assigned to the values would have been average of 7, 8 and 9 as 8, and the next value would have been ranked as 10.

Now the *H* statistic is calculated as

$$H = \frac{12}{18(8-1)}\left[\frac{34.0^2 + 54.5^2 + 82.5^2}{6}\right] - 3(18+1)$$

$$= \frac{12}{18(18-1)}\left(\frac{10932.5}{6}\right) - 57$$

$$= \frac{12}{342} \times \frac{10932.5}{6} - 57$$

$$= 63.93 - 57 = \mathbf{6.93}$$

The value of χ^2 at 2 d.f. and 5% level of significance is 5.99.

Since the calculated value of *H* is 14.45, and is greater than tabulated value, and it falls in the critical region, we reject the null hypothesis that all the three stores have equal sales.

13.10 TEST FOR GIVEN SAMPLES TO BE FROM THE SAME POPULATION—FRIEDMAN'S TEST (TWO-WAY ANOVA OR RANDOMISED BLOCK DESIGN)

Friedman's test is a non-parametric test for testing hypothesis that a given number of samples have been drawn from the same population. This test is similar to ANOVA but it does not require the assumption of normality and equal variance. Further, this test is carried out with the data in terms of ranks of observations rather than their actual values, like in ANOVA. It is used whenever the number of samples is greater than or equal to 3 (say k) and each of the sample size is equal (say n) like two-way analysis of variance. **In fact, it is referred to as Two-Way ANOVA.** The null hypothesis to be tested is that all the k samples have come from identical populations.

The use of the test is illustrated below through a numerical example.

Illustration 13.7

Following data gives the percentage growth of sales of three brands of refrigerators, say '*A*', '*B*' and '*C*' over a period of six years.

	Percentage Growth Rate of the Brands		
Year	*Brand 'A'*	*Brand 'B'*	*Brand 'C'*
1	15	14	32
2	18	15	30
3	15	11	27
4	13	19	38
5	20	18	33
6	27	20	22

In this case, the null hypothesis is that there is no significant difference among the growth rates of the three brands. The alternative hypothesis is that at least two samples (two brands) differ from each other.

Under null hypothesis, **the Friedman's test statistic is:**

$$F = \frac{12}{nk(k+1)}\left(\sum_{j=1}^{k} R_j^2\right) - 3n(k+1) \qquad \textbf{(13.5)}$$

where,

k = Number of samples (brands) = 3 (in the illustration)

n = Number of observations for each sample (brand) = 6 (in the illustration)

R_j = Sum of ranks of jth sample (brand)

It may be noted that this '*F*' is different from Fisher's '*F*'.

The statistical tables exist for the sampling distribution of Friedman's '*F*', but these are not readily available for various values of n and k. However, the sampling distribution of '*F*' can be approximated by a χ^2 (chi-square) distribution with $k - 1$ degrees of freedom. The chi-square distribution table shows, that with $3 - 1 = 2$ degrees of freedom, the chi-square value at 5% level of significance is $\chi^2 = 5.99$.

If the calculated value of '*F*' is less than or equal to the tabulated value of chi-square (χ^2 at 5% level of significance), growth rates of brands are considered statistically the same. In other words,

there is no significant difference in the growth rates of the brands. In case the calculated value exceeds the tabulated value, the difference is termed as significant.

For the above example, the following null hypothesis is framed:

H_o: There is no significant difference in the growth rates of the three brands of refrigerators.

For calculation of F, the following table is prepared. The figures in brackets indicate the rank of growth of a brand in a particular year—the lowest growth is ranked 1 and the highest growth is ranked 3.

The Growth Rates of Refrigerators of Different Brands

(Growth Rates of Refrigerators)

Year	*Brand 'A'*	*Brand 'B'*	*Brand 'C'*	*Total Ranks (Row Total)*
1	15 (2)	14 (1)	32 (3)	6
2	18 (2)	15 (1)	30 (3)	6
3	15 (2)	11 (1)	27 (3)	6
4	13 (l)	19 (2)	38 (3)	6
5	20 (2)	18 (1)	33 (3)	6
6	27 (3)	20 (1)	22 (2)	6
Total Ranks (R_j) (Column Total)	12	7	17	36

With reference to the above table, Friedman's test amounts to testing that sums of ranks (R_j) of various brands are all equal.

The value of F is calculated as,

$$F = \frac{12}{72}(12^2 + 7^2 + 17^2) - 3 \times 6(3 + 1)$$

$$= \frac{482}{6} - 72 = 80.3 - 72 = \mathbf{8.3}$$

It is observed that the calculated value of 'F' statistic is greater than the tabulated value of χ^2 (5.99 at 5% level of significance and 2 d.f.). Hence, the hypothesis that there is no significant difference in the growth rates of the three brands is rejected.

Therefore, we conclude that there is a significant difference in the growth rates of the three brands of refrigerators, during the period under study. The significant difference is due to the best growth rate of brand '*C*'.

In the above example, if the data were given for six showrooms instead of six years, the test would have remained the same.

13.11 TEST FOR SIGNIFICANCE OF SPEARMAN'S RANK CORRELATION

In Chapter 10 on Simple Correlation and Regression, the Spearman's rank correlation has been discussed in Section 10.4 and is defined as

$$r_s = \frac{\sum d_i^2}{n(n^2-1)}$$

where, n is the number of pairs of ranks given to individuals or units or objects, and d_i is the difference in the two ranks given to ith individual/unit/object

There is no statistic to be calculated for testing the significance of the rank correlation. The calculated value of r_s is itself compared with the tabulated value of r_s, given in Table T9, at 5% or 1% level of significance. If the calculated value is more than the tabulated value, the null hypothesis that there is no correlation in the two rankings is rejected.

Here, the hypotheses are as follows

H_0: $\rho_s = 0$
H_1: $\rho_s \neq 0$

In Example 10.1, of Chapter 10 on Correlation and Regression, the rank correlation between priorities of 'Job Commitment Drivers' among executives from India and Asia Pacific was found to be 0.9515. Comparing this value with the tabulated value of r_s at n, i.e.10 d.f. and 5% level of significance as 0.636, we find that the calculated value is more than the tabulated value, and hence we reject the null hypothesis that there is no correlation between priorities of 'Job Commitment Drivers' among executives from India and Asia Pacific.

13.11.1 Test for Significance of Spearman's Rank Correlation for Large Sample Size

If the number of pairs of ranks is more than 30, the distribution of the rank correlation r_s under the null hypothesis that $\rho_s = 0$, can be approximated by normal distribution with mean 0 and s.d. as $\frac{1}{\sqrt{n-1}}$. This can be expressed in symbolic form as follows:

$$\text{for } n > 30,\ r_s \sim N\left(0, \frac{1}{n-1}\right) \qquad \textbf{(13.6)}$$

It may be verified that the rank correlation between the rankings in Mumbai and Bangalore is 0.544. Thus the value of z i.e. standard normal variable is

$z = \frac{0.544 - 0}{1/7} = 0.544 \times 7 = 3.808$ which is more than 1.96, the value of z at 5% level of significance. Thus, it may be concluded that the ranking between Mumbai and Bangalore are significantly correlated.

13.12 TESTING EQUALITY OF SEVERAL RANK CORRELATIONS

Sometimes, more than 2 ranks, are given to an individual/entity/object, and we are required to test whether all the three rankings are equal. Consider the following situation wherein some cities have been ranked as per three criteria

Illustration 13.8

As per a study published in *Times of India* dt. 4th September 2006, the rankings of ten cities as per 'earning', 'investing' and 'living' criteria are as follows:

	Rankings		
City	*Earning*	*Investing*	*Living*
Bangalore	2	6	1
Coimbtore	5	1	5
Surat	1	2	10
Mumbai	7	5	2
Pune	3	4	7
Chennai	9	3	3
Delhi	4	7	8
Hyderabad	8	8	6
Kolkata	10	10	4
Ahmedabad	6	9	9

Source: City Skyline of India 2006 published by Indicus Analytics.

Here, the test statistic is

$$F = \frac{s_1^2}{s_2^2} \sim F_{(k-1),\, k(n-1)} \qquad \textbf{(13.7)}$$

which is distributed as Fisher's F with $\{k-1,\ k(n-1)\}$ d.f.

where k is the number of cities equal to 10 in the illustration, n is the number of criteria of rankings, equal to 3 in the illustration.

The calculations for s_1^2 and s_2^2 are indicated below

$$s_1^2 = \frac{s_d}{n(k-1)} \qquad \textbf{(13.8)}$$

where, s_d is the sum of squares of difference between mean rank of a city and the overall mean ranks of all cities.

$$s_2^2 = \frac{\left(s - \frac{s_d}{n}\right)}{k(n-1)} \qquad \textbf{(13.9)}$$

where,

$$s = \frac{nk(k^2-1)}{12} \qquad \textbf{(13.10)}$$

The required calculations are illustrated below:

City	Rankings			Sum of City Rankings	Mean of City Rankings	Difference from Grand Mean Ranking	Square of Difference from Grand Mean Ranking
	Earning	Investing	Living				
Bangalore	2	6	1	9	3	–2.5	6.25
Coimbtore	5	1	5	11	3.67	–1.83	3.36
Surat	1	2	10	13	4.33	–1.17	1.36
Mumbai	7	5	2	14	4.67	–0.83	0.69
Pune	3	4	7	14	4.67	–0.83	0.69
Chennai	9	3	3	15	5	–0.5	0.25
Delhi	4	7	8	19	6.33	0.83	0.69
Hyderabad	8	8	6	22	7.33	1.83	3.36
Kolkata	10	10	4	24	8	2.5	6.25
Ahmedabad	6	9	9	24	8	2.5	6.25
				Total =165	Mean = 5.5	Total	29.2

$$n = 3,\ k = 10$$

$$s = \frac{3 \times 10(100-1)}{12} = 247.5$$

$$s_d = 29.2$$

$$s_1^2 = \frac{29.2}{3(10-1)} = 1.08 \qquad \text{(d.f. = 9)}$$

$$s_2^2 = \frac{\left(247.5 - \frac{29.2}{3}\right)}{10(3-1)} = 11.89 \qquad \text{(d.f. = 20)}$$

$$F = \frac{s_2^2}{s_1^2} \quad (\text{Because } s_2^2 > s_1^2)$$

It may be recalled that in the Fisher's F ratio of two sample variances, the greater one is taken in the numerator, and d.f. of F are taken, accordingly vide Section 10.11

$$F = \frac{11.89}{1.08} = 11.00$$

Tabulated value of 'F' (20, 9) d.f. at $\alpha = 0.05$ as per Table T4.

$$F_{20,\,9} = 2.94$$

Since the calculated value is more than the tabulated value of F, we reject the null hypothesis, and conclude that rankings on the given parameter are not equal.

13.13 KENDALL'S RANK CORRELATION COEFFICIENT

In addition to the Spearman's correlation coefficient, there exists one more rank correlation coefficient, introduced by Maurice Kendall, in 1938. Like Spearman's correlation coefficient it is also

used when the data is available in ordinal form. It is more popular as Kendall's Tau, and denoted by the Greek letter 'tau'(corresponding to 't' in English). It measures the extent of agreement or association between rankings, and is defined as

$$\tau = \frac{(n_c - n_d)}{(n_c + n_d)} \tag{13.11}$$

where

n_c : Number of concordant pairs of rankings
n_d : Number of disconcordant pairs of rankings

The maximum value of $n_c + n_d$ could be the total number of pairs of ranks given by two different persons or by the same person on two different criteria. For example, if the number of observations is, say 3 (call them 'a', 'b' and 'c'), then the pairs of observations will be: ab, ac and bc. Similarly, if there are four pairs, the possible pairs are six in number as follows:

ab, ac, ad, bc, bd, cd

It may be noted that, more the number of concordant pairs, more will be the value of the numerator and more the value of the coefficient, indicating higher degree of consistency in the rankings.

A pair of subjects i.e. persons or units, is said to be concordant, if for a subject, the rank of both variables is higher than or equal to the corresponding rank of both the variables. On the other hand, if for a subject, the rank of one variable is higher or lower than the corresponding rank of the other variable and the rank of the other variable is opposite i.e. lower/higher than the corresponding rank of the other variable, then the pair of subjects is said to be discordant.

The concepts of concordant and discordant pairs, and the calculation of 'τ' are explained through an example given below.

As per a study published in Times of India dated 4th September 2006, several cities were ranked as per the criteria of 'Earning', 'Investing' and 'Living'. It is given in the CD in the chapter relating to 'Simple Correlation and Regression'. An extract from the full table is given as follows:

City	*Ranking as per 'Earning'*	*Ranking as per 'Investing'*
Chennai	3	2
Delhi	1	3
Kolkata	4	4
Mumbai	2	1

We rearrange the table by arranging cities as per 'Earning' ranking

City	*Ranking as per 'Earning'*	*Ranking as per 'Investing'*
Delhi	1	3
Mumbai	2	2
Chennai	3	1
Kolkata	4	4

Now, we form all possible pairs of cities. In this case, total possible pairs are $4C_2 = 6$, viz. DM, DC, DK, MC, MK and CK.

The status of each pair i.e. whether, it is concordant or discordant, along with reasoning is given in the following table:

Pair	Concordant (C)/ Disconcordant (D)	Reasoning
DM	D	Because both pairs 1,2 and 3,2 are in opposite order
DC	D	Because both pairs 1,3 and 3,1 are in opposite order
DK	C	Because both pairs 1,4 and 3,4 are in ascending order/same direction
MC	D	Because both pairs 2,3 and 2,1 are in opposite order
MK	C	Because both pairs 2,4 and 2,4 are ascending order/same direction
CK	C	Because both pairs 3,4 and 1,4 are in ascending order/same direction

It may be noted that, for a pair of subjects (cities in the above case), when a subject ranks higher on one variable also ranks higher on the variable or even equal to the rank of the other variable, then the pair is said to be concordant. On the other hand, if a subject ranks higher on one variable and lower on the other variable, it is said to be discordant.

From the previous table, we note that

$$n_c = 3 \quad \text{and} \quad n_d = 3$$

Thus, the Kendall's coefficient of correlation or concordance is

$$= (3 - 3)/5 = 0.$$

As regards the possible values and interpretation of the value of τ, the following results can be concluded:

- If the association between the two rankings is perfect (i.e., the two rankings are the same), the coefficient has value 1.
- If the association between the two rankings is perfect (i.e., one ranking is the reverse of the other), the coefficient has value −1.
- In other cases, the value lies between −1 and 1, and increasing values imply increasing association between the rankings. If the rankings are completely independent, the coefficient has value 0.

Incidentally, Spearman's correlation and Kendall's correlations are not comparable, and their values could be different.

The Kendall's Tau also measures the strength of association of the cross tabulations. It also tests the strength of association of the cross tabulations when both variables are measured at the ordinal level, and, in fact, is the only measure of association in ordinal form in a cross tabulation data available in ordinal form.

GLOSSARY

Non-Parametric Tests	Tests of significance used when certain assumptions about the usual tests of significance are not valid or doubtful. Also, these tests are for aspects like randomness, rank correlation, etc.
Run	A succession of values with the same sign or type (in case of qualitative data)
Signed Rank	The ranks assigned to observations are attached a sign, viz. + or –
Rank Sum	Sum of ranks

OBJECTIVE TYPE QUESTIONS

1. Most of the non-parametric methods utilise measurements on:
   (a) Interval scale (b) Ratio scale
   (c) Ordinal scale (d) Nominal scale
2. The most commonly used assumption about the distribution of a variable is:
   (a) Continuity of the distribution
   (b) Symmetry of the distribution
   (c) Both (a) and (b)
   (d) Neither (a) nor (b)
3. A 'run' is defined as
   (a) Succession of values with a '+'sign
   (b) Succession of values with a '–' sign
   (c) Succession of values with the same sign.
   (d) All of the above
4. In tests using rank methods, the null hypothesis is rejected if the calculated value:
   (a) Exceeds the tabulated value
   (b) Is less than the tabulated value
   (c) Is either less or more than the tabulated value
5. Kolmogorow–Smirnov have evolved tests for:
   (a) Goodness of fit of a distribution
   (b) Comparing two populations
   (c) Both of (a) and (b)
   (d) None of (a) and (b)
6. Mann–Whitney 'U' test is used for testing:
   (a) Equality of two means
   (b) Equality of more than two means
   (c) Equality of three means
   (d) Equality of two sets of rankings
7. Kruskal–Wallis Rank Sum Test is for testing:
   (a) Equality of two Means
   (b) Equality of more than two means
   (c) Equality of two sets of ranks
   (d) All of the above
8. A sample of consumers of coffee is selected and asked whether they had coffee with or without sugar. 'S' (with sugar) and 'N' (without sugar). Their responses are recorded as follows:
   How many runs are in this data? S N N N S N N S S S S N N S N N N S S N S S S S
   (a) 24 (b) 2 (c) 11 (d) 12
   (e) None of the above
9. The non-parametric test that is analogous to the One-Way ANOVA is the:
   (a) Kruskal-Wallis test (b) Friedman test
   (c) Mann-Whitney U test (d) None of the above

10. The zero value of rank correlation between two pairs of rankings about the same subjects: implies that:
    (a) The two rankings of subjects are not related to each other
    (b) $6\ \Sigma d_i^2 = n\ (n^2 - 1)$
    (c) $\Sigma d_i^2 = 0$
    (d) The rankings are biased

EXERCISES

1. The following data gives the increase (I) or decrease (D) in daily rate of return of a share on 30 consecutive days.

I	I	I	D	D	I	D	I	I	I
D	D	D	I	I	D	D	I	I	D
I	I	I	D	I	I	D	I	D	D

   Test whether the increase or decrease in the daily rate of return follows a random pattern.
2. A car manufacturer claims that its new car gives an average mileage of 12 kms. per litre (km.p.l) of petrol. A sample of 12 cars is taken, and their mileage recorded as follows (in km.p.l):

   13.2, 12.7, 13.3, 13.0, 12.8, 12.7, 12.6, 12.6, 12.7, 12.4, 11.8, 13.2

   Use the signed rank test to ascertain the claim of the manufacturer. Also use this test to ascertain whether the median mileage of the car is 12 km.p.l.
3. Following is the data for number of guest faculty visiting a management institute during the academic year 2006-07 comprising 280 academic days.

Number of Guest Faculty	*Number of Days*
0	100
1	105
2	55
3	15
4	5

   Use Kolmogorov–Smirnov Test to test whether the above data follows a Poisson distribution with mean equal to 1.
4. Following is the data of the percentage of MBA girl students in five similar level management institutes.

Management Institutes	*Percentage of MBA Girl Students*
A	21
B	20
C	19
D	18
E	22

Use Kolmogorov-Smirnov Test to test whether the percentage of MBA girl students in all the five institutes is the same.

5. The management of an organisation wanted to assess whether there is any difference in the officers who were recruited from campus (MBAs) and who were recruited from open market with respect to their knowledge and application relevant to organisation Accordingly, a suitable test was designed and conducted for 10 officers each from the two categories. Their scores on a scale of 10 are given below:

Srl. No.	*Campus Recruited Officers*	*Srl. No.*	*Open Market Recruited Officers*
1	9.7	1	9.2
2	9.2	2	8.9
3	8.9	3	7.2
4	7.9	4	8.5
5	9.5	5	8.1
6	9.6	6	7.8
7	9.0	7	8.5
8	9.5	8	9.0
9	8.8	9	9.3
10	8.5	10	8.0

Use Mann-Whitney test to test whether the mean scores of the two categories of officers are equal.

6. A car manufacturer is procuring car batteries from two companies. For testing whether the two brands of batteries, say 'A' and 'B', had the same life, the manufacturer collected data about the lives of both brands of batteries from 20 car owners–10 using 'A' brand and 10 using 'B' brand. The lives were reported as follows:

	Lives in Months									
Battery 'A' :	50	61	54	60	52	58	55	56	54	53
Battery 'B' :	65	57	60	55	58	59	62	67	56	61

Use Mann-Whitney test to test the equality of lives of the two types of batteries.

7. For the data given in Exercise 3 of Chapter 10, about the closing stock prices of three individual companies viz. ICICI Bank, L & T and Reliance Industries Ltd. test whether the ranks as daily rate of return are equal for all the above three companies, using Kruskall-Wallis test.
8. For the data given in Exercise 2 of Chapter 10, about ranks of some selected companies as per net profit, market capitalisation and overall rank; test whether the ranks are equal for all the three parameters of all the companies.
9. For the same data as in the Exercise 10 below, use the Wilcoxon-Wilcox Test for Comparison of Multiple Treatments to test equality of daily rate of returns for all the three companies as also the three pairs of companies.
10. Given the following monthly rates of return on BSE 30, BSE 100, BSE 200 and BSE 500, for the period January 2005 to November 2005, use Friedman's test to test whether the rates of return are equal for all the above indices of Bombay Stock Exchange.

2005	BSE 30	BSE 100	BSE 200	BSE 500
1	2.41	2.56	3.36	3.64
2	–3.29	–3.6	–3.38	–3.22
3	5.21	–4.84	–5.37	–4.54
4	9.11	8.7	8.32	8.38
5.	7.13	5.51	3.8	3.5
6.	6.14	7.16	5.98	6.71
7	2.23	2.77	3.48	4.74
8	10.62	9.12	8.47	7.6
9	–8.6	–8.91	8.97	–9.1
10	11.36	11.79	11.67	11.56
11	6.93	6.53	6.46	6.38

Multivariate Statistical Techniques

14

Contents

LEARNING OBJECTIVES

There are several situations in real life or work environment when we are required to collect data on several variables. The situation is analogous to the examination for selection in IAS and other Central services when a candidate is examined in several subjects and marks in these subjects are summarised for deciding the suitability of a candidate. The medical fitness of a candidate for entry to Defence Services is based on several health-related parameters. Similar is the situation relating to studies conducted for designing and marketing of physical as well as financial products and services. This chapter describes various techniques that are available for reducing the data on many variables and summarising it with few indicators that can be interpreted to derive meaningful conclusions relating to designing and marketing of products and services. Incidentally, these techniques are also used for many behavioural and social studies. In addition to providing a comprehensive understanding and applications of the techniques, the chapter also illustrates, with examples, the use of SPSS, step by step, in conducting the studies in their entirety. The chapter is aimed at generating confidence in using SPSS for arriving at final conclusions/solutions in a research study. This should provide motivation for learning these techniques that have, so far, not received due importance mainly due to the lack of confidence as well as awareness in using SPSS package.

Relevance

Nowadays, there are several features associated with any product or service. With time, these features are increasing and providing several options to the consumers. Like other mobile phone manufacturing companies, the MOCEL company was also coming out with new models and new features. However, the CEO of the company was not too satisfied with the moderate growth in business in view of the exponential growth in the industry. He, therefore, hired the services of a consultant to advise the company about the preferences of potential customers and suggestions of the existing customers. The consultant conducted a survey among the various strata of the society, to ascertain the needs, preferences, price sensitiveness, etc. He also collected similar information from dealers. The vast amount of data was analysed with the help of techniques such as factor analysis, cluster analysis, conjoint analysis and multidimensional scaling, described in this chapter, and suitable recommendations made to the CEO of MOCEL company. MOCEL could improve its market share, substantially.

Artificial intelligence software analyses the credentials of the candidate (contesting an election) and gives a rating on how successful he or she will be as a politician.

Dr APJ Abdul Kalam (in Hindustan Times dt. 25th March, 2007)

14.1 RELEVANCE AND INTRODUCTION

In general, what is true about predicting the success of a politician with the help of intelligence software, as pointed out above by the former President of India, is equally true for predicting the success of products and services with the help of statistical techniques. In this chapter, we discuss a number of statistical techniques which are especially useful in designing of products and services. The products and services could be physical, financial, promotional like advertisement, behavioural like motivational strategy through incentive package or even educational like training programmes/ seminars, etc. These techniques, basically involve reduction of data and its subsequent summarisation, presentation and interpretation. A classical example of data reduction and summarisation is provided by SENSEX (Bombay Stock Exchange) which is one number like 18, 000, but it represents movement in share prices listed in Bombay Stock Exchange. Yet another example is the Grade Point Average, used for assessment of MBA students, which 'reduces' and 'summarises' marks in all subjects to a single number.

In general, any problem in life whether relating to individual, like predicting the cause of an ailment or behavioural pattern, or relating to an entity, like forecasting its futuristic status in terms of products and services, needs collection of data on several parameters. These parameters are then analysed to summarise the entire set of data with a few indicators which are then used for drawing conclusions. The following techniques (with their abbreviations in brackets) coupled with the appropriate computer software like SPSS, play a very useful role in the endeavour of reduction and summarisation of data for easy comprehension:

- Multiple Regression Analysis (MRA)
- Discriminant Analysis (DA)
- Logistic Regression (LR)

- Multivariate Analysis of Variance (MANOVA)
- Factor Analysis (FA)
- Principal Component Analysis (PCA)
- Canonical Correlation Analysis (CRA)
- Cluster Analysis
- Conjoint Analysis
- Multidimensional Scaling (MDS)

Before describing these techniques in detail, we provide their brief description as also indicate their relevance and uses, in a tabular format given below. This is aimed at providing motivation for learning these techniques and generating confidence in using SPSS for arriving at final conclusions/solutions in a research study.

Statistical Techniques, Their Relevance and Uses for Designing and Marketing of Products and Services

(This Table may be referred to when begins to read the particular techniques described in various Sections of the Chapter)

Technique	*Relevance and Uses*
Multiple Regression Analysis (MRA) • It deals with the study of relationship between one metric dependent variable and more than one metric independent variables.	• One could assess the individual impact of the independent variables on the dependent variable. • Given the values of the independent variables, one could forecast the value of the dependent variable. For example, the sale of a product depends on expenditures on advertisements as well as on R&D. Given the values of these two variables, one could establish a relationship among these variables and the dependent variable, say, profit. Subsequently, if the relationship is found appropriate, it could be used to predict the profit with the knowledge of the two types of expenditure.
Discriminant Analysis It is a statistical technique for classification or determining a linear function, called discriminant function, of the variables which helps in discriminating between two groups of entities or individuals.	• The basic objective of discriminant analysis is to perform a classification function. • From the analysis of past data, it can classify a given group of entities or individuals into two categories—one those which would turn out to be successful and others which would not be so. For example, it can predict whether a company or an individual would turn out to be a good borrower. • With the help of financial parameters, a firm could be classified as worthy of extending credit or not. • With the help of financial and personal parameters, an individual could be classified as eligible for loan or not or whether he would be a buyer of a particular product/service or not. • Salesmen could be classified according to their age, health, sales aptitude score, communication ability score, etc.

(*Contd*)

(*Contd*)

Logistic Regression • It is a technique that assumes the errors are drawn from a binomial distribution. • In logistic regression, the dependent variable is the probability that an event will occur, hence it is constrained between 0 and 1. • All of the predictors can be binary, a mixture of categorical and continuous or just continuous.	• Logistic regression is highly useful in biometrics and health sciences. It is used frequently by epidemiologists for the probability (sometimes interpreted as risk) that an individual will acquire a disease during some specified period of vulnerability. • Credit Card Scoring: Various demographic and credit history variables could be used to predict if an individual will turn out to be 'good' or 'bad' customers. • Market Segmentation: Various demographic and purchasing information could be used to predict if an individual will purchase an item or not.
Multivariate Analysis of Variance (MANOVA) • It simultaneously explores the relationship between several non-metric independent variables (Treatments, say Fertilisers) and **two or more** metric dependant variables (say, Yield and Harvest Time). If there is only one dependent variable, MANOVA is the same as ANOVA.	• Determines whether statistically significant differences of means of several variables occur simultaneously between two levels of a variable. For example, assessing whether (i) a change in the compensation system has brought about changes in sales, profit and job satisfaction. (ii) geographic region (North, South, East, West) has any impact on consumers' preferences, purchase intentions or attitudes towards specified products or services. (iii) a number of fertilisers have equal impact on the yield of rice as also on the harvest time of the crop.
Principal Component Analysis (PCA) • Technique for forming set of new variables that are linear combinations of the original set of variables, and are uncorrelated. The new variables are called Principal Components. • These variables are fewer in number as compared to the original variables, but they extract most of the informant provided by the original variables.	• One could identify several financial parameters and ratios exceeding ten for determining the financial health of a company. Obviously, it would be extremely taxing to interpret all such pieces of information for assessing the financial health of a company. However, the task could be much simpler if these parameters and ratios could be reduced to a few indices, say two or three, which are linear combinations of the original parameters and ratios. • A multiple regression model may be derived to forecast a parameter like sales, profit, price, etc. However, the variables under consideration could be correlated among themselves indicating multicollinearity in the data. This could lead to misleading interpretation of regression coefficients as also increase in the standard errors of the estimates of parameters. It would be very useful, if the new uncorrelated variables could be formed which are linear combinations of the original variables. These new variables could then be used for developing the regression model, for appropriate interpretation and better forecast.
Common Factor Analysis (CFA) • It is a statistical approach that is used to analyse interrelationships among a large number of variables (indicators) and to explain these variables (indicators) in terms of a few unobservable constructs (factors). In fact, these factors impact the variables, and are reflective indicators	Helps in assessing • the image of a company/enterprise • attitudes of sales personnel and customers • preference or priority for the characteristics of – a product like television, mobile phone, etc. – a service like TV programme, air travel, etc.

(*Contd*)

(*Contd*)

of the factors. The statistical approach involves finding a way of condensing the information contained in a number of original variables into a smaller set of constructs (factors)—mostly one or two—with a minimum loss of information. • Identifies the smallest number of common factors that best explain or account for most of the correlation among the indicators. For example, intelligence quotient of a student might explain most of the marks obtained in Mathematics, Physics, Statistics, etc. Yet another example, when two variables *x* and y are highly correlated, only one of them could be used to represent the entire data.	
Canonical Correlation Analysis (CRA) • An extension of multiple regression analysis (MRA involving one dependant variable and several metric independent variables). It is used for situations wherein there are several dependent variables and several independent variables. • Involves developing linear combinations of the sets of variables (both dependant and independent) and studies the relationship between the two sets. The weights in the linear combination are derived based on the criterion that maximises the correlation between the two sets of variables.	• Used in studying relationship between types of products purchased and consumer life styles and personal traits. Also, for assessing impact of life styles and eating habits on health as measured by number of health-related parameters. • Given assets and liabilities of a set of banks/financial institutions, helps in examining interrelationship of variables on the asset and liability sides. • HRD department might like to study the relationship between set of behavioural, technological and social skills of a salesman with the set of variables representing sales performance, discipline and cordial relations with staff. • The Central Bank of a country might like to study the relationship between sets of variables representing several risk factors and the financial indicators arising out of a bank's operations. Similar analysis could be carried out for any organisation.
Cluster Analysis • It is an analytical technique that is used to develop meaningful subgroups of entities which are homogeneous or compact with respect to certain characteristics. Thus, observations in each group would be similar to each other. Further, each group should be different from each other with respect to the same characteristics, and therefore, observations of one group would be different from the observations of the other groups.	• It helps in classifying a given set of entities into a smaller set of distinct entities by analysing similarities among the given set of entities. Some situations where the technique could be used are: • A bank could classify its large network of branches into clusters (groups) of branches which are similar to each other with respect to specified parameters. • An investment bank could identify groups of firms that are vulnerable for takeover. • A marketing department could identify similar markets where products or services could be tested or used for target marketing. • An insurance company could identify groups of motor insurance policy holders with high claims.

(*Contd*)

(*Contd*)

Conjoint Analysis • Involves determining the contribution of variables (each of several levels) to the choice preference over combinations of variables that represent realistic choice sets (products, concepts, services, companies, etc.)	• Useful for analysing consumer responses, and use the same for designing of product and services. • Helps in determining the contributions of the predictor variables and their respective levels to the desirability of the combinations of variables. For example, how much does the quality of food contribute to continued loyalty of a traveller to an airline? Which type of food is liked most?
Multidimensional Scaling • It is a set of procedures, for drawing pictures of data so as to visualise and clarify relationships described by the data more clearly. • The requisite data is typically collected by having respondents give simple one-dimensional responses. • Transforms consumer judgments/perceptions of similarity or preferences in usually a two-dimensional space.	• Useful for designing of products and services. It helps in: • illustrating market segments based on indicated preferences. • identifying the products and services that are more competitive with each other. • understanding the criteria used by people while judging objects (products, services, companies, advertisements, etc.).

14.1.1 Multivariate Techniques

These techniques are classified in two types:

- Dependence Techniques
- Interdependence Techniques

Dependence Techniques These are the techniques, that define some of the variable/s as independent variable/s and some other as dependent variable/s. These techniques aim at finding the relationship of these variables and may, in turn, find the effect of independent variable on dependent variable.

The techniques to be used may differ as the type of independent/dependent variables change. For example, if all the independent and dependent variables are metric or numeric, Multiple Regression Analysis can be used, if dependent variable is metric, and independent variable is/are categorical, ANOVA can be used. If dependent variable is metric and some of the independent variables are metric, and some are qualitative ANACOVA (Analysis of co-variance) can be used. If the dependent variable is nonmetric or categorical, multiple discriminant analysis or logistic regression are the techniques used for the analysis.

All the above techniques require a single dependent variable.

If there is more than one dependent variable, the techniques used are MANOVA (Multivariate analysis of variance) or Canonical correlation.

MANOVA is used when there are more than one dependent variables and all independent variables are categorical. If some of the independent variables are categorical and some are metric, MANOCOVA (Multivariate analysis of covariance) can be used. If there is more than one dependent variable and all dependent and independent variables are metric, the best suited analysis is canonical correlation.

Interdependence Techniques Interdependence techniques do not assume any variable as independent/dependent variables or try to find the relationship. These techniques can be divided into variable interdependence and inter-object similarity.

The variable interdependence techniques can also be termed as data reduction techniques. Factor analysis is the example of the variable interdependence techniques. Factor analysis is used when there are many related variables and one wants to reduce the list of variables or find underlying factors that determine the variables.

The inter-object similarity is assessed with the help of cluster analysis, multidimensional scaling (MDS).

Brief descriptions of all the above techniques are given in subsequent sections of this chapter.

14.2 MULTIPLE REGRESSION ANALYSIS

In Chapter 10, we have discussed the correlation and regression analysis relating to two variables. Usually, one of them, called dependent variable, is of prime consideration, and depends on another variable called independent variable. Thus, we have one dependent and one independent variable. However, sometimes more than one variable may influence the dependent variable. For instance, the marks obtained by students in an examination could depend not only on their intelligent quotients (I.Qs) but also the time devoted for preparing for the examination. In agriculture, yield of a crop depends not only on the fertilizer used but also depends upon rainfall, temperature, etc. Similarly, in economic theory, the quantity demanded of a particular commodity may not only depend on its price but also on the prices of other substitute commodities and on the disposable incomes of households. Further, the price of a particular stock depends not only on the stock market, in general, but also on its dividend payout and the retained earnings by the concerned company. It may be noted that simple regression analysis helps in assessing the impact of a variable (independent) on another variable (dependent). With the help of such analysis, given a dependent variable, one could consider, one by one, individually, the impact of even several independent variables. However, with the help of multiple regression analysis, one could assess the impact of several independent variables, individually or jointly together, on the dependent variable.

Some of the situations wherein a multiple regression equation, giving the relationship between the dependent variable and independent variables, is used are as follows:

Manpower in a Sales Organisation	Number of Sales Offices + Business per Sales Office
EPS (time series) or EPS (cross-sectional)	Sales + Dividend + Price
Sales of a Company	Expenditure on Advertisement + Expenditure on R&D
Return on BSE SENSEX	Return on Stock of Reliance Industries + Return on Stock of Infosys Technologies

14.2.1 Estimation of Multiple Regression Equation and Calculation of Multiple Correlation Coefficient

In multiple regression analysis, the dependent variable is expressed as a function of independent variables. The equation giving such relationship is called regression equation. The correlation between

the dependent variable and independent variables is measured by multiple correlation coefficient. The methodology of deriving the multiple regression equation and calculating multiple correlation coefficient is illustrated below.

Illustration 14.1

Let the dependent variable of interest be y which depends on two independent variables, say x_1 and x_2.

The linear relationship, among y, x_1 and x_2 can be expressed in the form of the regression equation of y on x_1 and x_2, in the following form:

$$\boldsymbol{y = b_o + b_1 x_1 + b_2 x_2} \tag{14.1}$$

where b_o is referred to as 'intercept' and b_1 and b_2 are known as regression coefficients.

The sample comprises 'n' triplets of values of x_1, y and x_2, in the following format:

y	x_1	x_2
y_1	x_{11}	x_{21}
y_2	x_{12}	x_{22}
.	.	.
y_n	x_{1n}	x_{2n}

The values of constants, b_o, b_1 and b_2 are estimated with the help of Principle of Least Squares just like values of a and b were found while fitting the equation $y = a + bx$ in Chapter 10 on Simple Correlation and Regression Analysis. These are calculated by using the above sample observations/ values, and with the help of the formulas given below:

These formulas and manual calculations are given for illustration only. In real life, these are easily obtained with the help of personal computers wherein the formulas are already stored.

$$b_1 = \frac{(\Sigma y_i x_{1i} - \boldsymbol{n}\bar{y}\bar{x}_1)(\Sigma \bar{x}_{2i}^2 - \boldsymbol{n}\bar{x}_2^2) - (\Sigma y_i \bar{x}_{2i} - \boldsymbol{n}\bar{y}\,\bar{x}_2)(\Sigma x_{1i} x_{2i} - \boldsymbol{n}\bar{x}_1\bar{x}_2)}{(\Sigma x_{1i}^2 - \boldsymbol{n}x_1^2)(\Sigma x_{2i}^2 - \boldsymbol{n}\bar{x}_2^2) - (\Sigma x_{1i} x_{2i} - \boldsymbol{n}\bar{x}_1\bar{x}_2)^2}$$

$$b_2 = \frac{(\Sigma y_i x_{1i} - \boldsymbol{n}\bar{y}\,\bar{x}_2)(\Sigma x_{1i}^2 - \boldsymbol{n}\bar{x}_1^2) - (\Sigma y_i x_{1i}^2 - \boldsymbol{n}\bar{y}\bar{x}_1)(\Sigma x_{1i} x_{2i} - \boldsymbol{n}\bar{x}_1\bar{x}_2)}{(\Sigma x_{1i}^2 - \boldsymbol{n}\bar{x}_1^2)(\Sigma \bar{x}_{2i}^2 - \boldsymbol{n}\bar{x}_2^2) - (\Sigma x_{1i} x_{2i} - \boldsymbol{n}\bar{x}_1 x_2)^2} \tag{14.2}$$

$$b_0 = y - \mathbf{b}_1\bar{x}_1 - \mathbf{b}_2\bar{x}_2 \tag{14.3}$$

The calculations needed in the above formulas are facilitated by preparing the following table:

	y	x_1	x_2	yx_1	yx_2	x_1x_2	y^2	x_1^2	x_2^2
	y_1	x_{11}	x_{21}	y_1x_{11}	y_1x_{21}	$x_{11}x_{21}$	y_1^2	x_{11}^2	x_{21}^2
	.	.	.	.	.	.	.	.	.
	y_i	x_{1i}	x_{2i}	y_ix_{1i}	y_ix_{2i}	$x_{1i}x_{2i}$	y_i^2	x_{1i}^2	x_{2i}^2
	.	.	.	.	.	.	.	.	.
	y_n	x_{1n}	x_{2n}	y_ix_{1n}	y_ix_{2n}	$x_{1n}x_{2n}$	y_n^2	x_{1n}^2	x_{2n}^2
Sum	Σy_i	Σx_{1i}	Σx_{2i}	Σy_ix_{1i}	Σy_ix_{2i}	$\Sigma x_{1i}x_{2i}$	Σy_i^2	Σx_{1i}^2	Σx_{2i}^2

The effectiveness or the reliability of the relationship, thus, obtained is judged by the multiple coefficient of determination, usually denoted by $\boldsymbol{R^2}$, and is defined as the ratio of variation explained

by the regression equation 14.1 and total variation of the dependent variable y. Thus,

$$R^2 = \frac{\text{Explained Variation in } y}{\text{Total Variation in } y} \tag{14.4}$$

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} \tag{14.5}$$

$$= 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - y_i)^2} \tag{14.6}$$

It may be recalled from Chapter 10, that total variation in the variable y is equal to the variation explained by the regression equation plus unexplained variation by the regression equation. Mathematically, this is expressed as:

$$\underset{\text{Total Variation}}{\Sigma(y_i - \bar{y})^2} = \underset{\text{Explained Variation}}{\Sigma(\hat{y}_i - \bar{y})^2} + \underset{\text{Unexplained Variation}}{\Sigma(y_i - \hat{y}_i)^2}$$

where y_i is the observed value of $\bar{y}$, $\bar{y}$ is the mean of all, $\hat{y}$ is the estimate of the value y_i by the regression equation 14.1. It may be recalled that $\Sigma(\hat{y}_i - \bar{y})^2$ is the explained variation of y by the estimate of y, and $\Sigma(y_i - \hat{y}_i)^2$ is the unexplained variation of y by the estimate of $y(\hat{y})$. If y_i is equal to the estimate $\hat{y}_i$, then all the variation is explained by $\hat{y}_i$, and, therefore, unexplained variation is zero. In such case, total variation is fully explained by the regression equation, and R^2 is equal to 1.

The square root of R^2 viz. R is known as the **coefficient of multiple correlation** and is always between 0 and 1. In fact, R is the **correlation between the independent variable and its estimate** derived from the multiple regression equation, and as such it has to be positive.

All the calculations and interpretations for the multiple regression equation and coefficient of multiple correlation or determination have been explained with the help of an illustration given below:

Example 14.1

The owner of a chain of ten stores wishes to forecast net profit with the help of next year's projected sales of food and non-food items. The data about current year's sales of food items, sales of non-food items as also net profit for all the ten stores are available as follows:

Table 14.1 Sales of Food and Non-Food Items and Net Profit of a Chain of Stores

Supermarket No.	*Net Profit (Rs. Crore)* y	*Sales of Food Items (Rs. Crore)* x_1	*Sales of Non-Food Items (Rs. Crore)* x_2
1	5.6	20	5
2	4.7	15	5
3	5.4	18	6

(*Contd*)

(*Contd*)

4	5.5	20	5
5	5.1	16	6
6	6.8	25	6
7	5.8	22	4
8	8.2	30	7
9	5.8	24	3
10	6.2	25	4

In this case, the relationship is expressed by equation 14.1 reproduced below:

$$y = b_0 + b_1x_1 + b_2x_2$$

where, y denotes net profit, x_1 denotes sales of food items, and x_2 denotes sales of non-food items, and b_0, b_1 and b_2 are constants. Their values are obtained by the following formulas derived from the Principle of Least Squares:

The required calculations can be made with the help of the following table:

(Amounts in Rs. Crore)

Supermarket	*Net Profit (y)*	*Sales of Food Items (x_1)*	*Sales of Non-Food Items (x_2)*						
	y_i	x_{1i}	x_{2i}	x_{1i}^2	$y_i\,x_{1i}$	y_i^2	$y_i\,x_{2i}$	x_{2i}^2	$x_{1i}x_{2i}$
1	5.6	20	5	400	112	31.36	28	25	100
2	4.7	15	5	225	70.5	22.09	23.5	25	75
3	5.4	18	6	324	97.2	29.16	32.4	36	108
4	5.5	20	5	400	110	30.25	27.5	25	100
5	5.1	16	6	256	81.6	26.01	30.6	36	96
6	6.8	25	6	625	170	46.24	40.8	36	150
7	5.8	22	4	484	127.6	33.64	23.2	16	88
8	8.2	30	7	900	246	67.24	57.4	49	210
9	5.8	24	3	576	139.2	33.64	17.4	9	72
10	6.2	25	4	625	155	38.44	24.8	16	100
Sum	**59.1**	**215**	**51**	**4815**	**1309.1**	**358.07**	**305.6**	**273**	**1099**
Average	**5.91**	**21.5**	**5.1**						

Substituting the values of b_o, b_1 and b_2, the desired relationship is obtained as

$$y = 0.233 + 0.196\,x_1 + 0.287\,x_2 \qquad (14.7)$$

This equation is known as the multiple regression equation of y on x_1 and x_2, and it indicates as to how y changes with respect to changes in x_1 and x_2. The interpretation of the value of the coefficient of x_1 viz. 'b_1' i.e. 0.196, is that if x_2 (sales of non-food items) is held constant, then for every crore of sales of food items, the net profit increases by Rs. 0.196 crore i.e. Rs. 19.6 lakh. Similarly, the interpretation of the value of coefficient of x_2 viz. 'b_2' i.e. 0.287 is that if the sales of non-food items increases by one crore rupees, the net profit increases Rs. by 0.287 crore i.e. Rs. 28.7 lakh.

The effectiveness or the reliability of this relationship is judged by the multiple coefficient of determination**, usually denoted by R^2, and is defined as given in equation 14.4 as**:

$$R^2 = \frac{\text{Explained Variation in } y \text{ by the Regression Equation}}{\text{Total Variation in } y}$$

The above two quantities are calculated with the help of the following Table.
Column (3) gives the difference in the observed value $\bar{y}_i$ and its estimate $\hat{y}_i$ derived from the fitted regression equation by substituting corresponding values of x_1 and x_2

y_i (1)	$\hat{y}_i^*$ (2)	$y_i - \hat{y}_i$ (3)	$(y_i - \hat{y}_i)^2$ (4)	$(y_i - \bar{y}_i)^2$ (5)
5.6	5.587	0.0127	0.0002	0.0961
4.7	4.607	0.0928	0.0086	1.4641
5.4	5.482	–0.082	0.0067	0.2601
5.5	5.587	–0.087	0.0076	0.1681
5.1	5.09	0.0099	0.0001	0.6561
6.8	6.854	–0.054	0.0029	0.7921
5.8	5.693	0.1075	0.0116	0.0121
8.2	8.121	0.0789	0.0062	5.2441
5.8	5.798	0.0023	0.0000	0.0121
6.2	6.281	–0.081	0.0065	0.0841
Sum = 59.1	59.1	0	Sum = 0.0504	Sum = 8.789
$\bar{y}$ = 5.91			(Unexplained Variation)	(Total Variation)

* Derived from the earlier fitted equation, $\boldsymbol{y = 0.233 + 0.196\, x_1 + 0.287\, x_2}$

Substituting the respective values in equation 14.6, we get

$$\boldsymbol{R^2 = 1 - (0.0504/8.789) = 1 - 0.0057 = 0.9943}$$

The interpretation of the value of $R^2 = 0.9943$ is that 99.43% of the variation in net profit is explained jointly by variation in sales of food items and non-food items.

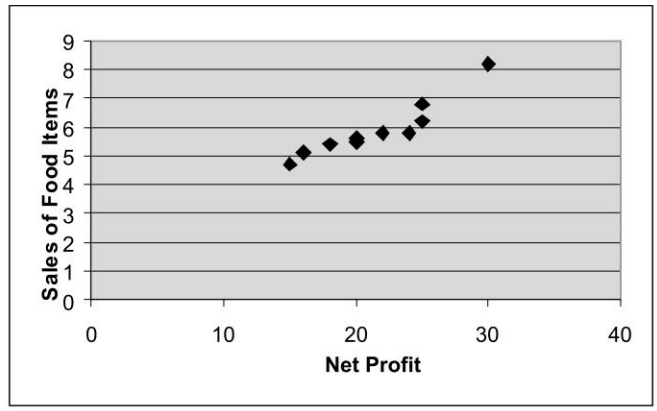
Incidentally, Explained Variation for the above example can be calculated by subtracting unexplained variation from total variation as 8.789 – 0.0504 = 8.7386.

It may be recalled that in Chapter 10 on Simple Correlation and Regression Analysis, we have discussed the impact of the change of variation in only one independent variable on the dependent variable. We shall now demonstrate the usefulness of two independent variables in explaining the variation in the dependent variable (net profit in this case).

Suppose, we consider only as one variable, say food items, then the basic data would be as follows:

Supermarket	*Net Profit (Rs. Crore) Y*	*Sales of Food Items (Rs. Crore) x_1*
1	5.6	20
2	4.7	15
3	5.4	18
4	5.5	20
5	5.1	16
6	6.8	25
7	5.8	22
8	8.2	30
9	5.8	24
10	6.2	25

The scatter diagram indicates a positive linear correlation between the net profit and the sales of food items.



The linear relationship is assumed as

$$y = a + bx_1 \tag{14.8}$$

which is the regression equation of y on x_1. While 'b' is the regression coefficient of y on x_1, 'a' is just a constant. In the given example, y is the 'net profit' and x_1 is the sales of food items.

The values of 'a' and 'b' are calculated from the following formulas given in Chapter 10.

$$b = \frac{\Sigma \bar{y}_i \bar{x}_{1i} - n\bar{y}\bar{x}_1}{\Sigma y_i^2 - n(\bar{x}_1)^2}$$

and,

$$a = \bar{y} - b\bar{x}_i$$

Thus, the desired regression equation can be obtained as

$$y = \mathbf{1.61 + 0.2\ x_1} \tag{14.9}$$

It tells us that how y changes with respect to changes in x_1 i.e. how the net profit increases with increase in sales of food-items. The interpretation of 'b' = 0.2 is that for every crore Rs. of sales of food items, the net profit increases by Rs. 0.2 crore i.e. Rs. 20 lakh.

As stated earlier, the effectiveness/reliability of the regression equation is judged by the coefficient of determination, can be obtained as

$$r^2 = \mathbf{0.876}$$

This value of $r^2 = 0.876$ indicates that 87.6% of the variation in net profit is explained by the variation in sales of food items, and thus one may feel quite confident in forecasting net profit with the help of the sales of food items. However, before doing so, it is desired that one examines the possibility of considering some other variables also either as an alternative or in addition to the variable (sales of food items) already considered, to improve the reliability of the forecast.

As mentioned in Chapter 10, the correlation coefficient is also defined as

$$r^2 = \frac{\text{Explained Variation in } y \text{ by the Regression Equation}}{\text{Total Variation in } y}$$

$$r^2 = 1 - \frac{\text{Unexplained Variation in } \bar{y}}{\text{Total Variation in } \bar{y}} 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}$$

In the above illustration,

Total variation = $\Sigma(y_i - \bar{y})^2$

Unexplained variation = $\Sigma(\hat{y}_i - \bar{y})^2$

These quantities can be calculated from the following table:

Supermarket	*Net Profit* y_i	*Sales of Food Items* x_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i = 1.61 + 0.2\, x_i$	$y_i - \hat{y}$	$(y_i - \hat{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
1	5.6	20	–0.31	0.0961	5.61	–0.01	0.0001	–0.3	0.09
2	4.7	15	–1.21	1.4641	4.61	0.09	0.0081	–1.3	1.69
3	5.4	18	–0.51	0.2601	5.21	0.19	0.0361	–0.7	0.49
4	5.5	20	–0.41	0.1681	5.61	–0.11	0.0121	–0.3	0.09
5	5.1	16	–0.81	0.6561	4.81	0.29	0.0841	–1.1	1.21
6	6.8	25	0.89	0.7921	6.61	0.19	0.0361	0.7	0.49
7	5.8	22	–0.11	0.0121	6.01	–0.21	0.0441	0.1	0.01
8	8.2	30	2.29	5.2441	7.61	0.59	0.3481	1.7	2.89
9	5.8	24	–0.11	0.0121	6.41	–0.61	0.3721	0.5	0.25
10	6.2	25	0.29	0.0841	6.61	–0.41	0.1681	0.7	0.49
Sum	**59.1**	**215**	**Sum**	**8.789**		**Sum**	**1.109**	**Sum**	**7.7**
Average	5.91	21.5							

It may be noted that the unexplained variation or residual error is **1.109** when the simple regression equation 14.9 of net profit on sales of food items is fitted but it was lower as reduced to 0.05044, when multiple regression equation 14.7 was used by taking into account adding one more variable as sale of non-food items (x_2).

Also it may be noted that only one variable viz. sales of food items is considered then r^2 is 0.876 i.e. 87.6% of variation in net profit is explained by variation in sales of food item but when both the variables viz. sales of food as well as non-food items are considered. R^2 is 0.9943 i.e. 99.43% of variation in net profit is explained by variation in both these variables.

14.2.2 Forecast with a Regression Equation

The multiple regression equation 14.1 can be used to forecast the value of the dependent variable at any point of time, given the values of the independent variables at that point of time. For illustration, in the above example about the net profit in a company, one may be interested in forecasting the net profit for the next year when the sales of food items is expected to increase to Rs. 30 crore and sales of sales of non-food items is expected to Rs. 7 crore. Substituting $\bar{x}_1 = 30$ and $\bar{x}_2 = 7$ in equation 14.7, we get

$$y = 0.233 + 0.196 \times 30 + 0.287 \times 7$$
$$= \mathbf{8.122}$$

Thus, the net profit for all the 10 stores, by the end of next year would be Rs. 8.122 crore.

Caution: It is important to note that a regression equation is valid for estimating the value of the dependent variable only within the range of independent variable(s) or only slightly beyond the range. However, it can be used even much beyond the range if no other better option is available, and it is supported by common sense.

14.2.3 Correlation Matrix

The multiple correlation coefficient can also be determined with the help of the total correlation coefficients between all pairs of dependent and independent variables. All the possible total correlations between any two pairs of x_1, x_2 and y and can be presented in the form of a matrix as follows:

$$\begin{matrix} r_{x_1x_1} & r_{x_1y} & r_{x_1x_2} \\ r_{yx_1} & r_{yy} & r_{yx_2} \\ r_{x_2x_1} & r_{x_2y} & r_{x_2x_2} \end{matrix}$$

Since $r_{x_1x_1}$, r_{yy} and $r_{x_2x_2}$ are all equal to 1, the matrix can be written as

$$\begin{matrix} 1 & r_{x_1y} & r_{x_1x_2} \\ r_{yx_1} & 1 & r_{yx_2} \\ r_{x_2x_1} & r_{x_1y} & 1 \end{matrix}$$

Further, since r_{xy} and r_{yx} are equal, and so are r_{xz} and r_{xz}, it is sufficient to write the matrix in the following form:

$$\begin{matrix} 1 & r_{x_1y} & r_{x_1x_2} \\ - & 1 & r_{yx_2} \\ - & - & 1 \end{matrix}$$

If there are three variables x_1, x_2 and y then simple correlation coefficient can be defined between all pairs of x_1, x_2 and y. However when there are more than two variables in a study, then the simple correlation between any two variables are known as total correlation. All nine of such possible pairs are represented in the form of a matrix given above.

14.2.4 Adjusted Multiple Coefficient of Determination

In a multiple regression equation, addition of an independent or explanatory variable increases the value of R^2. For comparing two values of R^2, it is necessary to take into consideration the number of independent variables on which it is based. This is done by calculating an adjusted R^2 denoted by $\overline{R}^2$ (R bar-squared). This adjusted value of R^2 is (number at variables) known as adjusted multiple coefficient of determination, takes into account n (number of observations) and k (number of variables) for comparison in two situations, and is calculated as

$$\overline{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2) \tag{14.10}$$

where n is the sample size or the number of observations on each of the variables, and k is the number of independent variables. For the above example,

$$\overline{R}^2 = 1 - \left(\frac{10-1}{10-2-1}\right) \quad (1 - 0.9943)$$

$$= 0.9927$$

To start with when an independent variable is added i.e. the value of k is increased, the value of $\overline{R}^2$ increases but when the addition of another variable does not contribute towards explaining the variability in the dependent variable, the value of $\overline{R}^2$ decreases. This implies that the addition of that variable is redundant.

The adjusted R^2 i.e. $\overline{R}^2$ is lesser than R^2 as number of observations per independent variable decreases. However, $\overline{R}^2$ tends to be equal to R^2 as sample size increases for the given number of independent variables.

$\overline{R}^2$ is useful in comparing two regression equations having different number of independent variables or when the number of variables is the same but both are based on different sample sizes.

14.2.5 Dummy Variable

So far, we have considered independent variables which are quantifiable and measurable like income, sales, profit, etc. However, sometimes the independent variables may not be quantifiable and measurable and be only qualitative and categorical, and could impact the independent variable under study. For example, the amount of insurance policy, a person takes, could depend on his/her marital status which is categorical i.e. married or unmarried. The sale of ice cream might depend on the seasons viz. summer or other seasons. The performance of a candidate at a competitive examination depends not only on his/her I.Q. but also on the categorical variable "coached" or "un-coached".

Dummy variables are very useful for capturing a variety of qualitative effects by indicating '0' and '1' as two states of qualitative or categorical data. The dummy variable is assigned the value '1' or '0' depending on whether it does or does not possess the specified characteristic. Some examples are male and female, married and unmarried, MBA executives and non-MBA executives, trained and not trained, advertisement - I and advertisement - II, strategy like financial discount or gift item for sales promotion. Thus, a dummy variable modifies the form of a non-numeric variable to a numeric one. They are used as explanatory variables in a regression equation. They act like 'switch' which turn various parameters 'on' and 'off ' in an equation. Another advantage of '0' and '1' dummy variables is that even though it is a nominal-level variable – it can be treated statistically just like 'interval-level' variable which takes the value 1 or 0. It marks or encodes a particular

attribute "Indicative Variable" to "Binary Variable". It is a form of coding to transform non-metric data to metric data. It facilitates in considering two levels of an independent variable, separately.

Example 14.2

It is normally expected that a person with high income will purchase life insurance policy for a higher amount. However, it may be worth examining whether there is any difference in the amounts of insurance purchased by married and unmarried persons. To answer these queries, an insurance agent collected the data about the policies purchased by his clients during the last month. The data is as follows:

Sr. No. of Clients	*Annual Income (in Thousand Rs.)*	*Amount of Stipulated Annual Insurance Premium (in Thousand Rs.)*	*Marital Status (Married/Single)*
1	800	85	M
2	450	50	M
3	350	50	S
4	1500	140	S
5	1000	100	M
6	500	50	S
7	250	40	M
8	60	10	S
9	800	70	S
10	1400	150	M
11	1300	150	M
12	1200	110	M

Note: The marital status is converted into a independent variable by substituting 'M' by 1 and 'S' by 0 for the purpose of fitting the regression equation.

It may be verified that the multiple regression equation with amount of insurance premium as dependent variable and income as well as marital status as independent variables is

$$\text{Premium} = 5.27 + 0.091 \text{ Income} + 8.95 \text{ Marital Status}$$

The interpretation of the coefficient 0.091 is that for every additional thousand rupees of income, the premium increases by $1000 \times 0.091 = \text{Rs. } 91$

The interpretation of the coefficient 8.95, is that a married person takes an additional premium of Rs. 8, 950 as compared to a single person.

14.2.6 Partial Correlation Coefficients

So far, we have discussed total correlation coefficient and multiple correlation coefficient. In the above case of net profit planning, we had three variables viz. x_1, y and x_2. The correlation coefficients between any two variables viz. r_{yx_1}, r_{yx_2} and $r_{x_1x_1}$ are called total correlation coefficients. The total correlation coefficients indicate the relationship between the two variables **ignoring** the presence or effect of the other third variable. The multiple correlation coefficient $R_y \cdot x_1 x_2$ indicates the correlation between y and the estimate of y obtained by the regression equation of y on x_1 and

x_2. The partial correlation coefficients are defined as correlation between any two variables when the effect of the third variable on these two variables is removed or when the third variable is held constant. For example, $r_{yx_1.x_2}$ means the correlation between y and x_1 when the effect of x_2 on y and x_1 is removed or x_2 is held constant. The various partial correlation coefficients viz. $r_{yx_1.x_2}$ $r_{yx_2.x_1}$ and $r_{x_1x_2.y}$ are calculated by their formulas as follows:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2)\,(1 - r_{x_1x_2}^2)}} \tag{14.11}$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_1}^2)\,(1 - r_{x_1x_2}^2)}} \tag{14.12}$$

$$r_{x_1x_2 \cdot y} = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1 - r_{yx_2}^2)\,(1 - r_{yx_1}^2)}} \tag{14.13}$$

The values of the above partial correlation coefficients, $r_{yx_1.x_2}$, $r_{yx_2.x_1}$ and $r_{x_1.x_2 \cdot y}$ are 0.997, 0.977 and 0.973, respectively.

The interpretation of $r_{yx_2.x_1}$ = 0.977 is that it indicates the extent of linear correlation between y and x_2 when x_1 is held constant or its impact on y and x_2 is removed.

Similarly, the interpretation of $r_{x_1x_2.y}$ = 0.973 is that it indicates the extent of linear correlation between x_1 and x_2 when y is held constant or its impact on x_1 and x_2 is removed.

14.2.7 Partial Regression Coefficients

The regression coefficients b_1 and b_2 in the regression equation 14.1 are known as partial regression coefficients. The value of b_1 indicates the change that will be caused in y with a unit change in x_1 when x_2 is held constant. Similarly, b_2 indicates the amount by which y will change given a unit change in x_2. For illustration, in the regression equation, the interpretation of the value of 'b_1', i.e. 0.186, is that if x_2 (sales of non-food items) is held constant, then for every increase of Rs. 1 crore rise in the sales of food items, on an average, the net profit will rise by Rs.18.6 lakh. Similarly, the interpretation of 'b_2' i.e. 0.287 is that if x_1 (sales of food items) is held constant, then for every increase of Rs. 1 crore rise in the sales of non-food items, on an average, net profit will rise by Rs. 28.7 lakh.

14.2.8 Beta Coefficients

If the independent variables are standardised i.e. they are measured from their means and divided by their standard deviations, then the corresponding regression coefficients are called beta coefficients. Their advantages, like in simple regression analysis vide Section 8.5.3 are that the correlation and regression between standardised variables solves the problem of dealing with different units of measurements of the variables. Thus, the magnitudes of these coefficients can be used to compare the relative contribution of each independent variable in the prediction of each dependent variable. Incidentally, for the data in Illustration 14.1, relating to sales and net profit of supermarkets is reproduced below:

Supermarket	Net Profit (Rs. Crore)	Sales of Food Items (Rs. Crore)	Sales of Non-Food Item		Standardised Variables = (Variable – Mean)/s.d.		
	y_i	x_i	z_i	x_i^2	Y	X_1	X_2
1	5.6	20	5	400	–0.331	–0.34	–0.09
2	4.7	15	5	225	–1.291	–1.48	–0.09
3	5.4	18	6	324	–0.544	–0.8	0.792
4	5.5	20	5	400	–0.437	–0.34	–0.09
5	5.1	16	6	256	–0.864	–1.25	0.792
6	6.8	25	6	625	0.949	0.798	0.792
7	5.8	22	4	484	–0.117	0.114	–0.97
8	8.2	30	7	900	2.443	1.937	1.673
9	5.8	24	3	576	–0.117	0.57	–1.85
10	6.2	25	4	625	0.309	0.798	–0.97
Sum	59.1	215	51	4815			
mean	5.91	21.5	5.1				
Variance	0.88	19.25	1.29				
s.d.	0.937	4.387	1.136				

The multiple regression equation is as follows:

$$\boldsymbol{Y = 0.917}\ X_1 + \boldsymbol{0.348}\ X_2$$

The interpretation of beta values is that the contribution by sales of food items in profit is 0.917 as compared to the contribution by sale of non-food items which is 0.348.

14.2.9 Properties of R^2

As mentioned earlier that the coefficient of multiple correlation R is the ordinary or total correlation between the dependent variable and its estimate as derived by the regression equation i.e. $R = r_{y\hat{y}_i}$, and as such is always positive. Further,

(i) $R^2 \geq$ each of the square of total correlation coefficients of y with any one of the variables, $x_2, \ldots, x_k$.

(ii) R^2 is high if correlation coefficients between independent variables viz. $r_{xi\,xj}$ s are all low.

(iii) If $r_{xi\,xj} = 0$ for each $i \neq j$, then $R^2 = r^2_{yx_1} + r^2_{yx_2} + r^2_{yx_3} + \ldots + r^2_{yxk}$

14.2.10 Multicollinearity

Multicollinearity refers to the existence of high correlation between independent variables. Even if the regression equation is significant for the equation as a whole, it could happen that due to multicollinearity, the individual regression coefficients could be insignificant indicating that they do not have much impact on the value of the dependent variable. When two independent variables are highly correlated, they basically convey the same information and logically appears that only one of the two variables could be used in the regression equation.

If the value of R^2 is high and the multicollinearity problem exists, the regression equation can still be used for prediction of dependent variables given values of independent variables. However, it should not be used for interpreting partial regression coefficients to indicate impact of independent variables on the dependent variable.

The multicollinearity among independent variables can be removed, with the help of Principal Component Analysis discussed in this chapter. It involves forming new set of independent variables which are linear combinations of the original variables in such a way that there is no multicollinearity among the new variables.

If there are two variables, sometimes the exclusion of one may result in an abnormal change in the regression coefficient of the other variable; sometimes even the sign of the regression coefficient may change from + to – or vice versa, as demonstrated for the data given below.

y	x_1	x_2
10	12	25
18	16	21
18	20	22
25	22	18
21	25	17
32	24	15

It may be verified that the correlation between x_1 and x_2 is 0.91 indicating the existence of multicollinearity.

It may be verified that the regression equation of y on x_1 is

$$y = -3.4 + 1.2\, x_1 \tag{i}$$

the regression equation of y on x_2 is

$$y = 58.0 - 1.9\, x_2 \tag{ii}$$

and the regression equation of y on x_1 and x_2 is

$$y = 70.9 - 0.3\, x_1 - 2.3\, x_2 \tag{iii}$$

It may be noted that the coefficient of x_1 (1.2) in (i) which was positive when the regression equation of y on x_1 was considered, has become negative (–0.3) in equation (iii), when x_2 is also included in the regression equation. This is due to high correlation of –0.91 between x_1 and x_2. It is, therefore, desirable to take adequate care of multicollinearity.

14.2.11 Tests for Significance of Regression Model and Regression Coefficients

In any given situation, one can always define a dependent variable and some independent variables, and thus define a regression model or equation. However, an important issue arises is whether all the defined variables in the model, as a whole, have a real influence on the dependent variable, and are able to explain the variation caused in the dependent variable. For example, one may postulate that the sales of a company manufacturing a paint (defined as dependent variable) is dependent on the expenditure on R&D, advertising expenditure, price of the paint, discount to the whole sellers and number of salesmen. While, these variables might be found to be significantly impacting the sales of the company, it could also happen that these variables might not influence the sales as the more important factors could be the quality of the paint, availability and pricing of another similar

type of paint. Further, even if the four variables mentioned above are found to be significantly contributing, as a whole, to the sales of the paint, but one or some of these might not be influencing the sales in a significant way.

For example, it might happen that the sales are insensitive to the advertising expenses i.e. increasing the expenditure on advertising might not be increasing the sales in a significant way. In such case, it is advisable to exclude this variable from the model, and use only the other three variables. As explained in the next section, it is not advisable to include a variable unless its contribution to variation in the dependent variable is significant. These issues will be explained with examples in subsequent sections.

14.2.12 Regression Model with More Than Two Independent Variables

So far we have discussed only the derivation of the regression equation and interpretations of the correlation and regression coefficients. Further, we have confined to only two independent variables. However, sometimes, it is advisable to have more than two independent variables. Also, for using the equation for interpreting and predicting the values of the dependent variable with the help of independent variables, there are certain assumptions to validate the regression equation. We also have to test whether all or some of the independent variables are really significant to have an impact on the dependent variable. In fact, we also have to ensure that only the optimum numbers of variables are used in the final regression equation. While details will be discussed later on, it may be mentioned for now that mere increase in the number of independent variables does not ensure better predictive capability of the regression equation. Each variable has to compete with the other to be included or retained in the regression equation.

14.2.13 Selection of Independent Variables

Following are three prime methods of selecting independent variables in a regression model:

- General Method
- Hierarchical Method
- Stepwise Method

These are described as follows:

14.2.13.1 General Method This method is used when a researcher knows exactly which independent variables contribute significantly in the regression equation. In this method, all the independent variables are considered together and the regression model is derived.

It is often difficult to identify the exact set of variables that are significant in the regression model and the process of finding these may have many steps or iterations as explained through an illustration in the next section. This is the limitation of this method. This limitation can be overcome in the stepwise regression method.

14.2.13.2 Hierarchical Method This method is used when a researcher has clearly identified three different types of variables namely dependent variable, independent variable/s and the control variable/s.

This method helps the researcher to find the relationship between the independent variables and the dependent variable, in the presence of some variables that are controlled in the experiment. Such variables are termed as control variables. The control variables are first entered in the hierarchy, and then the independent variables are entered. This method is available in most statistical software including SPSS.

14.2.13.3 Stepwise Method This method is used when a researcher wants to find out, which independent variables significantly contribute in the regression model, out of a set of independent variables. This method finds the best fit model, i.e. the model which has a set of independent variables that contribute significantly in the regression equation.

For example, if a researcher has identified some three independent variables that may affect the dependent variable, and wants to find the best combination of these three variables which contribute significantly in the regression model, he or she may use stepwise regression. The software would give the exact set of variables that contribute or are worth keeping in the model.

There are three most popular stepwise regression methods namely, forward regression, backward regression and stepwise regression. In forward regression, one independent variable is entered with dependent variable and the regression equation is arrived along with other tests like ANOVA, t tests etc.; in the next iteration, one more independent variable is added and is compared with the previous model. If the new variable significantly contributes in the model, it is kept, otherwise it is thrown out from the model. This process is repeated for each remaining independent variables, thus arriving at a model that is significant containing all contributing independent variables. The backward method is exactly opposite to this method. In case of backward method, initially all the variables are considered and they are removed one by one if they do not contribute in the model.

The stepwise regression method is a combination of the forward selection and backward elimination methods. The basic difference between this and the other two methods is that in this method, even if a variable is selected in the beginning or gets selected subsequently, it has to keep on competing with the other entering variables at every stage to justify its retention in the equation.

These steps are explained in the next section, with an example.

14.2.14 Selection of Independent Variables in a Regression Model

Whenever, we have several independent variables which influence a dependent variable, an issue arises whether it is worthwhile to retain all the independent variables or whether it is worthwhile to include only some of the variables which have relatively more influence on the dependent variables as compared to the others. There are several methods to select the most appropriate or significant variables out of the given set of variables. However, we shall describe one of the methods using $\overline{R}^2$ as the selection criteria. The method is illustrated with the help of data given in Example 14.2.

Example 14.2

The following Table gives certain parameters about some of the top rated companies in the ET 500 listings published in the issue of February 2006.

Sr. No	*Company*	*M-Cap Oct. '05*		*Net Sales Sept.' 05*		*Net Profit Sept' 05*		*P/E as on Oct. 31' 05*	
	Company	*Amount*	*Rank*	*Amount*	*Rank*	*Amount*	*Rank*	*Amount*	*Rank*
1	Infosys Technologies	68560	3	7836	29	2170.9	10	32	66
2	Tata Consultancy Services	67912	4	8051	27	1831.4	11	30	74
3	Wipro	52637	7	8211	25	1655.8	13	31	67
4	Bharti Tele-Ventures *	60923	5	9771	20	1753.5	12	128	3

(*Contd*)

(*Contd*)

5	ITC	44725	9	8422	24	2351.3	8	20	183
6	Hero Honda Motors	14171	24	8086	26	868.4	32	16	248
7	Satyam Computer Services	18878	19	3996	51	844.8	33	23	132
8	HDFC	23625	13	3758	55	1130.1	23	21	154
9	Tata Motors	18881	18	18363	10	1314.9	17	14	304
10	Siemens	7848	49	2753	75	254.7	80	38	45
11	ONGC	134571	1	37526	5	14748.1	1	9	390
12	Tata Steel	19659	17	14926	11	3768, 6	5	5	469
13	Steel Authority of India	21775	14	29556	7	6442.8	3	3	478
14	Nestle India	8080	48	2426	85	311.9	75	27	99
15	Bharat Gorge Co.	6862	55	1412	128	190.5	97	37	48
16	Reliance Industries	105634	2	74108	2	9174.0	2	13	319
17	HDFC Bank	19822	16	3563	58	756.5	37	27	98
18	Bharat Heavy Electricals	28006	12	11200	17	1210.1	21	25	116
19	ICICI Bank	36890	10	11195	18	2242.4	9	16	242
20	Maruti Udyog	15767	22	11601	16	988.2	26	17	213
21	Sun Pharmaceuticals	11413	29	1397	130	412.2	66	30	75

*The data about Bharti Tele-Ventures is not considered for analysis because its P/E ratio is exceptionally high.

In the above example, we take Market Capitalisation as the dependent variable, and Net Profit, P/E Ratio and Net Sales as independent variables.

We may add that this example is to be viewed as an illustration of selection of optimum number of independent variables, and not the concept of financial analysis.

The notations used for the variables are as follows:

Y	*Market Capitalisation*
x_1	Net Sales
x_2	Net Profit
x_3	P/E Ratio

Step I:

First of all we calculate the total correlation coefficients among all the dependent and independent variables. We also calculate the correlation coefficients of the dependent variable. These are tabulated below.

	1	*2*	*3*
	Net Sales	*Net Profit*	*P/E Ratio*
Net Sales	1.0000		
Net Profit	0.7978	1.0000	
P/E Ratio	–0.5760	–0.6004	1.0000
Market Cap	0.6874	0.8310	–0.2464

We note that the correlation of y with x_2 is the highest. We, therefore, start by taking only this variable in the regression equation.

Step II:

The regression equation of y on x_2 is

$$y = 15465 + 7.906\ x_2$$

The values of R^2 and $\overline{R}^2$ are : $R^2 = 0.6906$ $\overline{R}^2 = 0.6734$

Step III:

Now, we derive two regression equations, one by adding x_1 and one by adding x_3 to see which combination viz. x_2 and x_3 or x_2 and x_1 is better.

The regression equation of y on x_2 and x_1 is

$$Y = 14989 + 7.397\ x_2 + 0.135\ x_1$$

The values of R^2 and $\overline{R}^2$ are: $R^2 = 0.6922$ $\overline{R}^2 = 0.656$

The regression equation of y on x_2 and x_3 is

$$Y = -\ 19823 + 10.163\ x_2 + 1352.4\ x_3$$

The values of R^2 and $\overline{R}^2$ are : $R^2 = 0.7903$ $\overline{R}^2 = 0.7656$

Since $\overline{R}^2$ for the combination x_2 and x_3 (0.7656) is higher than $\overline{R}^2$ for the combination x_2 and x_1 (0.6734), we select x_3 as the additional variable along with x_2.

It may be noted that R^2 with variable x_2 and x_3 (0.7903) is more than value of R^2 with only the variable (0.6906). Thus, it is advisable to have x_3 along with x_2 in the model.

Step IV:

Now we include the last variable viz. x_1 to have the model as

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

The requisite calculations are too cumbersome to be carried out manually, and, therefore, we use Excel spreadsheet which yields the following regression equation:

$$Y = -\ 23532 + 0.363x_1 + 8.95x_2 + 1445.6x_3$$

The values of R^2 and $\overline{R}^2$ are: $R^2 = 0.8016$ $\overline{R}^2 = 0.7644$

It may be noted that inclusion of x_1 in the model has very marginally increased the value of R^2 from 0.7903 to 0.8016, but the adjusted value of R^2 i.e. $\overline{R}^2$ has come down from 0.7656 to 0.7644. Thus, it is not worthwhile to add the variable x_1 to the regression model having variables as x_2 and x_3.

Step V:

The advisable regression model is by including only x_2 and x_3

$$Y = -19823 + 10.163\, x_2 + 1352.4\, x_3 \quad (14.14)$$

This is the best regression equation fitted to the data on the basis of $\overline{R}^2$ criterion, as discussed above.

We have discussed the same example to illustrate the method using SPSS in Section 14.2.17.

14.2.15 Generalised Regression Model

In general, a regression equation, or also referred to as model, is written as follows:

$$y_i = b_o + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + \ldots\ldots + b_1 x_{ki} + e_i \quad (14.15)$$

where, there are k independent variables viz. x_1, x_2, x_3 ……. and x_k, and e_i is the error or residual term which is not explained by the regression model.

14.2.15.1 Assumptions for the Multiple Regression Model As in simple regression analysis, we need certain assumptions for validity of regression analysis of the model defined in the above equation 14.14.

(a) Each error term, e_i, is distributed normal with conditional mean zero and variance σ^2, for $i = 1, 2 \ldots n$.

(b) The error term e_i is independent of each of the k independent variables viz. $x_1, x_2, \ldots, x_k$.

(c) Error terms of any two variables, say x_i and x_j, viz. e_i and e_j are not correlated with one another; that is, correlation between any two of them is zero. This assumption means that there is no (serial correlation) among residual terms.

(d) There is no multicollinearity among the independent variables i.e. they are not linearly related to each other.

14.2.16 Applications in Finance

In this section, we indicate financial applications of regression analysis in some aspects relating to stock market.

(i) Individual Stock Rates of Return, Payout Ratio and Market Rates of Return

Let the relationship of rate of return of a stock with the payout ratio defined as the ratio of dividend per share to customer earnings per share, and the rate of return on BSE SENSEX stocks as a whole, be

$$y = b_0 + b_1 \text{ (payout ratio)} + b_2 \text{ (rate of return on Sensex)}$$

Let us assume that the relevant data is available, and is collected over a period of last 10 years yields the following equation

$$y = 1.23 - 0.22 \text{ payout ratio} + 0.49 \text{ rate of return}$$

The coefficient – 0.22 indicates that for a 1% increase in pay-out ratio, the return on the stock reduces by 0.22% when the rate of return is held constant. Further, the coefficient 0.49 implies that a 1% increase in the rate of return on BSE SENSEX, the return on the stock increases by 0.49 % when the payout ratio is held constant.

Further, let the calculations yield the value of R^2 as 0.66.

The value of $R^2 = 0.66$ implies that 66% of variation in the rate of return on the investment in the stock is explained by pay-out ratio and the return on BSE SENSEX.

(ii) Determination of Price per Share

To further demonstrate application of multiple regression techniques, let us assume that a cross-section regression equation is fitted with dependent variable being the price per share (y) of the 30 companies used to compile the SENSEX, and the independent variables being the dividend per share (x_1) and the retained earnings per share (x_2) for the 30 companies. As mentioned earlier, in a cross-section regression, all data come from a single period.

Let us assume that the relevant data is available, and the data is collected for SENSEX stocks in a year, yield the following regression equation:

$$y = 25.45 + 15.30x_1 + 3.55x_2$$

The regression equation could be used for interpreting regression coefficients and predicting average price per share given the values of dividend paid and earnings retained.

The coefficient 15.30 of x_1 (average price per share) indicates that the average price per share increases by 15.30 when the dividend per share increases by Re. 1 when the retained earnings are held constant.

The regression coefficient 3.55 of x_2 means that when the retained earnings increases by Re. 1.00, the price per share increases by Rs. 3.55 when dividend per share is held constant.

The use of multiple regression analysis in carrying out cost analysis was demonstrated by Bentsen in 1966. He collected data from a firm's accounting, production and shipping records to establish a multiple regression equation.

14.2.17 Multiple Regression Using SPSS

SPSS is the most commonly used statistical tool for performing Multiple Regression Analysis.

The method, terms used in SPSS, and the interpretations of the SPSS output shall be discussed in brief.

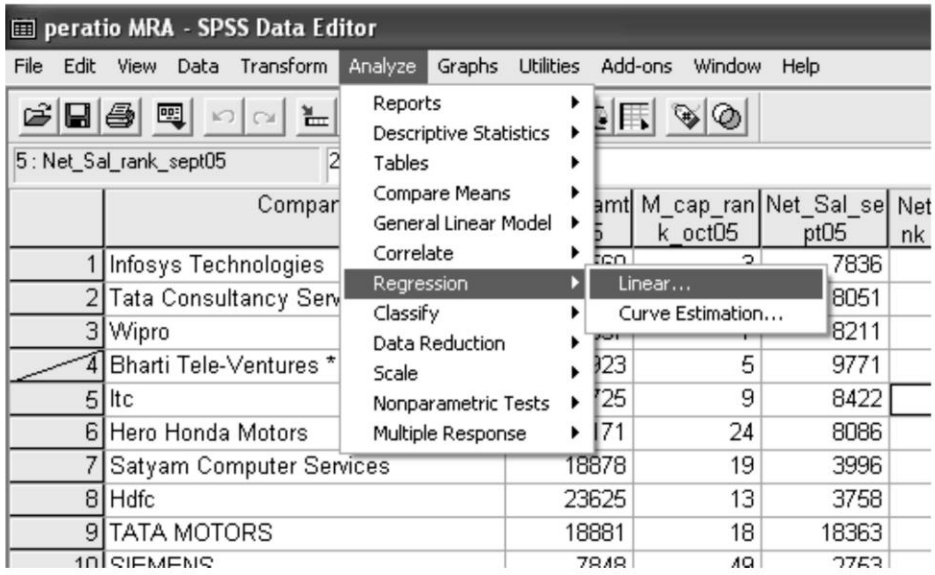
We have discussed in Section 14.2.14, three methods of entering variables for multiple regression analysis. SPSS allows selecting one of the three methods while entering the variables.

We shall illustrate the Example 14.2, given in Section 14.2.14, using SPSS. It may also be noted that as Bharti Telecom has too high P/E Ratio, it is therefore omitted from the analysis. A researcher may perform outlier analysis on the data and omit the cases that are outliers.

It may also be noted that we will discuss the output which is different than the previous method. The output that is similar to the previous method is not discussed in this method.

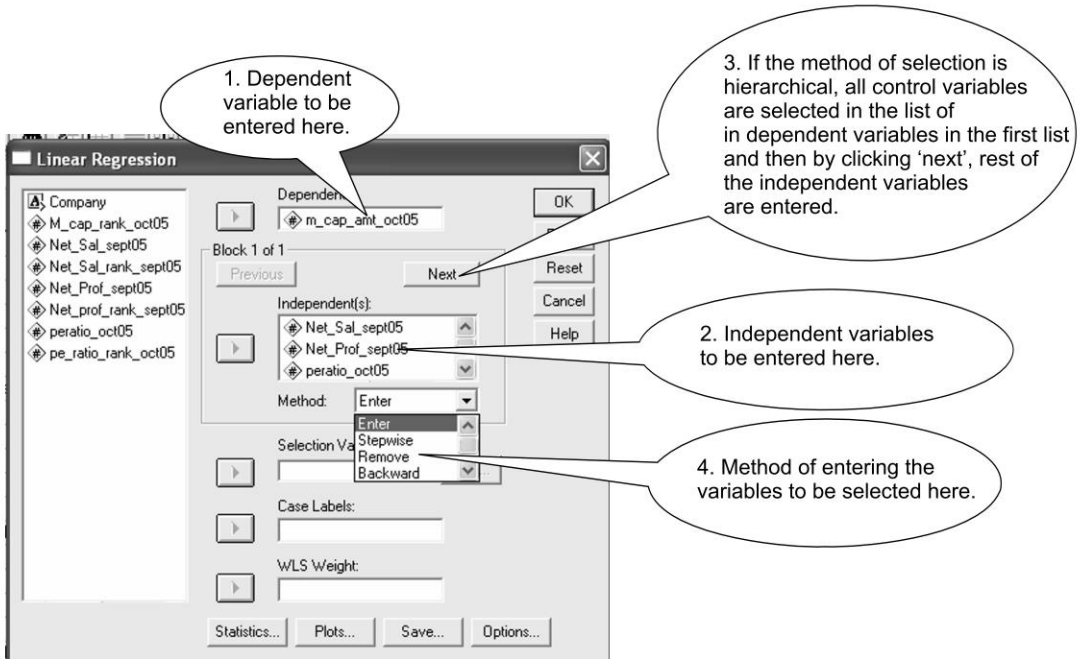
The regression analysis can be done in SPSS by using 'Analyse' option and selecting 'Linear regression' as shown in the following snapshot:

SPSS Snapshot MRA 1



The next box that appears is shown in the following SPSS snapshot:

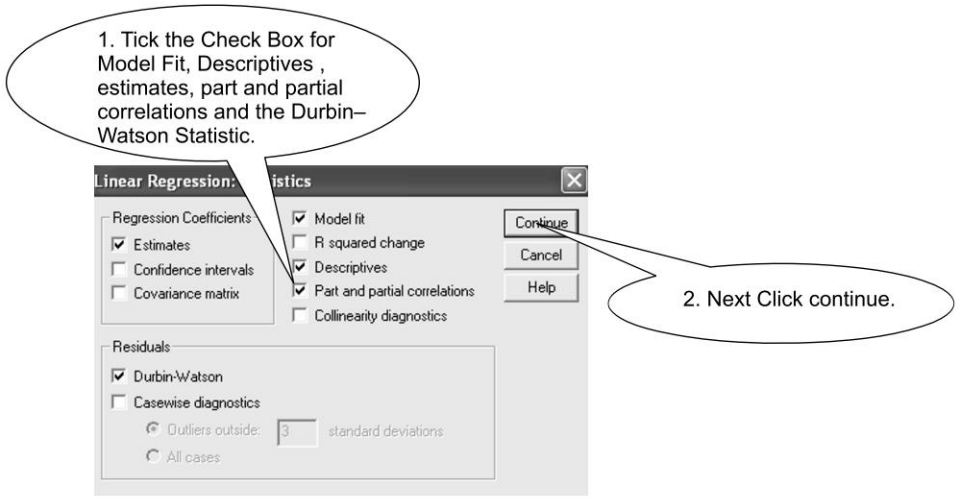
SPSS Snapshot MRA 2



If the method of selection is General method, 'Enter' should be selected from the drop box. If the method is stepwise, 'Stepwise' should be selected from the list. We have explained the criteria for selecting appropriate method in Section 14.2.14.

The next step is to click on 'Statistics' button in the bottom of the box. When one clicks on Statistics, the following box will appear.

SPSS Snapshot MRA 3

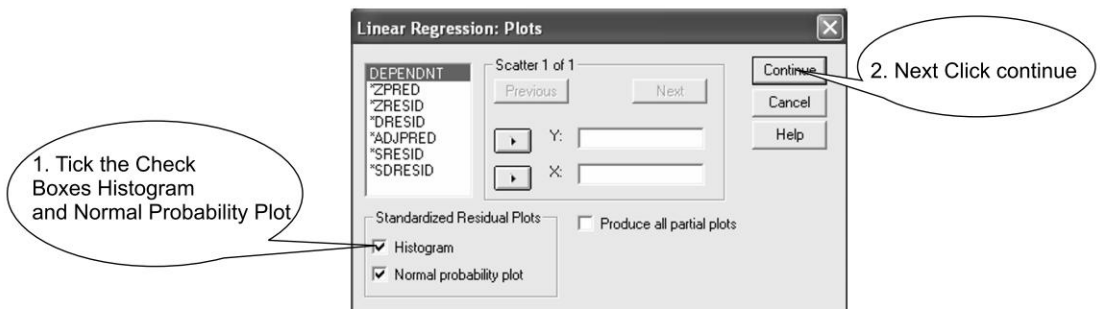


The **Durbin–Watson Statistic** is used to check the assumption of regression analysis which states that the error terms should be uncorrelated. While its desirable value is 2, the desirable range is 1.5 to 2.5.

After clicking 'Continue', SPSS will return to screen as in SPSS Snapshot MRA 2.

In this snapshot, click on the plots button at the bottom. The next box that would appear is given in the following Snapshot MRA 4.

SPSS Snapshot MRA 4



The residual analysis is done to check the assumptions of multiple regression that the residuals should be normally distributed. This assumption can be checked by viewing the histogram and normal probability plot.

After clicking 'Continue', SPSS will take back to the Snapshot MRA 3.

By clicking 'OK', SPSS will carry out the analysis and give the output in the Output View.

We will discuss two outputs using the same data. One, by using General method for entering variables, and the other by selecting stepwise method for entering variables.

General Method for Entering Variables – SPSS Output Regression

Descriptive Statistics

	Mean	*Std. Deviation*	*N*
m_cap_amt_oct05	3628.580	34367.670	20
Net_Sal_sept05	13419.30	17025.814	20
Net_Prof_sept05	2633.38	3612.171	20
peratio_oct05	21.70	10.037	20

Descriptive statistics is generally useful in understanding overall distributions of variables.

Correlations

		m_cap_amt_ oct05	Net_Sal_ sept05	Net_Prof_ sept05	peratio_oct05
Pearson Correlation	M_cap_amt_oct05	1.000	0.687	0.831	–0.246
	Net_Sal_sept05	0.687	1.000	0.798	–0.576
	Net_Prof_sept05	0.831	0.798	1.000	–0.600
	peratio_oct05	–0.246	–0.576	–0.600	1.000
Sig. (1-tailed)	M_cap_amt_oct05	.	0.000	0.000	0.148
	Net_Sal_sept05	0.000	.	0.000	0.004
	Net_Prof_sept05	0.000	0.000	.	0.003
	peratio_oct05	0.148	0.004	0.003	.
N	M_cap_amt_oct05	20	20	20	20
	Net_Sal_sept05	20	20	20	20
	Net_Prof_sept05	20	20	20	20
	peratio_oct05	20	20	20	20

'Part and partial correlations matrix' is useful in understanding the relationships between the independent and dependent variables. The regression analysis is valid only if the independent and dependent variables are not interrelated. If these are related to each other, they may lead to misinterpretation of the regression equation. This is termed as multicollinearity, and its impact is described in Section 14.2.10. The above correlation matrix is useful in checking the inter relationships between the independent variables. In the above table, the correlations in square are correlation of independent variables with dependent variables and are high (0.687 and 0.831) which means that the two variables are related. Whereas the correlations between the independent variables (0.798, –0.576 and –0.6) are high which means that this data may have multicollinearity. Generally, very high correlations between the independent variables like more than 0.9, may make the entire regression analysis unreliable for interpreting the regression coefficients.

Variables Entered/Removed[b]

Model	*Variable Entered*	*Variables Removed*	*Method*
1	peratio_oct05,Net_Sal_sept05,Net_Prof_sept05a	.	Enter

a. All requested variables entered.
b. Dependent Variable: m_cap_amt_oct05

Since the method selected was Enter method or General method, this table does not communicate any meaning.

Model Summary[b]

Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*	*Durbin–Watson*
1	0.895[a]	0.802	0.764	16682.585	0.982

a. Predictors: (Constant), peratio_oct05, Net_Sal_sept05, Net_Prof_sept05
b. Dependent Variable: m_cap_amt_oct05

This table gives the model summary for the set of independent and dependent variables. R^2 for the model is 0.802 which is high and means that around 80% of variation in dependent variable (market capitalisation) is explained by the three independent variables (net sale, net profit and P/E ratio). The Durbin–Watson statistic for this model is 0.982, which is very low. The desired value is in the range 1.5 to 2.5. It may, therefore, be appended as a caution that the assumption that the residuals are uncorrelated is not valid.

ANOVA[b]

Model		*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
1	Regression	1.80E + 10	3	5996219888	21.545	0.000[a]
	Residual	4.45E + 09	16	278308655.0		
	Total	2.24E + 10	19			

a. Predictors: (Constant), peratio_oct05, Net_Sal_sept05, Net_Prof_sept05
b. Dependent Variable: m_cap_amt_oct05

The ANOVA table for the regression analysis indicates whether the model is significant, and valid or not. The ANOVA is significant, if the 'Sig.' column in the above table is less than the level of significance (generally taken as 5% or 1%). Since 0.000 < 0.01, we conclude that this model is significant.

If the model is not significant, it implies that no relationship exists between the set of variables.

Coefficients[a]

Model	*Unstandardized Coefficients*		*Standardized Coefficients*			*Correlations*		
	B	*Std. Error*	*Beta*	*t*	*Sig.*	*Zero-order*	*Partial*	*Part*
1. (Constant)	–3531.5	13843.842		–1.700	0.109			
Net_Sal_sept05	0.363	0.381	0.180	0.953	0.355	0.687	0.232	0.106
Net_Prof_sept05	8.954	1.834	0.941	4.882	0.000	0.831	0.774	0.544
peratio_oct05	1445.613	486.760	0.422	2.970	0.009	–0.246	0.596	0.331

a. Dependent Variable: m_cap_amt_oct05

This table gives the regression coefficients and their significance. The equation can be considered as:

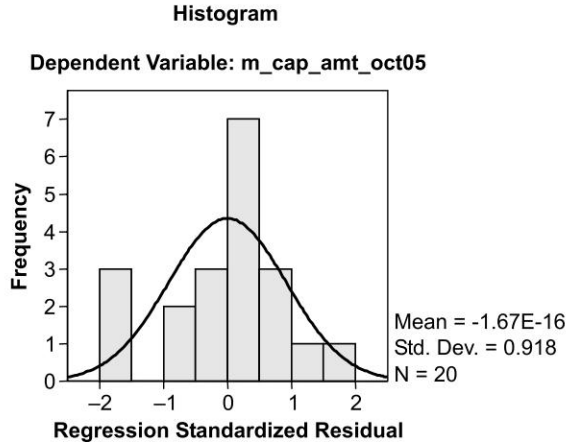
Market capitalisation = – 23531.5 + 0.363 × Net Sales + 8.954 × Net Profits + 1445.613 × P/E Ratio

Residuals Statistics[a]

	Minimum	*Maximum*	*Mean*	*Std. Deviation*	*N*
Predicted Value	10307.05	135150.28	36285.80	30769.653	20
Std. Predicted Value	–0.844	3.213	0.000	1.000	20
Standard Error of Predicted Value	4242.990	16055.796	6689.075	3390.080	20
Adjusted Predicted Value	9759.82	139674.61	36258.65	30055.284	20
Residual	–27441.7	28755.936	0.000	15308.990	20
Std. Residual	–1.645	1.724	0.000	0.918	20
Stud. Residual	–1.879	1.809	–0.009	0.999	20
Deleted Residual	–35824.0	31681.895	27.147	18618.671	20
Stud. Deleted Residual	–2.061	1.964	–0.021	1.060	20
Mahal. Distance	0.279	16.649	2.850	4.716 20	
Cook's Distance	0.000	0.277	0.059	0.091	20
Centered Leverage Value	0.015	0.876	0.150	0.248	20

a. Dependent Variable: m_cap_amt_oct05 a.

Charts



The above chart is to test the validity of the assumption that the residuals are normally distributed. Looking at the chart one may conclude that the residuals are more or less normal. This can be tested using Chi-square goodness of fit test.

Stepwise Method for Entering Variables – SPSS Output

It may be noted that as Bharti telecom has too high P/E ratio, it is omitted from the analysis.

Regression

Descriptive Statistics

	Mean	*Std. Deviation*	*N*
m_cap_amt_oct05	36285.80	34367.670	20
Net_Sal_sept05	13419.30	17025.814	20
Net_Prof_sept05	2633.38	3612.171	20
peratio_oct05	21.70	10.037	20

Correlations

		M_cap_amt_oct05	*Net_Sal_sept05*	*Net_Prof_sept05*	*peratio_oct05*
Pearson Correlation	M_cap_amt_oct05	1.000	0.687	0.831	0.246
	Net_Sal_sept05	0.687	1.000	0.798	–0.567
	Net_Prof_sept05	0.831	0.798	1.000	–0.600
	peratio_oct05	–0.246	–0.576	–0.600	1.000
Sig. (1-tailed)	M_cap_amt_oct05	.	0.000	0.000	0.148
	Net_Sal_sept05	0.000	.	0.000	0.004
	Net_Prof_sept05	0.000	0.000	.	0.003
	peratio_oct05	0.148	0.004	0.003	.
N	M_cap_amt_oct05	20	20	20	20
	Net_Sal_sept05	20	20	20	20
	Net_Prof_sept05	20	20	20	20
	peratio_oct05	20	20	20	20

Variables Entered/Removed[a]

Model	*Variables Entered*	*Variables Removed*	*Method*
1	Net_Prof_sept05		Stepwise (Criteria: Probability-of-F-to-enter <= 0.050, Probability-of-F-to-remove >= 0.100).
2	peratio_oct05		Stepwise (Criteria: Probability-of-F-to-enter <= 0.050, Probability-of-F-to-remove >= 0.100).

[a]Dependent Variable: m_cap_amt_oct05 a.

This table gives the summary of the entered variables in the model.

Model Summary[c]

Model	*R*	*R square*	*Adjusted R Square*	*Std. Error of the Estimate*	*Durbin-Watson*
1	0.831[a]	0.691	0.673	19641.726	
2	0.889[b]	0.790	0.766	16637.346	1.112

a. Predictors: (Constant), Net Net_Prof_sept05
b. Predictors: (Constant), Net_Pref_sept05, peratio_oct05
c. Dependent Variable: m_cap_amt_oct05

In the previous method, there was only one model. Since this is a stepwise method, it will give all models that are significant in each step. The Durbin–Watson statistic is improved from the previous model but is still less than the desired range (1.5 to 2.5). The second model (stepwise) model is generally the best model. This can be verified by the higher value of R^2. It consists of dependent variable, market capitalisation and independent variables, Net profit and P/E Ratio is the best model. The following table gives coefficients for the best model:

The following table gives ANOVA for all the iterations (in this case 2), and both are significant:

ANOVA[c]

Model		*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
1.	Regression	1.55E+10	1	1.550E+10	40.169	0.000[a]
	Residual	6.94E+09	18	385797411.3		
	Total	2.24E+10	19			
2.	Regression	1.77E+10	2	8867988179	32.037	0.000[b]
	Residual	4.71E+09	17	276801281.6		
	Total	2.24E+10	19			

a. Predictors: (Constant), Net_Prof_sept05
b. Predictors: (Constant), Net_Prof_sept05, peratio_oct05
c. Dependent Variable: m_cap_amt_oct05

Coefficients[a]

Model		*Unstandardized Coefficients*		*Standardized Coefficients*			*Correlations*		
		B	*Std. Error*	*Beta*	*t*	*Sig.*	*Zero-order*	*Partial*	*Part*
1.	(Constant)	15465.085	5484.681		2.820	0.011			
	Net_Prof_sept05	7.906	1.247	0.831	6.338	0.000	0.831	0.831	0.831
2.	(Constant)	–19822.8	13249.405		–1.496	.153			
	Net_Prof_sept05	10.163	1.321	1.068	7.691	0.000	0.831	0.881	0.854
	peratio_oct05	1352.358	475.527	0.395	2.844	.011	–0.246	0.568	0.316

a. Dependent Variable: m_cap_amt_oct05.

It may be noted in the above Table that the values of the constant and regression coefficients are the same as in equation 14.14, derived manually. The SPSS stepwise regression did this automatically, and the results we got are the same.

The following table gives summary of excluded variables in the two models:

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	Net_Sal_sept05	0.067[a]	0.301	0.767	0.073	0.363
	Peratio_oct05	0.395[a]	2.844	0.011	0.568	0.639
2	Net_Sal_sept05	0.180[b]	0.953	0.355	0.232	0.349

a. Predictors in the Model: (Constant), Net_Prof_sept05
b. Predictors in the Model: (Constant), Net_Prof_sept05, peratio_oct05
c. Dependent Variable: m_cap_amt_oct05

There may be a situation that a researcher would like to divide the data into two parts, and use one part to derive the model and other part to validate the model. SPSS allows to split the data into two groups termed as estimation group and validation group. The estimation group is used to fit the model, which is validated using validation group. This improves the validity of the model. This process is called as **cross validation**. This method can be used only if the data is large enough to fit the model. Random variable functions from SPSS can be used to select the data randomly from the SPSS file.

14.3 DISCRIMINANT ANALYSIS

Discriminant analysis is basically a classifying technique that is used for classifying a given set of objects, individuals, entities into two (or more) groups or categories based on the given data about their characteristics. It is the process of deriving an equation called **Discriminant Function** giving relationship between one dependent variable which is categorical i.e. it takes only two values, say, 'yes' or 'no', represented by '1' or '0', and several independent variables which are continuous. The independent variables, selected for the analysis, are such which contribute towards classifying an object, individual or entity in one of the two categories. For example, with the help of several financial indicators, one may decide to extend credit to a company or not. The classification could also be done in more than two categories.

Identifying a set of variables which discriminate 'Best' between the two groups is the first step in the discriminant analysis. These variables are called discriminating variables.

One of the simplest examples of **discriminating variable** is the 'height' in case of students of graduate students. Let there be a class of 50 students comprising boys and girls. Suppose we are given only roll numbers, and we are required to classify them by their sex or segregate boys and girls. One alternative is to take 'height' as the variable, and premise all those equal to or more than 5'6'' are boys and less than that height are girls. This classification should work well except in some cases where girls are taller than 5'6'' and boys are less than that height. In fact, one could work out from a large sample of students, the most appropriate value of the discriminating height. This example illustrates one fundamental aspect of discriminant analysis that in real life we cannot find discriminating variable(s) or function that can provide 100% accurate discrimination or classification. We can only attempt to find the best classification from a given set of data. Yet another

example is the variable 'marks' (percentage or percentile), in an examination which are used to classify students in two or more categories. As is well known even marks cannot guarantee 100% accurate classification.

Discriminant analysis is used to analyse relationships between a non-metric dependent variable and metric or dichotomous (Yes/No type or Dummy) independent variables. Discriminant analysis uses the independent variables to distinguish among the groups or categories of the dependent variable. The discriminant model can be valid or useful only if it is accurate. The accuracy of the model is measured on the basis of its ability to predict the known group memberships in the categories of the dependent variable.

Discriminant analysis works by creating a new variable called the **discriminant function score** which is used to predict to which group a case belongs. The computations find the coefficients for the independent variables that maximise the measure of distance between the groups defined by the dependent variable.

The discriminant function is similar to a regression equation in which the independent variables are multiplied by coefficients and summed to produce a score.

The general form of discriminant function is:

$$D = b_0 + b_1X_1 + b_2X_2 + \ldots \ldots + b_kX_k \qquad (14.16)$$

D = Discriminant Score

b_i = Discriminant coefficients or weights

X_i = Independent variables

The weights $\mathbf{b_i}$ s are calculated by using the criteria that the groups differ as much as possible on the basis of discriminant function.

If the dependent variable has only two categories, the analysis is termed as discriminant analysis. If the dependent variable has more than two categories, then the analysis is termed as **Multiple Discriminant Analysis**.

In case of multiple discriminant analysis, there will be more than one discriminant function. If the dependent variable has three categories like high risk, medium risk, low risk, there will be two discriminant functions. If dependent variable has four categories, there will be three discriminant functions. In general, the number of discriminant functions is one less than the number of categories of the dependent variable.

It may be noted that in case of multiple discriminant functions, each function needs to be significant to conclude the results.

The following illustrations explain the concepts and the technique of deriving a Discriminant function, and using it for classification. The objective in this example is to explain the concepts in a popular manner without mathematical rigour.

Illustration 14.2

Suppose, we want to predict whether a science graduate, studying *inter alia* the subjects of Physics and Mathematics, will turn out to be a successful scientist or not. Here, it is premised that the performance of a graduate in Physics and Mathematics, to a large extent, contributes in shaping up a successful scientist. The next step is to select some successful and some unsuccessful scientists, and record the marks obtained by them in Mathematics and Physics in their graduate examination. While in real life application, we have to select sufficient, say 10 or more number of students in both categories, just for the sake of simplicity, let the data on two successful and two unsuccessful scientists be as follows:

Successful Scientist		*Unsuccessful Scientist*	
Marks in Mathematics (M)	*Marks in Physics (P)*	*Marks in Mathematics (M)*	*Marks in Physics (P)*
12	8	11	7
8	10	5	9
Average : $\overline{M}_s = 10$	$\overline{P}_s = 9$	$\overline{M}_u = 8$	$\overline{P}_u = 8$
S: Successful		U: Unsuccessful	

It may be mentioned that the marks as 8, 10, etc. are taken just for the sake of ease in calculations.

The **discriminant function** assumed is

$$Z = w_1 M + w_2 P$$

The requisite calculations on the above data yield

$$w_1 = 9 \text{ and } w_2 = 23$$

Thus, the discriminant function works out to be

$$Z = 9 M + 23 P$$

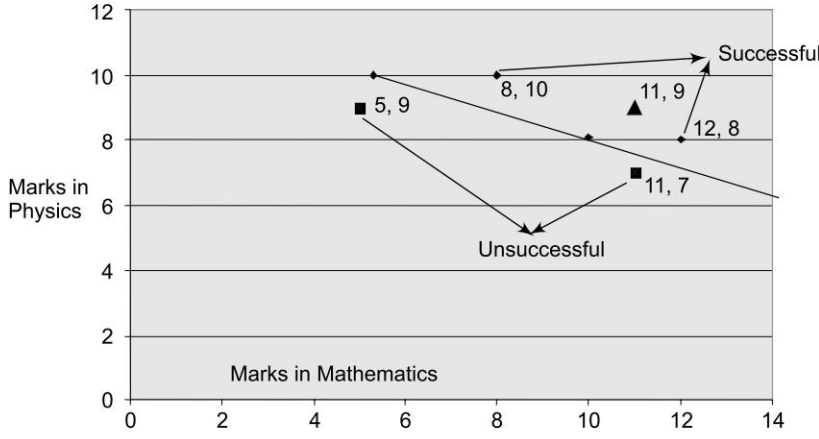
and the **discriminant score** works out to be

$$Z_C = \frac{(9 \times \overline{M}_s + 23\overline{P}_s) + (9 \times \overline{M}_U + 23\overline{P}_U)}{2}$$

$$= \frac{9 \times 10 + 23 \times 9 + 9 \times 8 + 23 \times 8}{2}$$

$$= \mathbf{276.5}$$

This discriminant score helps us to predict whether a graduate student will turn out to be a successful scientist or not. This score for the two successful scientists is 292 and 302, both being more than the discriminant score 276.5, the score is 214 and 270 for unsuccessful scientists, both being



less than 276.5. If a young graduate gets 11 marks in Mathematics and 9 marks in Physics, his or her score as per the discriminant function is $9 \times 11 + 23 \times 9 = 306$. Since this is more than the discriminant score of 276.5, we can predict that this graduate will turn out to be a successful scientist. This is pictorially depicted in the following:

It may be noted that the both the successful scientists' scores are above the discriminant line and the scores of both the unsuccessful scientists are below the discriminant line.

The student with assumed marks is classified in the category of successful scientist.

This example illustrates that with the help of past data, about objects including entities, individuals, etc., and their classification in two categories, one could derive the discriminant function and the discriminant scores. Subsequently, if the same type of data is given for some other object, the discriminant score could be worked out for that object and thus classify it in either of the two categories.

14.3.1 Some Other Applications of Discriminate Analysis

Some other applications of discriminant analysis are given below:

(i) Based on the past data available for a number of firms, about
- Current ratio (defined as Current Assets ÷ Current Liabilities)
- Debt/Asset Ratio (defined as Total Debt ÷ Total Assets)

and the information whether a firm succeeded or failed, a determinant function could be derived which could discriminate successful and failed firms based on their current ratio and debt/asset ratio.

As an example, the determinant function in a particular case could be

$$Z = -0.5 + 1.2x_1 - 0.07x_2$$

where,

x_1 represents the current ratio, and

x_2 represents the debt/asset ratio.

The function could be used to sanction or not sanction the credit to an approaching firm based on its current and debt asset ratios.

(ii) Suppose, one wants to have a comparative assessment of a number of factors (performances in various tests) responsible for effective performance of an executive, in an organisation. Let the factors and the corresponding variables, given in brackets:
- Score in admission test for MBA (x_1)
- Score in MBA (x_2)
- Score in the internal examination after initial induction training of one month (x_3)

Let the determinant function derived from the past data for 25 executives be:

$$Z = 3x_1 + 4x_2 + 6x_3$$

It may be noted that x_3 has maximum weightage as six. Thus, we may conclude that the most important factor for effective performance in the organization is the score in the internal examination after induction training.

14.3.2 Assumptions of Discriminant Analysis and Measure of Goodness of a Discriminant Function

For any statistical analysis to be valid, there are certain assumptions for the variables involved that must hold good. Further, as mentioned above, one cannot derive a discriminant function that would

ensure 100% accurate classification. It is, therefore, logical to measure the goodness of a function that would indicate the extent of confidence one could attach to the obtained results.

14.3.2.1 Assumptions of Discriminant Analysis The first requirement for using discriminant analysis is that the dependent variable should be non-metric and the independent variable should be metric or dummy.

The ability of discriminant analysis to derive discriminant function that provides accurate classifications is enhanced when the assumptions of normality, linearity and homogeneity of variance are satisfied. In discriminant analysis, the assumption of linearity applies to the relationships between pairs of independent variable. This can be verified from the correlation matrix, defined in Section 14.2.3. Like multiple regression, multicollinearity in discriminant analysis is identified by examining '**tolerance**' values. The multicollinearity problem can be resolved by removing or combining the variables with the help of Principal Component Analysis discussed in Section 14.6.3.

The assumption of homogeneity of variance is important in the classification stage of discriminant analysis. If one of the groups defined by the dependent variable has greater variance than the others, more cases will tend to be classified in that group. Homogeneity of variance is tested with Box's M test, which tests the null hypothesis that the group variance–covariance matrices are equal. If we fail to reject this null hypothesis and conclude that the variances are equal, we may use a pooled variance–covariance matrix in classification.

14.3.2.2 Tests Used for Measuring Goodness of a Discriminant Function There are two tests for judging goodness of a discriminant function:

1. An F test, Wilks' lambda Λ is used to test if the discriminant model as a whole is significant. Wilks' Λ for each independent variable is calculated using the formula:

$$\text{Wilks' } \Lambda = \frac{\text{Within Group Sum of Squares}}{\text{Total Sum of Squares}}$$

   It lies between 0 and 1. The large value of Λ indicates that there is no difference in the group means for the independent variable. Small values of Λ indicate group means are different for the independent variable. Smaller the value of Λ, more is the discriminating power of the variable, in the group.
2. If the F test shows significance, then the individual independent variables are assessed to see which of these differ significantly (in mean) by group and are subsequently used to classify the dependent variable.

14.3.3 Key Terms Related to Discriminant Analysis

Some key terms related to discriminant analysis are described below:

Term	*Description Glossary at end*
Discriminating Variables	These are the independent variables which are used as criteria for discrimination.
Discriminant Function	A discriminant function is a linear combination of discriminating (independent) variables as stated in equation 14.16. There are in general, k–1 discriminant functions if the dependent variable has k categories.

(*Contd*)

(*Contd*)

Eigenvalue	Eigenvalue for each discriminating function is defined as the ratio between groups to within group sum of squares. More the eigenvalue, more appropriate is the differentiation, hence the model. There is one eigenvalue for each discriminant function. For two-group DA, there is one discriminant function and one eigenvalue, which accounts for all of the explained variance. If there is more than one discriminant function, the first will be the largest and most important, the second next most important is explanatory power and so on.
Relative Percentage	The **relative percentage** of a discriminant function equals a function's eigenvalue divided by the sum of all eigenvalues of all discriminant functions in the model. Thus it is the percent of discriminating power for the model associated with a given discriminant function. Relative % is used to tell how many functions are important.
The Canonical Correlation, R*	It measures the extent of association between the discriminant scores and the groups. When R* is zero, there is no relation between the groups and the function. When the canonical correlation is large, there is a high correlation between the discriminant functions and the groups. It may be noted that for two-group DA, the canonical correlation is equivalent to the Pearson's correlation of the discriminant scores with the grouping variable.
Centroid	Mean values for discriminant scores for a particular group. The number of centroids equals the number of groups, being one for each group. Means for a group on all the functions are the group centroids.
Discriminant Score	The Discriminant Score, also called the DA score, is the value resulting from applying a discriminant function formula to the data for a given case. The *Z score* is the discriminant score for standardised data.
Cutoff	If the discriminant score of the function is less than or equal to the cutoff, the case is classified as 0, or if above the cutoff, it is classified as 1. When group sizes are equal, the cutoff is the mean of the two centroids (for two-group DA). If the groups are unequal, the cutoff is the weighted mean.
Standardised Discriminant Coefficients	Also termed as *standardised canonical discriminant function coefficients*, they are used to compare the relative importance of the independent variables, much as beta weights are used in regression. Note that the importance is assessed relative to the model being analysed. Addition or deletion of variables in the model can change discriminant coefficients markedly.
Functions at Group Centroids	The mean discriminant scores for each of the dependent variable categories for each of the discriminant functions in MDA. Two-group discriminant analysis has two centroids, one for each group. We want the means to be well apart to show the discriminant function is clearly discriminating. The closer the means, the more errors of classification.
(Model) Wilks' lambda	Used to test the significance of the discriminant function as a whole. The "Sig." level for this function is the significance level of the discriminant function as a whole. The larger the lambda, the more likely it is significant. A significant lambda means one can reject the null hypothesis that the two groups have the same mean discriminant function scores and conclude the model as discriminating.
ANOVA Table for Discriminant Scores	Another overall test of the DA model. It is an F test, where a "Sig." p value < 0.05 means the model differentiates discriminant scores between the groups significantly.

(*Contd*)

(*Contd*)

(Variable) Wilks' lambda	It can be used to test which independents contribute significantly to the discrimiinant function. The smaller the value of Wilks' lambda for an independent variable, the more that variable contributes to the discriminant function. Lambda varies from 0 to 1, with 0 meaning group means differ (thus, the more the variable differentiates the groups), and 1 meaning all group means are the same.
Classification Matrix or Confusion Matrix	Also called assignment, or prediction matrix or table, is used to assess the performance of DA. This is a table in which the rows are the observed categories of the dependent and the columns are the predicted categories of the dependents. When prediction is perfect, all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications. This percentage is called the **hit ratio**.
Expected hit ratio	The hit ratio is not relative to zero but to the percent that would have been correctly classified by chance alone. For two-group discriminant analysis with a 50–50 split in the dependent variable, the expected percentage is 50%. For unequally split 2-way groups of different sizes, the expected percent is computed in the "Prior Probabilities for Groups" table in SPSS, by multiplying the prior probabilities times the group size, summing for all groups, and dividing the sum by N. The best strategy is to pick the largest group for all cases, the expected percent is then the largest group size divided by N.
Cross-validation	Leave-one-out classification is available as a form of cross-validation of the classification table. Under this option, each case is classified using a discriminant function based on all cases except the given case. This is thought to give a better estimate of what classificiation results would be in the population.
Measures of association	Can be computed by the crosstabs procedure in SPSS if the researcher saves the predicted group membership for all cases.
Mahalanobis D-Square, Rao's V, Hotelling's trace, Pillai's trace, and Roy's gcr (greatest characteristic root)	Indices other than Wilks' lambda indicating the extent to which the discriminant functions discriminate between criterion groups. Each has an associated significance test. A measure from this group is sometimes used in stepwise discriminant analysis to determine if adding an independent variable to the model will significantly improve classification of the dependent variable. SPSS uses Wilks' lambda by default but also offers Mahalanobis distance, Rao's V, unexplained variance and smallest F ratio on selection.
Structure Correlations	These are also known as discriminant loadings, can be defined as simple correlations between the independent variables and the discriminant functions.

14.3.4 Discriminant Analysis Using SPSS

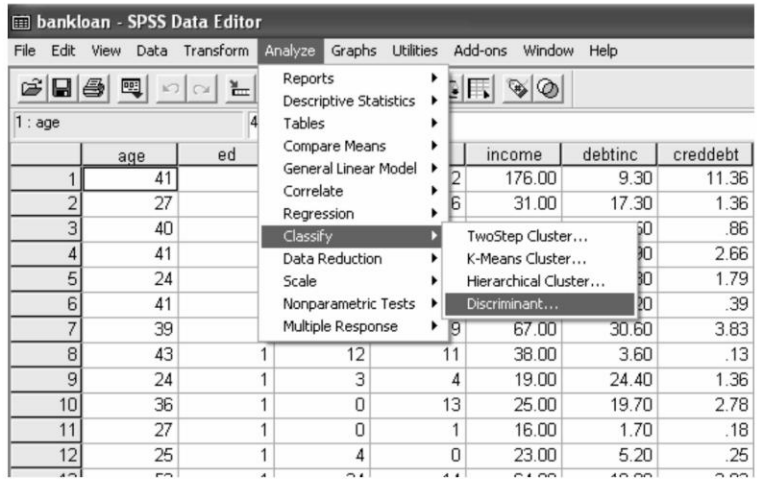
For illustration, the file **bankloan.sav** is used. This is a data file that concerns a bank's efforts to reduce the incidence of loan defaults. The file contains financial and demographic information on 850 past and prospective customers. The first 700 cases are customers who were previously given loans. The last 150 cases are prospective customers that the bank needs to classify as good or bad credit risks. This file is part of SPSS cases, and is in the tutorial folder of SPSS. Within tutorial folder, this file is in the sample files folder. For the convenience of readers this file has been provided in the CD with the book.

As in the case with multiple regression, discriminant analysis can be done by two principal methods, namely, enter method and stepwise method.

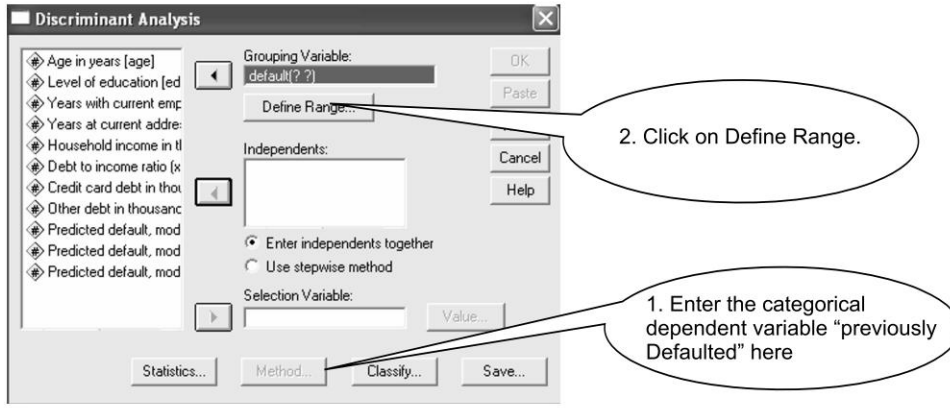
We will illustrate output for both the methods.

After opening the file bankloan.sav, one can click on 'Analyse' and 'Classify' as shown in the following snapshot:

SPSS Snapshot DA 1

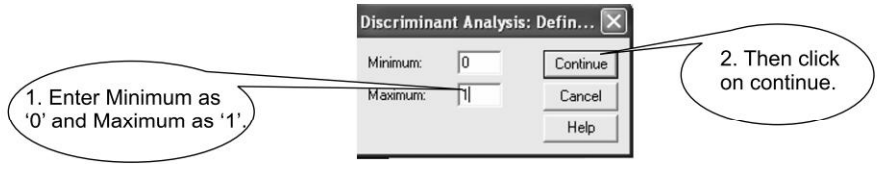


The next box that will appear is given in the following snapshot:



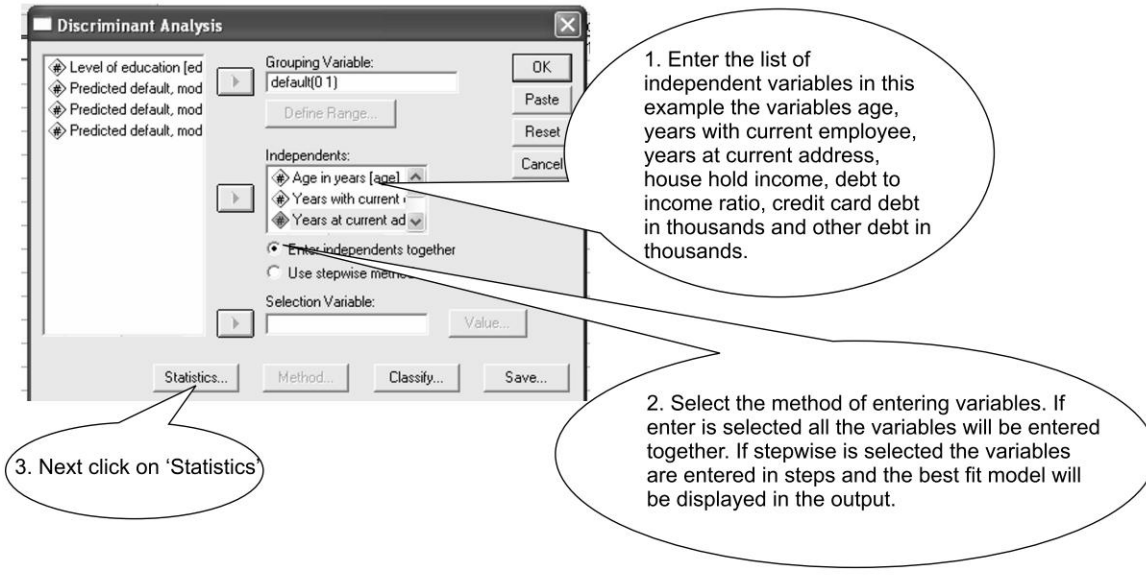
After entering the dependent variable and clicking on the 'Define Range' as shown above, SPSS will open the following box:

SPSS Snapshot DA 2



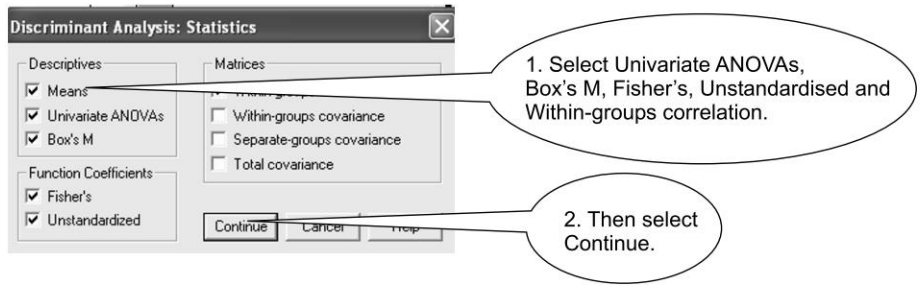
After defining the variable, one should click on 'Continue' button as shown above. SPSS will go back to the previous box shown below:

SPSS Snapshot DA 3



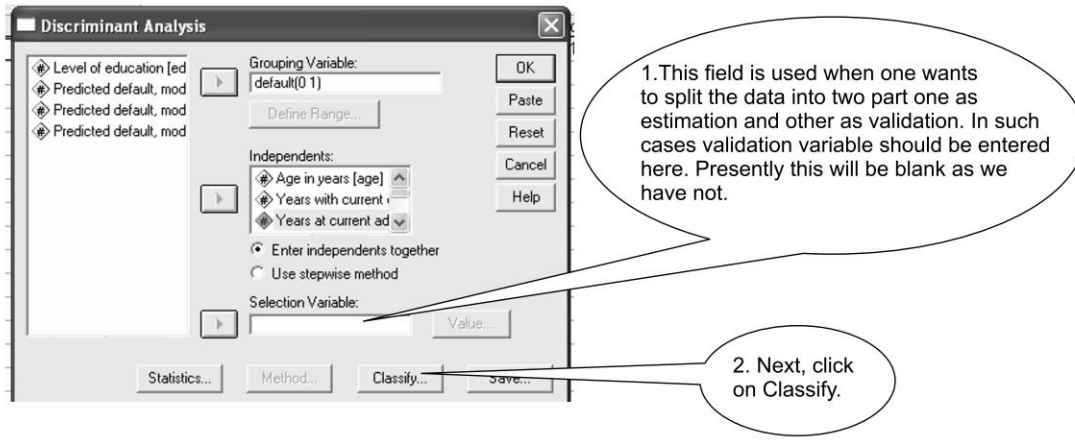
After selecting the dependent, independent variables and the method of entering variables one may click on Statistics, SPSS will open a box as shown below:

SPSS Snapshot DA 4



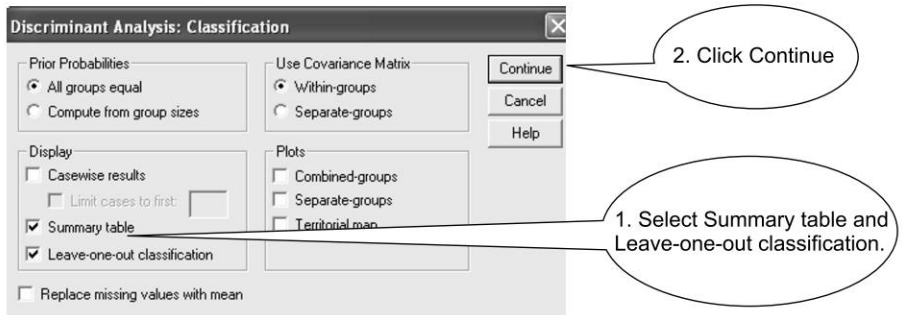
After selecting the descriptives SPSS will go back to the previous box shown below:

SPSS Snapshot DA 5



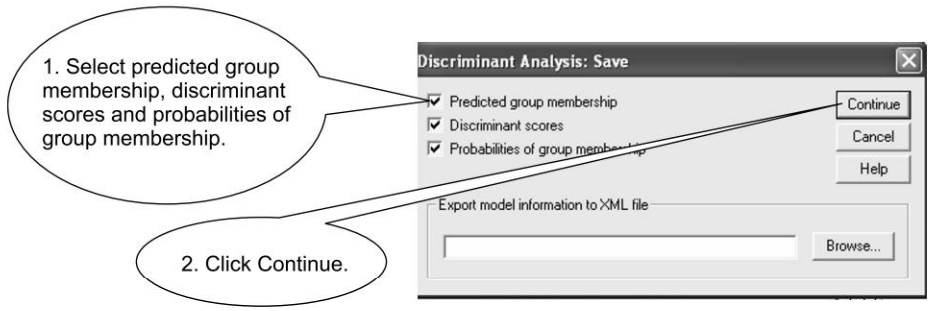
After clicking on classify SPSS will open a box as shown below:

SPSS Snapshot DA 6



After clicking 'Continue', SPSS will be back to the previous box as shown in the Snapshot DA 6, then click on the Save button at the bottom. SPSS will open a box as shown below:

SPSS Snapshot DA 7



After clicking 'Continue', SPSS will again go back to the previous window shown in Snapshot DA 6, at this stage one may click OK button. This will lead SPSS to analyse the data and the output will be displayed in the output view of SPSS.

Output for Enter Method

We will discuss interpretation for each output.

Discriminant

Analysis Case Processing Summary

Unweighted Cases		*N*	*Per cent*
Valid		700	82.4
Excluded	Missing or out-of-range group codes	150	17.6
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Total	150	176
Total		850	100.0

This table gives case processing summary, i.e. how may valid cases were selected, how many were excluded (due to missing data), total and their respective percentages.

Group Statistics

Previously defaulted		*Mean*	*Std. Deviation*	*Valid N (listwise)*	
				Unweighted	*Weighted*
No	Age in years	35.5145	7.70774	517	517.000
	Years with current employer	9.5087	6.66374	517	517.000
	Years at current address	8.9458	7.00062	517	517.000
	Household income in thousands	47.1547	34.22015	517	517.000
	Debt to income ratio (x100)	8.6793	5.61520	517	517.000
	Credit card debt in thousands	1.2455	1.42231	517	517.000
	Other debt in thousands	2.7734	2.81394	517	517.000
No	Age in years	33.0109	8.51759	183	183.000
	Years with current employer	5.2240	5.54295	183	183.000
	Years at current address	6.3934	5.92521	183	183.000
	Household income in thousands	41.2131	43.11553	183	183.000
	Debt to income ratio (x100)	14.7279	7.90280	183	183.000
	Credit card debt in thousands	2.4239	3.23252	183	183.000

(*Contd*)

(*Contd*)

	Other debt in thousands	3.8628	4.26368	183	183.000
Total	Age in years	34.8600	7.99734	700	700.000
	Years with current employer	8.3886	6.65804	700	700.000
	Years at current address	8.2786	6.82488	700	700.000
	Household income in thousands	45.6014	36.81423	700	700.000
	Debt to income ratio (x100)	10.2606	6.82723	700	700.000
	Credit card debt in thousands	1.5536	2.11720	700	700.000
	Other debt in thousands	3.0582	3.28755	700	700.000

This table gives the group statistics of independent variables, for each categories (here yes and no) of dependent variables.

Tests of Equality of Group Means

	Wilks' Lambda	*F*	*df1*	*df2*	*sig*
Age in years	0.981	13.482	1	698	0.000
Years with current employer	0.920	60.759	1	698	0.000
Years at current address	0.973	19.402	1	698	0.000
Household income in thousands	0.995	3.533	1	698	0.061
Debt to income ratio (x100)	0.848	124.889	1	698	0.000
Credit card debt in thousands	0.940	44.472	1	698	0.000
Other debt in thousands	0.979	15.142	1	698	0.000

This table gives the test for Wilks' Λ for each independent variable if this is significant (<0.05 or 0.01), it means that the respective variable, mean is different for the two groups (in this case previously defaulted and previously not defaulted). Any insignificant value will indicate that the variable is not different for different group or in other terms does not discriminate the dependent variable. In the above example, all the variables are significant except household income in thousands. This implies that the default of the loan does not depend on the household income.

Pooled Within-Groups Matrices

		Age in years	*Years with current employer*	*Years at current address*	*Household income in thousands*	*Debt to income ratio (x100)*	*Credit card debt in thousands*	*Other debt in thousands*
Correlation	Age in years	1.000	.524	.588	.475	.077	.342	.368
	Years with current employer	.524	1.000	.292	.627	.089	.509	.471
	Years at current address	.588	.292	1.000	.310	.083	.260	.257
	Household income in thousands	.475	.627	.310	1.000	.001	.608	.629
	Debt to income ratio (x100)	.077	.089	.083	.001	1.000	.455	.580
	Credit card debt in thousands	.342	.509	.260	.608	.455	1.000	.623
	Other debt in thousands	.368	.471	.257	.629	.580	.623	1.000

Analysis I

Box's Test of Equality of Covariance Matrices

Log Determinants

Previously defaulted	*Rank*	*Log Determinant*
No	7	21.292
Yes	7	24.046
Pooled within-groups	7	22.817

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		563.291
F	Approx.	19.819
	df1	28
	df2	431743.0
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

This table indicates that the Box's M is significant which means the assumption of equality of variance may not be true. This is a caution for interpreting results.

Summary of Canonical Discriminant Functions

Eigenvalues

Function	*Eigen value*	*% of Variance*	*Cumulative %*	*Canonical Correlation*
1	0.404[a]	100.0	100.0	0.536

a. First 1 canonical discriminant functions were used in the analysis.

This table gives summary of canonical discriminant function. This indicates eigenvalue for this model is 0.404 and canonical correlation is 0.536. Since there is single discriminant function all the explained variation is contributed by the function. 53.60% of variation in the dependent variable is explained by the model.

Wilks' Lambda

Test of Function(s)	*Wilks' Lambda*	*Chi-square*	*df*	*Sig.*
1	0.712	235.447	7	0.000

This table tests the significance of the model. As seen in the Sig. column, the model is significant.

Standardized Canonical Discriminant Function Coefficients

	Function 1
Age in years	.122
Years with current employer	–.829
Years at current address	–.310
Household income in thousands	.215
Debt to income ratio (×100)	.603
Credit card debt in thousands	.564
Other debt in thousands	–.178

Structure Matrix

	Function 1
Debt to income ratio (×100)	0.666
Years with current employer	–0.464
Credit card debt in thousands	0.397
Years at current address	–0.262
Other debt in thousands	–0.232
Age in years	–0.219
Household income in thousands	–0.112

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.

This table gives simple correlations between the independent variables and the discriminant function. High correlation will get translated to high discriminating power.

Canonical Discriminant Function Coefficients

	Function 1
Age in years	0.015
Years with current employer	–0.130
Years at current address	–0.046
Household income in thousands	0.006
Debt to income ratio (×100)	0.096
Credit card debt in thousands	0.275
Other debt in thousands	–.0.055
(Constant)	–0.576

Unstandardized Coefficients

This table gives the canonical correlations. Negative sign indicates inverse relation. For example, years at current employer is -0.130; it means that more the number of years spent at the current employer, lesser the chance that the person will default.

Functions at Group Centroids

	Function
Previously defaulted	*1*
No	–0.377
Yes	1.066

Unstandardized canonical discriminant functions evaluated at group means

Classification Statistics

Classification Processing Summary

Processed		850
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		850

Prior Probabilities for Groups

Previously defaulted	*Prior*	*Cases Used in Analysis*	
		Unweighted	*Weighted*
No	0.500	517	517.000
Yes	0.500	183	183.000
Total	1.000	700	700.000

Classification Function Coefficients

	Previously Defaulted	
	No	*Yes*
Age in years	0.803	0.825
Years with current employer	–0.102	–0.289
Years at current address	–0.294	–0.360
Household income in thousands	0.073	0.081
Debt to income ratio (×100)	0.639	0.777
Credit card debt in thousands	–1.004	–0.608
Other debt in thousands	–1.044	–1.124
(Constant)	–15.569	–16.898

Fisher's linear discriminant functions

The classification functions are used to assign cases to groups.

There is a separate function for each group. For each case, a classification score is computed for each function. The discriminant model assigns the case to the group whose classification function obtained the highest score. The coefficients for Years with current employer and Years at current address are smaller for the Yes classification function, which means that customers who have lived at the same address and worked at the same company for many years are less likely to default. Similarly, customers with greater debt are more likely to default.

Classification Results (b, c)

			Predicted Group Membership		
		Previously defaulted	*No*	*Yes*	*Total*
Original	Count	No	393	124	517
		Yes	44	139	183
		Ungrouped cases	99	51	150
	%	No	76.0	24.0	100.0
		Yes	24.0	76.0	100.0
		Ungrouped cases	66.0	34.0	100.0
Cross-validated(s)	Count	No	39.1	126	517
		Yes	47	136	183
	%	No	75.6	24.4	100.00
		Yes	25.7	74.3	100.0

a Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b 76.0% of original grouped cases correctly classified.

c 75.3% of cross-validated grouped cases correctly classified.

This is classification matrix or confusion matrix. This gives the percentage of cases that are classified correctly i.e. the hit ratio. This hit ratio should be at least 25% more than the random probability.

In the above example, 532 of the 700 cases are classified correctly. Overall, 76% of the cases are classified correctly. 139 out of 263 defaulters were identified correctly.

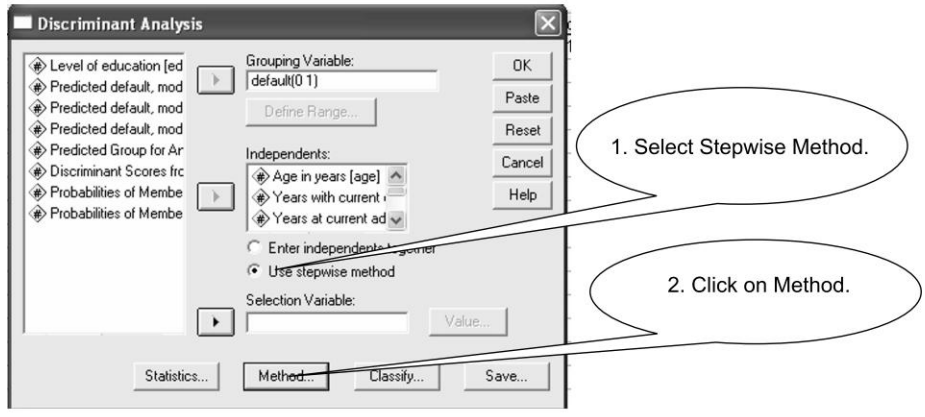
Output for Stepwise Method

All the steps of the 'Enter' method are very similar to stepwise method except that the method to be selected is stepwise. We have shown it in SPSS Snapshot **DA3.**

If one selects the stepwise method, one also needs to select the method which SPSS should use to select best set of independent variables.

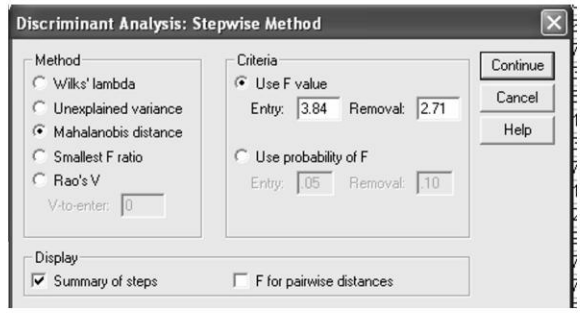
This can be selected by clicking method button from Snapshot **DA 8** shown in the following:

SPSS Snapshot DA 8



After clicking on method, SPSS will open a window as shown in the following:

SPSS Snapshot DA 9



There are five methods available in SPSS namely,

- Wilk's Λ
- Unexplained variance
- Mahalanobis distance
- Smallest 'F' ratio
- Rao's V

One may select any one of the methods. We have selected Mahalanobis distance method. We will now discuss the output using this method. It may be noted that we will discuss only the output that is different than the previous method.

Stepwise Statistics

Variables Entered/Removed a,b,c,d

Step	Entered	Min. D Squared Statistics	Between Groups	Exact F Statistic	df1	df2	df3
1	Debt to income ratio (×100)	0.924	No and Yes	124.889	1	698.000	0.000
2	Years with current employer	1.501	No and Yes	101.287	2	697.000	0.000
3	Credit card debt in thousands	1.926	No and Yes	86.502	3	696.000	0.000
4	Years at current address	2.038	No and Yes	68.572	4	695.000	0.000

At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.
(a) Maximum number of steps is 14.
(b) Minimum partial F to enter is 3.84.
(c) Maximum partial F to remove is 2.71.
(d) F level, tolerance, or VIN insufficient for further computation.

Variables in the Analysis

Step		Tolerance	F to Remove	Min. D Squared	Between Groups
1	Debt to income ratio (×100)	1.000	124.889		
2	Debt to income ratio (×100)	0.992	130.539	0.450	No and Yes
	Years with current employer	0.992	66.047	0.924	No and Yes
3	Debt to income ratio (×100)	0.766	35.888	1.578	No and Yes
	Years with current employer	0.716	111.390	0.947	No and Yes
	Credit card debt in thousands	0.572	44.336	1.501	No and Yes
4	Debt to income ratio (×100)	0.766	35.000	1.693	No and Yes
	Years with current employer	0.691	89.979	1.213	No and Yes
	Credit card debt in thousands	0.564	48.847	1.565	No and Yes
	Years at current address	0.898	11.039	1.926	No and Yes

Wilks' Lambda

Step	Number of Variables	Lambda	df1	df2	df3	Exact F Statistic	df1	df2	Sig.
1	1	0.848	1	1	698	124.889	1	698.000	0.000
2	2	0.775	2	1	698	101.287	2	697.000	0.000
3	3	0.728	3	1	698	86.502	3	696.000	0.000
4	4	0.717	4	1	698	68.572	4	695.000	0.000

These tables give summary of variables that are in analysis variables that are not in the analysis and the model at each step, its significance.

It can be concluded that variables Debt to income ratio (x100), Years with current employer, Credit card debt in thousands, Years at current address remain in the model and others are removed from the model. This means that only these variables contribute in the model.

14.4 LOGISTIC REGRESSION/MULTIPLE LOGISTIC REGRESSION

As discussed earlier, in a regression equation $y = a + bx$, both the dependent variable, y, and independent variable, x, are assumed to be normally distributed. As such, both x and y are continuous variables, and can take any value from $-\infty$ to $+\infty$. However, suppose, there is a situation when the distribution of variable y is binomial, which takes only one of the two possible values, say 0 and 1. In such a case, the regression equation will not be valid, as for a given value of x, y can assume any value, and not only either of the two values. This limitation of regression equation led to the development of **logistic regression** where the dependent variable takes only two values, say 0 and 1. As discussed above, in Discriminant Analysis also, the dependent variable takes only two 'Yes' and 'No' type possible values. However, therein, the limitation on the independent variable x is that it has to be a continuous variable. Logistic regression is the response to this limitation, also wherein there is no such constraint. Thus, logistic regression is preferred and used in both the following situations:

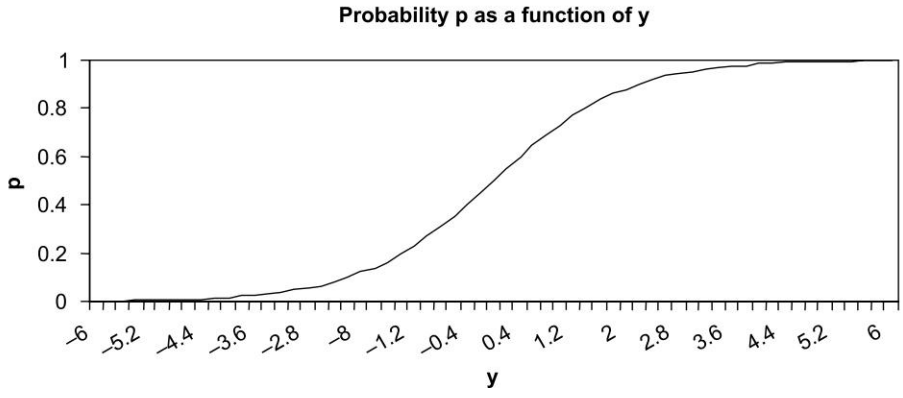
- when the dependent variable is dichotomous/binary/categorical
- when the independent variable is continuous or categorical. If there are more than one independent variable, these could be either continuous or dichotomous or a combination of continuous and dichotomous variables.

In logistic regression, the relationship between dependent variable and independent variable is not linear. It is of the type

$$p = \frac{1}{1 + e^{-y}}$$

where, p is the probability of 'success' i.e. dichotomous variable y taking the value 1, and (1 – p) is the probability of 'failure' i.e., y taking the value 0, and y = a + bx.

The graph of this relationship between p and y is depicted below:



The logistic equation (8) can be reduced to a linear form by converting the probability p into log of (p)/(1 – p)p or **logit** as follows:

$$y = \log [(p)/(1 - p))] = a + bx \quad (2)$$

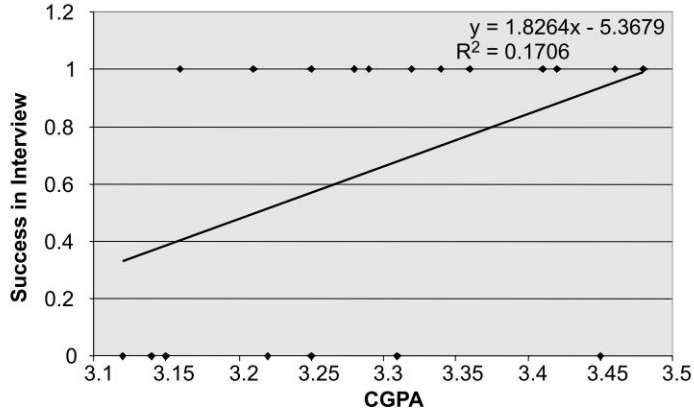
The logarithm, here, is the natural logarithm to the base 'e'. The logarithm of any number to this base is obtained by multiplying the logarithm to the base 10 by log of 10 to the base 'e' i.e. 2.303.

The equation (8) is similar to a regression equation. However, here, a unit change in the independent variable causes a change in the dependent variable, expressed as logit rather than the dependent variable p, directly. Such regression analysis is known as Logistic Regression.

The fitting of a logistic regression equation is explained through an illustration wherein data was recorded on the CGPA (up to first semester in the second year of MBA) of 20 MBA students, and their success in the first interview for placement. The data collected was as follows where Pass is indicated as "1" while Fail is indicated as "0".

Student (Srl. No.)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
CGPA	3.12	3.21	3.15	3.45	3.14	3.25	3.16	3.28	3.22	3.41
Result of First Interview	0	1	0	0	0	1	1	1	0	1
Student (Srl. No.)	11	12	13	14	15	16	17	18	19	20
CGPA	3.48	3.34	3.25	3.46	3.32	3.29	3.42	3.28	3.36	3.31
Result of First Interview	1	1	0	1	1	1	1	1	1	0

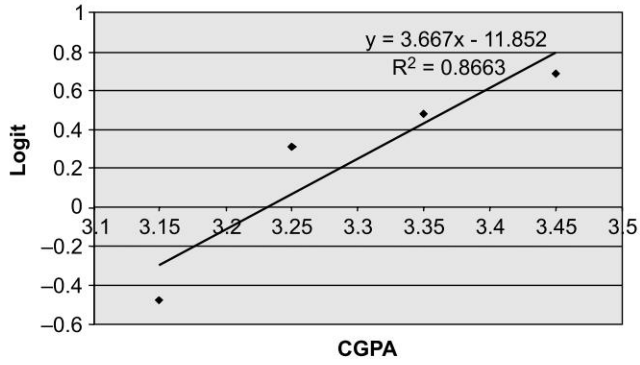
Now, given this data, can we find the probability of a student succeeding in the first interview given the CGPA?



$$y = 1.83\ x - 5.37$$

Instead, let us now attempt to fit a logistic regression to the student data. We will do this by computing the logits and then fitting a linear model to the logits. To compute the logits, we will regroup the data by CGPA into intervals, using the midpoint of each interval for the independent variable. We calculate the probability of success based on the number of students that passed the interview for each range of CGPAs. This results in the following data:

Class Interval (CGPA)	*Middle Point of Class Interval*	*Probability of Success*	*Logit {p/(1 – p)}*
3.1–3.2	3.15	1/4 = 0.25	– 0.477
3.2–3.3	3.25	4/6 = 0.67	0.308
3.3–3.4	3.35	3/4 = 0.75	0.477
3.4–3.5	3.45	5/6 = 0.83	0.689



We plot the Logit against the CGPA and then look for the linear fit which gives us the equation: y = 3.667 x – 11.852

Thus, if p is the probability of passing the interview and x is the CGPA, the logistic regression can be expressed as:

$$\ln\left(\frac{p}{1-p}\right) = 3.667x - 11.852$$

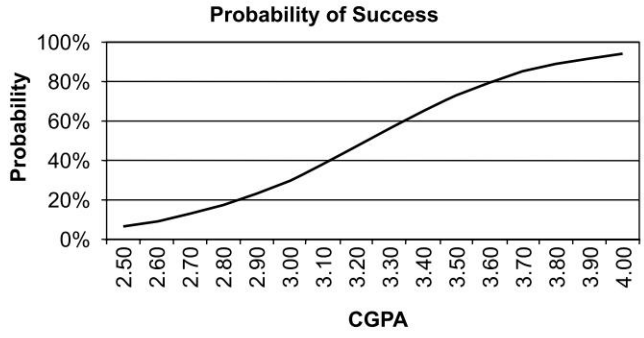
Converting the logarithm to an equivalent exponential form, this equation can also be expressed as expressed as:

$$p = \frac{e^{3.677x - 11.852}}{1 + e^{3.667x - 11.852}}$$

x	2.5	2.6	2.7	2.8	2.9	3	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4
y*	–2.68	–2.32	–1.95	–1.58	–1.22	–0.85	–0.48	–0.12	0.25	0.62	0.98	1.35	1.72	2.08	2.45	2.82
p	6%	9%	12%	17%	23%	30%	38%	47%	56%	65%	73%	79%	85%	89%	92%	94%

*Upto 2 decimal places

This can be displayed graphically as follows:



From this regression model, we can see that probability of success at the interview is below 25% for CGPAs below 2.90 but is above 75% for CGPAs above 3.60.

While one could apply logistic regression to a number of situations, it has been found useful particularly in the following situations:

- Credit – Study of creditworthiness of an individual or a company. Various demographic and credit history variables could be used to predict if an individual will turn out to be 'good' or 'bad' customers.
- Marketing/Market Segmentation – Study of purchasing behaviour of consumers. Various demographic and purchasing information could be used to predict if an individual will purchase an item or not.
- Customer loyalty – The analysis could be done to identify loyal or repeat customers using various demographic and purchasing information.
- Medical – Study of risk of diseases/body disorder.

14.4.1 Assumptions of Logistic Regressions

The multiple regression assumes assumptions like linearity, normality etc. These are not required for logistic regression. Discriminant Analysis requires the independent variables to be metric, which is not necessary for logistic regression. This makes the technique to be superior to discriminant analysis. The only care to be taken is that there are no extreme observations in the data.

14.4.2 Key Terms of Logistic Regressions

Following are the key terms used in logistic regression:

Factor	The independent variable in logistic regression is termed as factor. The factor is dichotomous in nature, and is usually converted into a dummy variable.
Covariate	The independent variable that is metric in nature is termed as covariate.
Maximum Likelihood Estimation	This method is used in logistic regression to predict the odd ratio for the dependent variable. In least square estimate, the square of error is minimised, but in maximum likelihood estimation, the log likelihood is maximised.
Significance Test	Hosmer and Lemeshow chi-square test is used to test the overall model of goodness-of-fit test. It is the modified chi-square test, which is better than the traditional chi-square test.

(*Contd*)

(*Contd*)

Stepwise logistic regression	In stepwise logistic regression, the three methods available are enter, backward and forward. In enter method, all variables are included in logistic regression, irrespective the variable is significant or insignificant. In backward method, the model starts with all variables and removes nonsignificant variables from the list. In forward method, logistic regression starts with single variable and adds one by one variable and tests significance and removes insignificant variables from the model.
Measures of Effect Size	In logistic regression, R^2 is no more accepted because R^2 tells us the variance extraction by the independent variable. The maximum value of the Cox and Snell r-squared statistic is actually somewhat less than 1; the Nagelkerke r-squared statistic is a "correction" of the Cox and Snell statistic so that its maximum value is 1.
Classification Table	The classification table shows the practical results of using the logistic regression model. It is useful to understand validity of the model.

14.4.3 Logistic Regressions Using SPSS

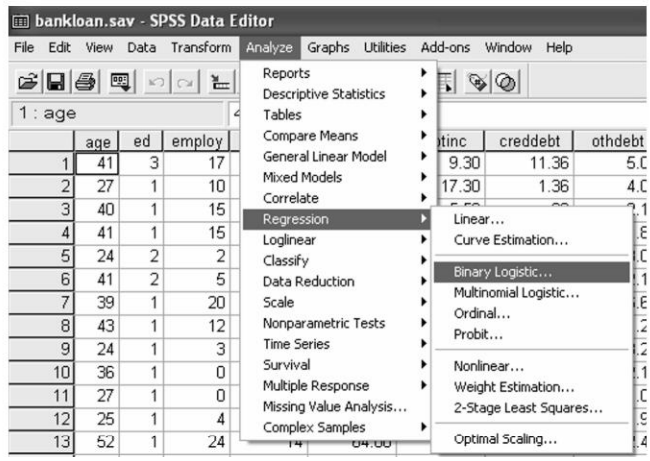
For illustration, the file bankloan.sav that was used in the Section 14.3 dealing with discriminant analysis has been used. This is a data file that concerns a bank's efforts to reduce the rate of loan defaults. The file contains financial and demographic information on 850 past and prospective customers. The first 700 cases are customers who were previously given loans. The last 150 cases are prospective customers that the bank needs to classify as good or bad credit risks. This file is part of SPSS cases and is in the tutorial folder of SPSS. Within tutorial folder, this file is in the sample files folder. For the convenience of readers we have provided this file in the CD with the book.

As in case with multiple regression, discriminant analysis can be done by two principle methods, namely, enter method and stepwise method.

We will illustrate output for both methods.

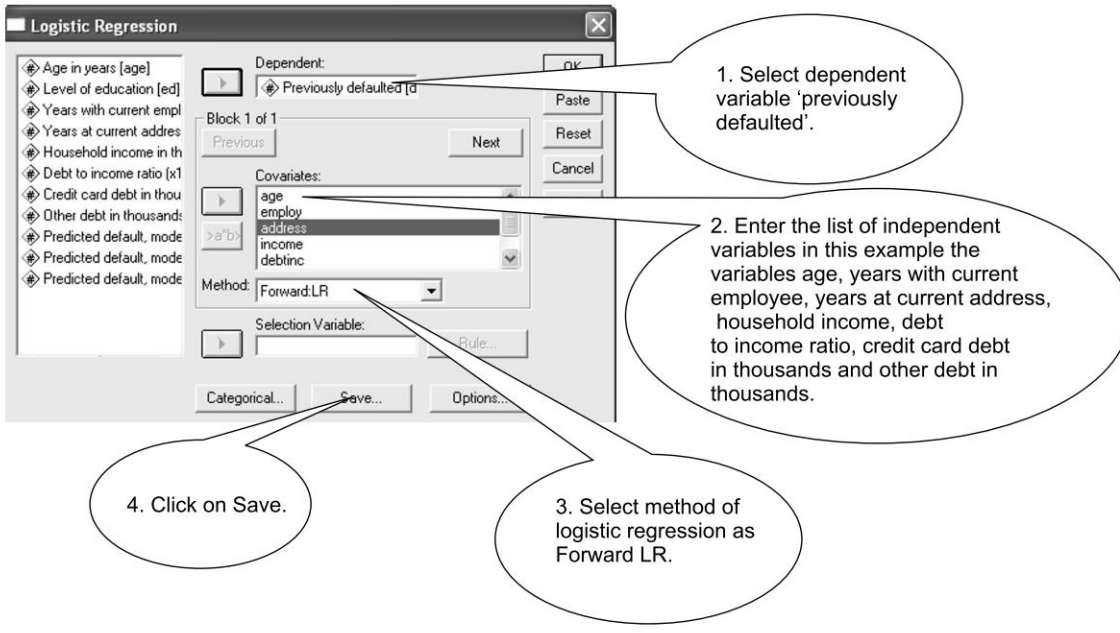
After opening the file bankloan.sav, one can click on Analyze and classify as shown in the following snapshot:

LR Snapshot 1



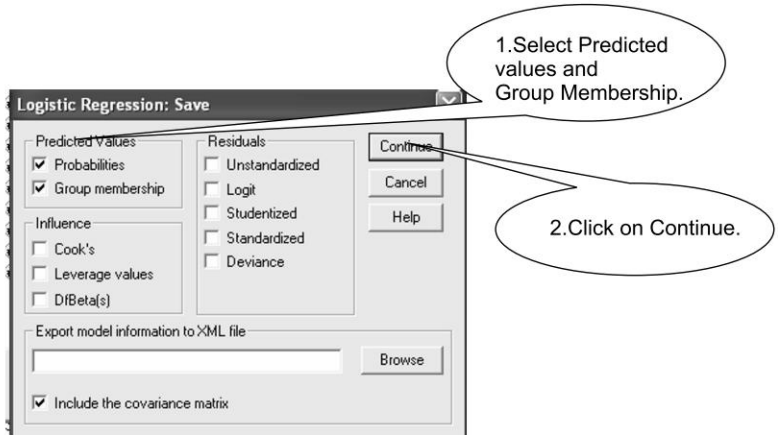
SPSS will open the following window:

LR Snapshot 2



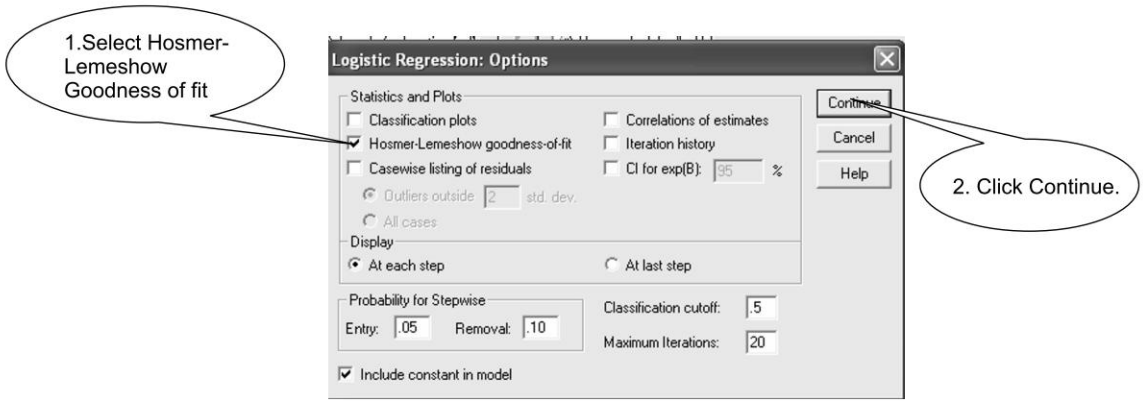
SPSS will open the following window:

LR Snapshot 3



SPSS will take back to the window as displayed in LR Snapshot 2, at this stage click on Options. The following window will be opened.

LR Snapshot 4



SPSS will be back to the window as shown in LA Snapshot 2. At this stage, click OK. Following output will be displayed.

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		*N*	*Percent*
Selected Cases	Included in Analysis	700	82.4
	Missing Cases	150	17.6
	Total	850	100.0
Unselected Cases		0	0.0
Total		850	100.0

a. If weight is in effect, see classification table for the total number of cases.

This table indicates the case processing summary 700 out of 850 cases are used for the analysis 150 are ignored as these have missing values.

Dependent Variable Encoding

Original Value	*Internal Value*
No	0
Yes	1

This table indicates the coding for the dependent variable 0=>not defaulted, 1=>defaulted

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			Previously defaulted		Percentage Correct
			No	Yes	
Step 0	Previously defaulted	No	517	0	100.0
		Yes	183	0	0.0
	Overall Percentage				73.9

a. Constant is included in the model.
b. The cut value is 0.500

Variables not in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step0	Constant	–1.039	0.086	145.782	1	0.000	0.354

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	age	13.265	1	0.000
		employ	56.054	1	0.000
		address	18.931	1	0.000
		income	3.526	1	0.060
		debtinc	106.238	1	0.000
		creddebt	41.928	1	0.000
		othdebt	14.863	1	0.000
Overall Statistics			201.271	7	0.000

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	102.935	1	0.000
	Block	102.935	1	0.000
	Model	102.935	1	0.000
Step 2	Step	70.346	1	0.000
	Block	173.282	2	0.000
	Model	173.282	2	0.000
Step 3	Step	55.446	1	0.000
	Block	228.728	3	0.000
	Model	228.728	3	0.000
Step 4	Step	18.905	1	0.000
	Block	247.633	4	0.000
	Model	247.633	4	0.000

Model Summary

Step	*–2 Log likelihood*	*Cox & Snell R Square*	*Nagelkerke R Square*
1	701.429[a]	0.137	0.200
2	631.083[b]	0.219	0.321
3	575.636[b]	0.279	0.408
4	556.732[c]	0.298	0.436

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than 0.001.
b. Estimation terminated at iteration number 5 because parameter estimates changed by less than 0.001.
c. Estimation terminated at iteration number 6 because parameter estimates changed by less than 0.001.

Hosmer and Lemeshow Test

Step	*Chi-square*	*off*	*Sig.*
1	3.160	8	0.924
2	4.158	8	0.843
3	6.418	8	0.600
4	8.556	8	0.381

The Hosmer–Lemeshow statistic indicates a poor fit if the significance value is less than 0.05. Here since the value is above 0.05, the model adequately fits the data.

Classification Table[a]

			Predicted		
			Previously defaulted		*Percentage Correct*
Observed			*No*	*Yes*	
Step 1	Previously	No	490	27	94.8
	defaulted	Yes	137	46	25.1
	Overall Percentage				76.6
Step 2	Previously	No	481	36	93.0
	defaulted	Yes	110	73	39.9
	Overall Percentage				79.1
Step 3	Previously	No	477	40	92.3
	defaulted	Yes	99	84	45.9
	Overall Percentage				80.1
Step 3	Previously	No	478	39	92.5
	defaulted	Yes	91	92	50.3
	Overall Percentage				81.4

a. The cut value is .500

This table is the classification table. It indicates the number of cases correctly classified as well as incorrectly classified. Diagonal elements represent correctly classified cases and non-diagonal elements represent incorrectly classified cases.

It may be noted that for each step, the number of correctly classified cases are improved than in the previous step. The last column gives the percentage of correctly classified cases, which is improved at each step.

Variables in the Equation

		B	*S.E.*	*Wald*	*df*	*Sig.*	*Exp(B)*
Step	debtnic	0.132	0.014	85.377	1	0.000	1.141
1[a]	Constant	–2.531	0.195	168.524	1	0.000	0.080
Step	employ	–0.141	0.019	53.755	1	0.000	0.868
2[b]	debtinc	0.145	0.016	87.231	1	0.000	1.156
	Constant	–1.693	0.219	59.771	1	0.000	0.184
Step	employ	–.244	0.027	80.262 1		0.000	0.783
3[c]	debtinc	0.088	0.018	23.328	1	0.000	1.092
	creddebt	0.503	0.081	38.652	1	0.000	1.653
	Constant	–1.227	0.231	28.144	1	0.000	0.293
Step	employ	–0.243	0.028	74.761	1	0.000	0.785
4[d]	address		0.020	17.183	1	0.000	0.922
	debtinc	0.088	0.019	22.659	1	0.000	1.092
	credebt	0.573	0.087	43.109	1	0.000	1.774
	Constant	–0.791	0.252	9.890	1	0.002	0.453

a. Variable(s) entered on step 1: debtinc.
b. Variable(s) entered on step 2: employ.
c. Variable(s) entered on step 3: creddebt.
d. Variable(s) entered on step 4: address.

The best model is usually the last model i.e. step 4. It contains variables: years to current employee, years at current address, debt to income ratio and credit card debt. All other variables are insignificant in the model.

Model if Term Removed

	Variable	*Model Log Likelihood*	*Change in –2 Log Likelihood*	*df*	*Sig. of the Change*
Step 1	debtinc	–402.182	102.935	1	0.000
Step 2	employ	–350.714	70.346	1	0.000
	debtinc	–369.708	108.332	1	0.000
Step 3	employ	–349.577	123.518	1	0.000
	debtinc	–299.710	23.783	1	0.000
	creddebt	–315.541	55.446	1	0.000
Step 4	employ	–333.611	110.490	1	0.000
	address	287.818	18.905	1	0.000
	debtinc	–290.006	23.281	1	0.000
	credebt	–311.176	65.621	1	0.000

Variables not in the Equation

			Score	*df*	*Sig.*
Step 1	Variables	age	16.478	1	0.000
		employ	60.934	1	0.000
		address	23.474	1	0.000
		income	3.219	1	0.073
		creddebt	2.261	1	0.133
		othdebt	6.631	1	0.010
	Overall Statistics		113.910	6	0.000
Step 2	Variables	age	0.006	1	0.939
		address	8.407	1	0.004
		income	21.437	1	0.000
		credebt	64.958	1	0.000
		othdebt	4.503	1	0.034
	Overall Statistics		84.064	5	0.000
Step 3	Variables	age	0.635	1	0.426
		address	17.851	1	0.000
		income	0.773	1	0.379
		othdebt	0.006	1	0.940
	Overall Statistics		22.221	4	0.000
Step 4	Variables	age	3.632	1	0.057
		income	0.012	1	0.912
		othdebt	0.320	1	0.572
	Overall Statistics		4.640	3	0.200

The above table gives the scores which can be used to predict whether the person having certain values of variable will default or not. In fact, the scores can be used to find the probability of default.

14.5 MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

In ANOVA, we study the impact of one or more factors on a single variable. For example, we could study the differences in yields of rice due to say, 3 fertilisers. However, if we wish to study the impact of fertilisers on more than one variable, say in yields as also on harvest times of rice crop, then MANOVA could be used to test the null hypothesis that the

(i) yields due to the use of three fertilisers are equal and
(ii) harvest times due to the use of these fertilisers are equal.

Another example could be to assess the impacts of two training programmes, conducted for a group of employees, on their knowledge as well as motivation relevant for their job. While one programme was mostly based on 'Classroom' training, the other was mostly based on the 'On Job' training.

The data collected could be as follows:

	Classroom		*No Training*		*Job Based*	
	K	*M*	*K*	*M*	*K*	*M*
1	92	98	70	75	83	90
2	88	77	56	66	65	76
3	91	88	89	90	93	91
4	85	82	87	84	90	85
5	88	85	72	71	77	73
6	81	82	74	71	89	81
7	92	83	75	75	84	78
8	88	90	80	68	85	76
9	80	79	78	65	73	80
10	84	87	72	75	83	81

In this case, one of the conclusions drawn was that both the programmes had positive impact on both knowledge and motivation but there was no significant difference between classrooms based and job based training programmes.

As another example, we could assess whether a change in Compensation System-1 to Compensation System-2 has brought about changes in sales, profit and job satisfaction in an organisation.

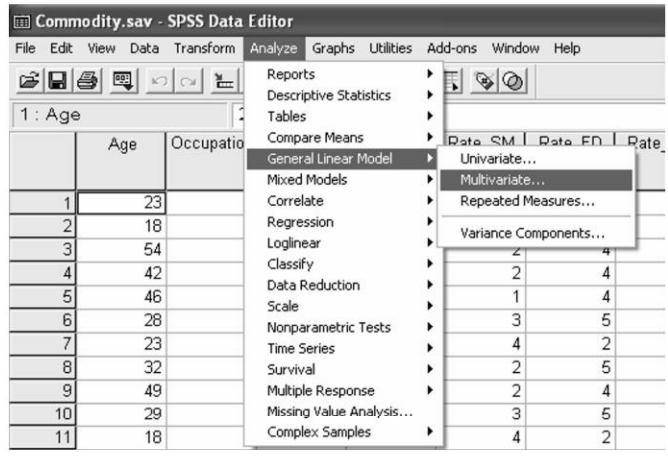
MANOVA is typically used when there are more than one dependent variables, and independent variables are qualitative/categorical.

14.5.1 MANOVA Using SPSS

We will use the case data on commodity market perceptions displayed at the end of this chapter.

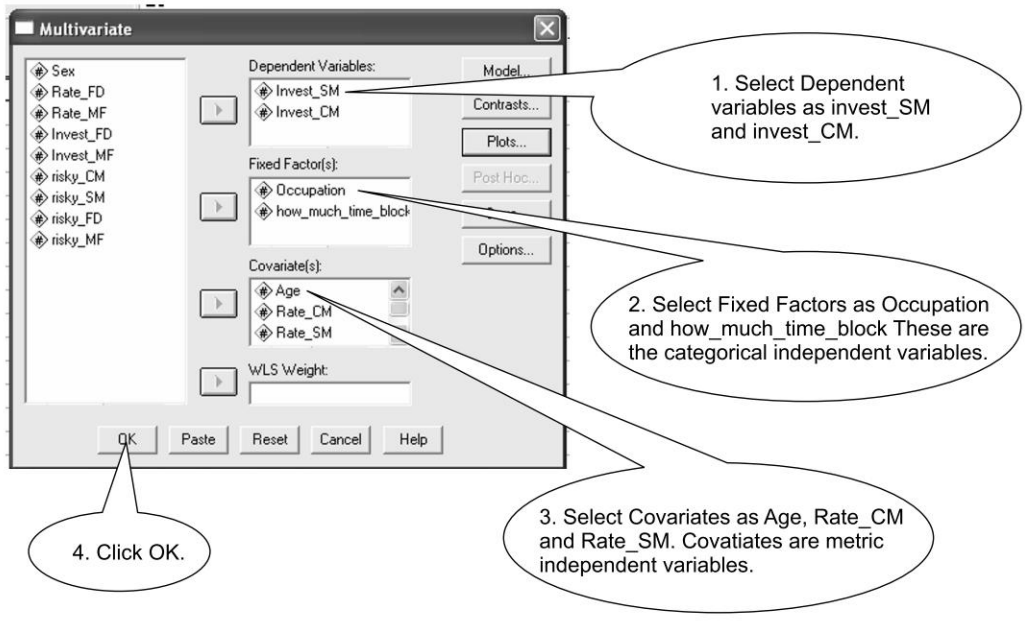
- Open the file Commodity.sav
- Select from the menu Analyze – General Linear Model – Multivariate as shown below:

MANOVA Snapshot 1



The following window will be displayed:

MANOVA Snapshot 2



It may be noted that the above example is of MANOCOVA as we have selected some categorical variables and some metric variables.

We are assuming in the above example that the dependent variables are the investments in commodity market and in share market and the categorical independent variables are occupation and how long the respondents block investments. The metric independent variables are age, respondent's rating for commodity market and share market. Here, we assume that their investments depend on their ratings, occupation, age and how long they block their investments.

The following output will be displayed:

General Linear Model

Between-Subjects Factors

		Value Label	*N*
Occupation	1	"Self Employed"	11
	2	Govt	15
	3	Student	4
	4	Housewife	14
how_much_time_	1	<6 months	6
block_your_money	2	6 to 12 months	8

(*Contd*)

(*Contd*)

3	1 to 3 years	5
4	> 3 years	10
5		4
6		6
7		3
8		2

This table gives summary of number of cases for the factors.

Multivariate Tests[c]

Effect		*Value*	*F*	*Hypothesis df*	*Error df*	*Sig.*
Intercept	Pillai's Trace	0.139	1.537[a]	2.000	19.000	0.241
	Wilks' Lambda	0.861	1.537[a]	2.000	19.000	0.241
	Hotelling's Trace	0.162	1.537[a]	2.000	19.000	0.241
	Roy's Largest Root	0.162	1.537[a]	2.000	19.000	0.241
Age	Pillai's Trace	0.157	1.770[a]	2.000	19.000	0.197
	Wilks' Lambda	0.843	1.770[a]	2.000	19.000	0.197
	Hotelling's Trace	0.186	1.770[a]	2.000	19.000	0.197
	Roy's Largest Root	0.186	1.770[a]	2.000	19.000	0.197
Rate_CM	Pillai's Trace	0.096	1.011[a]	2.000	19.000	0.383
	Wilks' Lambda	0.904	1.011[a]	2.000	19.000	0.383
	Hotelling's Trace	0.106	1.011[a]	2.000	19.000	0.383
	Roy's Largest Root	0.106	1.011[a]	2.000	19.000	0.383
Rate_SM	Pillai's Trace	0.027	0.268[a]	2.000	19.000	0.768
	Wilks' Lambda	0.973	0.268[a]	2.000	19.000	0.768
	Hotelling's Trace	0.028	0.268[a]	2.000	19.000	0.768
	Roy's Largest Root	0.028	0.268[a]	2.000	19.000	0.768
Occupation	Pillai's Trace	0.908	5.547	6.000	40.000	0.000
	Wilks' Lambda	0.250	6.333[a]	6.000	38.000	0.000
	Hotelling's Trace	2.366	7.099	6.000	36.000	0.000
	Roy's Largest Root	2.059	13.725[b]	3.000	20.000	0.000
how_much_time_block_your_money	Pillai's Trace	0.855	2.132	14.000	40.000	0.031
	Wilks' Lambda	0.318	2.102[a]	14.000	38.000	0.035
	Hotelling's Trace	1.607	2.066	14.000	36.000	0.040
	Roy's Largest Root	1.125	3.214[b]	7.000	20.000	0.019
Occupation * how_much_time_block_your_money	Pillai's Trace	0.823	1.399	20.000	40.000	0.180
	Wilks' Lambda	0.335	1.382[a]	20.000	38.000	0.191
	Hotelling's Trace	1.511	1.360	20.000	36.000	0.206
	Roy's Largest Root	1.069	2.137[b]	10.000	20.000	0.071

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+Age+Rate_CM+Rate_SM+Occupation+how_much_time_block_your_money+Occupation *how_much_time_block_your_money

This table indicates the null hypothesis that the investments are equal for all occupations is rejected since the significance value (p-value) is less than 0.05 as indicated by circles. Thus, we may conclude at 5% Level of Significance (LOS) that there is significant difference in the both investments (share market and commodity markets) and occupation of the respondents.

The null hypothesis that the investments are equal for different levels of time the investment blocked, is rejected since the significance value (p-value) is less than 0.05 as indicated by circles. Thus, we may conclude at 5% LOS that there is significant difference in the both investments (share market and commodity markets) and the period for which the respondents would likely to block their money.

The other hypothesis about age, ratings of CM and ratings of SM are not rejected (as p-value is greater than 0.05) this means there is no significant difference in the investments for these variables.

Tests of Between-Subjects Effects

Source	*Dependent Variable*	*Type III Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
Corrected Model	Invest_SM	1.014E+011[a]	23	4407960763	4.250	0.001
	Invest_CM	1.193E+010[b]	23	518647980.2	2.033	0.057
Intercept	Invest_SM	64681514.2	1	64681514.21	0.062	0.805
	Invest_CM	532489379	1	532489379.0	2.087	0.164
Age	Invest_SM	313752666	1	313752666.0	0.302	0.588
	Invest_CM	472901520	1	472901520.3	1.854	0.189
Rate_CM	Invest_SM	224388812	1	224388812.0	0.216	0.647
	Invest_CM	526600356	1	526600355.7	2.064	0.166
Rate_SM	Invest_SM	528516861	1	528516860.7	0.510	0.484
	Invest_CM	76132638.6	1	76132638.56	0.298	0.591
Occupation	Invest SM	4.254E+010	3	1.418E+010	13.672	0.000
	Invest_CM	3131507520	3	1043835840	4.092	0.020
how_much_time_block_your_money	Invest_SM	1.715E+010	7	2450532038	2.362	0.062
	Invest_CM	2659656810	7	379950972.8	1.489	0.227
Occupation * how_much_time_block_your_money	Invest_SM	2.190E+010	10	2190181184	2.111	0.074
	Invest_CM	2642792207	10	264279220.7	1.036	0.450
Error	Invest_SM	2.075E+010	20	1037276941		
	Invest_CM	5102386227	20	255119311.4		
Total	Invest_SM	2.109E+011	44			
	Invest_CM	2.780E+010	44			
Corrected Total	Invest_SM	1.221E+011	43			
	Invest_SM	1.703E+010	43			

a. R Squared = 0.830 (Adjusted R Squared = 0.635)
b. R Squared = 0.700 (Adjusted R Squared = 0.356)

14.6 FACTOR ANALYSIS

Factor Analysis is an interdependence technique. In interdependence techniques, the variables are not classified as independent or dependent variable, but their interrelationship is studied. Factor analysis is the general name for two different techniques namely, Principal Component Analysis (PCA) and Common Factor Analysis (CFA).

The Factor analysis originated about a century ago when Charles Spearman propounded that the results of a wide variety of mental tests could be explained by a single underlying intelligence factor.

The factor analysis is done mainly for two reasons:

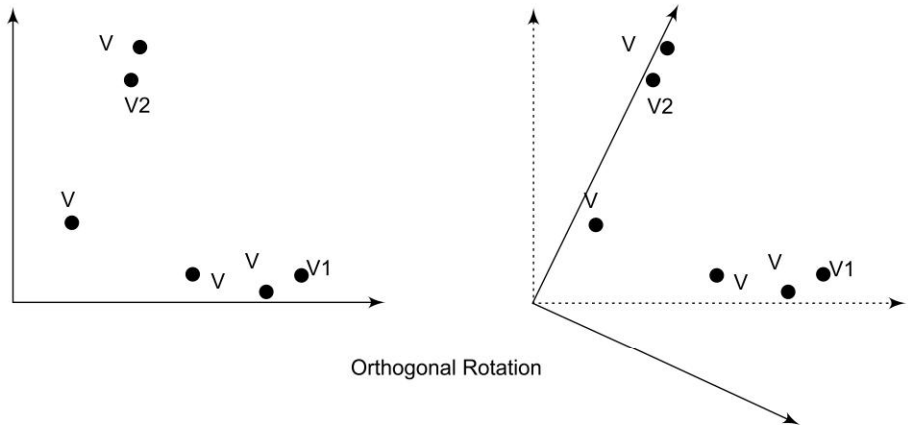
- To identify a new, smaller set of uncorrelated variables to be used in subsequent multiple regression analysis. In this situation, the Principal Component Analysis is performed on the data. PCA considers the total variance in the data while finding principle components from a given set of variables.
- To identify underlying dimensions/factors that are unobservable but explain correlations among a set of variables. In this situation, the Common Factor Analysis is performed on the data. FA considers only the common variance while finding common factors from a given set of variables. The common factor analysis is also termed as **Principal Axis Factoring**.

The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of few underlying, but unobservable, random quantities called factors. Basically, the factor model is motivated by the following argument. Suppose variables can be grouped by their correlations. That is, all variables, within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. In that case, it is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the correlations.

14.6.1 Rotation in Factor Analysis

If several factors have high loading with the same variable, it is difficult to interpret the factors clearly. This can be improved by using rotation.

Rotation does not affect communalities and percentage of total variance explained. However, the percentage of variance accounted by each factor changes. The variance explained by individual factors is redistributed by rotation.



There are basic two types of rotations viz.

- Orthogonal Rotation
- Oblique Rotation

The rotation is said to be orthogonal, if the axes maintain the right angle. This type of rotation is used when the factors are known to be non-correlated.

Varimax procedure is a type of orthogonal rotation wherein it maximises the variance of each of the factors, so that the amount of variance accounted for is redistributed over the extracted factors. This is the most popular method of rotation.

It may be noted that in the above diagram, the factors can be easily interpreted after the orthogonal rotation. Variables v2, v3 and v6 contribute to factor 1 and v1, v4 and v5 contribute to factor 2.

The rotation is said to be oblique, if the axes do not maintain right angle. This type of rotation is used when the factors are known to be correlated.

Mr Pankaj is a Director in one of the most prestigious companies in the world. His parents went from Pune to Delhi, for a visit when he was about 2 years old. After a few days after the visit, his grandfather, incidentally, enquired as to what he would like to become when he would grow old. Pankaj promptly replied "Bus Conductor". All those present on the occasion were taken aback with surprise, and asked the natural question "Why"? Prompt came the reply "Because he controls the people from entering the bus and allows only some of them to enter". After few years, when he was about 4, his grandfather repeated the question. Prompt came the reply "Jockey". When asked to explain "why"; he replied having seen a film wherein how well a jockey was controlling the horse and made him win a race. It was just a matter of chance that the grandfather repeated the question to Pankaj when he was about 7 years old. This time, the prompt reply without any hesitation was "Pilot". The justification given was that he observed, in an air show, how well a pilot controls the plane. One might infer that Pankaj in his thoughts laid lot of importance to the aspect of 'Control' that could be termed in BRM terminology as 'Construct' defined in Chapter 2. This unobserved abstract feature could be called as a 'Factor' that was responsible for the observed responses. Or, alternatively one could say that the observed responses reflected the inner desire to control. This is the essence of Factor Analysis. It attempts to find out the factors responsible for observed human responses.

14.6.2 Key Terms used in Factor Analysis

Following is the list of some key concepts used in factor analysis:

Exploratory Factor Analysis (EFA)	This technique is used when a researcher has no prior knowledge about the number of factors that the variables will be indicating. In such cases, computer-based techniques are used to indicate appropriate number of factors.
Confirmatory Factor Analysis (CFA)	This technique is used when the researcher has the prior knowledge (on the basis of some pre-established theory) about the number of factors the variables will be indicating. This makes it easy as there is no decision to be taken about the number of factors and the number is indicated in the computer based tool while conducting analysis.
Correlation Matrix	This is the matrix showing simple correlations between all possible pairs of variables. The diagonal element of this matrix is 1 and this is a symmetric matrix, since correlation between two variables *x* and y is same as between y and x.

(*Contd*)

(*Contd*)

Communality	The amount of variance, an original variable shares with all other variables included in the analysis. A relatively high communality indicates that a variable has much in common with the other variables taken as a group.
Eigen value	Eigenvalue for each factor is the total variance explained by each factor.
Factor	A linear combination of the original variables. Factor also represents the underlying dimensions (constructs) that summarises or accounts for the original set of observed variables.
Factor Loadings	The factor loadings, or component loadings in PCA, are the correlation coefficients between the variables (given in output as rows) and factors (given in output columns) These loadings are analogous to Pearson's correlation coefficient r, the squared factor loading is defined as the percent of variance in the respective variable explained by the factor.
Factor Matrix	This contains factor loadings on all the variables on all the factors extracted.
Factor Plot or Rotated Factor Space	This is a plot where the factors are on different axis and the variables are drawn on these axes. This plot can be interpreted only if the number of factors are 3 or less.
Factor Scores	Each individual observation has a score, or value, associated with each of the original variables. Factor analysis procedures derive factor scores that represent each observation's calculated values, or score, on each of the factors. The factor score will represent an individual's combined response to the several variables representing the factor. The component scores may be used in subsequent analysis in PCA. When the factors are to represent a new set of variables that they may predict or be dependent on some phenomenon, the new input may be factor scores.
Goodness of a Factor	How well can a factor account for the correlations among the indicators? One could examine the correlations among the indicators after the effect of the factor is removed. For a good factor solution, the resulting partial correlations should be near zero, because once the effect of the common factor is removed, there is nothing to link the indicators.
Bartlett's Test of specificity	This is the test statistics used to test the null hypothesis that there is no correlation between the variables.
Kaiser Meyer Olkin (KMO) Measure of Sampling Adequacy	This is an index used to test appropriateness of the factor analysis. High values of this index, generally, more than 0.5, may indicate that the factor analysis is an appropriate measure, where as the lower values (less than 0.5) indicate that factor analysis may not be appropriate.
Scree Plot	A plot of eigenvalues against the factors in the order of their extraction.
Trace	The sum of squares of the values on the diagonal of the correlation matrix used in the factor analysis. It represents the total amount of variance on which the factor solution is based.

14.6.3 Principal Component Analysis (PCA)

Suppose, in a particular situation, **k** variables are required to explain the entire system under study. Through PCA, the original variables are transformed into a new set of variables called **principal components**, numbering much less than k. These are formed in such a manner that they extract almost the entire information provided by the original variables. Thus, the original data of n observations on each of the k variables is reduced to a new data of n observations on each of the principal components. That is how PCA is referred to as one of the data reduction and interpretation techniques. Some indicative applications are given below:

There are a number of financial parameters/ratios for predicting health of a company. It would be useful if only a couple of indicators could be formed as linear combination of the original parameters/ratios in such a way that the few indicators extract most of the information contained in the data on original variables.

Further, in the regression model, if independent variables are correlated implying there is multicollinearity, then new variables could be formed as linear combinations of original variables which themselves are uncorrelated. The regression equation can then be derived with these new uncorrelated independent variables, and used for interpreting the regression coefficients as also for predicting the dependant variable with the help of these new independent variables. This is highly useful in marketing and financial applications involving forecasting, sales, profit, price, etc. with the help of regression equations.

Further, analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretations that would not be ordinarily understood. A good example of this is provided by stock market indices.

Incidentally, PCA is a means to an end and not the end in itself. PCA can be used for inputting principal components as variables for further analysing the data using other techniques such as cluster analysis, regression and discriminant analysis.

14.6.4 Common Factor Analysis

It is yet another example of a data reduction and summarisation technique. It is a statistical approach that is used to analyse inter relationships among a large number of variables (e.g., test scores, test items, questionnaire responses) and then explaining these variables in terms of their common underlying dimensions (factors). For example, a hypothetical survey questionnaire may consist of 20 or even more questions, but since not all of the questions are identical, they do not measure all the basic underlying dimensions to the same extent. By using factor analysis, we can identify the separate dimensions being measured by the survey and determine a factor loading for each variable (test item) on each factor.

Common Factor Analysis (unlike multiple regression, discriminant analysis or canonical correlation, in which one or more variables are explicitly considered as the criterion or dependant variable and all others the predictor or independent variables) is an interdependence technique in which all variables are simultaneously considered. In a sense, each of the observed (original) variables is considered as a dependant variable that is a function of some underlying, latent and **hypothetical/unobserved** set of factors (dimensions). One could also consider the original variables as reflective indicators of the factors. For example, marks (variable) in an examination reflect the intelligence (factor).

The statistical approach followed in factor analysis involves finding a way of condensing the information contained in a number of original variables into a smaller set of dimensions (factors) with a minimum loss of information.

Common Factor Analysis was originally developed to explain students' performance in various subjects and to understand the link between grades and intelligence. Thus, the marks obtained in an examination reflect the student's intelligence quotient. A salesman's performance in term of sales might reflect his attitude towards the job, and efforts made by him.

One of the studies relating to marks obtained by students in various subjects, led to the conclusion that students' marks are a function of two common factors viz. Quantitative and Verbal abilities.

The quantitative ability factor explains marks in subjects like Mathematics, Physics and Chemistry and verbal ability explains marks in subjects like Languages and History.

In another study, a detergent manufacturing company was interested in identifying the major underlying factors or dimensions that consumers used to evaluate various detergents. These factors are assumed to be latent; however, management believed that the various attributes or properties of detergents were indicators of these underlying factors. Factor analysis was used to identify these underlying factors. Data was collected on several product attributes using a five-point scale. The analysis of responses revealed existence of two factors viz. ability of the detergent to clean and its mildness.

In general, the factor analysis performs the following functions:

- Identifies the smallest number of common factors that best explain or account for the correlation among the indicators.
- Identifies a set of dimensions that are latent (not easily observed) in a large number of variables.
- Devises a method of combining or condensing a large number of consumers with varying preferences into distinctly different number of groups.
- Identifies and creates an entirely new smaller set of variables to partially or completely replace the original set of variables for subsequent regression or discriminant analysis from a large number of variables. It is especially useful in multiple regression analysis when multicollinearity is found to exist as the number of independent variables is reduced by using factors and thereby minimising or avoiding multicollinearity. In fact, factors are used in lieu of original variables in the regression equation.

Distinguishing Feature of Common Factor Analysis

Generally, the variables that we define in real life situations reflect the presence of unobservable factors. These factors impact the values of those variables. For example, the marks obtained in an examination reflect the student's intelligence quotient. A salesperson's performance in term of sales might reflect his or her attitude towards the job, and efforts made by him or her.

Each of the above examples requires a scale, or an instrument to measure the various constructs (i.e., attitudes, image, patriotism, sales aptitude and resistance to innovation). These are but a few examples of the type of measurements that are desired by various business disciplines. Factor analysis is one of the techniques that can be used to develop scales to measure these constructs.

14.6.4.1 Applications of Common Factor Analysis In one of the studies conducted by a group of the students of a management institute, they undertook a survey of 120 potential buyers outside retail outlets and at dealer counters. Their opinions were solicited through a questionnaire for each of the 20 parameters relating to a television.

Through the use of principal component analysis and factor analysis using computer software, the group concluded that the following five parameters are most important out of the twenty parameters on which their opinion was recorded. The five factors were:

- Price (price, schemes and other offers)
- Picture quality
- Brand ambassador (person of admiration)
- Wide range
- Information (website use, brochures, friends' recommendations)

In yet another study, another group of students of a management institute conducted a survey to identify the factors that influence the purchasing decision of a motorcycle in the 125 cc category. Through the use of Principal Component Analysis and factor analysis using computer software, the group concluded that the following three parameters are most important:

- Comfort
- Assurance
- Long-term Value

14.6.5 Factor Analysis on Data Using SPSS

We shall first explain PCA using SPSS and than Common Factor Analysis.

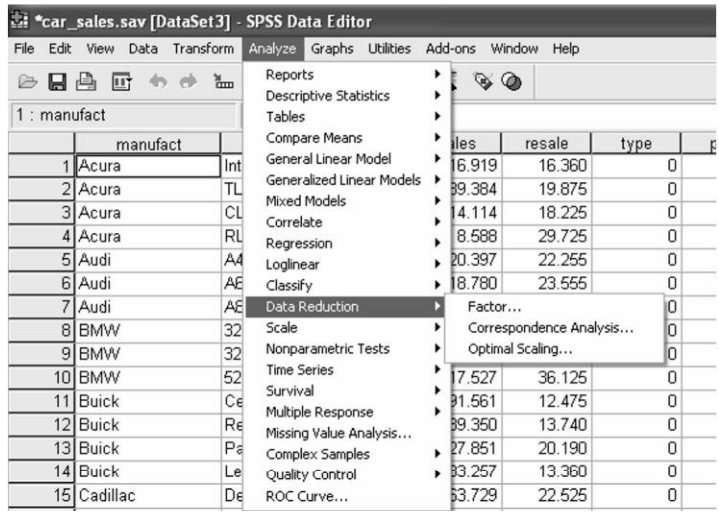
Principle Component Analysis Using SPSS

For illustration file car_sales.sav will be used. This file is part of SPSS cases and is in the tutorial folder of SPSS. Within tutorial folder, this file is in the sample_files folder. For the convenience of readers, this file has been provided in the CD with the book. This data file contains hypothetical sales estimates, list prices and physical specifications for various makes and models of vehicles. The list prices and physical specifications were obtained alternately from edmunds.com and manufacturer sites. Following is the list of the major variables in the file:

1. Manufacturer	6. Price in thousands	11. Length
2. Model	7. Engine size	12. Curb weight
3. Sales in thousands	8. Horsepower	13. Fuel capacity
4. 4-year resale value	9. Wheelbase	14. Fuel efficiency
5. Vehicle type	10. Width	

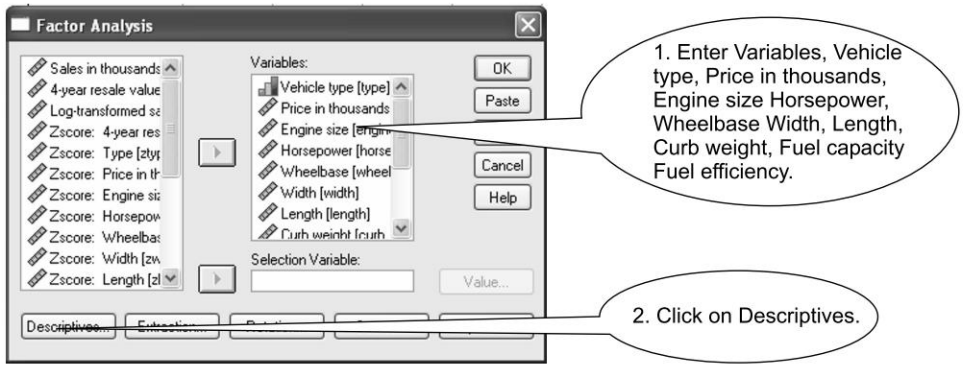
After opening the file Car_sales.sav, one can click on Analyze – Data Reduction and Factor as shown in the following snapshot:

FA Snapshot 1



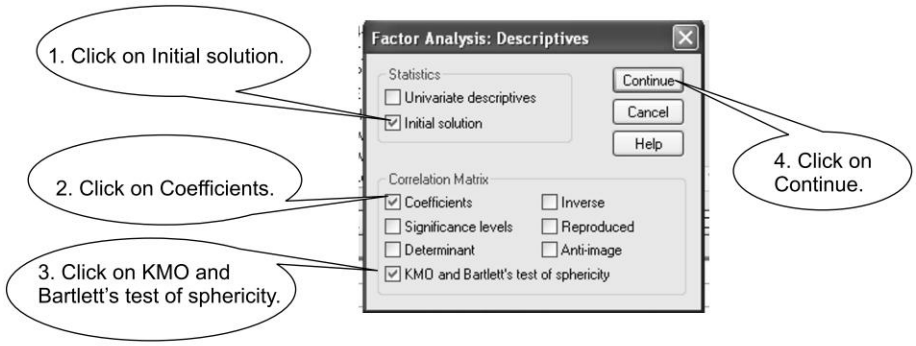
This command will open a new window as shown below:

FA Snapshot 2



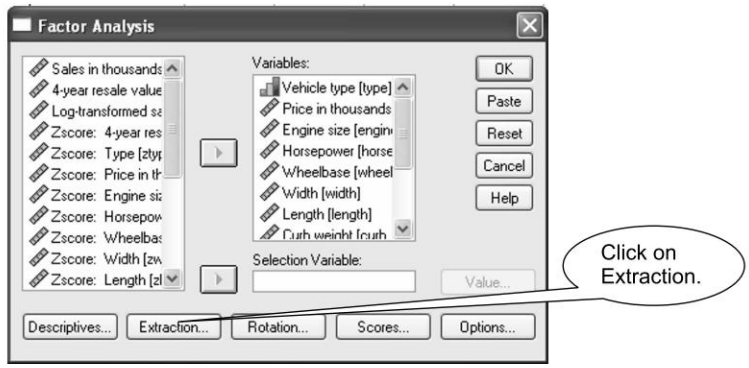
This will open a new window as shown below:

FA Snapshot 3



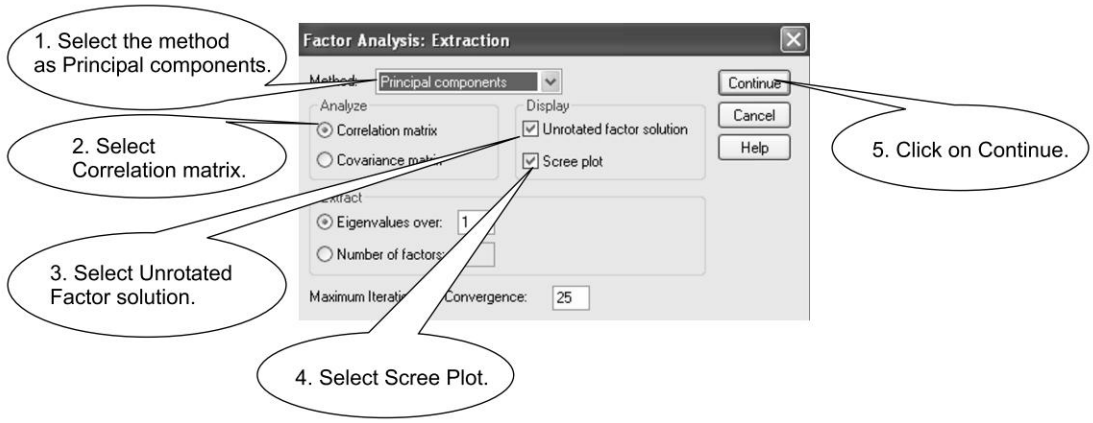
SPSS will take back to previous window as shown below:

FA Snapshot 4



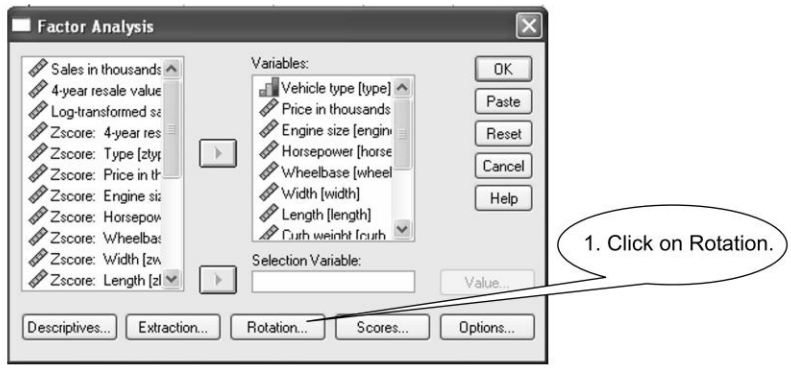
The new window that will appear is shown below:

FA Snapshot 5



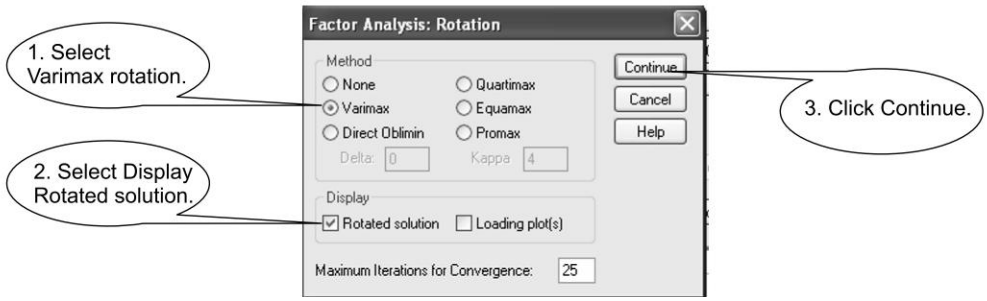
SPSS will take back to the window shown below:

FA Snapshot 6



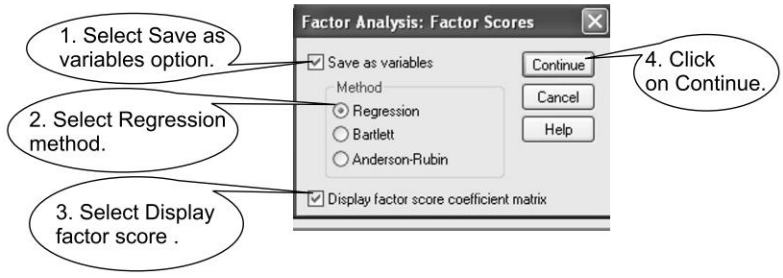
A window will open as follows:

FA Snapshot 7



SPSS will take back to the window as shown in FA Snapshot 6. Click on the button 'Scores'. This will open a new window as shown as follows:

FA Snapshot 8



This will take back to window as shown in FA Snapshot 7, in this window now click on 'Ok'. SPSS will give the following output. We shall explain each in brief.

Factor Analysis

Correlation Matrix

		Vehicle type	*Price in thousands*	*Engine size*	*Horsepower*	*Wheelbase*	*Width*	*Length*	*Curb weight*	*Fuel capacity*	*Fuel efficiency*
Correlation	Vehicle type	1.000	–0.042	0.269	0.017	0.397	0.260	0.150	0.526	0.599	–0.577
	Price in thousands	–0.042	1.000	0.624	0.841	0.108	0.328	0.155	0.527	0.424	–0.492
	Engine size	0.269	0.624	1.000	0.837	0.473	0.692	0.542	0.761	0.667	–0.737
	Horsepower	0.017	0.841	0.837	1.000	0.282	0.535	0.385	0.611	0.505	–0.616
	Wheelbase	0.397	0.108	0.473	0.282	1.000	0.681	0.840	0.651	0.657	–0.497
	Width	0.260	0.328	0.692	0.535	0.681	1.000	0.706	0.723	0.663	–0.602
	Length	0.150	0.155	0.542	0.385	0.840	0.706	1.000	0.629	0.571	–0.448
	Curb weight	0.526	0.527	0.761	0.611	0.651	0.723	0.629	1.000	0.865	–0.820
	Fuel capacity	0.599	0.424	0.667	0.505	0.657	0.663	0.571	0.865	1.000	–0.802
	Fuel efficiency	–0.577	–0.492	–0.737	–0.616	–0.497	–0.602	–0.448	–0.820	–0.802	1.000

This is the correlation matrix. The PCA can be carried out if the correlation matrix for the variables contains at least two correlations of 0.30 or greater. It may be noted that the correlations >0.3 are marked in circle.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.833
Bartlett's Test of Sphericity	Approx. Chi-Square	1578.819
	df	45
	Sig.	0.000

KMO–Bartlett measure of sampling adequacy is an index used to test appropriateness of the factor analysis. The minimum required KMO is 0.5. The above table shows that the index for this

data is 0.833 and the chi-square statistics is significant (<0.05). This means the principal component analysis is appropriate for this data.

Communalities

	Initial	*Extraction*
Vehicle type	1.000	.930
Price in thousands	1.000	.876
Engine size	1.000	.843
Horsepower	1.000	.933
Wheelbase	1.000	.881
Width	1.000	.776
Length	1.000	.919
Curb weight	1.000	.891
Fuel capacity	1.000	.861
Fuel efficiency	1.000	.860

Extraction Method: Principal Component Analysis.

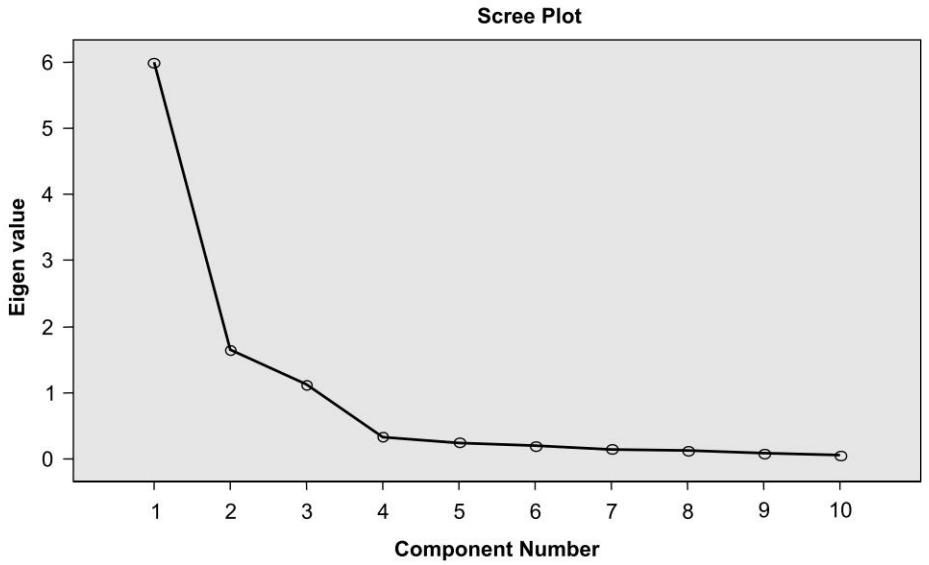
Extraction communalities are estimates of the variance in each variable accounted for by the components. The communalities in this table are all high, which indicates that the extracted components represent the variables well. If any communalities are very low in a principal components extraction, you may need to extract another component.

Total Variance Explained

	Initial Eigenvalues			*Extraction Sums of Squared Loadings*			*Rotation Sums of Squared Loadings*		
Component	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	5.994	59.938	59.938	5.994	59.938	59.938	3.220	32.199	32.199
2	1.654	16.545	76.482	1.654	16.545	76.482	3.134	31.344	63.543
3	1.123	11.227	87.709	1.123	11.227	87.709	2.417	24.166	87.709
4	.339	3.389	91.098						
5	.254	2.541	93.640						
6	.199	1.994	95.633						
7	.155	1.547	97.181						
8	.130	1.299	98.480						
9	.091	.905	99.385						
10	.061	.615	100.000						

Extraction Method: Principal Component Analysis.

This output gives total variance explained. This table gives the total variance contributed by each component. We can see that the percentage of total variance contributed by first component is 59.938, by second component is 16.545 and by third component is 11.2227. It is also clear from this table that there are total three distinct components for the given set of variables.



The scree plot gives the number of components against the eigenvalues and helps to determine the optimal number of components.

Incidentally, "scree" is the geological term referring to the debris which gets deposited on the lower part of a rocky slope.

The components having steep slope indicate that good percentage of total variance is explained by that component, hence the component is justified. The shallow slope indicates that the contribution of total variance is less, and the component is not justified. In the above plot, the first three components have steep slope and later the slope is shallow. This indicates the ideal number of components is three.

Component Matrix[a]

	Component		
	1	*2*	*3*
Vehicle type	.471	.533	–.651
Price in thousands	.580	–.729	–.092
Engine size	.871	–.290	.018
Horsepower	.740	–.618	.058
Wheelbase	.732	.480	.340
Width	.821	.114	.298
Length	.719	.304	.556
Curb weight	.934	.063	–.121
Fuel capacity	.885	.184	–.210
Fuel efficiency	–.863	.004	.339

Extraction Method: Principal Component Analysis.
a. 3 components extracted.

This table gives each variable component loadings but it is the next table, which is easy to interpret.

Rotated Component Matrix

	Component		
	1	*2*	*3*
Vehicle type	–0.101	0.095	0.954
Price in thousands	0.935	–0.003	0.041
Engine size	0.753	0.436	0.292
Horsepower	0.933	0.242	0.056
Wheelbase	0.036	0.884	0.314
Width	0.384	0.759	0.231
Length	0.155	0.943	0.069
Curb weight	0.519	0.533	0.581
Fuel capacity	0.398	0.495	0.676
Fuel efficiency	–0.543	–0.318	–0.681

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 4 iterations.

This table is the most important table for interpretation. The maximum of each row (ignoring sign) indicates that the respective variable belongs to the respective component. The variables 'price in thousands', 'engine size' and 'horsepower' are highly correlated and contribute to a single component. 'Wheelbase', 'width' and 'length' contribute to second component. And 'vehicle type', 'curb weight', 'fuel capacity' contribute to the third component.

Component Transformation Matrix

Component	*1*	*2*	*3*
1	0.601	0.627	0.495
2	–0.797	0.422	0.433
3	–0.063	0.655	–0.753

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Component Score Coefficient Matrix

	1	*2*	*3*
Vehicle type	–0.173	–0.194	0.615
Price in thousands	0.414	–0.179	–0.081
Engine size	0.226	0.028	–0.016
Horsepower	0.368	–0.046	–0.139

(*Contd*)

(*Contd*)

Wheelbase	–0.177	0.397	–0.042
Width	0.011	0.289	–0.102
Length	–0.105	0.477	–0.234
Curb weight	0.070	0.043	0.175
Fuel capacity	0.012	0.017	0.262
Fuel efficiency	–0.107	0.108	–0.298

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Component Scores.

This table gives the component scores for each variable. The component scores can be saved for each case in the SPSS file. These scores are useful to replace internally related variables in the regression analysis. In the above table, the scores are given component wise. The factor score for each component can be calculated as the linear combinations of the component scores of that component.

Component Score Covariance Matrix

Component	*1*	*2*	*3*
1	1.000	0.000	0.000
2	0.000	1.000	0.000
3	0.000	0.000	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Component Scores.

Common Factor Analysis Using SPSS

For illustration the following case is considered:
The file Telecom.sav is provided in the CD with the book.

Case 14.1

In the year 2009, the TRAI (Telephone Regulatory Authority of India) was assessing the requirements for number portability. Number portability is defined as switching a service provider, without changing the number. This year had seen fierce price war in the telecom sector. Some of the oldest service providers are still to certain extent immune to this war as most of their consumers would not like to change number. The number portability will increase the price war and give opportunity to relatively new service providers. The price war is so fierce that the industry experts comment that the future lies in how one differentiates in terms of services, than price.

With this background, a TELECOMM company conducted a research study to find the factors that affect consumers while selecting/switching a telecom service provider. The survey was conducted on 35 respondents. They were asked to rate 12 questions, about their perception of factors important to them while selecting a service provider, on 7-point scale (1= completely disagree, 7= completely agree).

The research design for data collection can be stated as follows:

35 telecom users were surveyed about their perceptions and image attributes of the service providers they owned. Twelve questions were asked to each of them, all answered on a scale of 1 to 7 (1= completely disagree, 7= completely agree).

I decide my telephone provider on the basis of following attributes. (1= completely disagree, 7= completely agree)

1. Availability of services (like drop boxes and different payment options, in case of postpaid, and recharge coupons in case of prepaid)
2. Good network connectivity all through the city.
3. Internet connection, with good speed.
4. Quick and appropriate response at customer care centre.
5. Connectivity while roaming (out of the state or out of country)
6. Call rates and Tariff plans.
7. Additional features like unlimited SMS, lifetime prepaid, 2 phones free calling, etc.
8. Quality of network service like minimum call drops, minimum down time, voice quality, etc.
9. SMS and Value Added Services charges.
10. Value Added Services like MMS, caller tunes, etc.
11. Roaming charges
12. Conferencing

The data collected is tabulated as follows:

SrNo	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Q7*	*Q8*	*Q9*	*Q10*	*Q11*	*Q12*
1	3	6	5	4	6	7	2	5	1	4	2	5
2	3	5	3	4	6	5	4	5	4	2	4	2
3	4	5	2	5	5	1	5	5	6	1	6	1
4	5	2	6	6	3	3	6	4	5	5	5	4
5	4	6	7	5	5	6	3	5	4	6	3	6
6	4	4	3	5	4	4	4	4	4	2	5	1
7	4	5	4	5	2	2	5	2	6	3	5	4
8	2	3	6	3	4	5	3	4	2	5	4	5
9	3	5	4	4	5	7	2	7	1	3	2	6
10	2	6	4	3	5	3	6	6	5	3	6	3
11	1	6	6	2	6	4	5	6	4	5	5	5
12	1	6	7	1	6	2	6	5	6	5	6	6
13	5	4	2	6	3	5	3	4	4	2	4	1
14	4	5	7	5	5	3	5	5	5	4	5	6
15	1	5	6	2	6	4	5	1	4	6	5	5
16	1	2	3	1	4	3	6	3	4	2	5	2
17	2	6	7	3	5	3	6	4	5	6	6	6

(*Contd*)

(*Contd*)

18	3	5	5	4	6	2	5	5	5	5	5	4
19	3	4	5	4	5	4	3	4	4	4	3	4
20	3	3	6	4	5	7	1	1	1	4	1	5
21	1	7	7	2	7	1	7	6	7	6	7	7
22	6	5	3	7	4	4	4	5	5	3	5	2
23	5	6	7	6	6	2	6	6	6	5	6	7
24	2	6	7	3	7	3	6	2	5	7	6	6
25	1	3	4	2	5	2	7	4	5	3	6	3
26	3	7	7	4	6	2	7	5	6	7	7	7
27	4	6	6	5	6	1	7	6	7	6	6	5
28	4	5	7	5	6	3	4	5	5	5	4	6
29	4	6	5	5	4	2	6	2	6	4	6	4
30	5	5	4	6	3	3	5	2	5	3	5	3
31	3	6	7	4	6	2	7	5	6	7	7	7
32	3	3	6	4	5	7	1	1	1	4	2	5
33	5	5	2	6	3	5	3	4	4	2	4	1
34	1	3	6	2	5	3	6	5	5	5	5	5
35	7	2	5	7	3	2	6	2	6	5	6	5

Carry out relevant analysis and write a report to discuss the findings for the above data.

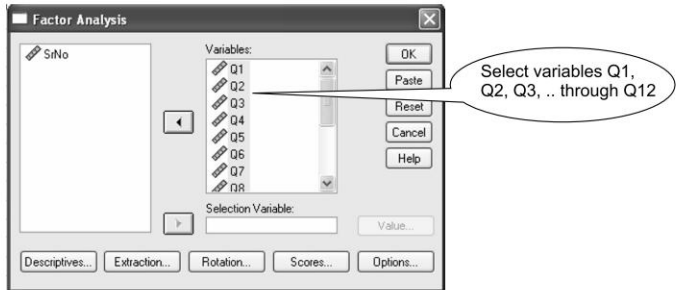
The initial process of conducting common factor analysis is exactly same as for principal component analysis except for the method of selection shown in FA Snapshot 5.

We will discuss only the steps that are different than the principal component analysis shown above.

Following steps are carried out to run factor analysis using SPSS:

1. Open file telcom.sav
2. Click on Analyse ->Data Reduction ->Factor as shown in FA Snapshot 1.
3. Following window will be opened by SPSS.

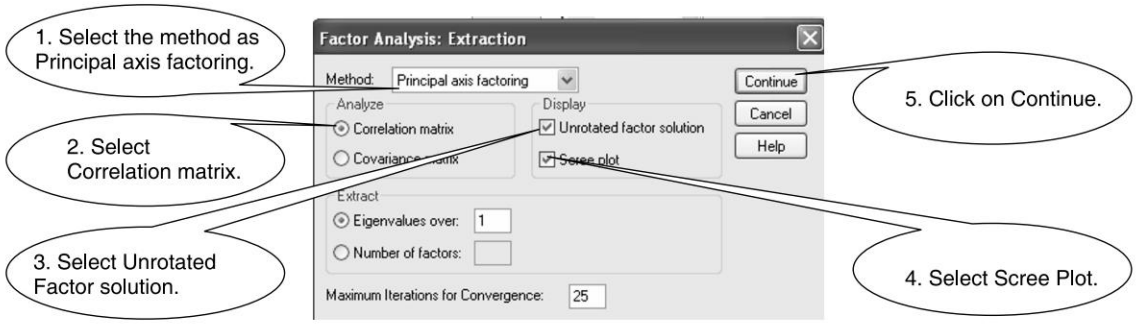
FA Snapshot 9



4. Click on descriptives, coefficients and click on initial solution, click on KMO and Bartlett's test of sphericity, and also select Anti-Image as shown in FA Snapshot 3.

   It may be noted that we did not select Anti-Image in PCA, but we are required to select it here.
5. Click on Extraction, following window will be opened by SPSS.

FA Snapshot 10



6. SPSS will take back to the window shown in FA Snapshot 9 at this stage. Click on Rotation, the window SPSS will open is shown in FA Snapshot.
7. Select Varimax rotation, select Display rotated solution and click continue, as shown in FA Snapshot 7.
8. It may be noted that in PCA of FA Snapshot 8 we selected to store some variables which is not required here.

Following output will be generated by SPSS:

Factor Analysis

Descriptive Statistics

	Mean	*Std. Deviation*	*Analysis N*
Q1	4.80	1.568	35
Q2	3.20	1.410	35
Q3	2.83	1.671	35
Q4	3.89	1.605	35
Q5	3.09	1.245	35
Q6	3.49	1.772	35
Q7	3.23	1.734	35
Q8	3.86	1.611	35
Q9	3.46	1.633	35
Q10	3.74	1.615	35
Q11	3.17	1.505	35
Q12	3.60	1.866	35

This is descriptive statistics given by the SPSS. This gives general understanding of the variables.

Correlation Matrix

		Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Q7*	*Q8*	*Q9*	*Q10*	*Q11*	*Q12*
Correlation	Q1	1.000	–0.128	–0.294	0.984	–0.548	–0.017	–0.188	–0.093	0.129	–0.242	–0.085	–0.259
	Q2	–0.128	1.000	0.302	–0.068	0.543	0.231	0.257	0.440	0.355	0.359	0.344	0.378
	Q3	–0.294	0.302	1.000	–0.282	0.558	0.148	0.258	0.056	0.148	0.898	0.164	0.930
	Q4	0.984	–0.068	–0.282	1.000	–0.510	–0.052	–0.223	–0.063	0.099	–0.227	–0.113	–0.251
	Q5	–0.548	0.543	0.558	–0.510	1.000	0.101	0.195	0.387	0.067	0.538	0.149	0.559
	Q6	–0.017	0.231	0.148	–0.052	0.101	1.000	0.901	0.159	0.937	0.230	0.906	0.096
	Q7	–0.188	0.257	0.258	–0.223	0.195	0.901	1.000	0.202	0.866	0.379	0.943	0.211
	Q8	–0.093	0.440	0.056	–0.063	0.387	0.159	0.202	1.000	0.204	0.042	0.192	0.156
	Q9	0.129	0.355	0.148	0.099	0.067	0.937	0.866	0.204	1.000	0.258	0.889	0.091
	Q10	–0.242	0.359	0.898	–0.227	0.538	0.230	0.379	0.042	0.258	1.000	0.309	0.853
	Q11	–0.085	0.344	0.164	–0.113	0.149	0.906	0.943	0.192	0.889	0.309	1.000	0.119
	Q12	–0.259	0.378	0.930	–0.251	0.559	0.096	0.211	0.156	0.091	0.853	0.119	1.000

This is the correlation matrix. The Common Factor Analysis can be carried out if the correlation matrix for the variables contains at least two correlations of 0.30 or greater. It may be noted that some of the correlations >0.3 are marked in circle.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measures of Sampling Adequacy		0.658
Bartlett's Test of Spherecity	Approx. Chi-Square	497.605
	df	66
	Sig.	0.000

KMO measure of sampling adequacy is an index used to test appropriateness of the factor analysis. The minimum required KMO is 0.5. The above table shows that the index for this data is 0.658 and the chi-square statistics is significant (0.000<0.05). This means the principal component analysis is appropriate for this data.

Communalities

	Initial	*Extraction*
Q1	0.980	0.977
Q2	0.730	0.607
Q3	0.926	0.975
Q4	0.978	0.996
Q5	0.684	0.753
Q6	0.942	0.917
Q7	0.942	0.941

(*Contd*)

(*Contd*)

Q8	0.396	0.379
Q9	0.949	0.942
Q10	0.872	0.873
Q11	0.934	0.924
Q12	0.916	0.882

Extraction Method: Principal Axis Factoring.

Initial communalities are the proportion of variance accounted for in each variable by the rest of the variables. Small communalities for a variable indicate that the proportion of variance that this variable shares with other variables is too small. Thus, this variable does not fit the factor solution. In the above table, most of the initial communalities are very high indicating that all the variables share a good amount of variance with each other, an ideal situation for factor analysis.

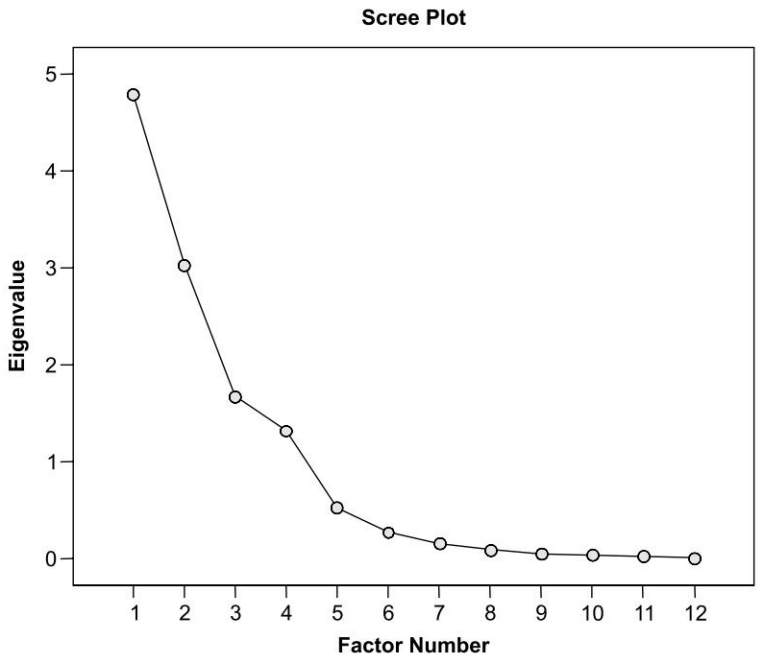
Extraction communalities are estimates of the variance in each variable accounted for by the factors in the factor solution. The communalities in this table are all high. It indicates that the extracted factors represent the variables well.

Total Variance Explained

	Initial Eigenvalues			*Extraction Sums of Squared Loadings*			*Rotation Sums of Squared Loadings*		
Factor	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	4.789	39.908	39.908	4.671	38.926	38.926	3.660	30.500	30.500
2	3.035	25.288	65.196	2.956	24.635	63.561	2.929	24.410	54.910
3	1.675	13.960	79.156	1.628	13.570	77.131	2.205	18.372	73.283
4	1.321	11.006	90.163	0.911	7.593	84.724	1.373	11.441	84.724
5	0.526	4.382	94.545						
6	0.275	2.291	96.836						
7	0.157	1.307	98.143						
8	0.093	0.774	98.917						
9	0.050	0.421	99.338						
10	0.042	0.353	99.691						
11	0.027	0.227	99.918						
12	0.010	0.082	100.000						

Extraction Method: Principal Axis Factoring.

This output gives the variance explained by the initial solution. This table gives the total variance contributed by each component. We may note that the percentage of total variance contributed by first component is 39.908, by second component is 25.288 and by third component is 19.960. It may be noted that the percentage of total variances is the highest for first factor and it decreases thereafter. It is also clear from this table that there are total three distinct factors for the given set of variables.



The scree plot gives the number of factors against the eigenvalues, and helps to determine the optimal number of factors. The factors having steep slope indicate that larger percentage of total variance is explained by that factor. The shallow slope indicates that the contribution to total variance is less. In the above plot, the first four factors have steep slope; and later on the slope is shallow. It may be noted from the above plot that the number of factors for eigenvalue greater than one are four. Hence, the ideal number of factors is four.

Factor Matrix[a]

	Factor			
	1	*2*	*3*	*4*
Q1	0.410	0.564	0.698	0.066
Q2	0.522	–0.065	0.139	0.557
Q3	0.687	–0.515	0.423	–0.245
Q4	–0.413	0.528	0.725	0.141
Q5	0.601	–0.500	–0.067	0.371
Q6	0.705	0.630	–0.111	–0.102
Q7	0.808	0.486	–0.176	–0.144
Q8	0.293	0.005	–0.031	0.540
Q9	0.690	0.682	0.039	0.019
Q10	0.727	–0.372	0.401	–0.213
Q11	0.757	0.575	–0.137	–0.040
Q12	0.644	–0.523	0.432	–0.090

Extraction Method: Principal Axis Factoring.
a. 4 factors extracted. 14 iterations required.

This table gives each variable factor loadings but it is the next table, which is easy to interpret.

Rotated Factor Matrix[a]

	Factor			
	1	*2*	*3*	*4*
Q1	0.001	–0.147	0.972	–0.109
Q2	0.195	0.253	–0.002	0.711
Q3	0.084	0.970	–0.146	0.074
Q4	–0.042	–0.137	0.987	–0.035
Q5	0.006	0.456	–0.430	0.600
Q6	0.953	0.053	–0.004	0.078
Q7	0.939	0.161	–0.162	0.086
Q8	0.117	–0.008	–0.040	0.603
Q9	0.936	0.068	0.164	0.186
Q10	0.210	0.899	–0.101	0.103
Q11	0.943	0.078	–0.059	0.158
Q12	0.022	0.909	–0.110	0.206

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

This table is the most important table for interpretation. The maximum in each row (ignoring sign) indicates that the respective variable belongs to the respective factor. For example, in the first row the maximum is 0.972 which is for factor 3; this indicates that the Q1 contributes to third factor. In the second row, maximum is 0.711; for factor 4 indicating that Q2 contributes to factor 4, and so on.

The variables 'Q6', 'Q7', 'Q9' and 'Q11' are highly correlated and contribute to a single factor which can be named as Factor 1 or '**Economy**'.

The variables 'Q3', 'Q10' and 'Q12' are highly correlated and contribute to a single factor which can be named as Factor 2 or '**Services beyond Calling**'.

The variables 'Q1' and 'Q4' are highly correlated and contribute to a single factor which can be named as Factor 3 or '**Customer Care**'.

The variables 'Q2', 'Q5' and 'Q8' are highly correlated and contribute to a single factor which can be named as Factor 4 or '**Anytime Anywhere Service'**.

We may summarise the above analysis in the following table:

Factors	*Questions*
Factor 1 Economy	Q.6. Call rates and Tariff plans Q.7. Additional features like unlimited SMS, lifetime prepaid, 2 phones free calling, etc. Q.9. SMS and Value Added Services charges Q.11. Roaming charges

(*Contd*)

(*Contd*)

Factor 2 Services beyond Calling	Q.3. Internet connection, with good speed Q.10. Value Added Services like MMS, caller tunes, etc. Q.12. Conferencing
Factor 3 Customer Care	Q.1. Availability of services (like drop boxes and different payment options, in case of post paid and recharge coupons in case of prepaid) Q.4. Quick and appropriate response at customer care centre.
Factor 4 Anytime Anywhere Service	Q.2. Good network connectivity all through the city Q.5. Connectivity while roaming (out of the state or out of country) Q.8. Quality of network service like minimum call drops, minimum down time, voice quality, etc.

It implies that the telecomm service provider should consider these four factors which customers feel are important, while selecting/switching a service provider.

14.7 CANONICAL CORRELATION ANALYSIS

The canonical correlation analysis, abbreviated as CRA is an extension of multiple regression analysis, abbreviated as MRA. While, in MRA, there is one metric (measurable or non-categorical) dependant variable, say y, and there are several metric independent variables, say x_1, x_2, ….., x_k, in CRA, there are several metric dependent variables, say y_1, y_2, ……, y_m .

CRA involves developing a linear combination of the two sets of above variables viz. y's and x's – one as a linear combination of dependent variables (also called predictor set) and the other as a linear combination of the set of independent variables (also called criterion set). **The two linear combinations are derived in such a way that the correlation between the two is maximum.** While the linear combinations are referred as **canonical variables**, the correlation between the two combinations is called **canonical correlation**. It measures the strength of the overall relationship between the linear combinations of the predictor and criterion sets of variables. In the next stage, the identification process involves choosing the second pair of linear combinations having the second largest correlation among all pairs but uncorrelated with the initial pair. The process continues for the third pair, and so on. **The practical significance of a canonical correlation is that it indicates as to how much variance in one set of dependent variables is accounted for by another set of independent variables**. **The weights in the linear combination are derived based on the criterion that maximises the correlation between the two sets.**
It can be represented as follows:

$$\underset{\text{(metric)}}{Y + Y2 + \ldots Y_p} = \underset{\text{(metric)}}{Y1 + Y2 + \ldots X_p}$$

Some Indicative Applications:

- A medical researcher could be interested in determining if individuals' lifestyle and personal habits have an impact on their health as measured by a number of health-related variables such as hypertension, weight, blood sugar, etc.
- The marketing manager of a consumer goods firm could be interested in determining if there is a relationship between types of products purchased and consumers' income and profession.

The practical significance of a canonical correlation is that it indicates as to how much variance in one set of variables is accounted for by another set of variables.

Squared canonical correlations are referred to as canonical roots or eigenvalues.

If $X_1, X_2, X_3, \ldots\ldots, X_p$ and $Y_1, Y_2, Y_3, \ldots\ldots, Y_q$ are the **observable variables** then canonical variables will be:

$$U_1 = a_1X_1 + a_2X_2 + \ldots\ldots\ldots + a_pX_p \qquad V_1 = b_1Y_1 + b_2Y_2 + \ldots\ldots\ldots + b_qY_q$$

$$U_2 = c_1X1 + c_2X2 + \ldots\ldots\ldots\ldots + c_pX_p \quad V_2 = d_1Y_1 + d_2Y_2 + \ldots\ldots\ldots + d_qY_q$$

and so on

Then Us and Vs are called **canonical variables** and coefficients are called **canonical coefficients**.

The first pair of sample canonical variables is obtained in such a way that

$$\text{Var } (U_1) = \text{Var } (V_1) = 1$$

and $\text{Corr } (U_1, V_1)$ is maximum.

The second pair U_2 and V_2 are selected in such a way that they are uncorrelated with U_1 and V_1, and the correlation between the two is maximum, and so on.

DA and CRA

CR^2 which is defined as the ratio of SSB to SST is a measure of the strength of the discriminant function. If its value is 0.84, it implies that 84% of the variation between the two groups is explained by the two discriminating variables.

Canonical Discriminant Function

Canonical correlation measures the association between the discriminant scores and the groups. For two groups, it is the usual person correlation between the scores and the groups coded as 0 and 1.

$$CR^2 = \frac{SSB}{SST} = \frac{SSB/SSW}{SST/SSW} = \frac{\text{Eigenvalue}}{1/\Lambda}$$

$$= \text{Eigenvalue} \times \lambda$$

Wilks' Λ is the proportion of the total variance in the discrimination scores not explained by differences among the groups. It is used to test H0 that the means of the variables of groups are equal.

λ is approximated by $\chi^2_{p,\,G-1} = -\{n\text{-}1 - (p + G/2)\} \times \log \lambda$

p : no. of variables

G : no. of groups

$0 \le \lambda \le 1$. If λ is small, H_0 is rejected, if it is high, H_0 is accepted.

MDA and PCA : Similarities and Differences

In both cases, a new axis is identified and a new variable is formed as a linear combination of the original variables. The new variable is given by the projection of the points onto this new axis.

The difference is with respect to the criterion used to identify the new axis.

In PCA, a new axis is formed such that the projection of the points onto this new axis account for maximum variance in the data. This is equivalent to maximising SST, because there is no criterion variable for dividing the sample into groups.

In MDA, the objective is not to account for maximum variance in the data (i.e. maximum SST), but to maximise the between-group to within-group sum of squares ratio (i.e. SSB/SSW) that results in the best discrimination between the groups. The new axis or the new linear combination that is identified is called Linear Discriminant Function. The projection of an observed point onto this discriminant function (i.e. the value of the new variable) is called the discriminant score.

Application: Asset – Liability Mismatch

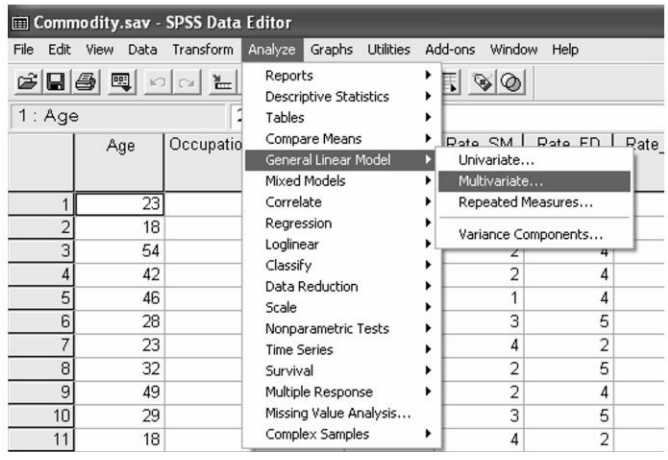
A study on Structural Changes and Asset Liability Mismatch in scheduled Commercial Banks in India was carried out in Reserve Bank of India. It was conducted as an empirical exercise to identify and explore the **relationships** and **structural changes**, including **hedging behaviour**, between asset and liability of cross-section of scheduled commercial banks **at two different time points** representing **pre- and post-liberalisation periods**. As there were two sets of dependent variables, instead of regression, the study used the canonical correlation technique to investigate the asset–liability relationship of the banks at the two time points.

14.7.1 Canonical Correlation Using SPSS

We will use the case data on commodity market perceptions given in Section 14.8.5.
Open the file Commodity.sav

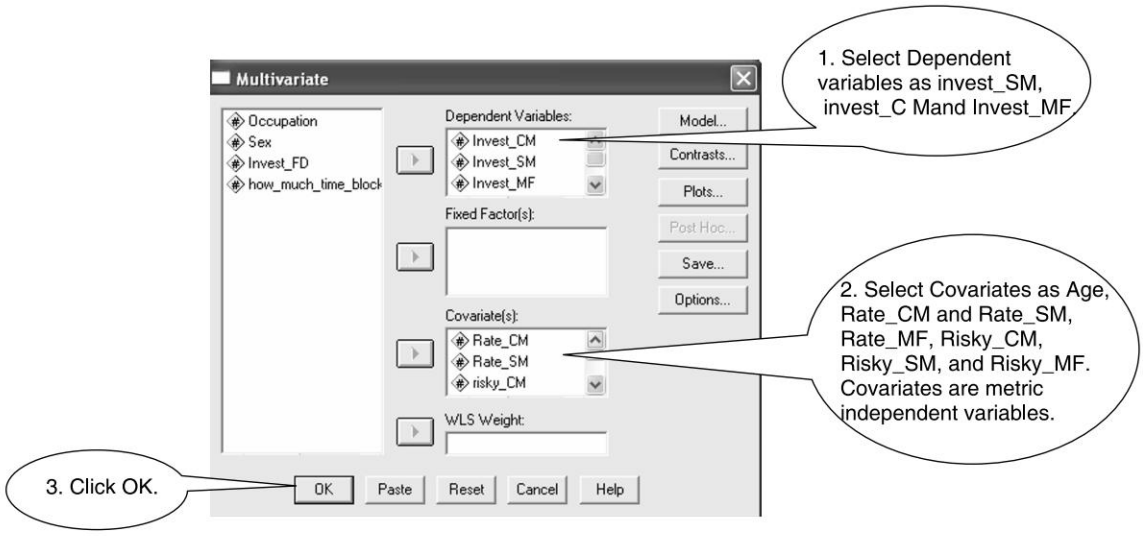
Select from the menu Analyze – General Linear Model – Multivariate as shown below:

Canonical Snapshot 1



The following window will be displayed:

Canonical Snapshot 2



It may be noted that the above example is discussed in Section 14.5.1. The difference between MANOCOVA and canonical correlation is that MANOCOVA can have both factors and metric independent variables, Canonical correlation can have only metric independent variables, factors (categorical independent variables) are not possible in canonical correlation.

We are assuming in the above example that the dependent variables are the investments in commodity market and in share market. The metric independent variables are age, respondent's rating for commodity market, share market and mutual funds and respondent's perception of risk for commodity market, share market and mutual funds. Here, we assume that their investments depend on their ratings, age and their risk perceptions for mutual funds, commodity markets and share markets.

The following output will be displayed:

General Linear Model

Multivariate Tests[b]

Effect		*Value*	*F*	*Hypothesis df*	*Error df*	*Sig.*
Intercept	Pillai's Trace	0.022	0.241[a]	3.000	32.000	0.867
	Wilks' Lambda	0.978	0.241[a]	3.000	32.000	0.867
	Hotelling's Trace	0.023	0.241[a]	3.000	32.000	0.867
	Roy's Largest Root	0.023	0.241[a]	3.000	32.000	0.867
Rate_CM	Pillai's Trace	0.024	0.267[a]	3.000	32.000	0.848
	Wilks' Lambda	0.976	0.267[a]	3.000	32.000	0.848
	Hotelling's Trace	0.025	0.267[a]	3.000	32.000	0.848
	Roy's Largest Root	0.025	0.267[a]	3.000	32.000	0.848

(*Contd*)

(*Contd*)

Rate_SM	Pillai's Trace	0.026	0.280[a]	3.000	32.000	0.839
	Wilks' Lambda	0.974	0.280[a]	3.000	32.000	0.839
	Hotelling's Trace	0.026	0.280[a]	3.000	32.000	0.839
	Roy's Largest Root	0.026	0.280[a]	3.000	32.000	0.839
Rate_SM	Pillai's Trace	0.123	1.497[a]	3.000	32.000	0.234
	Wilks' Lambda	0.877	1.497[a]	3.000	32.000	0.234
	Hotelling's Trace	0.140	1.497[a]	3.000	32.000	0.234
	Roy's Largest Root	0.140	1.497[a]	3.000	32.000	0.234
risky_SM	Pillai's Trace	0.044	0.490[a]	3.000	32.000	0.692
	Wilks' Lambda	0.956	0.490[a]	3.000	32.000	0.692
	Hotelling's Trace	0.046	0.490[a]	3.000	32.000	0.692
	Roy's Largest Root	0.046	0.490[a]	3.000	32.000	0.692
Age	Pillai's Trace	0.152	1.914[a]	3.000	32.000	0.147
	Wilks' Lambda	0.848	1.914[a]	3.000	32.000	0.147
	Hotelling's Trace	0.179	1.914[a]	3.000	32.000	0.147
	Roy's Largest Root	0.179	1.914[a]	3.000	32.000	0.147
Rate_FD	Pillai's Trace	0.031	0.338[a]	3.000	32.000	0.798
	Wilks' Lambda	0.969	0.338[a]	3.000	32.000	0.798
	Hotelling's Trace	0.032	0.338[a]	3.000	32.000	0.798
	Roy's Largest Root	0.032	0.338[a]	3.000	32.000	0.798
Rate_MF	Pillai's Trace	0.092	1.075[a]	3.000	32.000	0.373
	Wilks' Lambda	0.908	1.075[a]	3.000	32.000	0.373
	Hotelling's Trace	0.101	1.075[a]	3.000	32.000	0.373
	Roy's Largest Root	0.101	1.075[a]	3.000	32.000	0.373
Rate_FD	Pillai's Trace	0.145	1.814[a]	3.000	32.000	0.164
	Wilks' Lambda	0.855	1.814[a]	3.000	32.000	0.164
	Hotelling's Trace	0.170	1.814[a]	3.000	32.000	0.164
	Roy's Largest Root	0.170	1.814[a]	3.000	32.000	0.164
Rate_MF	Pillai's Trace	0.001	0.012[a]	3.000	32.000	0.998
	Wilks' Lambda	0.999	0.012[a]	3.000	32.000	0.998
	Hotelling's Trace	0.001	0.012[a]	3.000	32.000	0.998
	Roy's Largest Root	0.001	0.012[a]	3.000	32.000	0.998

a. Exact statistic

b. Design: Intercept+Rate_CM+Rate_SM+risky_CM+risky_SM+Age+Rate_FD+Rate_MF+risky_FD+risky_MF

This table indicates that the hypothesis about age, ratings of CM, SM and MF and Risky CM, SM and MF are not rejected (as p-value is greater than 0.05)—this means there is no significant difference in the investments for these variables.

Tests of Between-Subjects Effects

Source	*Dependent Variable*	*Type III Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
Corrected Model	Invest_CM	6554404686[a]	9	728267187.3	2.363	0.034
	Invest_SM	3.825E+010[b]	9	4250227452	1.723	0.122
	Invest_MF	3.458E+010[c]	9	3842114308	1.269	0.289
Intercept	Invest_CM	41236149.0	1	41236149.02	0.134	0.717
	Invest_SM	1820226683	1	1820226683	0.738	0.396
	Invest_MF	2063858495	1	2063858495	0.682	0.415
Rate_CM	Invest_CM	3146747.812	1	3146747.812	0.010	0.920
	Invest_SM	701640357	1	701640356.6	0.284	0.597
	Invest_MF	91570171.1	1	91570171.09	0.030	0.863
Rate_SM	Invest_CM	5827313.435	1	5827313.435	0.019	0.891
	Invest_SM	305120.840	1	305120.840	0.000	0.991
	Invest_MF	289045935	1	289045935.2	0.095	0.759
risky_CM	Invest_CM	1383505869	1	1383505869	4.490	0.041
	Invest_SM	3599660366	1	3599660366	1.459	0.235
	Invest_MF	4765133773	1	4765133773	1.574	0.218
risky_SM	Invest_CM	5068043.704	1	5068043.704	0.016	0.899
	Invest_SM	3129346168	1	3129346168	1.269	0.268
	Invest_MF	3115327283	1	3115327283	1.029	0.318
Age	Invest_CM	1759288000	1	1759288000	5.709	0.023
	Invest_SM	4483914274	1	4483914274	1.818	0.187
	Invest_MF	5852257582	1	5852257582	1.933	0.173
Rate_FD	Invest_CM	314194024	1	314194023.6	1.020	0.320
	Invest_SM	1038652882	1	1038652882	0.421	0.521
	Invest_MF	1418050354	1	1418050354	0.468	0.498
Rate_MF	Invest_CM	709118826	1	709118825.7	2.301	0.139
	Invest_SM	4469496231	1	4469496231	1.812	0.187
	Invest_MF	3743054315	1	3743054315	1.236	0.274
risky_FD	Invest_CM	925863328	1	925863327.7	3.005	0.092
	Invest_SM	7133105111	1	7133105111	2.891	0.098
	Invest_MF	4644806048	1	4644806048	1.534	0.224
risky_MF	Invest_CM	112028.404	1	112028.404	0.000	0.985
	Invest_SM	46590657.4	1	46590657.39	0.019	0.892
	Invest_MF	14383175.1	1	14383175.06	0.005	0.945
Error	Invest_CM	1.048E+010	34	308143679.0		
	Invest_SM	8.388E+010	34	2466958509		
	Invest_MF	1.029E+011	34	3027453031		

(*Contd*)

(*Contd*)

Total	Invest_CM	2.780E+010	44
	Invest_SM	2.109E+011	44
	Invest MF	2.302E+011	44
Corrected Total	Invest CM	1.703E+010	43
	Invest SM	1.221E+011	43
	Invest MF	1.375E+011	43

(a) R Squared = 0.385 (Adjusted R Squared = 0.222)
(b) R Squared = 0.313 (Adjusted R Squared = 0.131)
(c) R Squared = 0.251 (Adjusted R Squared = 0.053)

The above table gives three different models namely a, b and c. Model a is for the first dependent variable, invest CM, model b is for dependent variable invest SM and model c is for dependent variable invest MF.

The table also indicates the individual relationship between each dependent – independent variable pair. It is indicated above that only two pairs namely, Risky CM and Invest CM, and Age and Invest CM are significant (p value less than 0.05 indicated by circles). This indicates that the independent variable, perception of risk of commodity markets by consumers (variable name Risky CM) significantly affects the dependent variable, i.e. their investment in commodity markets indicating that the riskiness perceived by consumers affects their investments in the market. Similarly, variable AGE also impacts, their investments in commodity markets. All other combinations are not significant.

14.8 CLUSTER ANALYSIS

This type of analysis is used to divide a given number of entities or objects into groups called clusters. The objective is to classify a sample of entities into a small number of mutually exclusive clusters based on the premise that they are similar within the clusters but dissimilar among the clusters. The criterion for similarity is defined with reference to some characteristics of the entity. For example, for companies, it could be 'Sales', 'Paid up Capital', etc.

The basic methodological questions that are to be answered in the cluster analysis are:

- What are the relevant variables and descriptive measures of an entity?
- How do we measure the similarity between entities?
- Given that we have a measure of similarity between entities, how do we form clusters?
- How do we decide on how many clusters are to be formed?

For measuring similarity, let us consider the following data. The data has been collected on each of k characteristics for all the n entities under consideration of being divided into clusters. Let the k characteristics be measured by k variables as x_1, x_2, x_3 …, x_k, and the data presented in the following matrix form:

	Variables			
	x_1	x_2	…..	x_k
Entity 1	x_{11}	x_{12}	….	x_{1k}
Entity 2	x_{21}	x_{22}	…..	x_{2p}
		……..		
Entity n	x_{n1}	x_{n2}	…	x_{np}

The question is how to determine how similar or dissimilar each row of data is from the others?

This task of measuring similarity between entities is complicated by the fact that, in most cases, the data in its original form are measured in different units or/and scales. This problem is solved by standardising each variable by subtracting its mean from its value and then dividing by its standard deviation. This converts each variable to a pure number.

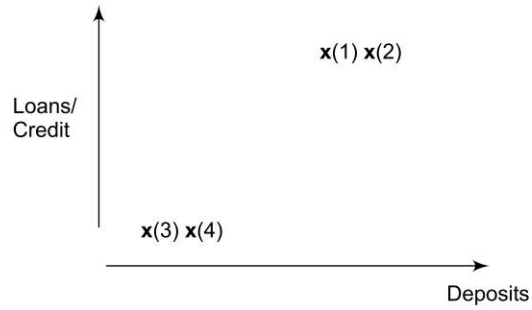
The measure to define similarity between two entities, i and j, is computed as

$$D_{ij} = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots\ldots\ldots + (x_{ik} - x_{jk})^2$$

Smaller the value of D_{ij}, more similar are the two entities.

The basic method of clustering is illustrated through a simple example given as follows:

Let there be four branches of a commercial bank each described by two variables viz. deposits and loans/credits. The following chart gives an idea of their deposits and loans/credits:



From the above chart, it is obvious that if we want two clusters, we should group the branches 1 and 2 (High Deposit, High Credit) into one cluster, and 3 and 4 (Low Deposit Low Credit) into another, since such grouping produces the clusters for which the entities (branches) within each other are most similar. However, this graphical approach is not convenient for more than two variables.

In order to develop a mathematical procedure for forming the clusters, we need a criterion upon which to judge alternative clustering patterns. This criterion defines the optimal number of entities within each cluster:

Now we shall illustrate the methodology of using distances among the entities from clusters. We assume the following distance similarity matrix among three entities:

Distance or Similarity Matrix

	1	*2*	*3*
1	0	5	10
2	5	0	8
3	10	8	0

The possible clusters and their respective distances are:

Cluster I	*Cluster II*	*Distance Within Two Clusters*	*Distance Between Two Clusters*	*Total Distance Among These Entities*
1	2 and 3	8	15 (= 5 + 10)	23
2	1 and 3	10	13 (= 5 + 8)	23
3	1 and 2	5	18 (= 8 + 10)	23

Thus, the best clustering would be to cluster entities 1 and 2 together. This would yield minimum distance within clusters (= 5), and simultaneously the maximum distance between clusters (=18). Obviously, if the number of entities is large, it is a prohibitive task to construct every possible cluster pattern, compute each 'within cluster distance' and select the pattern which yields the minimum. If the number of variables and dimensions are large, computers are needed.

The criterion of minimising the within cluster distances to form the best possible grouping to form 'k' clusters assumes that 'k' clusters are to be formed. If the number of clusters to be formed is not fixed a priori, the criterion will not specify the optimal number of clusters. However, if the objective is, to minimise the sum of the within cluster distances and the number of clusters is free to vary, then all that is required is to make each entity its own cluster, and the sum of the within cluster distances will be zero. Obviously, more the number of clusters lesser will be the sum of 'within cluster distances'. Thus, making each entity its own cluster is of no value. This issue is, therefore, resolved intuitively.

Discriminant and Cluster Analysis

It may be noted that though both discriminant analysis and cluster analysis are classification techniques. However, there is a basic difference between the two techniques. In DA, the data is classified in given set of categories using some prior information about the data. The entire rules of classification are based on the categorical dependent variable and the tolerance of the model. But the cluster analysis does not assume any dependent variable. It uses different methods of classification to classify the data into some groups without any prior information. The cases with similar data would be in the same group, and the cases with distinct data would be classified in different groups.

Most computer-oriented programs find the optimum number of clusters through their own algorithm. We have described the methods of forming clusters in Section 14.8.3, and the use of SPSS package in Section 14.8.5.

14.8.1 Some Applications of Cluster Analysis

Following are two applications of cluster analysis in the banking system:

(i) Commercial Bank In one of the banks in India, all its branches were formed into clusters by taking 15 variables representing various aspects of the functioning of the branches. The variables are: four types of deposits, four types of advances, miscellaneous business such as 'drafts issued', receipts on behalf of government, foreign exchange business, etc. The bank uses this grouping of branches into clusters for collecting information or conducting quick surveys to study any aspect, planning, analysing and monitoring. If any sample survey is to be conducted, it is ensured that samples are taken from branches in all the clusters so as to get a true representative of the entire bank. Since the branches in the same cluster are more or less similar to each other, only few branches are selected from each cluster.

(ii) Agricultural Clusters A study was conducted by one of the officers of the Reserve Bank of India, to form clusters of geographic regions of the country based on agricultural parameters like cropping pattern, rainfall, land holdings, productivity, fertility, use of fertilisers, irrigation facilities, etc. The whole country was divided into 9 clusters. Thus, all the 67 regions of the country were allocated to these clusters. Such classification is useful for making policies at the national level as also at regional/cluster levels.

14.8.2 Key Terms in Cluster Analysis

Agglomeration Schedule	While performing cluster analysis, the tool gives information on objects or cases being combined at each stage at hierarchical clustering process. This is in-depth table which indicates the clusters and the objects combined in the cluster, the table can be read from top to bottom. The table starts with any two cases combined together it also states 'Distance Coefficients' and 'Stage Cluster First Appears'. The distance coefficient is an important measure to identify the number of clusters for the data. Sudden jump in the coefficient indicates better grouping. The last row of the table represents one cluster solution, second last, two cluster solution, etc.
Cluster Centroid	Cluster centrioids are mean values of variables under consideration (variables given while clustering) for all the cases in a particular cluster. Each cluster will have different centroids for each variable.
Cluster Centre	These are initial starting points in non-hierarchical clusters. The clusters are built around these centres; these are also termed as seeds.
Cluster Membership	It is the cluster to which each case belongs. It is important to save cluster membership to analyse cluster and further perform ANOVA on the data.
Dendrogram	This is the graphical summary of the cluster solution. This is used more while interpreting results than the Agglomeration Schedule, as it is easier to interpret. The cases are listed along the left vertical axis. The horizontal axis shows the distance between clusters when they are joined. This graph gives an indication of the number of clusters the solution may have. The diagram is read from right to left. Rightmost is the single cluster solution, just before right is two cluster solution, and so on. The best solution is where the horizontal distance is maximum. This could be a subjective process.
Icicle Diagram	It displays information about how cases are combined into clusters at each iteration of the analysis.
Similarity/Distance Coefficient Matrix	It is a matrix containing the pair wise distances between the cases.

14.8.3 Clustering Procedures and Methods

The cluster analysis procedure could be:

- Hierarchical
- Non-hierarchical

Hierarchical methods develop a tree-like structure (dendrogram). These could be:

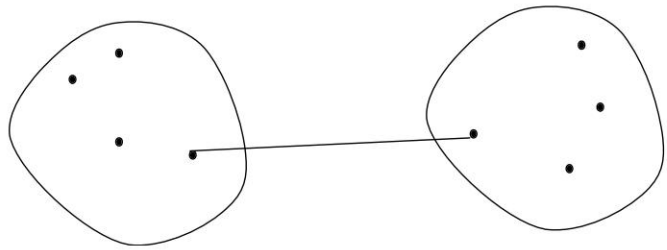
- Agglomerative – starts with each case as separate cluster and in every stage the similar clusters are combined. Ends with single cluster.
- Divisive – starts with all cases in a single cluster and then the clusters are divided on the basis of the difference between the cases. Ends with all clusters separate.

Most common methods of clustering are agglomerative methods. This could be further divided into:

- Linkage Methods – these distance measures. There are three linkage methods, Single linkage – minimum distance or nearest neighbourhood rule, Complete linkage – Maximum distance or furthest neighbourhood and Average linkage – average distance between all pair of objects. This is explained in the Diagram 14 provided in the following.
- Centroid Methods – this method considers distance between the two centroids. Centroid is the means for all the variables.
- Variance Methods – this is commonly termed as Ward's method; it uses the squared distance from the means.

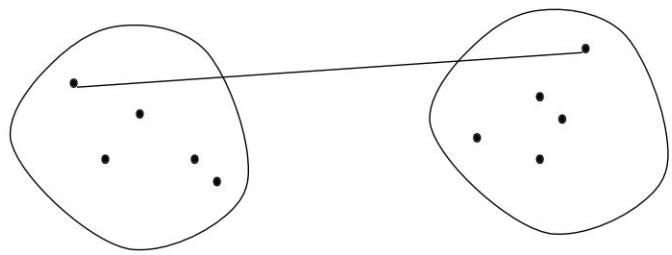
Diagram 14

(a) Single Linkage (Minimum Distance/Nearest Neighbourhood)



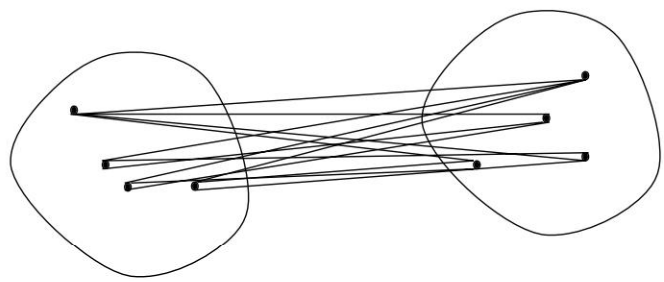
Cluster1 Cluster 2

(b) Complete Linkage (Maximum Distance/Furthest Neighbourhood)



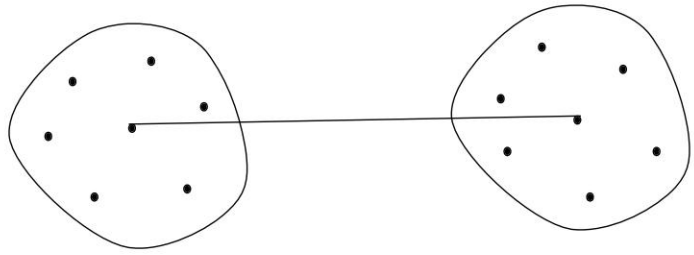
Cluster1 Cluster 2

(c) Average Linkage (Average Distance)



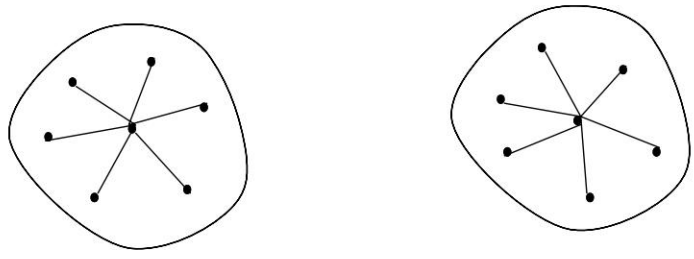
Cluster1 Cluster 2

(d) Centroid Method



Cluster1 Cluster 2

(e) Ward's Method (Variance Method)



Cluster1 Cluster 2

Non-hierarchical is frequently termed as K-means clustering.

It may be noted that each of the above methods may give different results and different set of clusters. The selection of method is generally done on the basis of the clarity of cluster solution. If a particular method does not give appropriate or clear results, other method may be tried. Cluster analysis is a trial and error process (especially if it is hierarchical cluster). There are no specific tests for finding the validity of the results. ANOVA can give the validity of distinction between the clusters.

14.8.4 Assumptions of Cluster Analysis

Since distance measures are used in cluster analysis, it assumes that the variables have similar means. i.e. the variables are on the same unit or dimension. If all the variables are rating questions and the scales of these ratings are same, then this assumption is satisfied. But if the variables have different dimensions, for example, one variable is salary, other is age, some are rating on 1 to 7 scale, then this difference may affect the clustering. This problem is solved through Standardisation. Standardisation allows one to equalise the effect of variables measured on different scales.

14.8.5 Cluster Analysis Using SPSS

For illustration we will be considering the following case study:

The file Commodity.sav is provided in the CD with the book.

CASE 14.2

INVESTMENT AWARENESS AND PERCEPTIONS

A study of investment awareness and perceptions was undertaken with the aim of achieving a better understanding of investor's behaviour. The main objective was to conduct an analysis of the various investment options available, behaviour patterns and perceptions of the investors in relation to the available options. For simplicity, four of the most popular investment options were selected for analysis, viz.

- Investment in Commodity Markets
- Investment in Stock Markets
- Investment in Fixed Deposits
- Investment in Mutual Funds

Special focus was on the study of the levels of awareness among the investors about the commodity markets and also perceptional ratings of the investment options by the investors.

This study was undertaken with an intention to gain a better perspective on investor behaviour patterns and to provide assistance to the general public, individual/small investors, broker's and portfolio managers to analyse the scope of investments and make informed decisions while investing in the above mentioned options. However, the limitation of the study is that it considers investors from Mumbai only, and hence, might not be representative of the entire country.

An *investment* is a commitment of funds made in the expectation of some positive rate of return. If properly undertaken, the return will be commensurate with the risk the investor assumes.

An analysis of the backgrounds and perceptions of the investors was undertaken in the report. The data used in the analysis was collected by e-mailing and distributing the questionnaire among friends, relatives and colleagues. 45 people were surveyed, and were asked various questions relating to their backgrounds and knowledge about the investment markets and options. The raw data contains a wide range of information, but only the data which is relevant to objective of the study was considered.

The questionnaire used for the study is as follows:

Questionnaire

Age: _____

Occupation:

- SELF EMPLOYED
- GOVERNMENT
- STUDENT
- HOUSEWIFE
- DOCTOR
- ENGINEER
- CORPORATE PROFESSIONAL
- OTHERS (PLEASE SPECIFY) : _____

Gender:

- MALE
- FEMALE

1. RATE FOLLOWING ON THE PREFERENCE OF INVESTMENT {**1- Least Preferred 5- Most Preferred (√ Appropriate Number)**}

Question	*1*	*2*	*3*	*4*	*5*
COMMODITY MARKET					
STOCK MARKET					
FIXED DEPOSITS					
Mutual Funds					

2. HOW MUCH ARE YOU READY TO INVEST IN COMMODITY MARKET? _____
3. HOW MUCH ARE YOU READY TO INVEST IN STOCK MARKET ? _____
4. HOW MUCH ARE YOU READY TO INVEST IN FIXED DEPOSITS? _____
5. HOW MUCH ARE YOU READY TO INVEST IN MUTUAL FUNDS? _____
6. FOR HOW MUCH TIME WOULD YOU BLOCK YOUR MONEY WITH INVESTMENTS?
   - UNDER 5 MONTHS
   - 6–12 MONTHS
   - 1–3 YEARS
   - MORE THAN 3 YEARS
7. ON A SCALE OF 1 –10 (1- LEAST RISKY and 10 - MOST RISKY), HOW RISKY DO YOU THINK IS THE COMMODITY MARKET? (Circle the appropriate number)

SAFE RISKY

8. ON A SCALE OF 1-10, HOW RISKY DO YOU THINK IS THE STOCK MARKET?

SAFE RISKY

9. ON A SCALE OF 1-10, HOW RISKY DO YOU THINK IS FIXED DEPOSITS?

SAFE RISKY

10. ON A SCALE OF 1-10 HOW RISKY DO YOU THINK ARE MUTUAL FUNDS?

SAFE RISKY

The collected data is given below:

Sr No	Age	Occu	Sex	Rate CM	Rate SM	Rate FD	Rate MF	Inv CM	Inv SM	Inv FD	Inv MF	Block Money	Risky CM	Risky SM	Risky FD	Risky MF
1	23	1	1	3	3	3	4	6000	8000	3000	5000	2	5	6	1	7
2	18	1	1	4	4	2	4	4000	5000	0	8000	2	7	9	3	6
3	54	1	2	1	2	4	3	50000	60000	200000	50000	8	8	8	1	4
4	42	1	1	2	2	4	2	4000	10000	8000	20000	6	8	7	2	4
5	46	1	2	1	1	4	2	.	35000	75000	15000	7	8	8	1	4
6	28	1	2	2	3	5	3	0	25000	50000	50000	2	6	8	2	7
7	23	1	1	3	4	2	5	50000	150000	70000	175000	4	3	3	1	3
8	32	1	1	2	2	5	3	35000	150000	75000	180000	4	6	6	1	6
9	49	1	2	1	2	4	3	25000	185000	100000	155000	7	6	7	1	7
10	29	2	1	2	3	5	3	5000	7000	4000	3000	2	4	6	1	5
11	18	2	1	4	4	2	5	0	5000	4000	8000	2	8	7	3	5
12	22	2	2	4	4	4	3	5000	3000	1000	2500	4	3	6	2	3
13	44	2	1	1	2	4	4	0	8000	2500	6000	6	9	3	2	4
14	18	2	2	5	5	2	5	1500	5000	2000	5000	2	7	4	2	2
15	21	2	2	3	3	3	4	0	5000	6000	7500	3	6	8	2	7
16	35	2	2	3	2	5	4	1500	2500	5000	3000	5	5	7	2	6
17	22	2	1	3	4	2	4	1500	5000	0	5000	1	4	8	2	6
18	40	2	1	1	3	5	2	7500	20000	11000	35000	6	3	6	2	5
19	21	2	1	3	3	4	4	20000	12500	15000	20000	2	7	7	2	4
20	25	2	1	2	3	3	5	3000	20000	5000	22000	3	5	1	2	3
21	22	2	2	3	2	3	3	2500	13500	27000	31000	4	3	9	2	5
22	25	2	2	2	2	3	2	0	1500	60000	3500	4	3	7	1	8
23	19	2	2	5	5	3	5	1500	24000	18000	26000	1	5	8	1	7
24	19	2	2	5	4	5	2	0	17500	25000	13500	2	7	5	1	10
25	29	3	2	1	1	5	3	0	1500	10000	3000	4	9	8	2	8
26	27	3	2	2	3	4	1	0	4500	7000	0	1	6	9	1	10
27	50	3	2	1	2	5	2	0	5000	3500	5000	7	7	7	1	9
28	24	3	2	3	3	2	4	1000	5000	2000	8000	1	9	8	1	3
29	19	4	2	4	4	3	2	3500	45000	6000	3000	1	4	2	1	7
30	36	4	2	1	2	5	3	0	7500	30000	15000	5	7	6	1	9
31	21	4	1	3	4	4	5	30000	50000	10000	55000	3	2	5	2	3
32	18	4	1	4	4	2	5	30000	35000	3000	75000	1	5	3	3	5
33	24	4	1	3	4	3	5	8000	60000	15000	45000	4	4	4	2	3

(Contd)

(*Contd*)

34	27	4	2	2	3	5	5	0	100000	25000	40000	3	7	7	1	4
35	34	4	2	2	2	4	3	23000	45000	25000	35000	5	6	9	1	9
36	48	4	1	2	2	5	3	10000	35000	45000	15000	7	6	9	1	8
37	33	4	1	2	3	5	3	75000	90000	55000	125000	5	4	7	1	8
38	25	4	1	2	3	3	5	55000	70000	60000	22000	4	2	6	1	5
39	55	4	1	1	2	4	2	45000	60000	75000	75000	8	8	8	1	7
40	18	4	1	4	5	2	2	0	30000	7500	9000	3	4	2	3	5
41	50	4	1	2	2	4	3	40000	55000	70000	75000	6	7	8	1	7
42	52	4	1	2	3	4	2	35000	50000	70000	70000	6	6	8	1	8
43	23	1	1	3	4	2	5	50000	150000	70000	175000	4	3	3	1	3
44	32	1	1	2	2	4	3	35000	150000	75000	180000	4	6	6	1	6
45	45	1	2	1	2	4	3	25000	185000	100000	155000	6	6	7	2	7

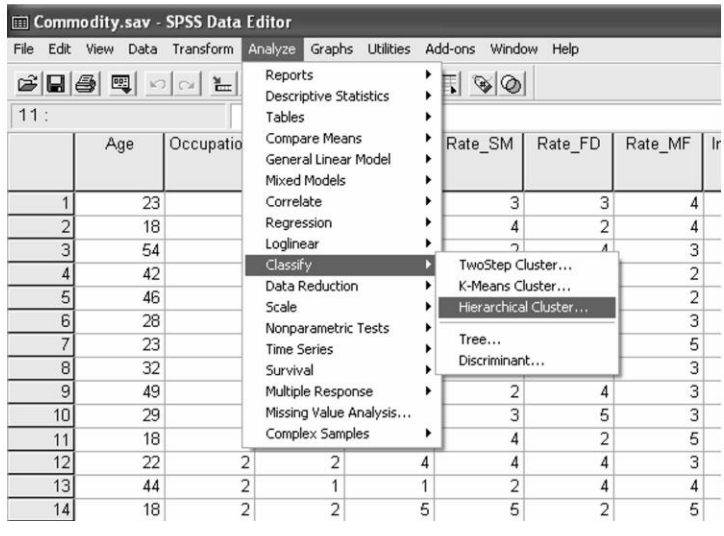
Perform the appropriate analysis to find different groups of investors.

We shall use the Commodity.sav file to conduct cluster analysis.

We will start with the hierarchical cluster analysis. K-means cluster will be explained later.

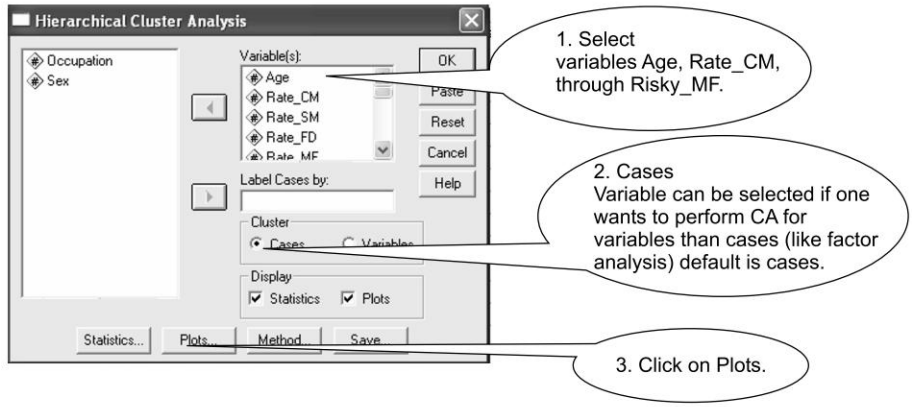
After opening the file Commodity.sav, one can click on Analyze – Classify and Hierarchical Cluster as shown in the following snapshot:

CA Snapshot 1



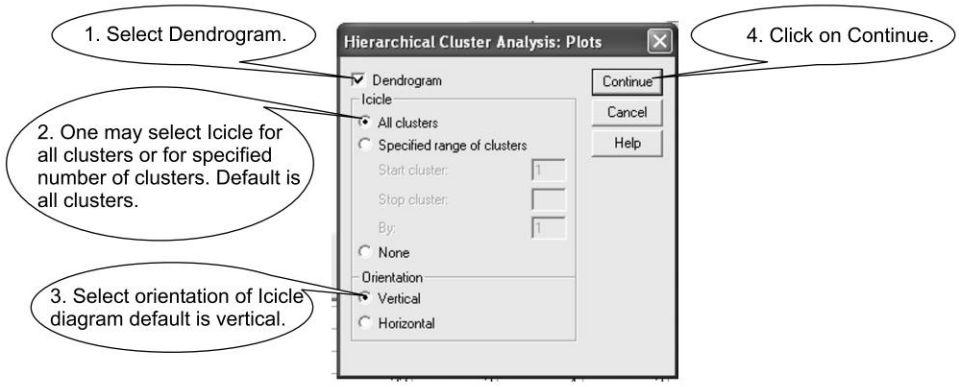
The window that will be opened is shown below:

CA Snapshot 2



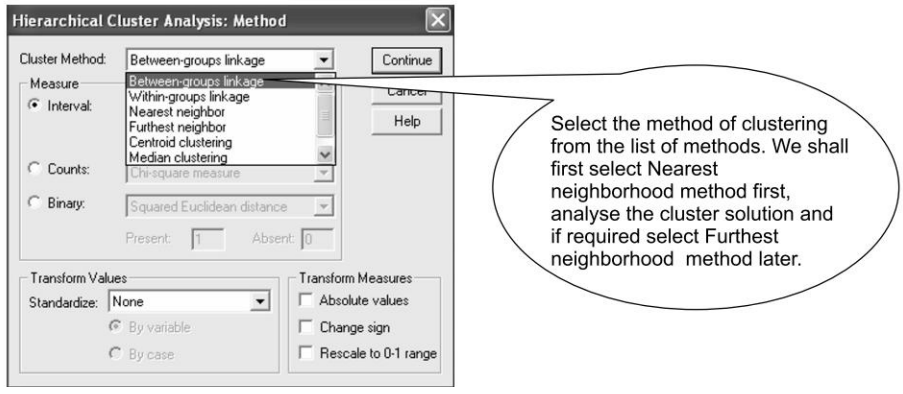
The following window will be opened:

CA Snapshot 3



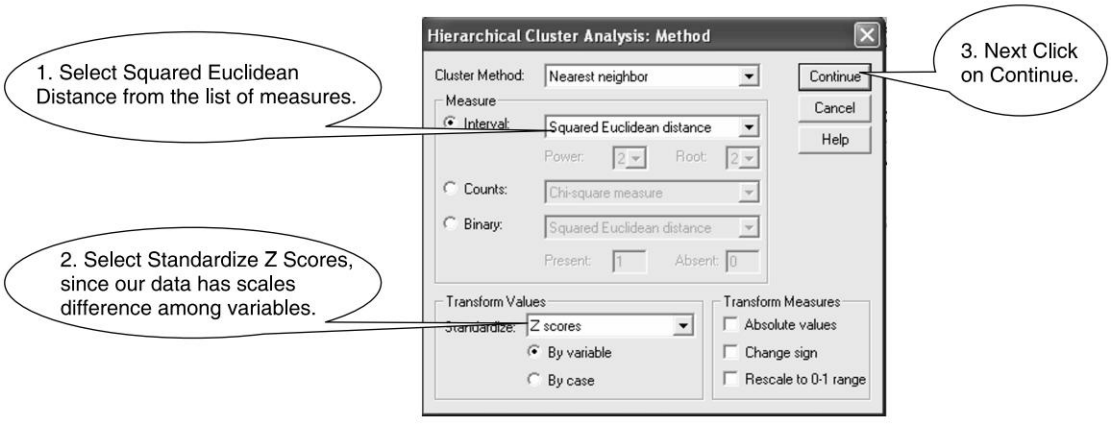
SPSS will take back to the window as displayed in CA Snapshot 2. At this stage, click on 'Method'. SPSS will open the following window:

CA Snapshot 4



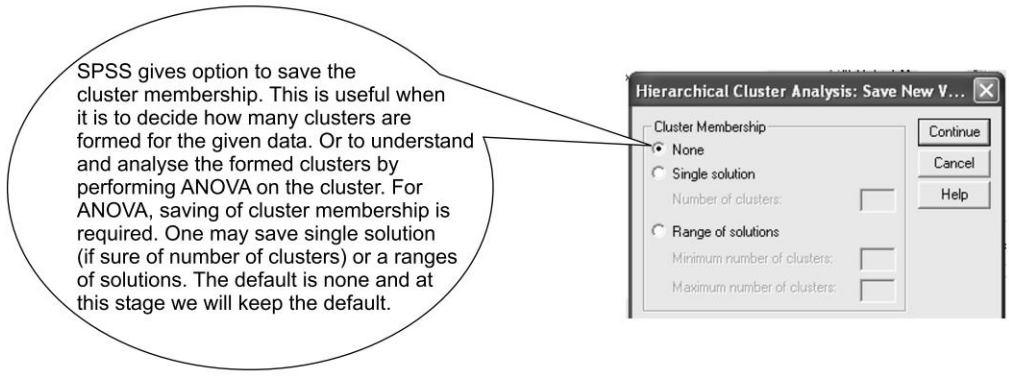
Next step is to select the clustering measure. The most common measure is the squared Euclidean distance.

CA Snapshot 5



SPSS will be back to the window as shown in CA Snapshot 2. At this stage, click on Save, the following window will be displayed.

CA Snapshot 6



Click on continue, and SPSS will be back as shown in CA Snapshot 2. At this stage, click OK. The following output will be displayed.
We shall discuss this output in detail.

Proximities

Case Processing Summary[a]

Cases					
Valid		*Missing*		*Total*	
N	*Per cent*	*N*	*Per cent*	*N*	*Per cent*
44	97.8%	1	2.2%	45	100.0%

a. Squared Euclidean Distance used

This table gives the case processing summary and its percentages. The above table indicates there are 44 out of 45 valid cases. Since one case has some missing values, it is ignored from the analysis.

Cluster

Single Linkage

This is the method we selected for cluster analysis.

Agglomeration Schedule

	Cluster Combined			*Stage Cluster First Appears*		
Stage	*Cluster 1*	*Cluster 2*	*Coefficients*	*Cluster 1*	*Cluster 2*	*Next Stage*
1	6	42	.000	0	0	43
2	7	43	.829	0	0	20
3	2	10	2.190	0	0	16
4	40	41	2.361	0	0	6
5	8	44	2.636	0	0	20
6	38	40	3.002	0	4	15
7	30	32	3.749	0	0	14
8	1	14	3.808	0	0	10
9	26	29	3.891	0	0	12
10	1	16	4.145	8	0	13
11	34	35	4.326	0	0	12
12	26	34	4.587	9	11	15
13	1	18	4.698	10	0	16
14	19	30	5.105	0	7	23
15	26	38	5.751	12	6	19
16	1	2	5.921	13	3	17
17	1	20	6.052	16	0	18
18	1	9	6.236	17	0	21
19	24	26	6.389	0	15	32
20	7	8	6.791	2	5	37
21	1	5	7.298	18	0	22
22	1	15	7.481	21	0	24
23	11	19	7.631	0	14	31
24	1	21	7.711	22	0	26
25	4	12	7.735	0	0	28
26	1	13	8.289	24	0	27
27	1	27	8.511	26	0	28
28	1	4	8.656	27	25	29
29	1	22	8.807	28	0	30

(*Contd*)

(*Contd*)

30	1	33	8.994	29	0	31
31	1	11	9.066	30	23	32
32	1	24	9.071	31	19	33
33	1	17	9.245	32	0	34
34	1	28	9.451	33	0	35
35	1	31	9.483	34	0	36
36	1	37	9.946	35	0	38
37	7	36	9.953	20	0	38
38	1	7	10.561	36	37	39
39	1	25	10.705	38	0	40
40	1	23	11.289	39	0	41
41	1	39	12.785	40	0	42
42	1	3	12.900	41	0	43
43	1	6	15.449	42	1	0

This table gives the agglomeration schedule or the details of the clusters formed in each stage. This table indicates that the cases 6 and 42 were combined at first stage. Cases 7 and 43 were combined at second stage, 2 and 10 were combined at third stage, and so on. The last stage (stage 43) indicates two cluster solution. One above last stage (stage 42) indicates three cluster solution, and so on. The column Coefficients indicates the distance coefficient. Sudden increase in the coefficient indicates that the combining at that stage is more appropriate. This is one of the indicators for deciding the number of clusters.

Agglomeration Schedule

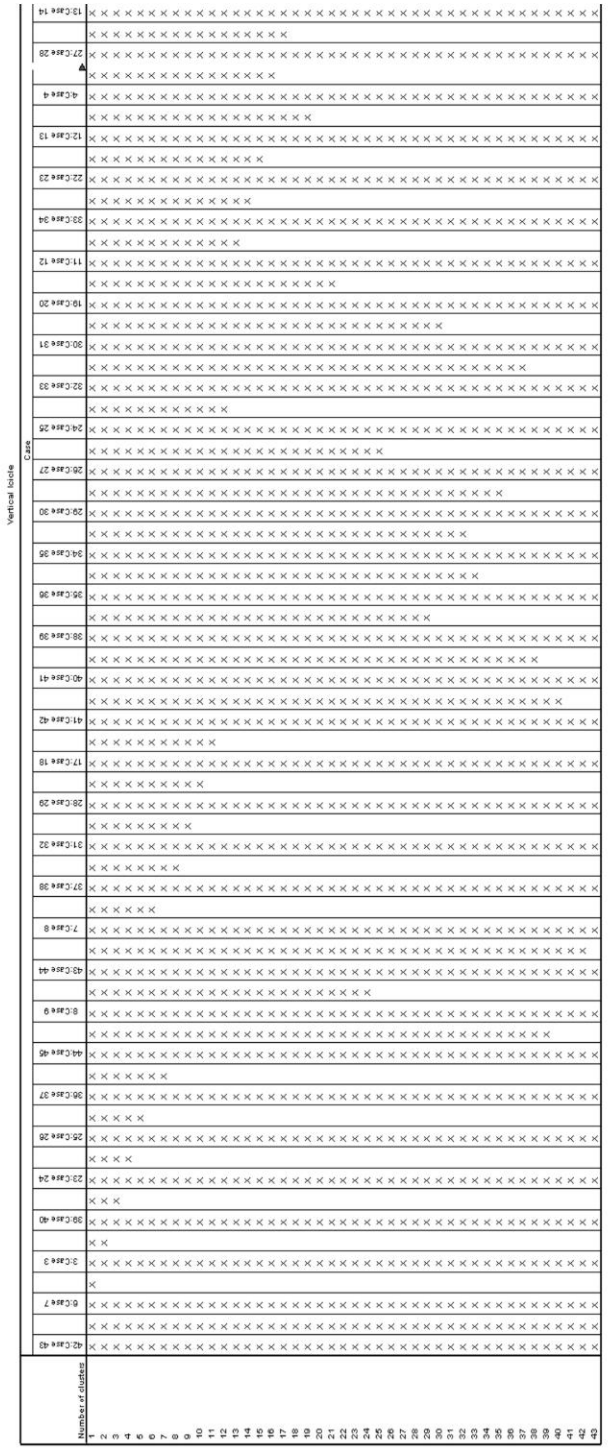
Stage	*Cluster Combined*		*Coefficients*	*Stage Cluster First Appears*		*Next Stage*	*Difference in Coefficients*
	Cluster 1	*Cluster 2*		*Cluster 1*	*Cluster 2*		
1	6	42	0	0	0	43	
2	7	43	0.828734	0	0	20	0.828734
3	2	10	2.189939	0	0	16	1.361204
4	40	41	2.360897	0	0	6	0.170958
5	8	44	2.636238	0	0	20	0.275341
6	38	40	3.002467	0	4	15	0.366229
7	30	32	3.749321	0	0	14	0.746854
8	1	14	3.808047	0	0	10	0.058726
9	26	29	3.89128	0	0	12	0.083232
10	1	16	4.1447	8	0	13	0.253421
11	34	35	4.325831	0	0	12	0.181131
12	26	34	4.587371	9	11	15	0.26154

(*Contd*)

(*Contd*)

13	1	18	4.697703	10	0	16	0.110332
14	19	30	5.105442	0	7	23	0.407739
15	26	38	5.750915	12	6	19	0.645473
16	1	2	5.921352	13	3	17	0.170437
17	1	20	6.052442	16	0	18	0.13109
18	1	9	6.236206	17	0	21	0.183764
19	24	26	6.389243	0	15	32	0.153037
20	7	8	6.790893	2	5	37	0.40165
21	1	5	7.297921	18	0	22	0.507028
22	1	15	7.480892	21	0	24	0.182971
23	11	19	7.631185	0	14	31	0.150293
24	1	21	7.710566	22	0	26	0.079381
25	4	12	7.735374	0	0	28	0.024808
26	1	13	8.288569	24	0	27	0.553195
27	1	27	8.510957	26	0	28	0.222388
28	1	4	8.656467	27	25	29	0.14551
29	1	22	8.807405	28	0	30	0.150937
30	1	33	8.994409	29	0	31	0.187004
31	1	11	9.066141	30	23	32	0.071733
32	1	24	9.070588	31	19	33	0.004447
33	1	17	9.244565	32	0	34	0.173977
34	1	28	9.450741	33	0	35	0.206176
35	1	31	9.483015	34	0	36	0.032274
36	1	37	9.946286	35	0	38	0.463272
37	7	36	9.953277	20	0	38	0.00699
38	1	7	10.56085	36	37	39	0.607572
39	1	25	10.70496	38	0	40	0.144109
40	1	23	11.28888	39	0	41	0.583924
41	1	39	12.78464	40	0	42	1.495759
42	1	3	12.90033	41	0	43	0.115693
43	1	6	15.44882	42	1	0	2.548489

We have replicated the table with one more column added called “Difference in the coefficients”, this is the difference in the coefficient between the current solution and the previous solution. The highest difference gives the most likely clusters. In the above table, the highest difference is 2.548 which is for 2-cluster solution. The next highest difference 1.4956 and is for 3-cluster solution. This indicates that there could be 3 clusters for the data.


Vertical Icicle
Case
Number of clusters

The icicle table also gives the summery of the cluster formation. It is read from bottom to top. The topmost is the single cluster solution and bottommost is all cases separate. The cases in the table are in the columns. The first column indicates the number of clusters for that stage. Each case is separated by an empty column. A 'cross' in the empty column means the two cases are combined. A 'gap' means the two cases are in separate clusters.

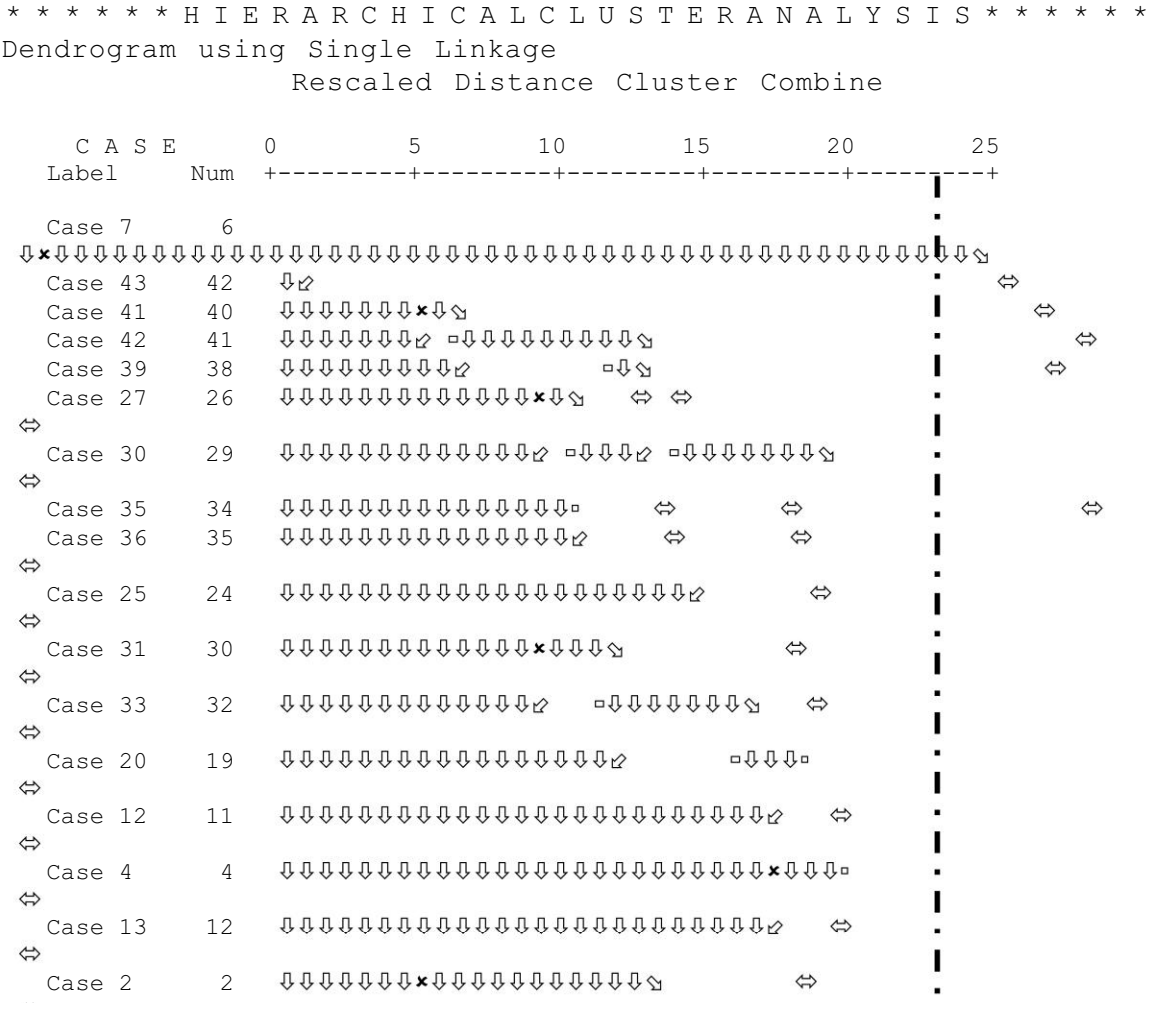
If the number of cases is huge, this table becomes difficult to interpret.

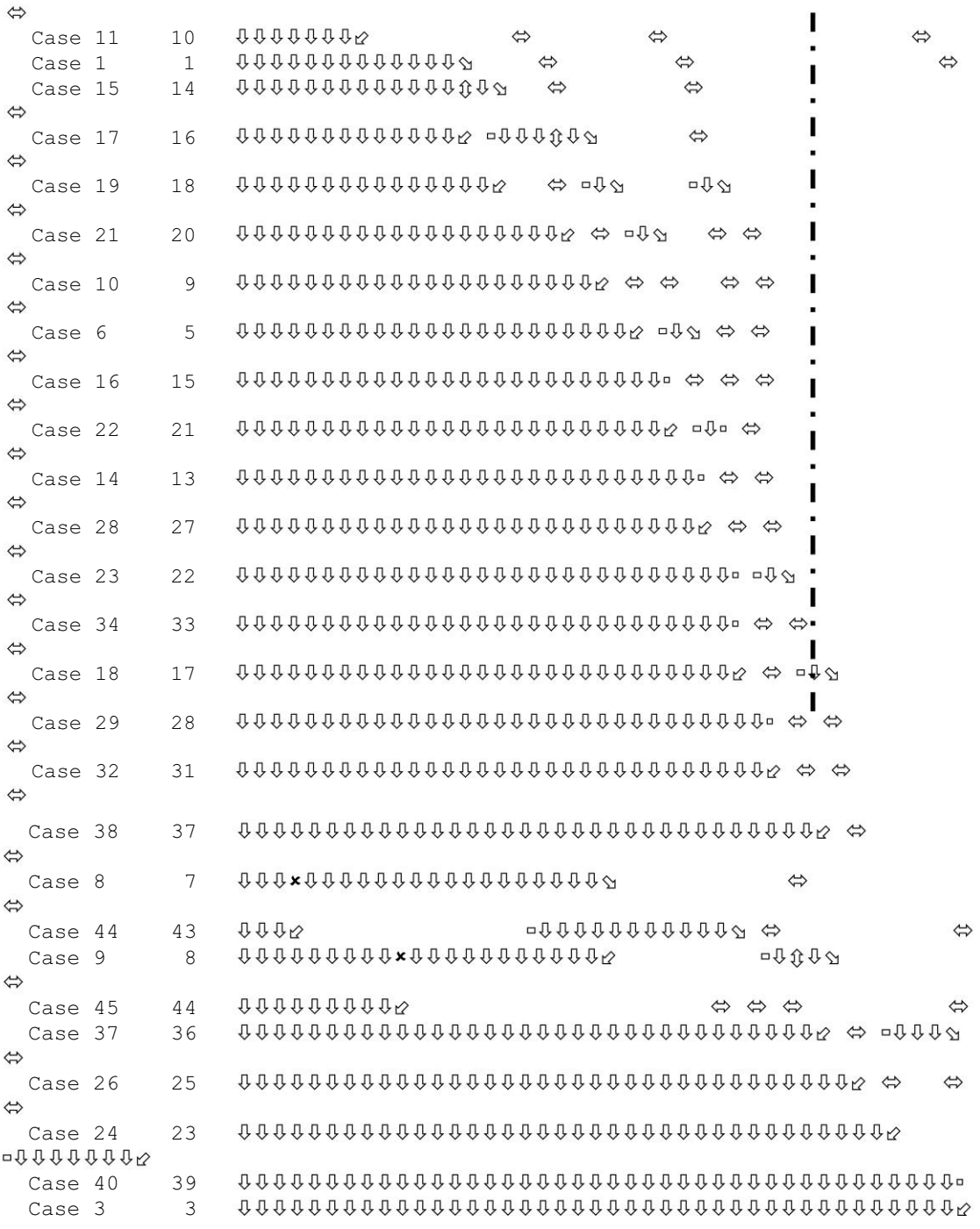
The diagram given below is called the dendrogram. A dendrogram is the most used tool to understand the number of clusters and cluster memberships. The cases are in the first column and they are connected by lines for each stage of clustering. This graph is from left to right—the leftmost is all cluster solution and rightmost is the one cluster solution.

This graph also has the distance line from 0 to 25. More is the width of the horizontal line for the cluster, more appropriate is the cluster.

The graph shows that 2-cluster solution is a better solution indicated by the thick dotted line.

Dendrogram




Case 11 10
Case 1 1
Case 15 14
Case 17 16
Case 19 18
Case 21 20
Case 10 9
Case 6 5
Case 16 15
Case 22 21
Case 14 13
Case 28 27
Case 23 22
Case 34 33
Case 18 17
Case 29 28
Case 32 31
Case 38 37
Case 8 7
Case 44 43
Case 9 8
Case 45 44
Case 37 36
Case 26 25
Case 24 23
Case 40 39
Case 3 3

The above solution is not decisive as the differences are very close. Hence, we shall try a different method, i.e. furthest neighbourhood.

The entire process is repeated and this time the method (as shown in CA Snapshot 4) selected is the furthest neighbourhood.

The output is as follows:

Proximities

Case Processing Summary[a]

Cases					
Valid		*Missing*		*Total*	
N	*Per cent*	*N*	*Per cent*	*N*	*Per cent*
44	97.8%	1	2.2%	45	100.0%

a. Squared Euclidean Distance used

Cluster

Complete Linkage

Agglomeration Schedule

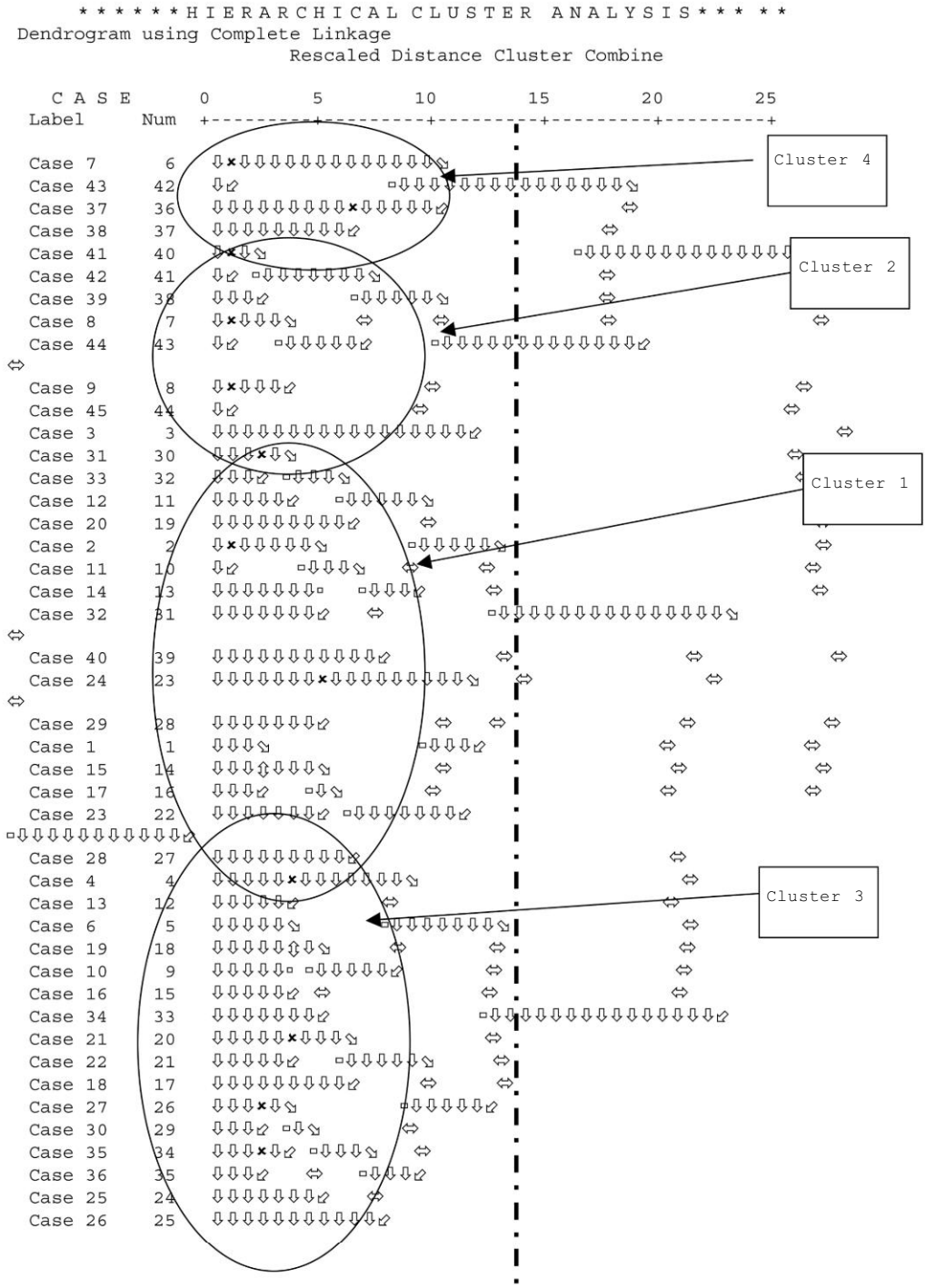
	Cluster Combined			*Stage Cluster First Appears*		
Stage	*Cluster 1*	*Cluster 2*	*Coefficients*	*Cluster 1*	*Cluster 2*	*Next Stage*
1	6	42	.000	0	0	35
2	7	43	.829	0	0	16
3	2	10	2.190	0	0	25
4	40	41	2.361	0	0	10
5	8	44	2.636	0	0	16
6	30	32	3.749	0	0	15
7	1	14	3.808	0	0	11
8	26	29	3.891	0	0	19
9	34	35	4.326	0	0	19
10	38	40	4.386	0	4	31
11	1	16	5.796	7	0	22
12	5	18	7.298	0	0	18
13	20	21	7.711	0	0	26
14	4	12	7.735	0	0	33
15	11	30	7.770	0	6	28

(*Contd*)

(*Contd*)

16	7	8	7.784	2	5	31
17	9	15	7.968	0	0	18
18	5	9	8.697	12	17	23
19	26	34	8.830	8	9	24
20	23	28	11.289	0	0	37
21	13	31	11.457	0	0	25
22	1	22	11.552	11	0	29
23	5	33	12.218	18	0	33
24	24	26	13.013	0	19	32
25	2	13	13.898	3	21	30
26	17	20	14.285	0	13	36
27	36	37	14.858	0	0	35
28	11	19	14.963	15	0	34
29	1	27	16.980	22	0	37
30	2	39	18.176	25	0	34
31	7	38	18.897	16	10	38
32	24	25	19.342	24	0	36
33	4	5	21.241	14	23	40
34	2	11	25.851	30	28	39
35	6	36	26.366	1	27	41
36	17	24	26.691	26	32	40
37	1	23	29.225	29	20	39
38	3	7	30.498	0	31	41
39	1	2	36.323	37	34	42
40	4	17	37.523	33	36	42
41	3	6	55.294	38	35	43
42	1	4	63.846	39	40	43
43	1	3	87.611	42	41	0

Dendrogram


******HIERARCHICAL CLUSTER ANALYSIS*** **
Dendrogram using Complete Linkage
Rescaled Distance Cluster Combine
CASE
Label Num
0 5 10 15 20 25
Case 7 6
Case 43 42
Case 37 36
Case 38 37
Case 41 40
Case 42 41
Case 39 38
Case 8 7
Case 44 43
Case 9 8
Case 45 44
Case 3 3
Case 31 30
Case 33 32
Case 12 11
Case 20 19
Case 2 2
Case 11 10
Case 14 13
Case 32 31
Case 40 39
Case 24 23
Case 29 28
Case 1 1
Case 15 14
Case 17 16
Case 23 22
Case 28 27
Case 4 4
Case 13 12
Case 6 5
Case 19 18
Case 10 9
Case 16 15
Case 34 33
Case 21 20
Case 22 21
Case 18 17
Case 27 26
Case 30 29
Case 35 34
Case 36 35
Case 25 24
Case 26 25
Cluster 4
Cluster 2
Cluster 1
Cluster 3

The above dendrogram clearly shows that the **longest horizontal lines** for clusters are for 4-cluster solution, shown by thick dotted line (the dotted line intersects four horizontal lines). It indicates that the cluster containing cases 7, 43, 37 and 38, named as cluster 4 the cluster containing 41, 42, 39, 8, 44, 9, 45 and 3, named as cluster 2, and so on.

We shall run the cluster analysis again with same method and this time we shall save the cluster membership for single solution = '4' clusters as indicated in CA Snapshot 6.

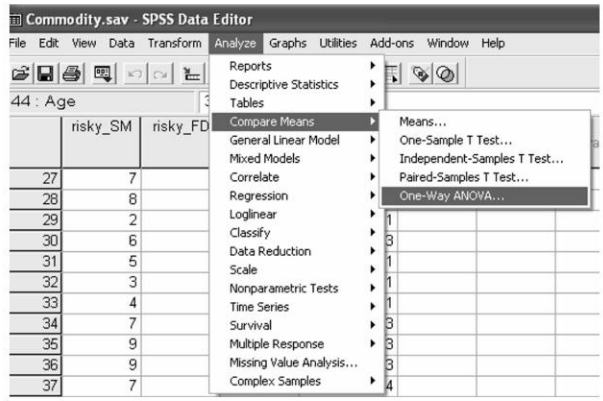
The output will be same as discussed except a new variable is added in the SPSS file with name '**CLU4_1**'. This variable takes value between 1 to 4 each value indicates the cluster membership.

We shall conduct ANOVA test on the data where the dependent variables are taken as all the variables that were included while performing cluster analysis, and the factor is the cluster membership indicated by variable **CLU4_1.** This ANOVA will indicate if the clusters really distinguish on the basis of the list of variables, which variables significantly distinguish the clusters and which do not distinguish.

The ANOVA procedure is as follows:

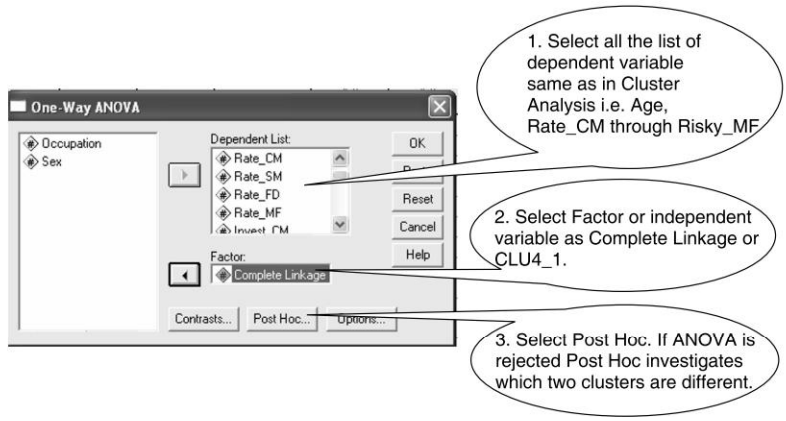
Select 'Analyze' – Compare Means – One way ANOVA from the menu as shown below.

CA Snapshot 7



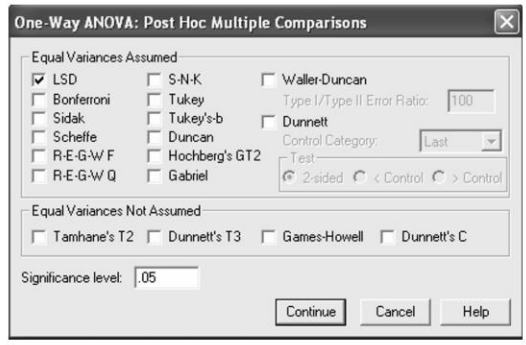
SPSS will open the following window:

CA Snapshot 8



The following window will be opened:

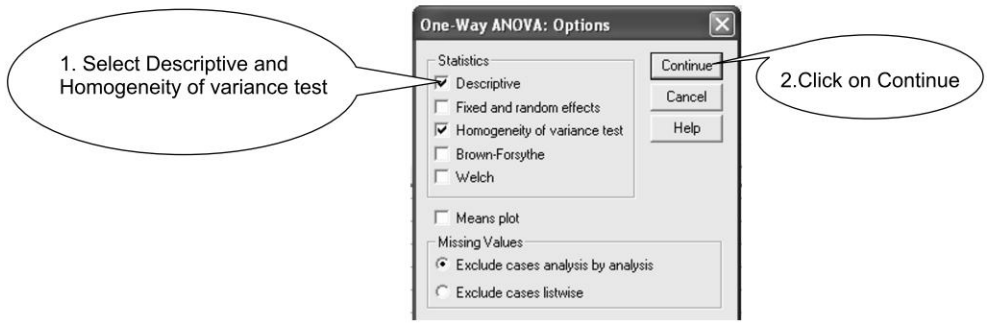
CA Snapshot 9



This gives list of Post Hoc tests for ANOVA. Most common are LSD and HSD (discussed in Chapter 12) we shall select LSD and click on continue.

SPSS will take back to CA Snapshot 8. Click on Options and the following window will be opened.

CA Snapshot 10



SPSS will take back to window as shown in CA Snapshot 8, at this stage click on OK.

The following output will be displayed:

Oneway

Descriptives

						95% Confidence Interval for Mean			
		N	*Mean*	*Std. Deviation*	*Std. Error*	*Lower Bound*	*Upper Bound*	*Mini-mum*	*Maxi-mum*
Age	1	16	20.56	2.502	0.626	19.23	21.90	18	25
	2	8	46.13	9.250	3.270	38.39	53.86	32	55

(*Contd*)

(*Contd*)

	3	16	33.56	9.063	2.266	28.73	38.39	21	50
	4	4	26.00	4.761	2.380	18.42	33.58	23	33
	Total	44	30.43	11.571	1.744	26.91	33.95	18	55
Rate_CM	1	16	3.69	0.873	0.218	3.22	4.15	2	5
	2	8	1.50	0.535	0.189	1.05	1.95	1	2
	3	16	1.88	0.719	0.180	1.49	2.26	1	3
	4	4	2.50	0.577	0.289	1.58	3.42	2	3
	Total	44	2.52	1.171	0.177	2.17	2.88	1	5
Rate_SM	1	16	3.94	0.680	0.170	3.58	4.30	3	5
	2	8	2.13	0.354	0.125	1.83	2.42	2	3
	3	16	2.31	0.602	0.151	1.99	2.63	1	3
	4	4	3.50	0.577	0.289	2.58	4.42	3	4
	Total	44	2.98	1.000	0.151	2.67	3.28	1	5
Rate_FD	1	16	2.81	0.911	0.228	2.33	3.30	2	5
	2	8	4.13	0.354	0.125	3.83	4.42	4	5
	3	16	4.44	0.727	0.182	4.05	4.83	3	5
	4	4	3.00	1.414	0.707	0.75	5.25	2	5
	Total	44	3.66	1.098	0.166	3.33	3.99	2	5
Rate_MF	1	16	4.00	1.155	0.289	3.38	4.62	2	5
	2	8	2.75	0.463	0.164	2.36	3.14	2	3
	3	16	2.94	0.998	0.249	2.41	3.47	1	5
	4	4	4.50	1.000	0.500	2.91	6.09	3	5
	Total	44	3.43	1.149	0.173	3.08	3.78	1	5
Invest_CM	1	16	5937.50	9681.382	2420.346	778.66	11096.34	0	30000
	2	8	36250.00	8762.746	3098.098	28924.16	43575.84	25000	50000
	3	16	4593.75	7289.762	1822.440	709.31	8478.19	0	23000
	4	4	57500.00	11902.381	5951.190	38560.66	76439.34	50000	75000
	Total	44	15647.73	19901.671	3000.290	9597.07	21698.39	0	75000
Invest_SM	1	16	20156.25	18716.943	4679.236	10182.70	30129.80	3000	60000
	2	8	111875.00	60999.854	21566.705	60877.85	162872.15	50000	185000
	3	16	18656.25	24993.145	6248.286	5338.34	31974.16	1500	100000
	4	4	115000.00	41231.056	20615.528	49392.19	180607.81	70000	150000
	Total	44	44909.09	53293.535	8034.303	28706.38	61111.81	1500	185000
Invest_FD	1	16	6718.75	7061.560	1765.390	2955.91	10481.59	0	25000
	2	8	95625.00	43951.069	15539.049	58880.99	132369.01	70000	200000
	3	16	20500.00	18071.156	4517.789	10870.56	30129.44	2500	60000
	4	4	63750.00	7500.000	3750.000	51815.83	75684.17	55000	70000
	Total	44	33079.55	39780.056	5997.069	20985.30	45173.79	0	200000

(*Contd*)

(*Contd*)

Invest_MF	1	16	18593.75	21455.550	5363.887	7160.89	30026.61	2500	75000
	2	8	117500.00	54837.422	19387.956	71654.77	163345.23	50000	180000
	3	16	17781.25	15921.651	3980.413	9297.20	26265.30	0	50000
	4	4	124250.00	72126.625	36063.312	9480.44	239019.56	22000	175000
	Total	44	45886.36	56550.540	8525.315	28693.43	63079.30	0	180000
how_much_time_	1	16	2.19	1.047	0.262	1.63	2.75	1	4
block_your_money	2	8	6.13	1.553	0.549	4.83	7.42	4	8
	3	16	4.31	1.887	0.472	3.31	5.32	1	7
	4	4	4.25	0.500	0.250	3.45	5.05	4	5
	Total	44	3.86	2.030	0.306	3.25	4.48	1	8
risky_CM	1	16	5.31	1.887	0.472	4.31	6.32	2	9
	2	8	6.63	0.916	0.324	5.86	7.39	6	8
	3	16	6.00	1.966	0.492	4.95	7.05	3	9
	4	4	3.00	0.816	0.408	1.70	4.30	2	4
	Total	44	5.59	1.921	0.290	5.01	6.17	2	9
risky_SM	1	16	5.38	2.527	0.632	4.03	6.72	1	9
	2	8	7.25	0.886	0.313	6.51	7.99	6	8
	3	16	7.19	1.559	0.390	6.36	8.02	3	9
	4	4	4.75	2.062	1.031	1.47	8.03	3	7
	Total	44	6.32	2.122	0.320	5.67	6.96	1	9
risky_FD	1	16	1.94	0.772	0.193	1.53	2.35	1	3
	2	8	1.13	0.354	0.125	0.83	1.42	1	2
	3	16	1.50	0.516	0.129	1.22	1.78	1	2
	4	4	1.00	0.000	0.000	1.00	1.00	1	1
	Total	44	1.55	0.663	0.100	1.34	1.75	1	3
risky_MF	1	16	5.13	2.187	0.547	3.96	6.29	2	10
	2	8	6.50	1.195	0.423	5.50	7.50	4	8
	3	16	6.56	2.159	0.540	5.41	7.71	4	10
	4	4	4.75	2.363	1.181	0.99	8.51	3	8
	Total	44	5.86	2.120	0.320	5.22	6.51	2	10

The above table gives descriptive statistics for the dependent variables for each cluster the short summary of above table is displayed below:

Descriptives

			Mean							*risky_mutual_funds_one_to_ten*
	N	*Age*	*Rate CM*	*Rate SM*	*Rate FD*	*Rate MF*	*Spend CM*	*Spend SM*	*Spend FD*	
1	16	20.5625	3.6875	3.9375	2.8125	4	5937.5	20156.25	6718.75	5.125
2	8	46.125	1.5	2.125	4.125	2.75	36250	111875	95625	6.5
3	16	33.5625	1.875	2.3125	4.4375	2.9375	4593.75	18656.25	20500	6.5625

(*Contd*)

(*Contd*)

4	4	26	2.5	3.5	3	4.5	57500	115000	63750	4.75
Total	44	30.43182	2.522727	2.977273	3.659091	3.431818	15647.73	44909.09	33079.55	5.863636

	N	*Spend MF*	*Bolck Time*	*Risky CM*	*Risky SM*	*Risky FD*	*Risky MF*
		18593.75	2.1875	5.3125	5.375	1.9375	5.125
1	16	117500	6.125	6.625	7.25	1.125	6.5
2	8	17781.25	4.3125	6	7.1875	1.5	6.5625
3	16	124250	4.25	3	4.75	1	4.75
4	4	45886.36	3.863636	5.590909	6.318182	1.545455	5.863636
Total	44						

It may be noted that these four clusters have average age as 20.56, 46.13, 33.56 and 26. This clearly forms four different age groups. The other descriptive is summarised as follows:

Cluster 1 Sr No: 31, 33, 12, 20, 2, 11, 14, 32, 40, 24, 29, 1, 15, 17, 23, 28	Average age 20.56 (Young non-working) Prefers investing in commodity market, share market and mutual funds; does not prefer to invest in fixed deposits; invests less money, blocks money for lesser period and finds share market and commodity market investments as least risky (among other people).
Cluster 2 Sr No: 41, 42, 39, 8, 44, 9, 45, 3	Average age 46.13 (Oldest) Least prefers investing in commodity market, share market and mutual funds; prefers to invest in fixed deposits; invests high money, blocks money for more period and finds share market and commodity market investments as most risky (among other people).
Cluster 3 Sr No: 4, 13, 6, 19, 10, 16, 34, 21, 22, 18, 27, 30, 35, 36, 25, 26	Average age 33.56 (Middle age) Less prefers investing in commodity market, share market and mutual funds; prefers to invest in fixed deposits; invests less money, blocks money for lesser period (but not lesser than people of cluster 1) and finds share market and commodity market investments as more risky.
Cluster 4 Sr No: 7, 43, 37, 38	Average age 26 (Young working) Less prefers investing in commodity market, share market but prefers to invest in mutual funds and fixed deposits; invests more money, blocks money for moderate period and finds share market and commodity market investments as more risky.

It may be noted that these clusters are named in the dendrogram on the basis of above criteria.

Test of Homogeneity of Variances

	Levene Statistic	*df1*	*df2*	*Sig.*
Age	7.243	3	40	0.001
Rate_CM	0.943	3	40	0.429
Rate_SM	1.335	3	40	0.277
Rate_FD	3.186	3	40	0.034

(*Contd*)

(*Contd*)

Rate_MF	1.136	3	40	0.346
Invest_CM	0.369	3	40	0.775
Invest_SM	17.591	3	40	0.000
Invest_FD	4.630	3	40	0.007
Invest_MF	15.069	3	40	0.000
how_much_time_block_your_money	3.390	3	40	0.027
risky_CM	1.995	3	40	0.130
risky_SM	4.282	3	40	0.010
risky_FD	5.294	3	40	0.004
risky_MF	2.118	3	40	0.113

This table gives Leven's Homogeneity test which is a must for ANOVA as ANOVA assumes that the different groups have equal variance. If the significance is less than 5% (LOS), the null hypothesis which states that the variances are equal is rejected. i.e. the assumption is not followed. In such a case, ANOVA cannot be used. The above table rejection of the assumption is indicated by circles, which means ANOVA could be invalid for those variables.

It may be noted that when ANOVA is invalid, the test that can be performed is non parametric test, Kruskal–Wallis test discussed in Chapter 13.

Anova

		Sum of squares	*df*	*Mean Square*	*F*	*Sig.*
Age	Between Groups	3764.045	3	1254.682	25.185	0.000
	Within Groups	1992.750	40	49.819		
	Total	5756.795	43			
Rate_CM	Between Groups	36.790	3	12.263	22.108	0.000
	Within Groups	22.188	40	0.555		
	Total	58.977	43			
Rate_SM	Between Groups	28.727	3	9.576	26.879	0.000
	Within Groups	14.250	40	0.356		
	Total	42.977	43			
Rate_FD	Between Groups	24.636	3	8.212	12.054	0.000
	Within Groups	27.250	40	0.681		
	Total	51.886	43			
Rate_MF	Between Groups	17.358	3	5.786	5.869	0.002
	Within Groups	39.438	40	0.986		
	Total	56.795	43			

(*Contd*)

(*Contd*)

Invest_CM	Between Groups	1E+010	3	4621914299	58.403	0.000
	Within Groups	3E+009	40	79138671.88		
	Total	2E+010	43			
Invest_SM	Between Groups	8E+010	3	2.545E+010	22.243	0.000
	Within Groups	5E+010	40	1144289844		
	Total	1E+011	43			
Invest_FD	Between Groups	5E+010	3	1.624E+010	33.585	0.000
	Within Groups	2E+010	40	483427734.4		
	Total	7E+010	43			
Invest_ME	Between Groups	9E+010	3	3.005E+010	25.377	0.000
	Within Groups	5E+010	40	1184108594		
	Total	1E+011	43			
how_much_time_block_your_money	Between Groups	89.682	3	29.894	13.666	0.000
	Within Groups	87.500	40	2.188		
	Total	177.182	43			
risky_CM	Between Groups	39.324	3	13.108	4.394	0.009
	Within Groups	119.313	40	2.983		
	Total	158.636	43			
risky_SM	Between Groups	43.108	3	14.369	3.821	0.017
	Within Groups	150.438	40	3.761		
	Total	193.545	43			
risky_FD	Between Groups	5.097	3	1.699	4.920	0.005
	Within Groups	13.813	40	0.345		
	Total	18.909	43			
risky_MF	Between Groups	24.744	3	8.248	1.959	0.136
	Within Groups	168.438	40	4.211		
	Total	193.182	43			

ANOVA not rejected as > 0.05

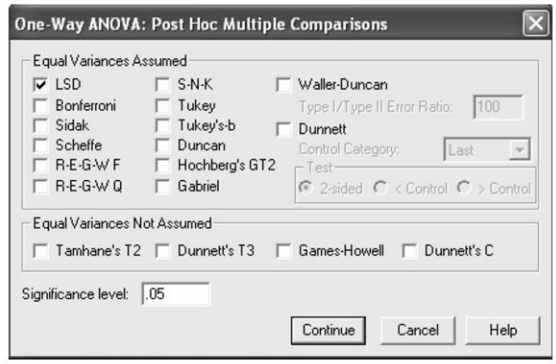
The above ANOVA table tests the difference between means for the different clusters. The null hypothesis states that there is no difference between the clusters for given variable. If significance is less than 5% (p value less than 0.05), the null hypothesis is rejected.

It may be noted that for above table, null hypothesis that the variable is equal for all clusters is rejected for all variables except for Risky_MF. This means all other variables significantly vary for different clusters. It also indicates that the four cluster solution is a good solution.

K-means Cluster

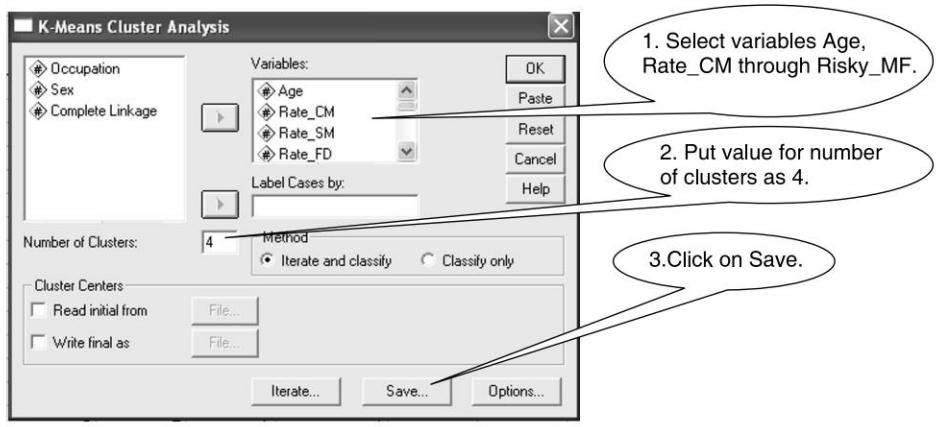
This method is used when one knows in advance, how many clusters to be formed. The procedure for k-means cluster is as follows:

CA Snapshot 11



The following window will be displayed:

CA Snapshot 12

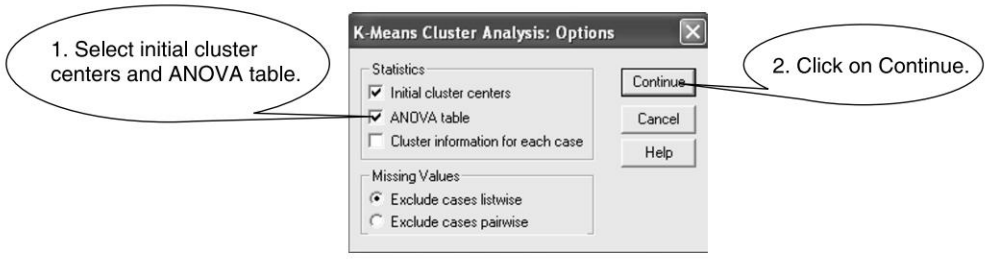


The following window will appear:

CA Snapshot 13

SPSS will take back to the window as shown in CA Snapshot 12. Click on Options and the following window will appear:

CA Snapshot 14



SPSS will take back to CA Snapshot 12 at this stage click OK.
The following output will be displayed:

Quick Cluster

Initial Cluster Centres

	Cluster			
	1	*2*	*3*	*4*
Age	55	22	54	45
Rate_CM	1	4	1	1
Rate_SM	2	4	2	2
Rate_FD	4	4	4	4
Rate_MF	2	3	3	3
Invest_CM	45000	5000	50000	25000
Invest_SM	60000	3000	60000	185000
Invest_FD	75000	1000	200000	100000
Invest_MF	75000	2500	50000	155000
how_much_time_block_your_money	8	4	8	6
risky_CM	8	3	8	6
risky_SM	8	6	8	7
risky_FD	1	2	1	2
risky_MF	7	3	4	7

This table gives initial cluster centres. The initial cluster centres are the variable values of the k well-spaced observations.

Iteration History[a]

	Change in Cluster Centres			
Iteration	*1*	*2*	*3*	*4*
1	31980.518	19406.551	0.000	35237.293
2	7589.872	3733.790	0.000	0.000
3	0.000	0.000	0.000	0.000

(a) Convergence achieved due to no or small change in cluster centres. The maximum absolute coordinate change for any centre is .000. The current iteration is 3. The minimum distance between initial centres is 124824.882.

The iteration history shows the progress of the clustering process at each step.
This table has only three steps as the process has stopped due to no change in cluster centres.

Final Cluster Centres

	Cluster			
	1	*2*	*3*	*4*
Age	33	27	54	34
Rate_CM	2	3	1	2
Rate_SM	3	3	2	3
Rate_FD	4	4	4	4
Rate_MF	4	3	3	4
Invest_CM	31000	2981	50000	36667
Invest_SM	58182	11769	60000	161667
Invest_FD	41636	11827	200000	81667
Invest_MF	60636	10846	50000	170000
how_much_time_block_your_money	4	3	8	5
risky_CM	5	6	8	5
risky_SM	7	6	8	5
risky_FD	1	2	1	1
risky_MF	6	6	4	5

This table gives final cluster centres.

ANOVA

	Cluster		*Error*			
	Mean Square	*df*	*Mean Square*	*df*	*F*	*Sig.*
Age	318.596	3	120.025	40	2.654	0.062
Rate_CM	2.252	3	1.306	40	1.725	0.177
Rate_SM	0.599	3	1.029	40	0.582	0.630
Rate_FD	0.377	3	1.269	40	0.297	0.827
Rate_MF	0.722	3	1.366	40	0.528	0.665
Invest_CM	3531738685	3	160901842.9	40	21.950	0.000
Invest_SM	3.750E+010	3	240364627.0	40	156.032	0.000
Invest_FD	1.819E+010	3	336746248.5	40	54.022	0.000
Invest_MF	4.225E+010	3	268848251.7	40	157.162	0.000
how_much_time_block_your_money	10.876	3	3.614	40	3.010	0.041
risky_CM	3.654	3	3.692	40	0.990	0.407
risky_SM	3.261	3	4.594	40	0.710	0.552
risky_FD	0.603	3	0.427	40	1.411	0.254
risky_MF	1.949	3	4.683	40	0.416	0.742

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

The ANOVA indicates that the clusters are different only for different investment options like invest in CM, invest in SM, invest in FD and invest in MF as also block money, as the significance is less than 0.05 only for these variables.

Number of Cases in each Cluster

Cluster	1	11
	2	26
	3	1
	4	6
Valid	44	
Missing	1	

The above table gives the number of cases for each cluster.

It may be noted that this solution is different than hierarchal solution and hierarchal cluster is more valid for this data as it considers standardised scores and this method does not consider standardisation.

14.9 CONJOINT ANALYSIS

The name "Conjoint Analysis" implies the study of the joint effects. In marketing applications, it helps in the study of joint effects of multiple product attributes on product choice. Conjoint analysis involves the measurement of psychological judgements such as consumer preferences, or perceived similarities or differences between choice alternatives.

In fact, conjoint analysis is a versatile marketing research technique which provides valuable information for new product development, assessment of demand, evolving market segmentation strategies and pricing decisions. This technique is used to assess a wide number of issues including:

- The profitability and/or market share for proposed new product concepts given the existing competition.
- The impact of new competitors' products on profits or market share of a company if status quo is maintained with respect to it's products and services.
- Customers' switch rates either from a company's existing products to the company's new products or from competitors' products to the company's new products.
- Competitive reaction to the company's strategies of introducing a new product.
- The differential response to alternative advertising strategies and/or advertising themes.
- The customer response to alternative pricing strategies, specific price levels, and proposed price changes.

Conjoint analysis examines the trade-offs that consumers make in purchasing a product. In evaluating products, consumers make trade-offs. A TV viewer may like to enjoy the programmes on a LCD TV but might not go for it because of the high cost. In this case, cost is said to have a high

utility value. Utility can be defined as a number which represents the value that consumers place on specific attributes. A low utility indicates less value; a high utility indicates more value. In other words, it represents the relative 'worth' of the attribute. This helps in designing products/services that are most appealing to a specific market. In addition, because conjoint analysis identifies important attributes, it can be used to create advertising messages that are most appealing.

The process of data collection involves showing respondents a series of cards that contain a written description of the product or service. If a consumer product is being tested, then a picture of the product can be included along with a written description. Several cards are prepared describing the combination of various alternative set of features of a product or service. A consumer's response is collected by his/her selection of number between 1 and 10. While '1' indicates strongest dislike, '10' indicates strongest like for the combination of features on the card. Such data becomes the input for final analysis which is carried out through computer software.

The concepts and methodology are elaborated in the case study given below.

14.9.1 Conjoint Analysis Using SPSS

CASE 14.3 CREDIT CARDS

The new head of the credit card division in a bank wanted to revamp the credit card business of the bank and convert it from loss making business to profit making business. He was given freedom to experiment with various options that he considered as relevant. Accordingly, he organised a focus group discussion for assessing the preference of the customers for various parameters associated with the credit card business. Thereafter, he selected the following parameters for study:

1. **Transaction Time** – This is the time taken for credit card transaction
2. **Fees –** The annual fees charged by the credit card company
3. **Interest rate** – The interest rate charged by the credit card company for the customers who revolve the credits (customers who do not pay full bill amount but use partial payment option and pay at their convenience)

The levels of the above mentioned attributes were as follows:

- **Transaction Time** – 1 minute, 1.5 minutes, 2 minutes
- **Fees** – 0, Rs. 1000, Rs. 2000
- **Interest rate** 1.5%, 2%, 2.5% (per month)

This led to a total of 3×3×3 = 27 combinations. Twenty-seven cards were prepared representing each combination and the customers were asked to arrange these cards in order of their preference.

The following table shows all the possible combinations and the order given by the customer:

Input Data for Credit Card

Sr. No	*Transaction Time (min)* *1, 1.5, 2*	*Card Fees (Rs)* *0, 1000, 2000*	*Interest Rate* *1.5%, 2.0%, 2.5%*	*Rating ** *27 to 1*
1	1	0	1.5	27
2	1.5	0	1.5	26

(*Contd*)

(*Contd*)

3	1	1000	1.5	25
4	1.5	1000	1.5	24
5	2	0	1.5	23
6	2	1000	1.5	22
7	1	0	2	21
8	1.5	0	2	20
9	1	2000	1.5	19
10	1.5	2000	1.5	18
11	1	1000	2	17
12	1.5	1000	2	16
13	1	2000	2	15
14	2	2000	1.5	14
15	1.5	2000	2	13
16	2	0	2	12
17	2	1000	2	11
18	1	0	2	10
19	1.5	0	2.5	9
20	1	1000	2.5	8
21	2	1000	2.5	7
22	2	2000	2	6
23	2	0	2.5	5
24	2	1000	2.5	4
25	1	2000	2.5	3
26	1.5	2000	2.5	2
27	2	2000	2.5	1

* rating 27 indicates most preferred and rating 1 indicates lest preferred option by customer.

Conduct appropriate analysis to find the utility for these three factors.
The data is available in credit card.sav file, given in the CD.

Running Conjoint as a Regression Model: Introduction of Dummy Variables

Representing dummy variables:
X_1, X_2 = Transaction time
X_3, X_4 = Annual Fees
X_5, X_6 = Interest Rates
The 3 **levels of life** are coded as follows:

Transaction Time	X_1	X_2
1	1	0
1.5	0	1
2	–1	–1

The 3 **levels of price** are coded as follows:

Fees	X_3	X_4
0	1	0
1000	0	1
2000	–1	–1

The 3 **levels of colour** are coded as follows:

Interest rates	X_5	X_6
1.5	1	0
2	0	1
1.5	–1	–1

Thus, 6 variables, ie. X_1 to X_6 are used to represent the 3 levels of life of the transaction time (1, 1.5, 2), 3 levels of fees (0, 1000, 2000) and 3 levels of interest rates (1.5, 2, 2.5). All the six variables are **independent variables** in the regression run. Another variable Y which is the rating of each combination given by the respondent forms the **dependent variable** of the regression curve.

Thus, we generate the regression equation as: $\mathbf{Y= a +b_1X_1+ b_2X_2+ b_3X_3+ b_4X_4+ b_5X_5+ b_6X_6}$

Input data for the regression model:

Sr No	*Transaction time*	*Fees*	*Interest Rate*	*Y*	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*
1	1	0	1.5	27	1	0	1	0	1	0
2	1.5	0	1.5	26	0	1	1	0	1	0
3	1	1000	1.5	25	1	0	0	1	1	0
4	1.5	1000	1.5	24	0	1	0	1	1	0
5	2	0	1.5	23	–1	–1	1	0	1	0
6	2	1000	1.5	22	–1	–1	0	1	1	0
7	1	0	2	21	1	0	1	0	0	1
8	1.5	0	2	20	0	1	1	0	0	1
9	1	2000	1.5	19	1	0	–1	–1	1	0
10	1.5	2000	1.5	18	0	1	–1	–1	1	0
11	1	1000	2	17	1	0	0	1	0	1
12	1.5	1000	2	16	0	1	0	1	0	1
13	1	2000	2	15	1	0	–1	–1	0	1
14	2	2000	1.5	14	–1	–1	–1	–1	1	0
15	1.5	2000	2	13	0	1	–1	–1	0	1
16	2	0	2	12	–1	–1	1	0	0	1
17	2	1000	2	11	–1	–1	0	1	0	1
18	1	0	2	10	1	0	1	0	0	1
19	1.5	0	2.5	9	0	1	1	0	–1	–1
20	1	1000	2.5	8	1	0	0	1	–1	–1

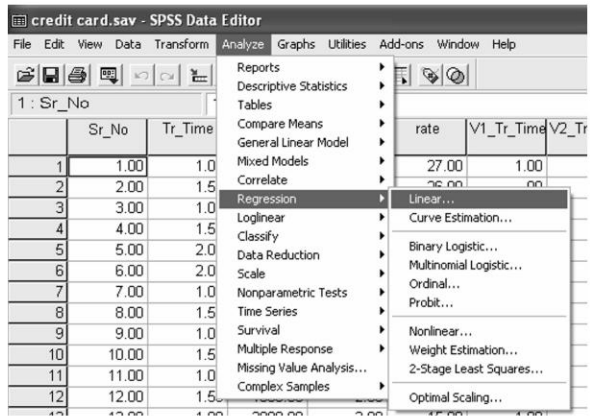
(*Contd*)

(*Contd*)

21	2	1000	2.5	7	–1	–1	0	1	–1	–1
22	2	2000	2	6	–1	–1	–1	–1	0	1
23	2	0	2.5	5	–1	–1	1	0	–1	–1
24	2	1000	2.5	4	–1	–1	0	1	–1	–1
25	1	2000	2.5	3	1	0	–1	–1	–1	–1
26	1.5	2000	2.5	2	0	1	–1	–1	–1	–1
27	2	2000	2.5	1	–1	–1	–1	–1	–1	–1

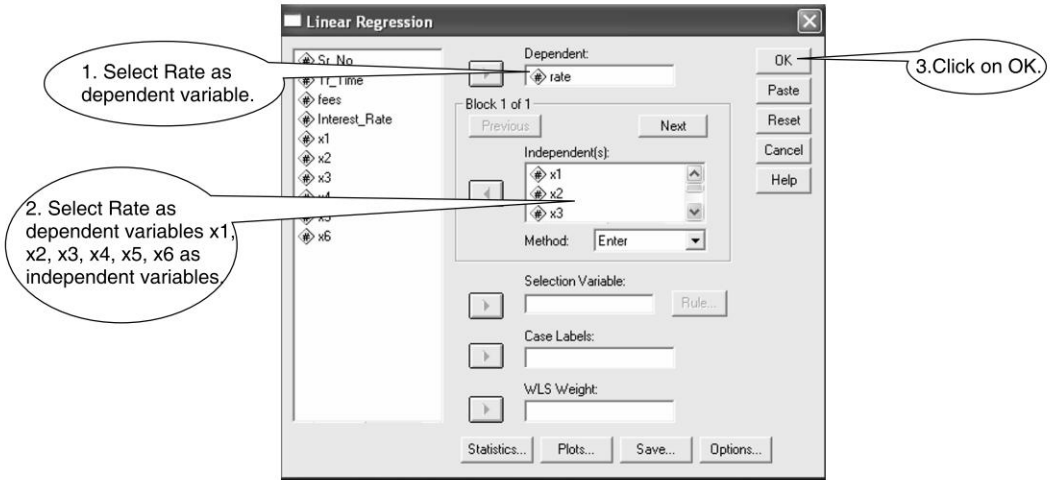
This can be processed using SPSS package as follows:
Open the file credit card.sav
Select 'Analyse' – Regression – Linear option from the menu as shown below:

Conjoint Snapshot 1



The following window will be displayed:

Conjoint Snapshot 2



The output generated is as follows:

Regression

Variables Entered/Removed[b]

Model	Variable Entered	Variable Removed	Method
1	x6, x4, x2, x5, x3, x1[a]	.	Enter

a. All requested variables entered.
b. Dependent Variable: rate

This table summarises the regression model:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.963[a]	0.927	0.905	2.45038

a. Predictors: (Constant), x6, x4, x2, x5, x3, x1

This table indicates that r square for the above model is 0.963, which is close to one. This indicates that 96.3% variation in the rate is attributed by the six independent variables (x_1 to x_6).

We conclude that the regression model is fit and explains the variations in the dependent variables quite well.

ANOVA[b]

Model		Sum of Square	df	Mean Square	F	Sig.
1	Regression	1517.912	6	252.985	42.133	0.000[a]
	Residual	120.088	20	6.004		
	Total	1638.000	26			

a. Predictors: (Constant), x6, x4, x2, x5, x3, x1
b. Dependent Variable: rate

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	13.857	0.476		29.104	0.000
	x1	1.377	0.673	0.148	2.044	0.054
	x2	1.326	0.695	0.138	1.908	0.071
	x3	2.265	0.673	0.237	3.364	0.003
	x4	1.480	0.673	0.155	2.198	0.040
	x5	8.143	0.670	0.829	12.152	0.000
	x6	–0.121	0.661	–0.013	–0.183	0.857

a. Dependent Variable: rate

The coefficients are circled; this indicates the utility values for each variable.
The Regression equation is as follows:

$$Y = 13.857 + 1.377X_1 + 1.326X_2 + 2.265X_3 - 1.48X_4 + 8.143X_5 - 0.121X_6$$

Output and Interpretation

Utility (U_{ij}) – The utility or the part worth contribution associated with the j^{th} level (j, j=1, 2, 3) of the i^{th} attribute (i, i=1, 2, 3) for ex- U_{21} in our example means utility associated with the life of 4 years.

Importance of an attribute (I_i)–is defined as the range of the part worth U_{ij} across the levels of that attribute. $I_i = \{\text{Max } (U_{ij}) - \text{Min } (U_{ij})\}$ for each attribute (i)

Normalisation: The attributes importance is normalised to desire its relative importance among all attributes.

$$W_i = \frac{I_i}{\Sigma_{i=1}^{2} (I_i)} \qquad \text{so that } \Sigma Wi = 1$$

The output provides the part utility of each level of attribute which is shown below:

X1 = 1.377 (partial utility for 1 min transaction)
X2 = 0.11 (partial utility for 1.5 min transaction)

For 2 min transaction partial utility = –2.703 (as all the utilities for a given attribute should sum to 0; hence –1.377 – 1.326 =–2.703)

X3 = 2.265 (partial utility for 0 fees)
X4 = 1.48 (partial utility for 1000 fees)

For 2 min transaction partial utility = –3.745 (as all the utilities for a given attribute should sum to 0; hence –2.265 – 1.48 = –3.745)

X5 = 8.143 (partial utility for 1.5% interest)
X6 = –0.121 (partial utility for 2% interest)

For 2% interest transaction partial utility = –8.022 (as all the utilities for a given attribute should sum to 0; hence –8.143 + 0.121 = –8.022)

Utilities Table for Conjoint Analysis

Attributes	*Levels*	*Part Utility*	*Range of Utility (Max – Min)*	*Percentage Utility*
Transaction Time	1 min	1.377		15.54%
	1.5 min	1.326	= 1.377 – (–2.703) = 4.08	
	2. min	–2.703		
Annual Fees in Rupees	0	2.265		22.89%
	1000	1.480	= 2.265– (–3.745) = 6.01	
	2000	–3.745		
Interest Rate	1.5%	8.143		61.57%
	2.0%	–0.121	= 8.143– (–8.022) = 16.165	
	2.5%	–8.022		

From the above table, we can interpret the following:

- The **Interest rate is the most important attribute** for the customer. This is indicated by following:
  (a) The range of utility value is the highest (16.165) for the interest rate (refer range of utility column). This contributes to 61.57% of total utility.
  (b) The highest individual utility value of this attribute is at the 1st level i.e. 8.143.
- The **Annual Fees is the second most important attribute**, as its range of utilities is 6.01 and it contributes to 22.89% of the total.
- The **last attribute in relative importance is the Transaction Time**, with the utility range of 4.08, contributing to 15.54% of the total.

Combination Utilities

The total utility of any combination can be calculated by picking up the attribute levels of our choice.
For example,
The combined utility of the combination of 1.5 min + 1000 Fees + 2% Interest
= 1.326 + 1.480 – 0.121
= 2.685
To know the **BEST COMBINATION,** it is advisable to pick the highest utilities from each attribute and then add them.

The best combination here is: 1 min + 0 Fees + 1.5% Interest
1.377 + 2.265 + 8.143
11.785

Individual Attributes

The difference in utility with the change of one level in one attribute can also be checked.

1. Transaction Time
   - For the time 1 min to 1.5 min – There in decrease in utility value of 0.051 units.
   - But the next level, that is, 1.5 min to 2 min has decrease in utility of 4.029 units.
2. Annual Fees
   - Increase fees from 0 to Rs.1000 induces a utility drop of 0.785
   - Whereas, Rs.1000 to Rs. 2000, there is an decrease in utility of 5.225
3. Interest Rates
   - Interest rate increase from 1.5% to 2.0% induces 8.264 units drop in utility.
   - Interest rate increase from 2.0% to 2.5% induces 7.901units drop in utility.

14.10 MULTIDIMENSIONAL SCALING

Multidimensional Scaling transforms consumer judgments/perceptions of similarity or preferences in a multidimensional space (usually 2 or 3 dimensions). It is useful for designing products and services. In fact, MDS is a set of procedures for drawing pictures of data so that the researcher can:

- Visualise relationships described by the data more clearly
- Offer clearer explanations of those relationships

Thus, MDS reveals relationships that appear to be obscure when one examines only the numbers resulting from a study.

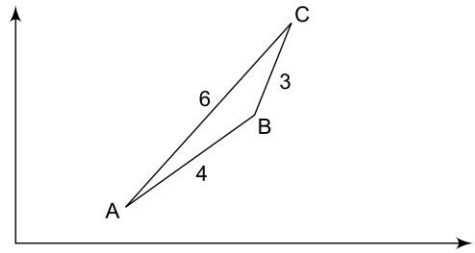
It attempts to find the structure in a set of distance measures between objects. This is done by assigning observations to specific locations in a conceptual space (2 to 3 dimensions) such that the distances between points in the space match the given dissimilarities as closely as possible.

If objects A and B are judged by the respondents as being most similar compared to all other possible pairs of objects, multidimensional technique positions these objects in the space in such a manner that the distance between them is smaller than that between any other two objects.

Suppose, data is collected for perceiving the differences or distances among three objects say A, B and C, and the following distance matrix emerges:

	A	*B*	*C*
A	0	4	6
B	4	0	3
C	6	3	0

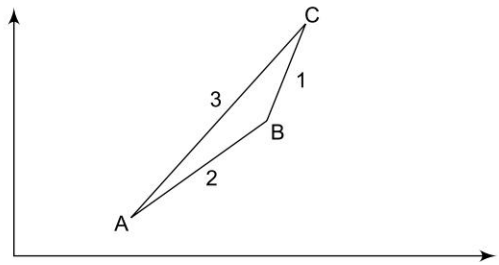
This matrix can be represented by a two-dimensional diagram as follows:



However, if the data comprises only ordinal or rank data, then the same distance matrix could be written as:

	A	*B*	*C*
A	0	2	3
B	1	0	3
C	3	1	0

and can be depicted as:



If the actual magnitudes of the original similarities (distances) are used to obtain a geometric representation, the process is called **Metric Multidimensional Scaling**.

When only this ordinal information in terms of ranks is used to obtain a geometric representation, the process is called **Non-metric Multidimensional Scaling**.

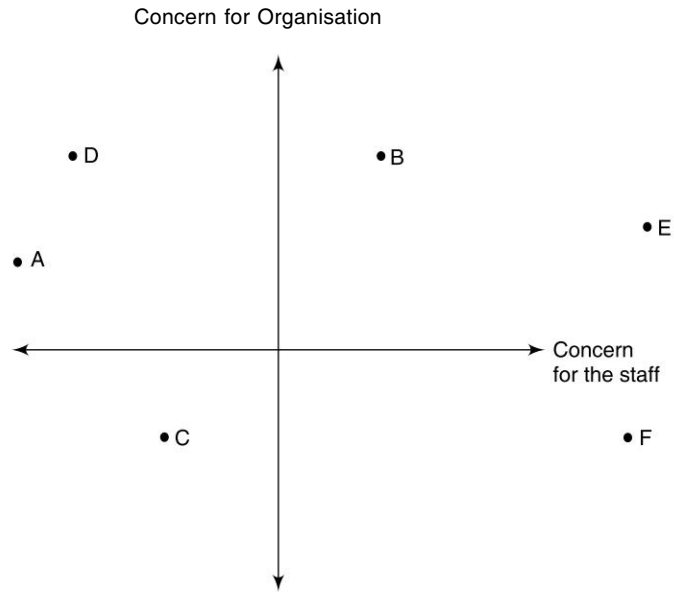
14.10.1 Uses of MDS

(i) Illustrating market segments based on preference and judgments.
(ii) Determining which products are more competitive with each other.
(iii) Deriving the criteria used by people while judging objects (products, brands, advertisements, etc.).

Example 14.4

An all-India organisation had six zonal offices, each headed by a zonal manager. The top management of the organisation wanted to have a detailed assessment of all the zonal managers for selecting two of them for higher positions in the Head Office. They approached a consultant for helping them in the selection. The management indicated that they would like to have assessment on several parameters associated with the functioning of a zonal manager. The management also briefed the consultant that they laid great emphasis on the staff with a view to developing and retaining them.

The consultants collected a lot of relevant data, analysed it and offered their recommendations. In one of the presentations, they showed the following diagram obtained through Multi Dimensional Scaling technique. The diagram shows the concerns of various zonal managers, indicated by letters A to F, towards the organisation and also towards the staff working under them.



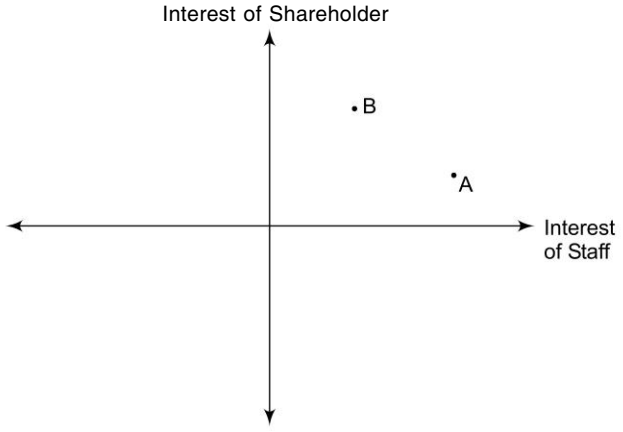
It is observed that two zonal managers viz. B and E exhibit high concern for both the organisation as well as staff. If these criteria are critical to the organisation, then these two zonal managers could be the right candidates for higher positions in the Head Office.

Example 14.5

Similar study could be conducted for a group of companies to have an assessment of the perception of investors about the attitude of companies towards interest of their shareholders and vis-à-vis

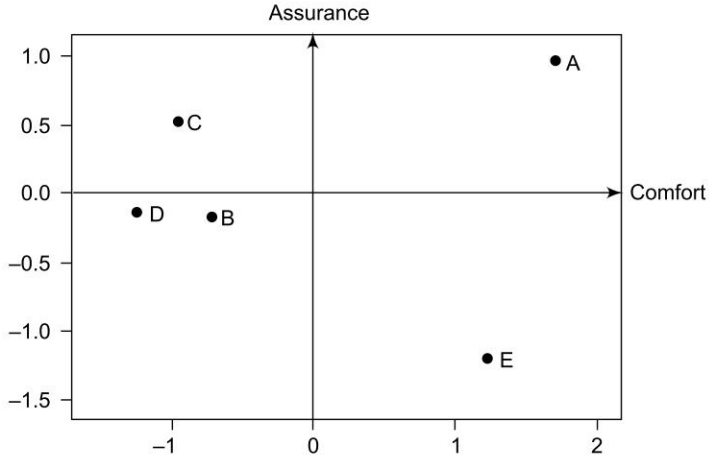
interest of their staff.

For example, from the following MDS graph, it is observed that company A is perceived to be taking more interest in the welfare of the staff than company B.



Example 14.5

A small team of students of a management institute conducted a study to decide upon the positioning of the new brand vis-à-vis existing brands of 125 cc motorcycles. They collected data through a questionnaire from 40 users of such motorcycles. They followed the multidimensional scaling (MDS) approach for positioning of a new brand. Through the use of SPSS software, they derived MDS – stimulus configuration as depicted by perceptual mapping in the following two-dimensional graph:



SUMMARY

A number of statistical techniques are especially useful in designing of products and services. These techniques, basically involve reduction of data and subsequent its summarisation, presentation and interpretation.

These techniques (with their abbreviations in brackets) coupled with the appropriate computer software like SPSS, play a very useful role in the endeavour of reduction and summarisation of data for easy comprehension.

A brief idea about these techniques is as follows:

Multiple Regression Analysis (MRA): It deals with the study of relationship between one metric dependent variable and more than one metric independent variables.

Principal Component Analysis (PCA): Technique for forming set of new variables that are linear combinations of the original set of variables, and are uncorrelated. The new variables are called Principal Components.

Canonical Correlation Analysis (CRA): An extension of multiple regression analysis (MRA involving one dependant variable and several metric independent variables). It is used for situations wherein there are several dependant variables and several independent variables.

Discriminant Analysis: It is a statistical technique for classification or determining a linear function, called discriminant function, of the variables which helps in discriminating between two groups of entities or individuals.

Multivariate Analysis of Variance (MANOVA): It explores, simultaneously, the relationship between several non-metric independent variables and two or more metric dependant variables.

Factor Analysis (FA): It is a statistical approach that is used to analyse inter-relationships among a large number of variables (indicators) and to explain these variables(indicators) in terms of a few unobservable constructs (factors). In fact, these factors impact the variables, and are reflective indicators of the factors.

Cluster Analysis: It is an analytical technique that is used to develop meaningful subgroups of entities which are homogeneous or compact with respect to certain characteristics.

Conjoint Analysis: Involves determining the contribution of variables (each of several levels) to the choice preference over combinations of variables that represent realistic choice sets (products, concepts, services, companies, etc.)

Logistic Regression: In logistic regression, the dependent variable is the probability that an event will occur, hence it is constrained between 0 and 1. All of the predictors can be binary, a mixture of categorical and continuous or just continuous.

Canonical Correlation: It relates a set of dependent variables with a set of independent variables. It involves developing linear combinations of the sets of variables (both dependant and independent) and studies the relationship between the two sets.

Multidimensional Scaling: It is a set of procedures drawing pictures of data so as to visualise and clarify relationships described by the data more clearly. It transforms consumer judgments/perceptions of similarity or preferences in usually a two-dimensional space.

DISCUSSION QUESTIONS

1. Distinguish between the dependence and interdependence techniques, with suitable examples.
2. Write short notes on the following bringing out their relevance and applications:

(i) Multiple Regression
(ii) Discriminant Analysis
(iii) Logistic Regression
(iv) Multivariate Analysis of Variance (MANOVA)
(v) Principal Component Analysis
(vi) Common Factor Analysis (Principal Axis Factoring)
(vii) Canonical Correlation Analysis
(viii) Cluster Analysis
(ix) Conjoint Analysis
(x) Multidimensional Scaling

3. Describe three situations for the applicability of each of the above techniques.
4. Work out the solution using SPSS, for the example relating to MANOVA in Section 14.5.

Report Writing

15

Contents

LEARNING OBJECTIVES

The objective of this chapter is to provide comprehensive guidelines for preparing and presenting a research study report.

15.1 INTRODUCTION AND RELEVANCE

Report writing is one of the most important activities of the whole assignment of conducting a research study. It is through this report that the researcher conveys in writing *inter alia* about

- Why the research study was conducted?
- How the research was conducted?
- What was achieved by conducting the study?
- What resources were used–literature, men, money and time?

It is through the report that the researcher wants to emphasise the contribution made by him in the pursuit of knowledge in the area. The reader may not have an idea about the credibility and competence of the researcher. The onus lies on the researcher to demonstrate these, as also the efforts he has put in the study through the report. Thus, the report should reflect the researcher's

- Knowledge and expertise in the area of research
- Logical and analytical skills
- Creativity right from the design of the first page of the report
- Credibility for acceptance of the report

However, for appreciating the above traits of the researcher, the report has to be read first. Here, again the onus lies on the researcher to **arouse interest** in the reader. The first **motivation for reading** the report is the title of the report. In general, the following are the criteria for the report to fulfill its objective:

- Title should be of **topical interest** to arouse the interest
- Summary of the report, given in the beginning, has to hold the **promise of knowledge enrichment** for the reader

- Format of the report should be such so as to hold **sustained interest** in reading the entire report
- **Scope and limitations** of the approaches followed and of the findings should be outlined. If the study relates to either estimation of parameters like sales, profit or, estimating association or relationship like relationship between advertising and sales, testing of some assumption like lives of two brands of car batteries are the same, then an idea about the faith one can place in the findings has to be provided.

15.2 FORMAT OF A REPORT

There is no prescribed format for writing reports. It varies according to the type and purpose of the research study. However, there are general guidelines in the form of a format that could be suitably modified at the discretion of the researcher, depending on the type and purpose of the study. It comprises the following components:

- **Initial Parts of the Report**
- **Main Body of the Report**
- **Concluding Parts of the Report**

These are detailed below.

A. Initial Parts of the Report

- Cover sheet with Title and Researcher's name(s) and other relevant details
- Title page
- Letter of transmittal to the sponsoring authority (This is required when the study is assigned to the researcher, through a formal letter)
- Index of contents with page numbers
- Summary or Executive Summary, if the report is submitted to higher authorities
- Acknowledgement for guidance

B. Main Body of the Report

- Preamble/Preface/Introduction
- Research Design
- Findings
- Interpretations/Conclusions
- Recommendations (if required)
- Glossary of technical terms (if needed)
- Epilogue

C. Concluding Parts of the Report

- References/Bibliography
- Acknowledgements
- Appendices

The above components are elaborated below to serve as a guideline.

15.2.1 Initial Parts of the Report

• Cover Sheet

The cover sheet includes title and researcher's name(s). The title should arouse the interest of the reader.

One may use creativity to come with unique design indicating the basic theme of the research. The idea is that the reader should start with a pleasant feeling. It assumes greater significance, if the readers do not know the researcher(s).

A few samples are given at the end of the chapter.

For MBA Students:

- The Institute's logo may be given at the top
- Title and Researcher's or Team Members' names
- 'MBA Batch Year (e.g. 2010)' may be mentioned

- **Title Page**

The title page covers

(i) Title of the Project (to arouse interest)—should be exactly the same as on the cover sheet.

(ii) Broad Description of the project, background or logic for selection—whether by the researcher or assigned by the faculty

- **Letter of Transmittal**

If the project is given to the researcher by an internal authority or external agency through an official letter, called **Letter of Authorisation**, then along with the title indicating its relevance to the assignment, the researcher has to submit what is called '**Letter of Transmittal**', forwarding the report to the concerned authority giving reference of the letter received from them. These types of letters are given on the next two pages after the title page.

- **Index of Contents with Pages**

It includes page numbers of heading/title of sections and subsections like 2, 2.1, 2.1.1, and their corresponding page numbers.

- **Summary** (**Executive Summary**, if the report is submitted to higher authorities)

Elaborate a bit about the title of the project bringing out the background and objective of the study, and to the extent of fulfillment through the conclusions drawn from the study. If the study requires the researcher to make recommendations, the recommendations are to be highlighted. The scope of the conclusions/recommendations and their limitations are to be mentioned. If the limitations are of serious nature, the report may not invoke the interest in the full report.

As mentioned in Chapter 8, a picture is equal to thousand words. So is true about charts, graphs or sometimes even tables. Therefore, the summary may contain a couple of these, if relevant.

Executive Summary is prepared when the report is submitted to the top management. It has to provide them a sense of receiving important inputs that could be useful for their decision-making. But due care has to be taken so that the summary is unbiased and without any exaggeration.

- **Acknowledgement for Guidance**

The names of the persons who suggested the topic or/and provided guidance are to me mentioned here.

15.2.2 Main Body of the Report

This is the main part of the report, and contains the entire research process.

It comprises introduction, methodology of conducting study, analysis, etc., indicated below. Before describing these, it is prudent to mention the general considerations that bind all the sections. These are:

(i) All sections to fulfill some objective and specific purpose

(ii) Cohesiveness among various sections. As far as possible, each one following from the previous and leading to the next section
(iii) All the statements, not flowing from analysis, should be validated by giving references

- **Preamble/Preface/Introduction**

Normally, the details of the conduct of a study start with '**Introduction**' which means the '**first section of a communication'**. However, it is preferable to title the first section as **Preamble** as it means '**stating the reasons and intent of what follows'.** This section is to contain reasons that led to the conduct of the study. Thus, one has to give reference to the available literature containing earlier work in this area, bring out the need for extending that work or even suggest a new approach to deal with the current situation. Incidentally, it could also be termed as **'Relevance'** of the study conducted.

The section contains literature review, earlier development, etc. While describing these, care has to be taken that only that literature and development have to be included that has direct bearing to the topic of the study. The focus should not be lost by the enthusiasm of the researcher to impress the vast literature surveyed by him/her or efforts made by him/her, by discussing broader issues.

- Appropriate Methodology or Research Design for the conduct for the study, justifying its use in the context of the research topic
- How it is considered superior to the methodology used in the earlier researches, if any
- Sampling Design—with justification
- Collection of data—Primary and Secondary: Questionnaire, Telephone/Mail/On-line Survey/Interviews/Focus Group
- Presentation of data through Tables, Charts, Graphs, etc., and their interpretation
- Qualitative/Quantitative analysis of the data
- **Conclusion** (and **recommendations** if applicable)

The interpretations and conclusions should **flow** from the presentation and analysis, without even an iota of bias on the part of the researcher.

However, the recommendations could be based on the implications of the conclusions for the organisation which had sponsored the study, and could contain some aspects not covered in the analysis. Some of these could relate to computerisation, setting website, management information system, etc. The specific recommendations could be followed by suggesting an action plan for the sponsor. Care is to be taken that the action plan is practical, and should be presented in the order of importance to the organisation. The action plan could also be divided into short- and long-term plans.

It may be added that resentation/conclusions/interpretation/recommendation could encompass the following points:

(i) The objective and scope of the study
(ii) Findings and results with reference to the analysis
(iii) Justifying validity through the findings/results
(iv) Viability or practicability of recommendations

- **Epilogue**

This is the concluding section relating to the research study. It reiterates the main findings in popular terms, highlighting the main 'achievement'/contribution by the researcher. It also indicates the scope and limitations, and suggests approaches that could be followed for broadening the scope and/or reducing the limitations, in future. It may also indicate the need for a further study in the related area. For example, if the study has been conducted for criteria for preference of features of motorcycles of different brands in one segment, one could suggest that a similar study could be conducted for motorcycles in the other segment. Further, suppose the study has been conducted in Mumbai, one could suggest a similar study being conducted in other metro cities. One could also add that the sample size of customers was restricted to only 100 users in and around a certain area in Mumbai due to paucity of resources. More reliable results could be obtained by taking a larger sample from all over Mumbai.

15.2.3 Concluding Part of The Report

- **Glossary** (if needed)

The technical terms, in alphabetical order, with which the intended reader may not be well aware of, should be defined/explained.

- **References**

The references contain the names of authors (alphabetical order) of books, articles, etc., publications/websites' addresses, used explicitly in the report.

- **Acknowledgements**

As a matter of courtesy, the researcher must acknowledge the type of guidance in the overall conduct of the study or a specific part like conducting the survey or use of computer package, etc., or support like facilitation in collection of data, computer support, etc. For the researcher, it could be an oversight but for the guiding/supporting person, it hurts a lot. In fact, over-generosity is to be preferred to omission.

- **Appendices**

The appendices include the data collected, the detailed analysis and Tables/Graphs/Charts, etc. The appendices are to be numbered according to the order in which these are referred in the text.

It is interesting to note that how the terminologies are used in describing various components of the report.

One such classification is as follows:

15.2.4 Formats for Various Types of Reports for Different Types of Research Studies

While the formats of the report comprising the preliminary components from Cover sheet to Table of contents, and the final components from Glossary to Appendices are about the same, but the formats of Preamble/Introduction, Research Design and Findings, Conclusions and Recommendations could vary depending on the type of research/study. Thus, the reports for each of the following types of studies would be different:

Table 15.1 Types of Research and Corresponding Relevance Report

Types of Research	*Relevance for Research Report*
• Analysis of performance of a company, as a whole and also of its zonal/regional office, for the last, say 5 years.	• This is a descriptive type of research, and the report's format would be simple. The study would involve collection of quantitative data from the sources within the company. No sampling involved. There is only simple quantitative analysis, and description of results/conclusions. No recommendations.
• However, if the study also includes an analysis vis-a-vis the developments that might have taken place in the economy, in general, and in the concerned industry.	• The report will change to include the analysis of developments in the industry and economy having implications for the company by collecting relevant data that would be quantitative as well as qualitative.
• Further, if the study were also to include feedback and suggestions from the users, as also the suggestions from the employees to improve.	• The study will also include qualitative study within the company, and the report will incorporate these aspects as well.
• Analysis of the performance of a company vis-a-vis the performances of other companies, and suggest a short term as well as long strategy to improve market share.	• This would involve conducting qualitative as well as quantitative studies, and the report would incorporate the findings of both the types of studies.
• Analysis of the impact of certain measures, taken in a company, on its business as also on the work environment within the company during the last six months with a view to review the measures.	• The study of the impact on business would be a quantitative study. • The impact on the work environment would have to be studied through a qualitative study.
• Whether a certain type of training imparted to sales personnel has resulted in the envisaged increase in the sales?	• The study would involve comparing the performance after the training with the earlier performance, and test the hypothesis that the increase is significant.
• Whether two different types of training imparted to the sales personnel have brought out the same impact?	• The study would involve comparing the performance after the training with the earlier performance, and test the hypothesis that the increase is equal for sales personnel receiving the two types of training.
• Analysing outstanding dues from the clients of a company	• The study would be a quantitative one.
• Suggestions about the measures to reduce overdues	• The study would be an exploratory one.
• Assessing the efficacy of direct selling (present) vs. through distributor (suggested), for the company.	• Exploratory study would have to be undertaken first before planning for further study.
• Assessment of the global sales of the concerned industry in the next three years.	• The study would be quantitative. However, it would be qualitative in nature, if relevant economic scenario at global level as also developments are to be studied.

(*Contd*)

(*Contd*)

• Career progress of MBA graduates (Women versus Men) • Factors leading to less progress for women	• Basic Research* • Descriptive—Comparative—Fact finding • Exploratory
• How to improve profitability of credit card business in a bank	• This would involve extensive collection and analysis of qualitative as well as quantitative data, setting up and testing of hypotheses, etc.

*If the same study is to be done for a **particular** company, it would be called 'Applied Research.'

15.3 A CLASSIFICATION OF THE SECTIONS OF THE REPORT

It is interesting to note that how the various sections of a report have been grouped in three broad groups in one of the classifications of the sections of a report.

Prefatory

It comprises:

- Title Page giving main objective of the research study
- Letter of Transmittal (where relevant as explained earlier)
- Letter of Authorisation (where relevant as explained earlier)
- Table of Contents
- Summary/Executive Summary containing main results/conclusions/recommendations

Main Body

It comprises:

- Relevance/Introduction
- Methodology—Methods used
- Results, Conclusions, Recommendations
- Scope and Limitations of Data, Methods and Conclusions

Appended Part

It comprises:

- Terminologies used in the Report
- Presentation of Data in the form of Tables/Charts/Graphs
- Calculations used for drawing conclusions
- Bibliography and References

In yet another classification, some portions of the report could be termed **'Cosmetic'** parts wherein the researcher can use his/her artistic creativity to make the appearance of the report attractive enough for the reader. However, it is no substitute for the quality of the report, but like cosmetics, it gives a pleasant feeling to the reader, and motivates the reader to read on.

15.4 POWER POINT PRESENTATIONS

Nowadays, a report is presented through power point presentations.

Therefore, the preparation of such presentation is similar, in a way, to the report writing, and therefore demands as much attention as to the report writing. Some tips for this purpose are as follows:

- It demands more of creativity as compared to report writing.
- Apart from font type and size, colour combination is quite important as is use of contrast—light on dark or dark on light.
- Font size should be easily readable even from the last row of the audience.
- First page should start with 'Welcome to the Presentation of …..'
- Title of the report may be condensed in 2 to 3 words, and may be on all pages. The pages may also contain logo of the Institute (for students), Company (for consultants and executives).
- A slide should contain only few points and lines containing only most distinguishing words. Charts and graphs may be preferred over text, for better grasping.
- The points may come on the screen, one after the other, next point coming only after the previous one is discussed.
- Verbs and adjectives may be avoided; could be mentioned by the presenter.
- The last slide should express 'Thanks' to the audience.

A sample is given in the CD provided with the book.

15.5 POWER OF REVISION

For improving the quality of a research report, it is advisable to revise it a couple of times. There are some persons bestowed with the capability of writing 'Best', in the first attempt. For all others, it is recommended that they should re-read and revise it at least once. This would, in all probability, result in the improved quality of the report in language, presentation and contents. It is due to the general premise that all good ideas do not come at one time—these might come during revision as well. This is illustrated through a live case as follows:

One of the Ph.D. students went to show his first paper to his Ph.D. guide viz. Late Padmabhushan Dr. V. S. Huzurbazar. Dr. Huzurbazar took the paper in his hands, and even without opening the pages, returned the paper, and asked the student to revise the paper, and meet him after two weeks. The student was a bit shaken for a moment. However, as directed, he worked on the paper again, and revised it. When he showed the paper to Dr. Huzurbazar, he took the paper in his hands, turned a few pages at random, and retuned the paper saying that the paper needed further revision. The student was rather disappointed but as a sincere student worked on the paper once again, and presented the paper to his guide after improving the language and presentation of contents. Dr. Huzurbazar asked the student about the difference in the first and third version of the paper. The student admitted that the third version was much better than the first one. Dr. Huzurbazar, then read the entire paper, and complimented the student for the good work, advised to make a few changes and then sent the paper for publication to one of the most prestigious journals in USA. The student was thrilled to receive acceptance of the paper for publication in the journal.

However, the revisions should not come in the way of timely submissions of the report.

PROJECT WORK REPORT

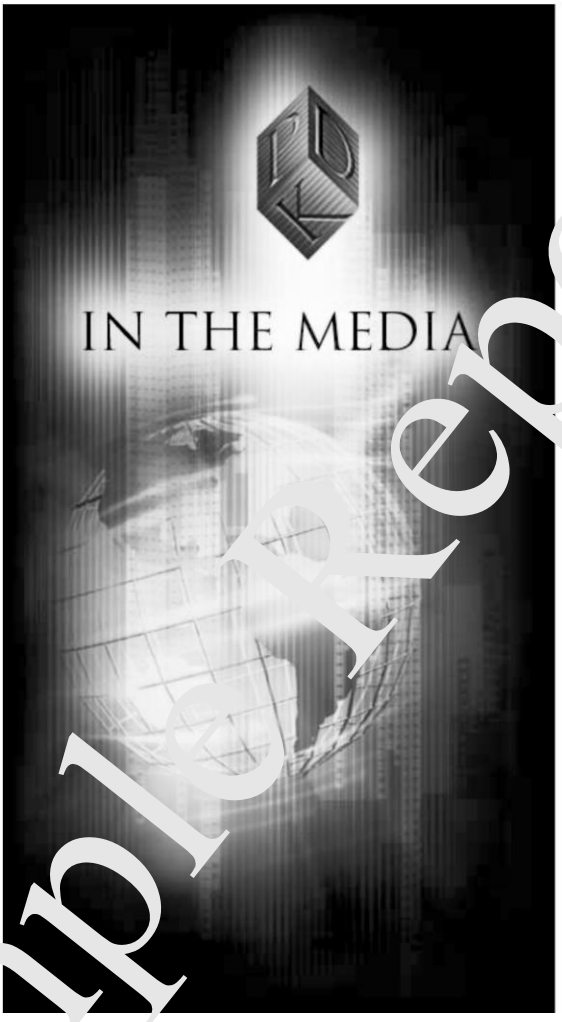
SOFT DRINK – You Prefer

Prepared by :

Mayank Aggarwal #03, Shabnam Charaniya #12, Akshay Kant #26, Suhaib Sayeed #46, Arpit Shah #49

NMIMS University

PROJECT WORK



VISHESH AGRAWAL: 4

ADITI AGARWAL: 8

SHREYANSH DEDHIA: 17

PRATEEK CHAMARIA: 18

AKANSHA MANDORA: 43

Research Annalysis Project Work

Consumers Perception of Glucose Biscuits

NICHEL ALEXANDER 05
ASHISH BORAR 10
NAYAN RANJAN 27
KAUSTAV MAITI 30
ISHAN SOMAIYA 54

NMIMS University

SUMMARY

Comprehensive guidelines for preparing and presenting a research study report are provided.

It starts with outlining the relevance of a research study report and its important criteria for fulfilling its objectives.

The format of a report is there with guidelines for each of the components of the report.

In addition to the general features of a report, the specific relevance for the research report for various types of research studies, has been provided.

The chapter concludes with guidance for preparing a power point presentation.

DISCUSSION QUESTIONS

1. Describe the format of a typical research study report with the help of a hypothetical study.
2. Describe various types of research studies and their corresponding relevance for a research report.
3. Prepare a list of ten points as a guidance for writing a research report.

Ethics in Business Research

16

Contents

LEARNING OBJECTIVES

The objective of this chapter is to acquaint the readers with all the relevant issues associated with business research. A tabular presentation of ethical issues involved at each step of research process has been provided to facilitate easy comprehension of various issues associated with the conduct of research.

"The truly wise man will know what is right, do what is good, and therefore be happy."

— Socrates (Greek Philosopher)

(Incidentally, Socrates was the first to draw attention to the intrinsic qualities of a person)

"For the right person, virtue denotes doing the right thing, at the right time, to the proper extent, in the correct fashion."

— Aristotle (Greek Philosopher)

"If you have integrity, nothing else matters. If you do not have integrity, nothing else matters."

—Alan K. Simpson (Statesman; USA)

"Commerce is as a heaven, whose sun is trustworthiness and whose moon is truthfulness."

—Baha'u'llah, Persian founder of the Baha'i religion

16.1 INTRODUCTION: ETHICS AT INDIVIDUAL LEVEL

It may be intriguing to read quotations about ethics relating to an individual in the chapter discussing ethics relating to business research. This is because the authors strongly believe that ultimately it is an individual's ethics which are crucial in any type of ethics whether it relates to business or research or even other types of ethics. There have been several instances in the recent past to justify this conviction. One of the instances is the scam by Mr. Ramalingam Raju – the ex-owner of Satyam Computers. Another instance, reported, is the UN Health Department which created undue hype about H1N1 Flu, leading to an oversupply of vaccines by those companies in which the concerned officials had vested interests.

We mention the following case wherein the recommendations based on professionally conducted business research, in an organisation, was bypassed while taking the decision by the top executive. The case is just to illustrate the role of an individual in business research ethics. This is further enforced, by the fact that all the steps/components of research process, described in Chapter 1, are performed by individuals.

The case relates to a company which was being run professionally and had good reputation. It was getting its properties insured with an insurance agency. As per the prescribed procedure, every year a note was initiated by the 'Organisation and Methods' Department recommending the name of an insurance agency for insuring all the properties. Based on the note, and the comments/suggestions of senior officials, the final decision was taken by the Chairman.

One year, the job of initiating the note was entrusted to a newly recruited officer who had requisite specialisation and experience. He noted that for the last 3 years, the award was being given to the same insurance company. He collected the relevant data for all the properties of the company and from all the eligible insurance companies. The analysis revealed that a substantial premium amount could be saved if the contract was given to another insurance company. The note was duly endorsed at all the levels, and finally reached the Chairman. The Chairman thought that he would discuss with a few top level executives the next day before taking the decision. It so happened, that the same evening, he went to a party with his wife. At the party, his wife was pleasantly surprised to meet her old friend, who incidentally was the wife of the Chairman of the insurance company which was being awarded the contract for the last 3 years. Next day, the Chairman came to the office, discussed the matter with a couple of top executives, and ruled that there was no need to change the insuring company!

However, in view of the emphasis of this chapter on ethics relating to conducting business research, we would like to highlight some of the reported unethical research behaviours by some individuals that led to developing conscientiousness about ethics in research are as follows:

- Researcher-altered data
- Reporting of research findings for research studies that were not conducted
- Non-existent co-author
- Fabricating scientific data about drug tests on hyperactive children
- Based on tests, recommended a medical device for smooth functioning of heart; the scientist had good fortune invested in the company making that device
- Not reporting risks to heart patients about use of a medicine

- Conducting tests on human beings without indicating the true reason for conducting test that involved serious side affects

Against the above backdrop, we shall discuss the ethical issues relating to conduct of business research.

16.2 ETHICS—DEFINITIONS AND NORMS

In addition to the quotations relating to ethics, cited above, ethics has been defined as follows:

Ethics (also known as moral philosophy) is defined as a branch of philosophy which seeks to address questions about morality i.e., about concepts like good and bad, right and wrong, justice, virtue, etc.

Ethics is also defined as norms or standards of behaviour that guide moral choices about our behaviour and our relationship with others.

Ethics, in day to day life, may also be defined as the action or behaviour of a person as per the accepted social norms.

16.2.1 Ethical Norms for Professionals

As regards the ethical norms for conducting business research, as of now, there is no prescribed code of conduct or ethical norms in India. However, in USA, AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH has prescribed code of professional ethics and practices. The code comprises:

- Principles of Professional Practice
- Principles of Professional Responsibility
- Standards for Minimal Disclosure
- Ethics in Publishing the Findings of a Research

We would like to add that the researcher while making the findings/conclusions public either orally or in writing must indicate the

- Scope and limitations of the study as well as findings
- Justification of sampling design and sample size
- Methodology of conduct of the research including questionnaire, type of analysis, assumptions made for analysis, etc.
- Scope and limitations of conclusions on account of data availability, assumptions made, analytical methods used, etc.

The researcher should keep in mind that quite often, the papers and magazines quote the findings of certain studies/researches. Some of these relate to the lifestyle, family relationships and food habits and could impact their lives to varying extent. He should, therefore, assume the moral responsibility for the results/findings stated by him.

Further, the researcher should refrain from attracting attention, gaining popularity or attempting to solicit further research from Government, corporate world or some other institutions, and release the results/findings only after due diligence.

16.3 ETHICAL ISSUES IN BUSINESS RESEARCH

The ethical issues in business research pervade:

- all those who are involved in research at various levels of hierarchy
- all components of a research process

Usually there are four levels of hierarchy, as described in Chapter 3, associated with the conduct of a research study, as follows:

(i) Sponsor/Solicitor
(ii) Consultant/Researcher
(iii) Researcher Team Members
(iv) Participants/Subjects/Respondents

Before we proceed further to discuss various aspects, in detail, we would like to explain the above levels in the context of a research study conducted by students of a management programme, as an integral part of the MBA curriculum.

The Faculty Member, who assigns the subject/topic for the study, could be termed as the 'sponsor'. He could decide the topic:

- On his own perception of the subject being of topical interest and relevant to the learning
- As directed by the management of the institute
- As a part of the consultancy assignment that he might have been given by some outside organisation

The group of students, as a team, who are assigned the topic could be considered as researcher. The group decides on the overall strategy encompassing research process, time schedule, allocation of responsibilities to individual students (team members), etc.

The individual students who actually conduct the study could be considered as member of the research team.

If the research involves collecting primary data by contacting companies or some other entities like retail outlets, individuals like consumers/customers, by interacting with them, then these entities/individuals could be termed as subjects/participants/respondents.

Now, we shall discuss the ethical issues at each of the above four levels as also the activities which the research team members carry out.

16.3.1 Sponsoring Research

As mentioned in Section 3.10 of Chapter 3, a research could be sponsored by:

- Government/Corporate body like CII/International Agency like U.N.
- Top Management of an organisation/company
- An Individual or a Group
- Faculty (applicable in a training institute)

The ethical issues relevant for the sponsor of the research are:

Dos:

- **Clarity of purpose without any hidden agenda**

A Director of an institute wanted to remove one faculty and appoint another (identified) faculty in his place. He knew that the faculty was taking much less lectures than others at his level but he was publishing papers and writing a book. He ordered a study to assess the work load

of faculty members, and suggest norms for lectures at the Institute. As a part of the study, data was collected about the lectures taken by different faculty members. He used that data about lectures only to convince the governing board that the concerned faculty should be removed.

- **Transparency in settlement of terms and conditions**

— With outside consultants

e.g.

- The sponsor should be transparent during project price negotiation, and should not seek to accept unfair concessions, based on the promise of future projects, which they are unlikely to offer.
- The project brief should not misrepresent the stated purpose of the research study. There should be clarity on the constraints desired, if any, by the sponsor, and should be informed of the schedule restrictions in order that a fair assessment of resource allocation is enabled.

— Within the organisation

e.g. availability of resources as per commitment, promise of reward subject to specified findings

- **Acceptance of Report**

The report should not be accepted as per convenience of the sponsor e.g. if the report envisaged evolving a new system for compensation to the employees along with more responsibilities and increased work load, the sponsor should not accept only the report relating to responsibilities and work load, in full, and marginal with respect to compensation.

Don'ts:

- Research study to bring out favourable points that could help in boosting the stock price of a company
- Designing and conducting a survey to contrive the findings justifying the thinking/ideas or action already taken by the decision-maker.
- Manipulating the data by removing/truncating data so as to influence the results in the desired direction.

In one particular year, there was a phenomenal growth in the business of a company. However, the figures were ordered to be toned down to show only good growth—the balance was to be used next year to show good growth even if there was decline or marginal growth.

- Encouraging/directing inappropriate analysis even if it is not the get the desired results.

16.3.2 Consultant/Researcher Level

- The entire research to be organised and conducted in a professional and detached manner ensuring privacy and confidentiality of the research data and findings
- Organise the research process in consultation with the management and concerned employees at all levels
- Ensuring that the sponsoring organisation provides fair and true state of affairs as also the relevant data without any distortion
- Soliciting full freedom to interact and obtain views and data from employees at all levels without any inhibition

- Using appropriate analytical tools relevant for the scope and objective of the research
- Bringing out fair and true picture in the organisation, and making recommendations without any inhibition
- There are instances where it could be obvious to the consultant, during exploratory study, that the stated objective of research could be achieved by a study that requires scaled-down resources as compared to the original assignment, and, therefore, deserving lesser compensation. The ethical consultants would inform the sponsor about this, irrespective of the reduced profits and would continue with the original assignment, only if mandated by the sponsor after deliberating on the advice

16.3.3 Individual/Group Sponsored Research

Such research is sponsored by an individual or a group within an organisation. Incidentally, such research is also called **solicited** research.

- The entire research to be organised and conducted in a professional i.e. objective and detached manner ensuring privacy and confidentiality of the research data and findings to the extent considered desirable by the top management
- Organise the research process in consultation with the management and concerned employees at all levels thus, taking into confidence the concerned departments/personnel
- Soliciting from the concerned departments to provide fair and true state of affairs and also the relevant data without any distortion
- Soliciting full freedom to interact and obtain views, data from employees at all levels without any iota of sense of 'superiority' because of the expertise in conducting of a study. In fact, paying due respect to the employees and their respective jobs ensures full co-operation in the conduct of the study
- Using appropriate analytical tools relevant for the scope and objective of the research
- Bringing out fair and true picture in the organisation, and make recommendations without any inhibition

16.3.4 Research Team

- Allocating the responsibilities matching the expertise needed and available
- Meeting at regular intervals to share the progress and sort out hindrances, if any, to ensure quality of research as per resources available
- Collecting internal data from the organisation, establishing full rapport with the sources of data (observation or collection), striving for the desired data without any compromise
- Collecting data from external sources that are reliable, taking note of their relevance, scope and limitations

16.4 ETHICAL STANDARDS IN QUALITATIVE AND QUANTITATIVE RESEARCH

As mentioned in Chapter 1, a research could be classified in two categories viz. Quantitative and Qualitative. The ethical standards and obligations in both the types of research are described below.

In qualitative research, there is a close interaction between the researcher and the participants/subjects/respondents, and so the ethical obligations to be met by both the researchers and the participants or subjects or respondents in a study, assume greater significance. We have discussed the ethical studies for both types of research separately.

16.4.1 Qualitative Research

In qualitative research, there is a close interaction between the research and the respondent. There are certain norms to be followed by both of them. These are described as follows:

16.4.1.1 Ethical Obligations for Researchers As per the accepted norms, the welfare of the participants is of primary importance as compared to the research per se. This is the underlying principle of researches conducted with the co-operation of the participants.

The three fundamentals of research ethics are:

- Respect for Participants
- Beneficence (Benefits of Research should accrue to the participants)
- Justice (Equitable distribution of benefits as well as risk, if any)

In general, while collecting data through in-depth interviews and focus groups, the participants should be informed and assured about the following aspects:

- The objective and purpose of the research and its benefits either, in general, or to the participants. There should not be any covert motive that could be detrimental to the interest of the participants.
- The researcher should keep in mind that the success of the research would depend upon his/her honesty and the confidence that the participants repose in him/her.
- Consent should be obtained from the respondents about their willingness to participate in the study.
- Expectation from the research participants, including the type of information being sought and the amount of time required for participation.
- Voluntary nature of participation i.e. option to withdraw at any time without any obligation
- Protection of their privacy and confidentiality
- Assurance of taking cognisance of their feelings and egos
- Assurance of the results and findings being shared with the participants.

16.4.1.2 Ethical Obligations for Respondents There are certain obligations on the part of the respondents also, as indicated in the following:

- Before volunteering for participating in the study, they should seek all the clarifications about various aspects of the research including the background/relevance and purpose intended to be achieved.
- Fully co-operate with the researcher in providing honest and truthful response.
- Should not withdraw from the study unless it is absolutely essential.

16.4.2 Quantitative Research

As mentioned in Chapter 2, quantitative research encompasses *inter alia* selection of the suitable study, setting up hypothesis, sampling plan, collection of data, statistical analysis, drawing conclusions, etc. There are ethical issues at each level, and are described, in details, in Section 16.6. However, we indicate below some non-ethical practices used in quantitative research

- **Hidden Agenda in the Purpose of the Study**

Designing and conducting a survey to contrive the findings for justifying the thinking/ideas or action already taken by the decision-maker.

- Manipulating the data by removing/truncating data so as to impact the results in the desired direction. This could be at any stage viz. data collection, coding, inputting, processing or even at decision-making.
- Choosing a type of analysis just for getting the desired result even if that is not appropriate.

16.5 RESEARCH ETHICS IN AN ORGANISATION

The following ethical issues are relevant for internal research within an organization:

(i) Inform all the concerned employees about the conduct of the study—its relevance and usefulness to the employees/organisation, and solicit their co-operation

(ii) If the research involves seeking information and suggestions from the employees, they have to be ensured that the information supplied by them will be kept confidential, and it will be used only for the purpose it is sought for.

In an all India organisation, an exhaustive database was to be prepared about the employees comprising factual data about their postings in various departments and various cities, training attended, etc. as also the data about the training required by them, area of responsibilities for which they could be well utilised, their preference for postings for the next 5 years, etc.

After a couple of years, the employees noticed that the only information being utilised by the organisation was for ordering place (city) of postings for those employees who had served in a city for more than 5 years. It caused a lot of discontentment among the employees resulting in the issue being taken up by their association with the top management. Ultimately, the management was 'forced' to evolve transparent policies about their assignments and training.

(iii) Salient aspects of the findings—both quantitative and qualitative—should be made known to the respondents. This would give them the satisfaction that the information and suggestions made by them have been found useful, and motivates the respondents to co-operate with further studies.

16.6 ETHICAL ISSUES AT VARIOUS LEVELS OF A RESEARCH PROCESS

In this section, we have described ethical issues at each level, in a tabular form given as follows:

Steps of a Research Process	*Some Ethical Issues/Unethical Practices*
Objective	• The objective should be transparent without any covert or hidden agenda. Ideally, it should evoke interest and enthusiasm among the researchers and the respondents, if any • There should not be any hidden agenda or covert objective of conducting a study.
Defining Problem	• The problem should be defined so as to flow directly from the objective, and should generate a sense of challenge for the researcher as well as respondents, if any • The problem should not be exaggerated and later on contrived to collect such data that supports the exaggerated view
Research Design	• The researcher should not select inappropriate design just for the ease of conducting study
Sampling	• The sample should be the representative of the population for which the result or conclusion is to be derived. The sample should not be selected so as to get the desired response. For example, in one of the newspapers, the following item was published relating to the film "Lage Raho Munna Bhai". "Media has exaggerated the role of Gandhi in the film. The success of the film is actually because of the script and the punch lines between Circuit and Munna bhai". Their finding was claimed to be based on 350 students from across the metro cities, and went on to give percentages of students agreeing or not agreeing to certain questions. It might have been better to take a much larger and representative sample of cinema goers (only those who have seen the film) when the issue revolved a serious matter as mentioned earlier involving reference to the father of our nation. • Similarly, a sample size has to be adequate to draw any valid conclusion. In fact, a sample can be so selected so as to prove anything. Sampling, thus, provides maximum scope for unethical research. Therefore, one has to be extremely cautious before believing in any findings/results.
Designing of Questionnaire	• The scaling technique used to record any respondent's opinion should be ethical, in a way that the scale does not include undue negative or positive text description, which leads the respondent to avoid those questions • It is unethical for the researcher to generate a research document that allow, for collecting data to serve the need of more than one client, without informing the parties concerned, of the method executed. Seeking to make accelerated profit by this method is unethical
Collection of Data	• A lot of sincerity and integrity is required in collection of data. Some of the common unethical practices are: ▪ Filling questionnaire without interviewing a Respondent (fictitious names given) ▪ Bias in collecting data—only those values that seem to support any pre-conceived notion ▪ Collecting only the data that is convenient to record or observe • Data collection should be such that it does not in any way reflect negatively on the respondents and does not invade their privacy.

(*Contd*)

(*Contd*)

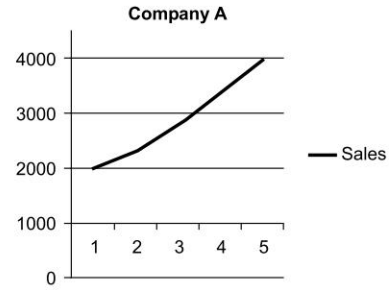
	• The need to disguise the exact purpose of research, in order to prevent a bias in the response, may to some extent require camouflaging using a cover. It is ethical to inform the respondents, in advance that the procedure is indirect to enable an unbiased response. It is also preferred that the true purpose is clarified after receiving the responses • Any use of video or voice recording equipment during the interviews should be informed to the respondents before hand • It is unethical to conduct the interview with the respondent in a manner, to receive the desired response • Use of leading questions, which bias the respondents fair and appropriate response, is unethical. The research document should be designed in to elicit a genuine response, rather than what is convenient for the researcher • It is unethical for the researcher to disclose the names of the respondents especially if they have been assured anonymity. It will be correct to inform the respondents in advance and seek their approval if their identity is planned to be disclosed. The names of the respondents should also not be used to promote sales without their explicit permission • Special care should be taken specially when the population size is small, since it is easier to guess as opposed to samples drawn from a larger population It is also desirable that quotations and opinions by respondents are delicately included or avoided in the report when sample is small, to avoid compromising the respondent's identity • It is also unethical for the researcher to inflate the standard deviation, thereby increase the sample size to enhance revenue. It is a standard practice to calculate the sample size based on an estimate of the standard deviation, it is possible to increase the sample size by inflating the standard deviation.
Ethics in Information Technology	There are many ethical issues that arise from the use of information technology. The most important traits of IT Department personnel are that they should be honest and trustworthy. It assumes more significance because what they do is not easily verifiable. • Coding data • Inputting data • Maintaining security and confidentiality of data • Software for generating reports It is easy to gain access to information. It could lead to privacy invasion. • There is an ethical obligation for the researcher to use only those data that is relevant to the research. The decision to use secondary or primary data to arrive at the conclusion should be judicious and not driven by profit alone. As primary data can be more expensive, secondary data should not be used just to reduce costs.
Presentation of Data:	The data should not be presented so as to deliberately cause visual deception.

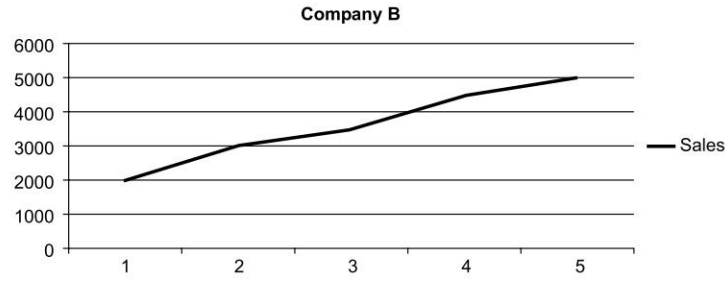
Company A		Company B	
Year	Sales	Year	Sales
2005	2000	2005	2000
2006	2300	2006	3000
2007	2800	2007	3500
2008	3400	2008	4500
2009	4000	2009	5000

(*Contd*)

(*Contd*)

Data





It may be noted just by increasing the width of the time interval on the x-axis, the company 'A's performance appears to be better than B.

Analysis of data

- The analysis of data has to be as per the planned methodology. If any change is contemplated after collecting the data, it has to be justified.
- It is an ethical responsibility of the researcher to ensure that the correct method is utilised to control the errors caused by variables. A pilot study is recommended wherever possible. This will allow an initial evaluation, before the bulk of the research is conducted, a continuous interval monitoring of research work is also recommended to allow improvement where necessary.
- The assumptions made in the model used for the data needs to be validated before arriving at conclusions. For example, regression analysis requires the assumptions of linearity, normality, etc. These should be validated before using the regression equation for prediction.
- The data values, that might lead to results contrary to the findings being contemplated, should not be ignored

Conclusions

Interpreting Conclusions and Report Writing

- Interpretations have to flow directly from the conclusions, giving valid arguments. Technical language and jargons should not be used to contrive the desired results and conclusions.
- It is ethical for a researcher to ignore data that is irrelevant and of questionable quality, before conducting an analysis. The basis for ignoring such data, and the percentage of the sample affected must preferably be included in the report.
- Conclusions have to be factual giving margin of error and confidence that can be placed therein. It is judicious to take that part of analysis that supports the intended hypothesis

(*Contd*)

(*Contd.*)

	• It is important to note that since confidence interval can be impacted by standard deviation. If in a study, the standard deviation is too large, it may affect the confidence interval. It is ethical for the researcher to share and jointly decide with the client the course of action, instead of miscommunication of the confidence interval.
Web-Assisted Research	With the advent of information technology, research work has been assisted by data available on the net. Data warehousing and mining has largely contributed to the ease in research work. The present ability of the user to subscribe to cloud computing, that offers the user an option to request additional resources on demand, and just as easily release those resources when no longer needed, have allowed for greater control and flexibility on research expenses. While the traditional data offerings can also include this facility, in cloud computing it is automated. There are several instances when data could be collected from routine transactions. Biometrics data such as fingerprints, digital computer monitoring, smart cards, etc. are sources of data. The other sources of data are the application forms submitted for credit cards and shopping mall memberships, etc. It should be an individual's choice, to have the right to online privacy and expect that they have a choice on how information is collected, used and shared. It is, therefore, important that raw data is protected from prying eyes. Data security, quality and measures to maximise the accuracy of personal information is an ethical obligation of those that put together information used by others. While filling the data online, routine pop-ups could be generated by hosting agencies, advising that the data could be accessed by others. This is a way of securing an informed consent from the user.

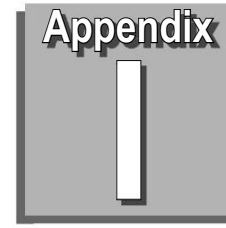
SUMMARY

The ethical issues at individual as well as organisational levels have been highlighted. Further, there are ethical issues that are relevant in business research at various hierarchical levels, ethical standards in qualitative and quantitative research, and ethical issues relevant at every step in the research process. A tabular presentation indicating the extent of variation in research reports that are suitable for different types of research studies is given.

DISCUSSION QUESTIONS

1. Highlight the importance of ethics, in general, and in business research, in particular.
2. Why the standards are different in Qualitative and Quantitative research? Explain with reference to a particular study.
3. Describe the ethical issues in business research at various hierarchical levels, with reference to a particular research study.
4. Describe, in brief, the ethical issues at various steps of a research process.
5. Write short notes on:
   (i) Ethics at individual level
   (ii) Ethics at organisation level
   (iii) Ethical norms for professionals

Indicative Topics for Business Research



Several research studies have been reported in books and other media. Equally relevant are the research studies conducted by students at Management Institutes. Here is a list of research studies which have been or could be conducted by students. The list is only indicative and not comprehensive.

Areas	*Topics for Research*
Marketing	Assessing demand of a product or service—at the national, regional or company level
	Brands—how has it changed the perspective of an Indian Consumer
	Competitive intelligence in FMCG sector or a company
	Consumer preferences on mobile Internet
	Consumption pattern of Media Mix for Retail Chain
	Customer Behaviour—Preference and patterns (for various products)
	Customer Profiling
	Customer Relationship Management (CRM)
	Differentiate customers on the basis of three independent variables viz. frequency, average purchase and number of years the customer has been purchasing from a showroom
	Digital marketing—Pros and cons
	Distribution planning in FMCG industry
	Distributor business model, lubricants for B2B customers
	Effectiveness of advertisement campaign
	Effectiveness of launching a product
	Effectiveness of various promotional activities
	Effectiveness of word-of-mouth marketing
	Effects of marketing on sales for a particular industry
	Growth of digital marketing vs. traditional marketing
	How consumer behaviour affects product failure?
	Impact of celebrity endorsement of consumer buying behaviour
	International marketing challenges in textile industry—current international trends
	International marketing
	Is a customer really treated as a king?

(*Contd*)

(*Contd*)

	Market development in joint replacement implants
	Marketing 'Home Automations' in India the cost divide
	Marketing in remittance hub (money transfer)
	Marketing potential of LED lighting and emerging technologies
	Petrochemical marketing
	Relationship between advertising expenditure and sales
	Sales force automation
	Social networking websites as a marketing tool
	Effectiveness of out-of-home advertising in Indian market
	Assessing brand awareness in the market—for specific brand
	Trends in FMCG industry
Finance/Economics and Banking	Bailout, a right solution by the government for the economy
	Bank's money transfer service in India: Qualitative study of two banks
	Basel II: Advantages and challenges for the Indian Banking Industry
	Brand Equity Analysis of acquisition strategies of a bank
	Credit derivatives—Risk mitigation or amplification
	Credit rating procedure of banks/financial institutions
	Credit Risk Management—Credit ratings in banking industry evolved with the era
	Currency derivatives and its impact on the corporate India
	Current trends and challenges in the mortgage industry
	Analysis of debt to equity ratio of a selected group of companies
	Derivatives/futures and options
	Derivative products in international markets
	Different mortgage products offered by Indian Mortgage Industry and the options of introducing new mortgage products in India
	Economics of data acquisition and information security spending in the financial sector in India
	Effect of subprime lending on financial institutions
	Effects of dollar fluctuation on Indian Trade (EXIM)
	Equity research
	Evaluation of performance of credit cards in India
	Factors affecting users choice of bank
	FDI or FII—which is a better option for Indian economy
	Feasibility of implementing mobile banking systems for urban India
	Feasibility of V.L.S.I. design services in India
	Feasibility study: should RBI introduce Credit Default Swaps (CDS) in India?

(*Contd*)

(*Contd*)

	Financial re-engineering of a company
	Fixed income and money market in India
	Forex management
	Further liberalisation of financial markets in India
	Game theory: applications in commerce and industry
	Government treasury bills in an economy
	Identifying critical factors responsible for the rising level of NPA in Indian banking system
	Identifying key factors and financial ratios for successful and unsuccessful companies
	Identifying traits of good and defaulting borrowers
	— auto loan
	— credit cards
	— home loan
	Identifying relevant factors for success in mergers and acquisitions
	Impact of transportation system on economic progress of a country
	Implementation of Basel II in Indian banks with particular focus on credit risk management
	Indian Mutual Funds Performance 2001-2010
	India's export policy and its impact on the country's economic development
	IPO: means of finance for SMEs: benefits vs. disadvantages
	Issue management (investment banking) and launch of NFO scheme analysis
	Issues concerning valuation of venture capital and private equity funds
	Key account management
	Mergers and acquisitions—Opportunities and challenges
	Recovery of education loans
	Mutual funds for retail/individual investor
	Study of mutual funds in India
	Operational asset effectiveness
	Process innovation in retail banking
	Project finance for major infrastructure projects
	Quality of assets
	Receivables and fixed asset management
	Relationship between interest rates and bank deposit patterns of customers
	Risk in international trade
	Risk management
	Setting up private equity venture capital funding for small to medium enterprises
	Short-term trading strategies for equity markets

(*Contd*)

(*Contd*)

	Significance of volatilities in derivative markets
	Study of credit ratings in India
	Study on currencies of emerging market as a potential alternative to dollar
	Study of security issues and risk management in online banking
	Technical analysis of stocks and commodities
	Application of clustering techniques to physical representation in asset performance benchmarking systems
	Applications of modern portfolio theory to multi-asset management efficiency in property portfolios
	Distortion of India FDI policies in the new era: coalition between foreign investor and local government
	Economic effects of India's foreign policies
	Impact of Internet on banking
	The planning, evaluation, financing and implementation of projects: a critical review of experience in India
	Trends in billing cycle management of SME contracting firms
	Trends in working capital management of SME contracting firms
	Venture capital for real estate funds
	What are the consequences of the Internet for banks and brokerage firms?
Operations	Study of import purchasing decision behaviour
	Comparative analyses of transport provision, land-use patterns and travel conditions in cities in developing and developed countries
	Decision-making framework for managers: Profit by forecasting, costs and price management
	Determining the procurement quantity and time of procurement
	EAI—Enterprise Application Integration
	Effective Resource Management in projects
	Evaluation of the organisation with reference to Organisational Project Management Maturity Model (OPM3)
	Impact of GST implementation on logistics cost for pesticides industry: Bayer perspective
	Inventory Management through Kanban systems
	Logistics management of chemicals in refineries
	Monitoring and analysis of road traffic speeds using GPS technology
	Operations research
	Optimising operational efficiency
	Process management
	Process migrations and outsourcing
	Reduction in the turnaround time for a mortgage offer

(*Contd*)

(*Contd*)

	Role of logistics in supply chain management
	Scope of supply chain and procurement in retail
	Service innovations
	Six Sigma—Controlling and improving production process and quality
	Six Sigma Methodology in service sector
	Supply chain management
	Technology absorption
	Temperature control warehousing
	Development of transportation modes in India
	Evolution of ERP in our new global economy
	Influence of highway geometry and traffic conditions on vehicle speeds and driver behaviour.
	Provision, impacts and funding of concessionary public transport fares for elderly and disabled people
	Role of market forces and regulations on transport policy
	Track record of Build, Operate and Transfer (BOT) Projects in India
	Third party inspection services in engineering industry in India
	Impact of on time performance on passenger preference in airline industry
	TQM implementation by training initiatives
	Transportation issues in less developed countries like India
	Use of remote process monitoring as an optimisation tool
HRD	A positive progressive approach to training and learning development
	Absenteeism and motivation in production units
	Attracting, selecting and retaining IT professionals
	Conflict management in the workplace
	Contingent workforces in the hospitality industry
	Diversity of workforce contributes more to organisational efficiency
	Does increased fitness cause a greater reduction in stress?
	Employers' experiences of shortages of skilled workers in India
	Exploring the linkage between managerial cognitions and organisational effectiveness
	Facilitation of action learning groups: An action research and grounded theory investigation
	Factors affecting performance of employees
	HR challenges during mergers and acquisitions
	Identifying traits of successful and not so successful employees
	Impact of downsizing
	Impact of training on employee performance
	Leadership: being an effective project manager

(*Contd*)

(*Contd*)

	Management-subordinate relationships in call centre environments
	More men than women are whistle blowers
	Motivating factors for call centre employees
	Organisational behaviour: development of flexible expertise in problem-solving by managers
	Outsourcing: managing inter-organisational relations
	Performance appraisal and reward system
	Performance/productivity measurement
	Recruitment and retention policies
	Recruitment practices of Indian family business
	Studying human resource management (HRM) practices across firms: implications for practice
	The role of emotional intelligence, personality, moral reasoning in ethical decision-making process
	Relationship between age and job satisfaction
	Women are more motivated than men
IT	An analysis of internet security
	An analysis of outsourcing information systems
	Animations business opportunities, challenges
	Application of Six Sigma for software quality improvement
	Business intelligence—A holistic view
	Business intelligence and data mining
	Challenges in patenting of information technology in India
	Comparing two websites of companies
	Cost modelling and neural networks
	Critical analysis: The mind of a hacker
	Data mining and business intelligence
	Decision Support System
	E-Governance strategies—Review and recommendations
	Effect of information technology on productivity
	Ethics in the information age: Privacy versus freedom of information
	Examining the causes and impacts of delays and disruption to large-scale software projects
	Feasibility study of cloud computing
	Global financial meltdown: Challenges for IT sector in India
	Harnessing potential of knowledge management to achieve sustainable competitive advantage within organisations
	Identification of issues and problems facing information systems professionals

(*Contd*)

(*Contd*)

	Impact of information technology on governance i.e. e-Governance
	Impact on QMS of a software company
	Internet privacy and institutions
	IT in BPO/KPO space
	IT security and risk management
	Moving IT infrastructure offshore
	Network management
	Neural networks/fuzzy logic as decision support systems
	Online auction assisting potential bidders through information systems
	Online banking: advantages and disadvantages
	QA activities on in-house financial software
	Remote infrastructure management in IT industry
	Safety and privacy: Internet security for the private user
	Sailing through rough weather: Journey of Indian 3D animation industry
	Study of online gaming in India
	The business of open source software
	The growth of E-commerce
	Trends in the IT industry
	Trends in IT outsourcing
	Users'/client's perception of a good web design
	Value addition of quality assurance processes to software testing
	Web 2.0—opportunity or a fad
	Website design and management
Strategy	BPO boom in India
	Enterprise application and integration
	Evaluation of cost cutting strategy
	Feasibility of setting up computer education centres
	Impact of US recession on Indian BPO industry
	Innovation as part of business strategy
	Knowledge management in BPO space—opportunities and challenges
	Setting up an enterprise in India
	Strategy formulation in IT/ITES industry in response to different economic/business scenarios
	Study of Indian outsourcing industry
	Study of process improvement tools for BPO industries in India
	The BPO industry—The call routing process and the factors affecting its revenue
	The impact of high attrition in BPO industry

(*Contd*)

(*Contd*)

	Supplier partnership for quality management
	Transition management in BPO's—A review
Insurance	Indian Life Insurance Companies foray into health and pensions products
	Technological breakthroughs as a source of competitive advantage for Indian insurance companies
	Analysis of the life insurance industry in India
	Bancassurance—Global trends in insurance
	Designing insurance policies of various types
	Future of insurance in india
	Growing insurance industry in India
	Impact of different factors on health and life
	Impact of PFRDA regulation on insurance pension policies
	Market leaders in insurance—ULIP products of insurance company
	Pros and cons of centralisation/decentralisation in general insurance back office operations
	Service excellence—A source of competitive advantage for life insurance companies in India
	Understanding the modern insurance to design better consumer stimuli
Telecom	Criteria for selection of phone and service provider among different age/income/professional groups
	Declining ARPU's—A concern for Indian telecom industry
	Effects of changing economic scenario/business environment on telecom industry
	FDI in India and its effect on telecom sector
	Managed services prospects in telecom industry in India
	M-Commerce—landscape, applications, trends and future
	Opportunities with 3G technology for Mobile VAS in India
	Sustainability in Indian telecom market
	The growth of the wireless communications industry
	The Indian Telecom Industry—Focus on cellular market growth strategy and upcoming technologies
Retail Management	Evaluating declining sales of a retail outlet
	Growth of luxury retail in India
	Identifying customer buying behaviour: Preferences and patterns
	Impact of economic slowdown on organised retail in India
	Impact of Foreign Direct Investment on the retail industry
	Research on consumer behaviour in malls
	Organised retailing in India—Opportunities and challenges ahead

(*Contd*)

(*Contd*)

Typical Research Studies	Banks failure to achieve higher growth rate Efficiency of operation and work climate
	Environmental scanning studies to assess business opportunities
	Feasibility studies for a launch of a product
	How to increase deposit growth?
	How to make enterprise competitive?
	How to reduce cost?
	Reducing NPAs in banks
	Use of new managerial tools at industry level
Pharmaceutical/ Medicine/Biotech	Impact of new varieties of seeds on the yield of a crop Study of corporate social responsibility in pharmaceutical industry
	Delivering better patient care: promoting well-being and performance of health care professionals
	Healthcare scenarios in India—Awareness and correct treatment modalities
	Impact of IPR on Indian pharmaceutical industry or growth of Indian pharmaceutical industry in last decade
	Issues in clinical trial
	Strategy for new pharmaceutical product launch
	Study of strategic techniques for outsourcing pharmaceutical services
	Surgical management of brain stroke patient with co-relation of coiling product
	The impact of promotional material on sales of pharmaceutical prescription drugs
	To find market potential for bulk and generic drugs in East European countries, Latin America and Premium South Asian countries
	Use of new technology in pharmaceutical manufacturing industry
Miscellaneous	Vocational training for rural Indian youths
	"To find out the awareness about HIV Aids in young population in India"
	An analysis of the power (electricity) crises in the country and the sources of fulfilling the needs
	Challenges and future prospects of Indian nuclear industry
	Consultant-client relationship In India
	Cricket vis-a-vis other sports in our country
	Eco-tourism—trends and prospects in India
	Effect of technology on the development of the lottery industry
	Effects of economic slowdown on airlines in India
	Facilitating independent mobility for elderly and disabled people in the context of changing social expectations, economic circumstances and demographic trends
	Failure of initial public offer—reliance power
	Feasibility study of having a central body for maintaining KYC details

(*Contd*)

(*Contd*)

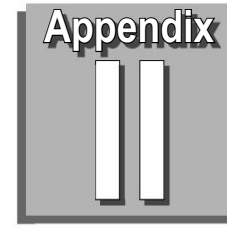
	Fund raising: India v/s other countries
	Future of background screening in India
	Future of batteries in Industrial applications
	Future of entertainment industry in India
	Future of gamma radiation processing in India
	Industrial applications of nuclear radiation
	Future of public transport system in Mumbai
	Future of SMART GRID Solution in India—with special focus on Pvt. Distribution Co. in metro cities
	Future of BTL as a mode of communication with the Consumer in India
	Global warming: Implications on the world environment
	Growth of India's Contract Research Organisation industry
	Holding the line: Call centres, resistance, accommodation and self-impact of newspaper on children
	Impact of research and development on real estate industry
	India housing market analysis
	Indian processed foods industry: A vision
	India's green revolution: Forecasting its agriculture success
	Investment in gold during recessions
	Issues and prospects of the real estate sector in India
	Media—Its avenues and effects
	New Topic: Evolution of ticketing in aviation industry
	Past, present and future of cancer treatment in India
	Private equity real estate market in India
	Public private partnership for infrastructure development in India
	Real estate: Project feasibility study
	Research in job portal industry
	Reuse of STP treated water for nonpotable application
	Setting of energy in rural India
	Sources of power generation and its effectiveness in the Indian context
	Study of a gold rated green building in Mumbai under the category core and shell
	The current and future status of women in India
	The dynamics of an "Industry" called cricket
	The emergence of stronger individualism in India
	Assessing casual clothing preference of youth
	Popularity of Indian fast food v/s continental fast food in Indian market
	Bottlenecks in iron-ore exports from Indian sea ports

(*Contd*)

(*Contd*)

	Bottlenecks in using "Renewable Electric Generation" resources
	Trends in clinical research outsourcing
	Trends in Indian automotive paint industry
	Wine market in India
B-School	Association between educational background and performance in terms of grades for PGD students
	Creativity in training systems
	Economic growth in India: The role of human capital and education
	Factors contributing to the satisfaction/dissatisfaction levels in MBA programme
	Future of e-learning in India
	Work stress experienced by students in management institutes

Excel—A Tool for Statistical Analysis



A.1 INTRODUCTION

Microsoft Excel, commonly known as MS Excel, is one of the popular application softwares included in the MS Office package of Microsoft Corporation. Excel is popular mainly for the use in financial analysis. Excel also comes with in-built statistical functions and various other applications that help statisticians carry out statistical analysis. However, Excel's capability to perform statistical analysis is not as popular as financial analysis. Here we focus on the use of Excel for various types of statistical analysis, described in this book.

A.1.1 Opening, Saving and Closing Files

Excel is one of the programs of the MS Office software. To use Excel, it must be first installed on the computer. If it is already installed, it can be opened either by selecting Start → Programs → Microsoft Excel, from the task bar, or by clicking on the Excel Short Cut which is at the appropriate location. We have described the various uses, in steps, detailed below.

Opening an Excel Sheet

A file in Excel is termed as a workbook. A workbook may have several worksheets in it. A workbook can be opened by following two methods:

- By clicking on File-Open to open/retrieve an existing workbook and changing the folder or drive to look for files in other locations
- By creating a new workbook and by clicking on File-New-Blank workbook.

Saving and Closing a Workbook

To save the workbook with its current file name, location and file format, one could click on File - Save. If one is saving for the first time, one could click on File-Save; choose/type a name for the workbook; then click OK. Also use File-Save, if one wants to save to a different filename/location. One may also save the file by clicking on the file save icon of the toolbar.

When one finishes working on a document, one needs to close it. One can do it by selecting File → Close option from the file menu. If one has made any changes since the file was last saved, Excel would ask, if one wishes to save them. One may also close the file by clicking on the file close icon of the toolbar.

Below is a snapshot of a typical Excel workbook. We have indicated different parts by numbers. We elaborate each of these parts, as follows:

1. The Title Bar: It shows the name of the workbook one is working in.
2. Menu Bar: Has different options and is used to access and execute commands like file-> save, discussed above.

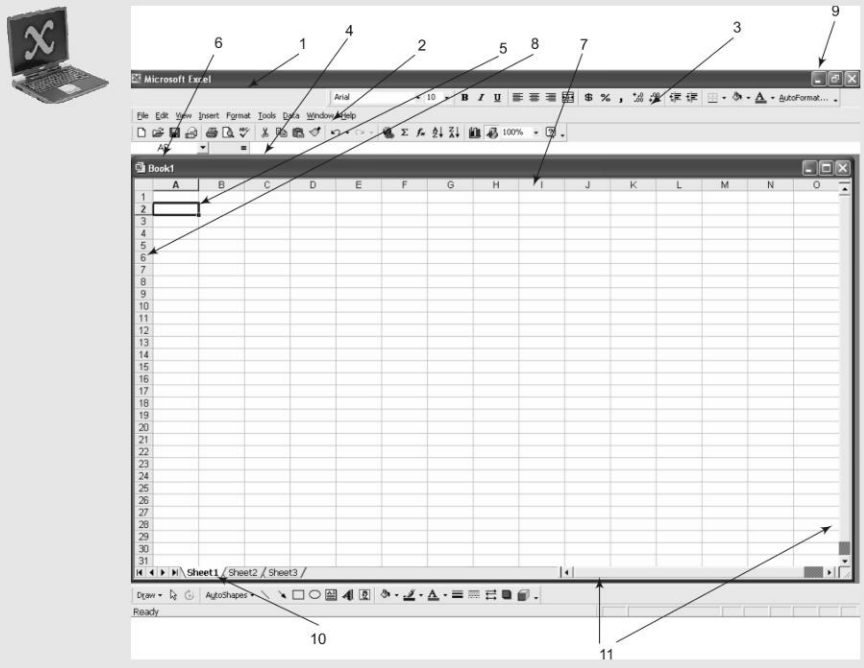


Fig. A.1 Understanding the Excel Screen

3. Tool Bar: Easy access to commonly used functions. One can work on excel by using tool bar and avoid using menu bar. This makes working easy as well as it saves time.
4. Formula Bar: It is useful to write formula in the worksheet. If functions are added using the menu bar, the respective formula is displayed in the formula bar.
5. Active Cell: A cell is the intersection of rows and columns. Each box denotes a cell and has an address, e.g. A1 denotes cell from first column and first row A2 denotes from first column and second row, B1 denotes cell from second column and first row, etc. Active cell is the cell where the Excel cursor is. Any entry made will be added to the active cell.
6. Cell Address: It gives the cell address of the active cell.
7. Column number: This gives the column address of the cells.
8. Row number: This gives the row address of the cells.
9. Minimise/Maximise/Close Button: It has three icons: first is minimise, second, maximise/restore, and third, close, which could be used to carry out respective functions.
10. Worksheet Tabs: These are useful for moving between the different worksheets.
11. Scroll Bars: Move vertically or horizontally to different areas of worksheet.

Workbooks and Worksheets

When the Excel is started, a blank worksheet is displayed which consists of a multiple grid of cells with numbered rows down the page and alphabetically numbered columns across the page. Each cell is referenced by its respective row address and the column address, as indicated in the fifth point above.

The current worksheet is called the active worksheet. To view a different worksheet in a workbook, one may click the appropriate worksheet tab (given as point 10 above).

One can access and execute commands from the main menu or by pointing to one of the toolbar buttons. When one places the cursor over these tool bar buttons, the Excel displays the name/action of the button.

Moving Around the Worksheet

It is important to be able to move around the worksheet effectively because one can only enter or change data at the position of the cursor. One can move the cursor by using the arrow keys or by moving the mouse to the required cell and clicking. Once selected, the cell becomes the active cell and is identified by a thick border; only one cell can be active at a time.

To move from one worksheet to another one may click the worksheet tabs. The name of the active sheet is shown in bold.

Moving Between Cells

These are some commonly used keyboard shortcuts to move the active cell. These are:

- Home—moves to the first column in the current row
- Ctrl+Home—moves to the top left corner of the document
- End and then Home—moves to the last cell in the document
- Ctrl+End—moves to the last cell in the document
- Home and then End—moves to the last column of the current row

To move between cells on a worksheet, one may click any cell or use the arrow keys. To see a different area of the sheet, one may use the scroll bars and click on the arrows or the area above/below the scroll box (indicated in point 11) in either the vertical or horizontal scroll bars.

A.1.2 Entering Data

A worksheet is a grid of **rows** and **columns**. Intersection of a row and a column is a **cell**. Each cell has an **address**, which is the column letter and the row number. The arrow on the worksheet to the right points to cell A2, which is currently **highlighted**, indicating that it is an **active cell**. A cell must be active to enter information into it. To highlight (select) a cell, one may click on it.

To select more than one cell one may use any one of the following:

- Click on a cell, then while pressing and holding the shift key, click on some other cell to select all the cells between and including the two cells.
- Click on a cell, drag the mouse across the desired range, and click on another cell to select all the cells between and including the two cells.

Many a times, one comes across a situation when one needs to select several cells which are not adjacent. In such a case, one may press and hold 'control' key, and click on the cells one want to select.

To select entire row or column, one could click a number or letter labelling a row or column.

Each cell can contain a text, also termed as label, a number or a value, logical value i.e. true/false, or a formula.

- Labels can contain any combination of letters, numbers, or symbols.
- Values are numbers. Calculations can be done only with the help of values. Excel considers date or time also as a value.
- Logical values are 'true' or 'false', which are either output of a formula or are used in a formula.

- Formulae automatically do calculations on the values in other specified cells and display the result in the cell in which the formula is entered. It may be noted that the formula would not be displayed in the cell but just the result of the formula will be displayed. One could look into formula bar (part 4 mentioned above) to read the formula.

A snapshot of a excel workbook is given below:

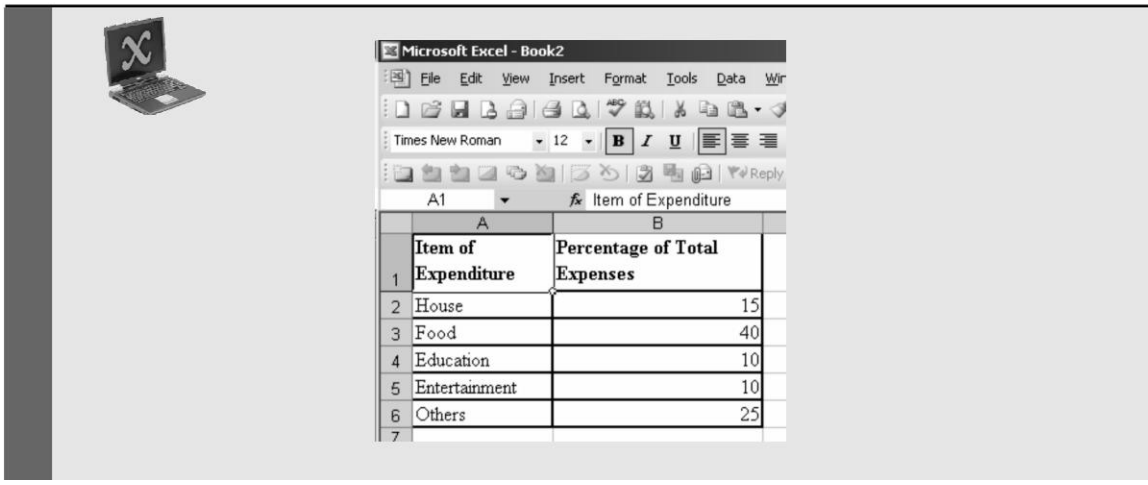


Fig. A.2 Entering Data in Excel

In this workbook, the active cell is A1 which is indicated by highlighting of Column A and Row 1. To enter information into a cell, one may select the cell and begin typing. As seen above, one may select the cell A1 and type "Item of Expenditure" in the cell.

It may be noted that as one types some information into the cell, it is also displayed in the formula bar.

After one completes typing in the cell, one could either Press "Enter" to move to the next cell below (in the above case, A2) or this could be done with the help of Tab key which would take cursor to cell B1. One could also select any cell by clicking on it.

Unless the information one enters is classified as a value or a formula, Excel interprets it as a label. The default alignment for a text or label is 'left'.

Entering Values:

A value is a number, date, or time, plus a few symbols if necessary to further define the numbers (like., +, –, ( ), %, $, /, etc.).

Numbers are assumed to be positive; to enter a negative number, one may use a minus sign "-" or enclose the number in parentheses "( )".

Dates are stored as MM/DD/YYYY, but one does not have to enter it precisely in that format. If you enter "jan 1" or "jan-1", Excel would recognise it at January 1 of the current year, and store it as 1/1/2007.

The default alignment for number is 'right'. Excel identifies any value as a number, if it does not contain any characters in it. E.g. '34a' would not be considered as a number.

A.1.3 Entering Formula and Functions

A formula starts with '=' sign. There are formulae available in excel for various functionalities. A formula may contain a number or the cell reference of some other cell. For example, if one wants to total the 'percentage of total expenses' in the above Fig. A.2, one could use 'sum' function in either of the following two ways.

- = sum (15 + 40 + 10 + 10 + 25), or
- = sum (B2:B6)

It may be clear from the above example that using cell reference is better than using numbers in the formula. Some of the advantages, other than ease of use, of cell referencing are,

- If the formula contains cell reference rather than the value, and if the values of cells are changed, the formula automatically updates the result.
- If similar formula is required in any other cell, one can directly copy this formula and paste it in the required cell. Excel would copy the formula but not the value.

We shall discuss more about formulae and functions in the subsequent Sections.

A.1.4 Absolute and Relative Referencing

The default copying in excel is with relative referencing.

Excel identifies two types of referencing, viz.:

- Relative Address, and
- Absolute Address.

Relative address is the addressing system where the formula, when copied to another cell, copies the formula with the reference. We explain this with the help of the Excel snapshot below.

Microsoft Excel - Book2

File Edit View Insert Format Tools Data Window Help

Times New Roman 12 B I U $ % , Reply with Changes... End R

B7 =SUM(B2:B6)

	A	B	C
1	**Item of Expenditure**	**Actual Expenses**	**Percentage of Total Expenses**
2	House	1500	15
3	Food	4000	40
4	Education	1000	10
5	Entertainment	1000	10
6	Others	2500	25
7	Total	10000	100
8			

In the above snapshot, the active cell is B7 and the formula entered in the cell is

=SUM (B2:B6)

If this formula is copied to the next cell at the right, i.e. to the cell C7, it would get copied relatively. It would be now

=SUM (C2:C6)

and the value of this formula would become 100.

If the same formula is copied down B7, i.e. in the cell B8, it would be now,

=SUM(B3:B7)

And the value of this formula would become 18500 (excludes 1500 and includes 10000).

Excel shifts the reference of the cell as it is copied. It would increase the column address, if copied at right, and would increase the row address, if copied down. This is very useful especially when one wants to use the same formula for different rows and columns.

Absolute address is the address which remains the same irrespective of the direction (left, right, up or down) in which it is copied. To make an address as absolute, one may assign a $ sign before the row as well as column. For example, in the above case, if the formula entered in the cell B7 was,

= SUM($B$2:$B$6)

then, if one copies this formula anywhere in the worksheet, it would remain the same, and the value at that place would also be same, i.e. 10000.

One could also have partial absolute address. When one puts $ for either row or column address, it is called partial absolute address. For example,

$B2 – means the column is absolute, and the row is relative. This formula when copied to next column would remain same i.e. $B2, but when copied to next row, would become, $B3.

B$2 – means the column is relative, and the row is absolute. This formula when copied to next row would remain same i.e. B$2, but when copied to next column, would become, C$2.

A.1.5 Editing Data

Many a times, we come across a situation where one needs to change or edit the existing data. This may involve adding, deleting, replacing etc. To add to the existing cell, one could go to the cell, and press F2. This will edit that cell. If one wants to replace the existing contents of a cell with new contents, one could directly enter the new contents, it would overwrite the existing contents.

To insert row, one could click on row address and right-click and select the option ‘insert’ row.

To insert column, one could click on column address and right-click and select the option ‘insert’ column.

The same method is to be followed for deleting a row or column.

A.2 USING EXCEL EFFECTIVELY

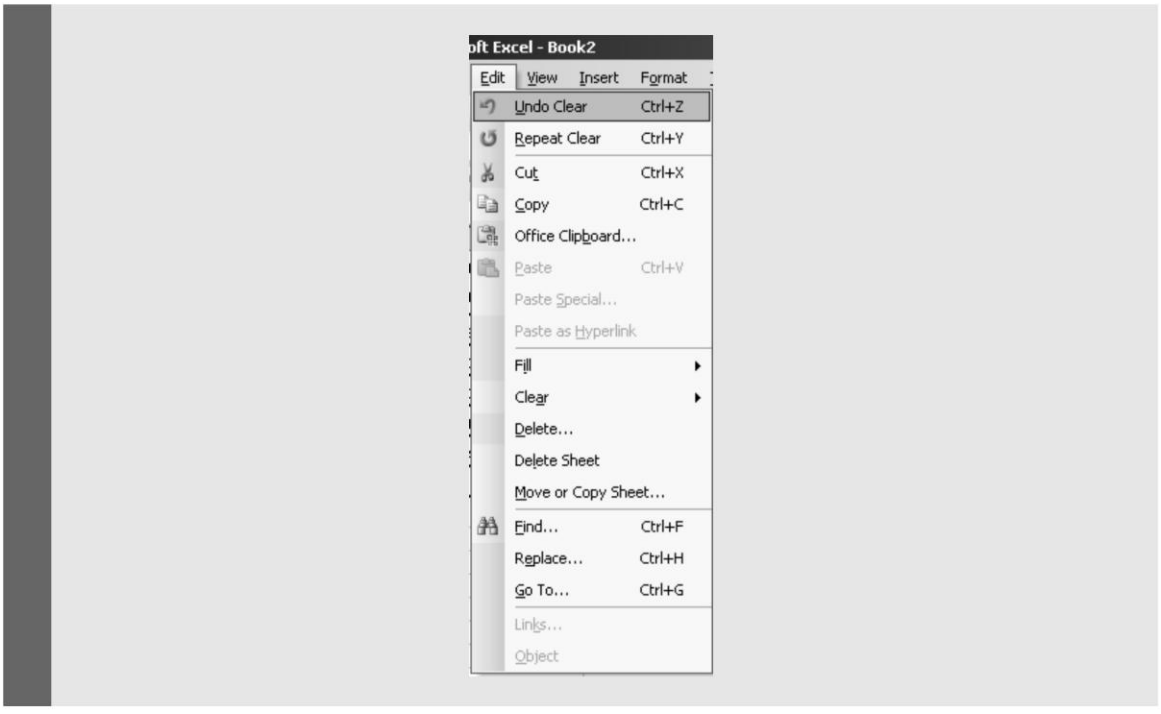
There are three types of shortcuts for effective use of Excel. These are:

- Shortcut keys using keyboard
- Shortcuts using tool bars
- Mouse shortcuts

We shall discuss each of these, in the subsequent sections.

A.2.1 Shortcut Keys Using Keyboard

Keyboard is the most efficient tool to be used in any programme. The keyboard is standard, and entering keys is faster than searching an option with the help of a mouse and then clicking it. Excel provides many keyboard shortcuts for the menu options. In fact, if one clicks on the menu, against some option, one would find some keywords as shown in the snapshot below.

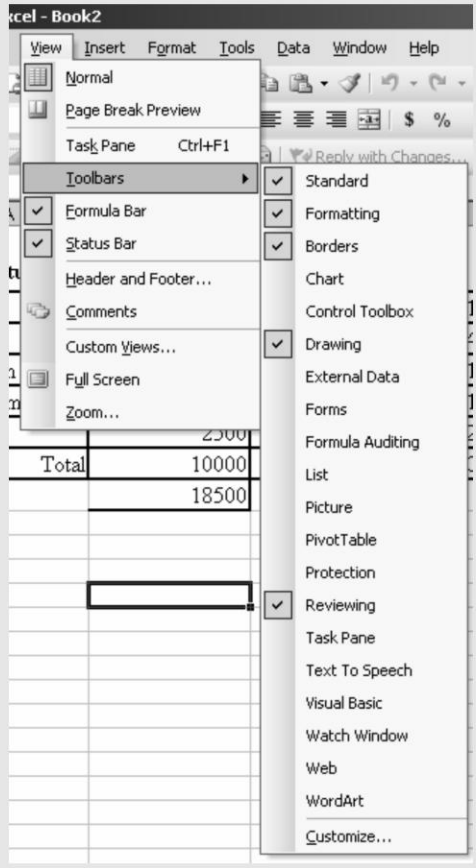


The above snapshot shows the edit menu of the menu bar. It shows the options as well as some keywords. For example, in the first option 'Undo Clear', the key word is Ctrl + Z. This keyword is the shortcut key for that option. Excel has provided keywords or shortcuts for most options in the menu bar. We have provided a list towards the end of this Chapter.

A.2.2 Shortcuts Using Toolbars

Excel has provided different toolbars that allow direct selection of the options. The advantage of toolbar over keyboard is that one need not remember the command. Every option (in the form of icon) is available on the screen, and one needs to just click on the required option. Sometimes, the required toolbar would not be on the screen, in such cases, one could select the required toolbar from the menu option 'View → toolbars'.

In the snapshot, given below, the different toolbars with ✓ sign are already present on the excel screen. For adding a toolbar, one could click once on the respective toolbar. If one wants to remove some toolbar from the screen, one could click on the ✓ sign against the toolbar, and the toolbar would disappear from the screen. For knowing more about each toolbar, one could take the mouse to the tool icon (button), and Excel would display the name/action of the button/icon.



A.2.3 Mouse Shortcuts

At any location in Excel, one could use right click of mouse to get a small option list specific to that location. This is quite convenient as the options in the list are less, and it is easier to pick them.

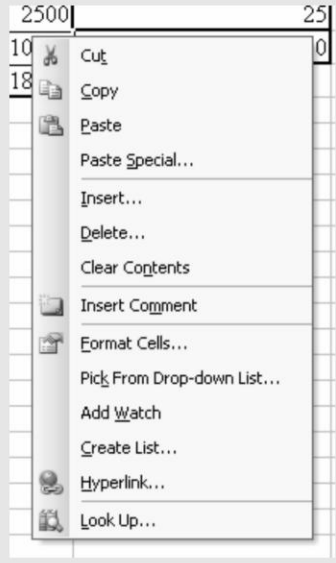
The above snapshot shows a short list that would appear if one would right clicks inside any cell. One could select the relevant option to complete the task.

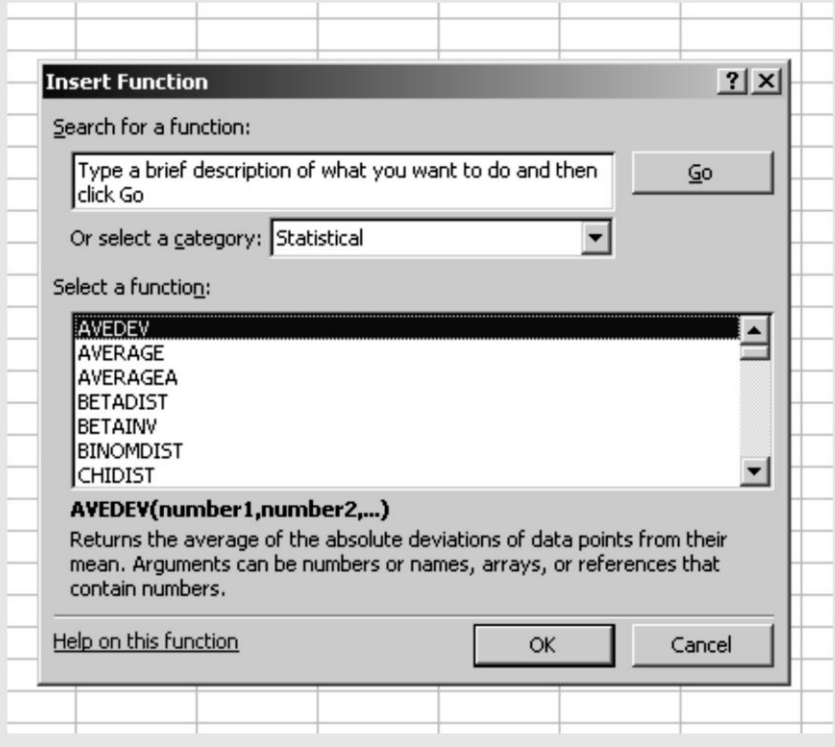
A.3 SOME IMPORTANT FUNCTIONS USED IN STATISTICS

Excel provides these functions in the category of 'Statistical Functions'. One can insert a function by either of the following two methods:

- By typing directly the function with = sign (if one knows the function name), or
- By selecting Insert → Function option from the menu. The insert function will open a window as given below.

One could select the required function from a long list of the statistical functions. Some of these are given below in a tabular form.


2500
25
10
0
18
Cut
Copy
Paste
Paste Special...
Insert...
Delete...
Clear Contents
Insert Comment
Format Cells...
Pick From Drop-down List...
Add Watch
Create List...
Hyperlink...
Look Up...


Insert Function
Search for a function:
Type a brief description of what you want to do and then click Go
Go
Or select a category: Statistical
Select a function:
AVEDEV
AVERAGE
AVERAGEA
BETADIST
BETAINV
BINOMDIST
CHIDIST
AVEDEV(number1,number2,...)
Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.
Help on this function
OK
Cancel

Statistical Functions (Arranged in alphabetic order)

Function Name	*Function Utility*
AVEDEV	Returns (gives) the average of the absolute deviations of data points from their mean
AVERAGE	Returns the average of its arguments
AVERAGEA	Returns the average of its arguments, including numbers, text, and logical values
BINOMDIST	Returns the individual term of a binomial distribution
CHIDIST	Returns the one-tailed probability of the chi-squared distribution
CHITEST	Returns the test for independence
CONFIDENCE	Returns the confidence interval for a population mean
CORREL	Returns the correlation coefficient between two data sets
COUNT	Counts how many numbers are in the list of arguments
COUNTA	Counts how many values are in the list of arguments
COVAR	Returns covariance, the average of the products of paired deviations
DEVSQ	Returns the sum of squares of deviations
EXPONDIST	Returns the exponential distribution
FDIST	Returns the F probability distribution
FISHER	Returns the Fisher transformation
FORECAST	Returns a value along a linear trend
FREQUENCY	Returns a frequency distribution as a vertical array
FTEST	Returns the result of an F-test
GEOMEAN	Returns the geometric mean
GROWTH	Returns values along an exponential trend
HARMEAN	Returns the harmonic mean
INTERCEPT	Returns the intercept of the linear regression line
KURT	Returns the kurtosis of a data set
LARGE	Returns the k-th largest value in a data set
LINEST	Returns the parameters of a linear trend
LOGEST	Returns the parameters of an exponential trend
LOGINV	Returns the inverse of the lognormal distribution
LOGNORMDIST	Returns the cumulative lognormal distribution
MAX	Returns the maximum value in a list of arguments
MEDIAN	Returns the median of the given numbers
MIN	Returns the minimum value in a list of arguments
MODE	Returns the most common value in a data set
NORMDIST	Returns the normal cumulative distribution
NORMSDIST	Returns the standard normal cumulative distribution
PEARSON	Returns the Pearson product moment correlation coefficient
PERCENTILE	Returns the k-th percentile of values in a range
PERCENTRANK	Returns the percentage rank of a value in a data set
PERMUT	Returns the number of permutations for a given number of objects
POISSON	Returns the Poisson distribution
PROB	Returns the probability that values in a range are between two limits
QUARTILE	Returns the quartile of a data set
RANK	Returns the rank of a number in a list of numbers
RSQ	Returns the square of the Pearson product moment correlation coefficient
SKEW	Returns the skewness of a distribution
SLOPE	Returns the slope of a linear regression line

(*Contd*)

(*Contd*)

SMALL	Returns the k-th smallest value in a data set
STANDARDIZE	Returns a normalized value
STDEV	Estimates standard deviation based on a sample
STDEVA	Estimates standard deviation based on a sample, including numbers, text, and logical values
STDEVP	Calculates standard deviation based on the entire population
STDEVPA	Calculates standard deviation based on the entire population, including numbers, text, and logical values
STEYX	Returns the standard error of the predicted y-value for each value of x in the Regression equation
TDIST	Returns the Student's t-distribution
TREND	Returns values along a linear trend
TRIMMEAN	Returns the mean of the interior of a data set
TTEST	Returns the probability associated with a Student's t-test
VAR	Estimates variance based on a sample
VARA	Estimates variance based on a sample, including numbers, text, and logical values
VARP	Calculates variance based on the entire population
VARPA	Calculates variance based on the entire population, including numbers, text, and logical values
ZTEST	Returns the two-tailed p-value of a z-test

Source: *Microsoft Excel Help available with the software. For using the above functions, one can refer to tutorials, which are integrated parts of the Excel software.*

A.4 KEYBOARD SHORTCUTS

Some of the keyboard operations and their corresponding functions are given in this section.

Keyboard Operation	*Corresponding Function*
CTRL+C	Copy
CTRL+X	Cut
CTRL+V	Paste
CTRL+Z	Undo
DELETE	Delete
SHIFT+DELETE	Delete the selected item permanently without placing the item in the Recycle Bin
CTRL while dragging an item	Copy the selected item
CTRL+SHIFT while dragging an item	Create short cut to the selected item
F2 key	Rename the selected item
CTRL+RIGHT ARROW	Move the insertion point to the beginning of the next word
CTRL+LEFT ARROW	Move the insertion point to the beginning of the previous word
CTRL+DOWN ARROW	Move the insertion point to the beginning of the next Paragraph
CTRL+UP ARROW	Move the insertion point to the beginning of the previous Paragraph
CTRL+SHIFT with any of the arrow keys	Highlight a block of text
SHIFT with any of the arrow keys	Select more than one item in a window or on the desktop, or select text in a document
CTRL+A	Select all

(*Contd*)

(*Contd*)

F3 key	Search for a file or a folder
ALT+ENTER	View the properties for the selected item
ALT+F4	Close the active item, or quit the active program
ALT+ENTER	Display the properties of the selected object
ALT+SPACEBAR	Open the shortcut menu for the active window
CTRL+F4	Close the active document in programs that enable you to have multiple documents open simultaneously
ALT+TAB	Switch between the open items
ALT+ESC	Cycle through items in the order that they had been opened
F6 key	Cycle through the screen elements in a window or on the desktop
F4 key	Display the Address bar list in My Computer or Windows Explorer
SHIFT+F10	Display the shortcut menu for the selected item
ALT+SPACEBAR	Display the System menu for the active window
CTRL+ESC	Display the Start menu
ALT+Underlined letter in a menu name	Display the corresponding menu—Underlined letter in a command name on an open menu (Perform the corresponding command)
F10 key	Activate the menu bar in the active program
RIGHT ARROW	Open the next menu to the right, or open a submenu
LEFT ARROW	Open the next menu to the left, or close a submenu
F5 key	Update the active window
BACKSPACE	View the folder one level up in My Computer or Windows Explorer
ESC	Cancel the current task
SHIFT while inserting CD-ROM drive to prevent it from automatically playing	

A.4.1 Dialog Box Keyboard Shortcuts

Keyboard Operation	*Corresponding Function*
CTRL+TAB	Move forward through the tabs
CTRL+SHIFT+TAB	Move backward through the tabs
TAB	Move forward through the options
SHIFT+TAB	Move backward through the options
ALT+Underlined letter	Perform the corresponding command or select the corresponding option
ENTER	Perform the command for the active option or button
SPACEBAR	Select or clear the check box if the active option is a check box
Arrow keys	Select a button if the active option is a group of option buttons
F1 key	Display Help
F4 key	Display the items in the active list
BACKSPACE	Open a folder one level up if a folder is selected in the Save As or Open dialog box
Microsoft Natural Keyboard Shortcuts	
Windows Logo	Display or hide the Start menu
Windows Logo+BREAK	Display the System Properties dialog box
Windows Logo	Display the desktop
Windows Logo+M	Minimize all of the windows
Windows Logo+SHIFT+M	Restore the minimised windows

(*Contd*)

(*Contd*)

Windows Logo+E	Open My Computer
Windows Logo+F	Search for a file or a folder
CTRL+Windows Logo+F	Search for computers
Windows Logo+F1	Display Windows Help
Windows Logo+ L	Lock the keyboard
Windows Logo+R	Open the Run dialog box
Windows Logo+U	Open Utility Manager

A.4.2 Accessibility Keyboard Shortcuts

Keyboard Operation	*Corresponding Function*
Right SHIFT for eight seconds	Switch FilterKeys either on or off
Left ALT+left SHIFT+PRINT SCREEN	Switch High Contrast either on or off
Left ALT+left SHIFT+NUM LOCK	Switch the MouseKeys either on or off
SHIFT five times	Switch the StickyKeys either on or off
NUM LOCK for five seconds	Switch the ToggleKeys either on or off
Windows Logo +U	Open Utility Manager
Windows Explorer Keyboard Shortcuts	
END	Display the bottom of the active window
HOME	Display the top of the active window
NUM LOCK+Asterisk sign (*)	Display all of the subfolders that are under the selected folder
NUM LOCK+Plus sign (+)	Display the contents of the selected folder
NUM LOCK+Minus sign (-)	Collapse the selected folder
LEFT ARROW	Collapse the current selection if it is expanded, or select the parent folder
RIGHT ARROW	Display the current selection if it is collapsed, or select the first subfolder

A.4.3 Shortcut Keys for Character Map

After double-clicking a character on the grid of characters, one can move through the grid by using the keyboard shortcuts:

Keyboard Operation	*Corresponding Function*
RIGHT ARROW	Move to the right or to the beginning of the next line
LEFT ARROW	Move to the left or to the end of the previous line
UP ARROW	Move up one row
DOWN ARROW	Move down one row
PAGE UP	Move up one screen at a time
PAGE DOWN	Move down one screen at a time
HOME	Move to the beginning of the line
END	Move to the end of the line
CTRL+HOME	Move to the first character
CTRL+END	Move to the last character
SPACEBAR	Switch between Enlarged and Normal mode when a character is selected
CTRL+R	Update the current Web page

(*Contd*)

(*Contd*)

CTRL+W	Close the current window
CTRL+ALT+Minus sign (-)	Place a snapshot of the active window in the client on the Terminal server clipboard and provide the same functionality as pressing PRINT SCREEN on a local computer.
CTRL+ALT+Plus sign (+)	Place a snapshot of the entire client window area on the Terminal server clipboard, and provide the same functionality as pressing ALT+PRINT SCREEN on computer.

A.5 EPILOGUE

We have provided a brief overview of Excel, which is considered necessary to understand the use of excel templates for statistical calculations, discussed in this book. The contents and coverage of Excel, here, have been decided in consonance with this objective. For detailed coverage of Excel, one may like to refer to the books written exclusively on Excel. A few of these are included in the list of books given under the title 'Some Other Useful Books…' for all the topics covered in this book.

Introduction to IBM SPSS Statistics 18

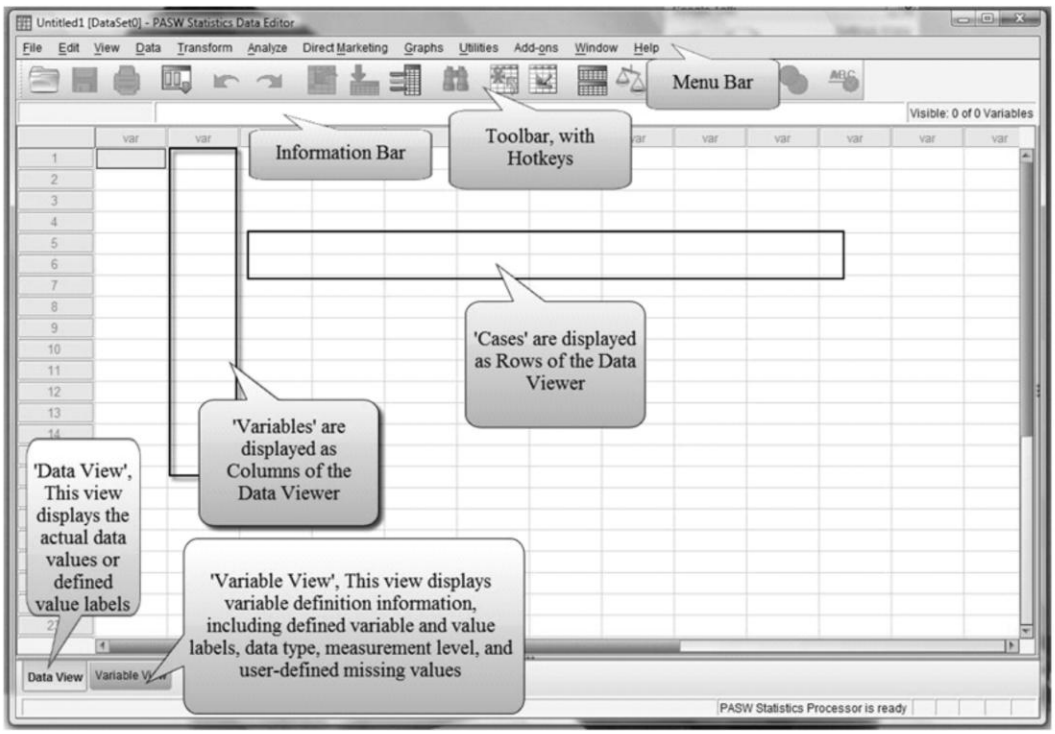
Appendix III

IBM SPSS Statistics is the preferred choice of most Government, Academic, Research and Corporate organisations in India for Data Analysis and Reporting. SPSS can be used for performing basic as well as advanced analysis on data. In this chapter the interface of the newly launched version 18 of IBM SPSS Statistics is explained in brief.

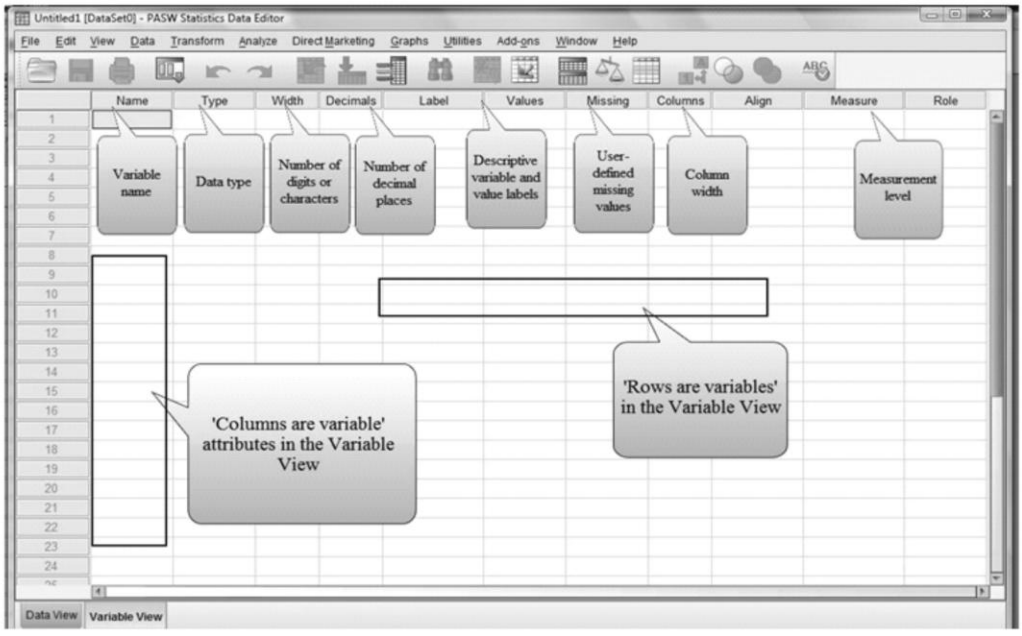
SPSS Statistics has three views, Data View, Variable View and the Output View. The data view shows the data entered into the file. The data is entered in a way that each row represents a record, and each column represents the variable. The variable view is used for defining and labelling variables, while the output window displays the results of all analysis. These are shown in the following snapshots.

1. INTRODUCTION TO IBM SPSS STATISTICS 18 INTERFACE

Data View

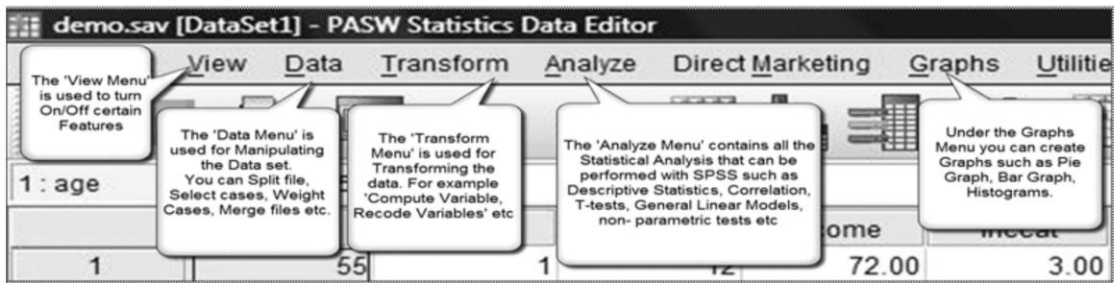


Variable View



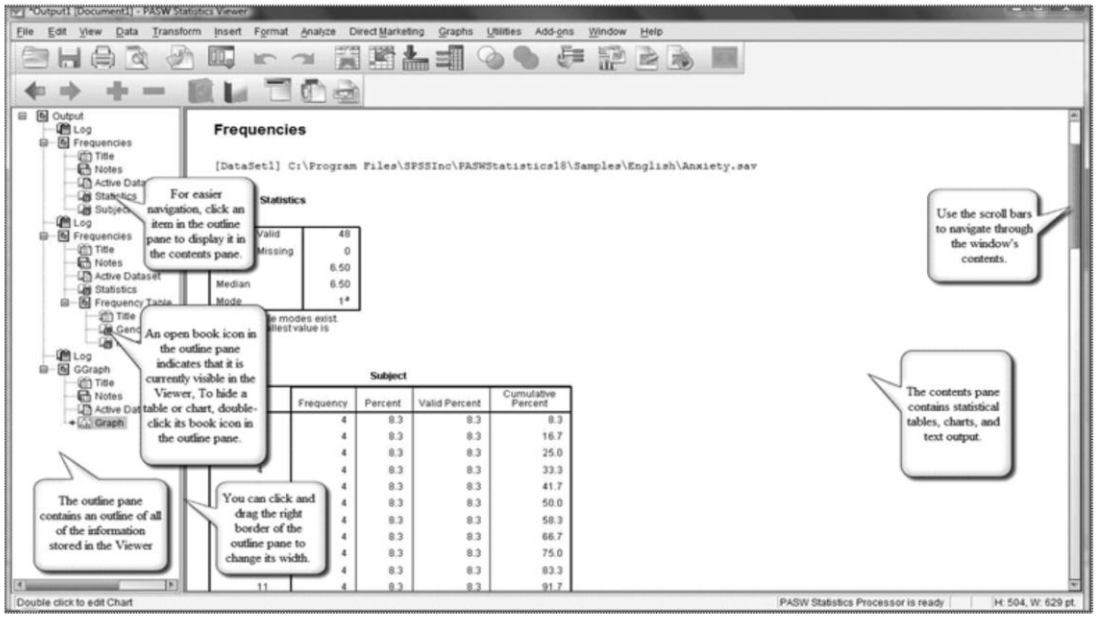
SPSS is a menu-based software. The menu has following broad options:

Menu Bar



The result of any analysis done is displayed separately in the output window. A typical output window is displayed as follows:

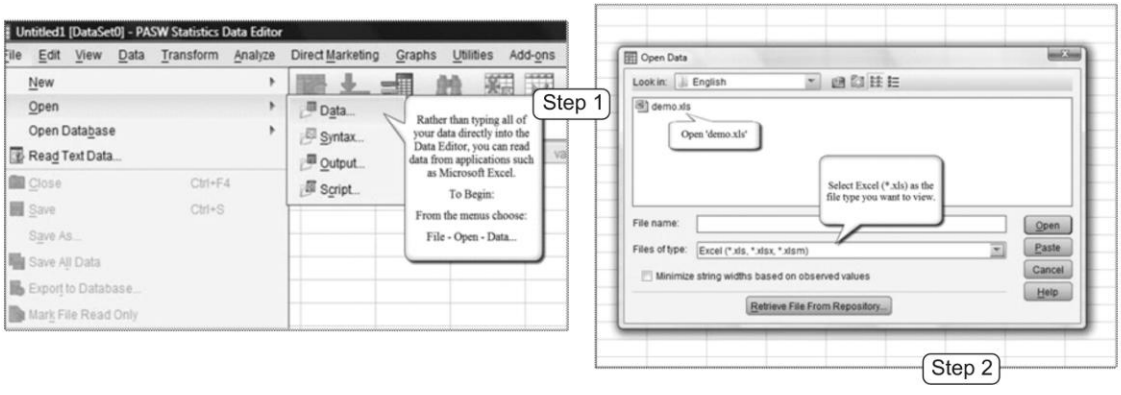
Output Window

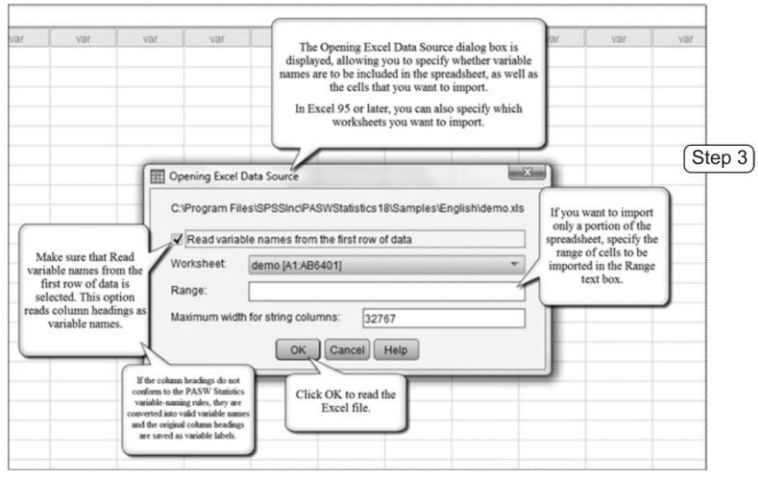


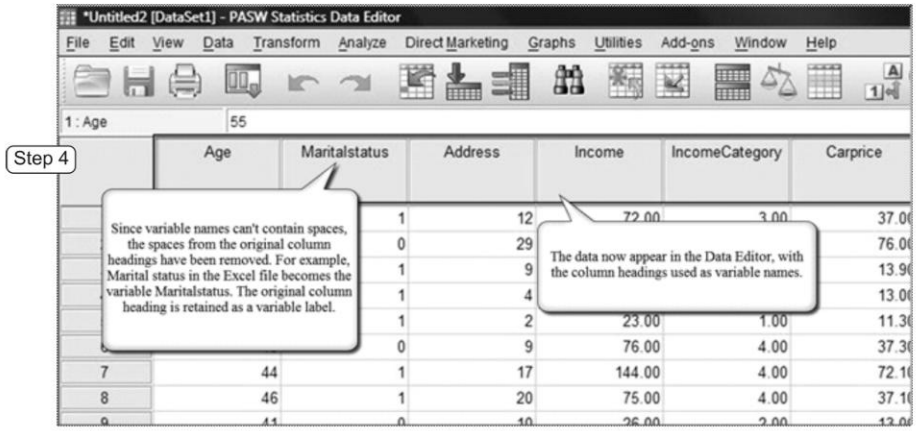
The data in SPSS can be either entered directly or can be imported through Excel or a database file. Following snapshots explains importing data from Excel file:

2. GETTING DATA INTO IBM SPSS STATISTICS 18

Reading Data from Excel

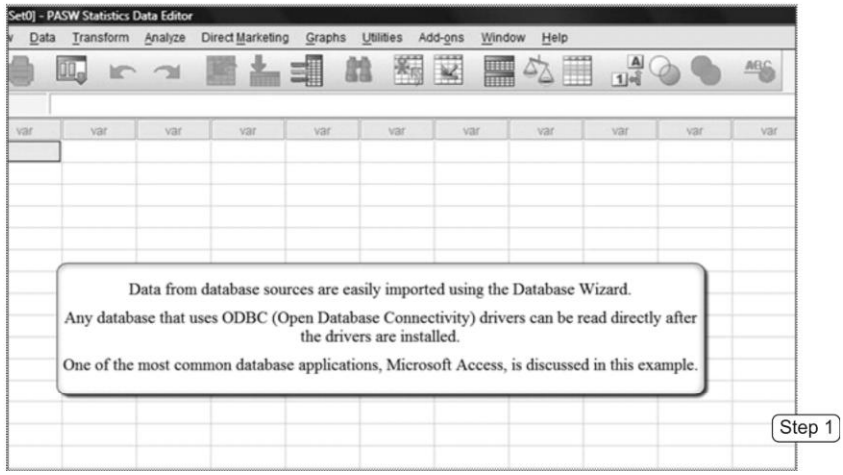


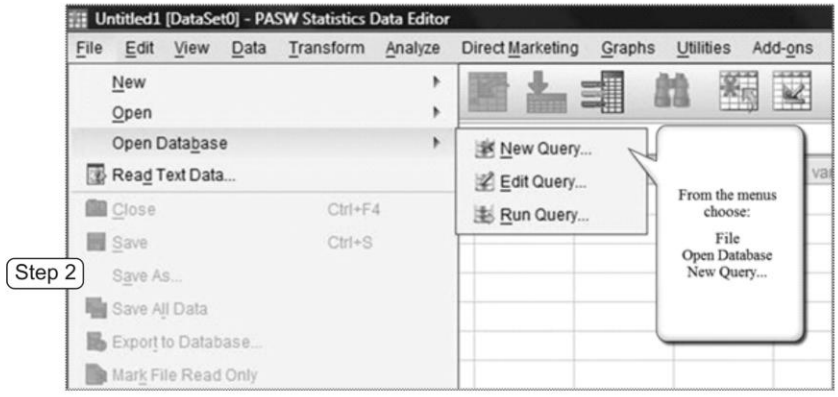


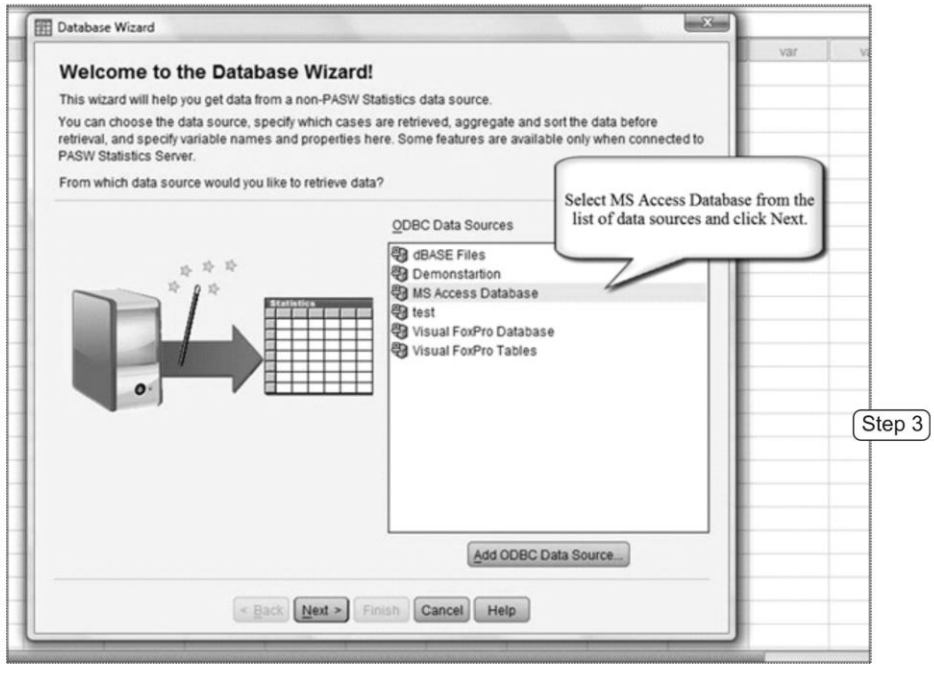


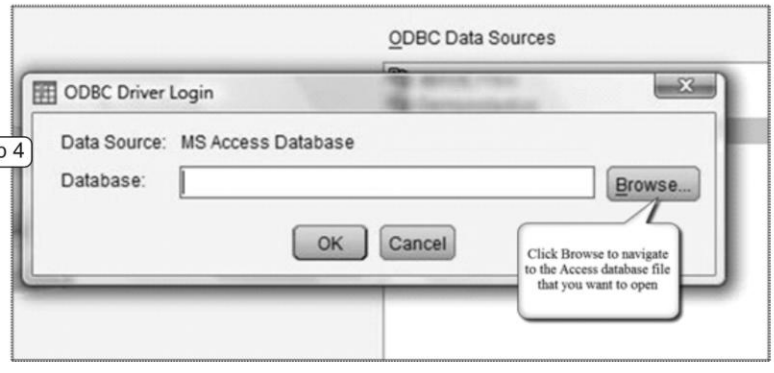
Following snapshots explain the importing of database file into SPSS:

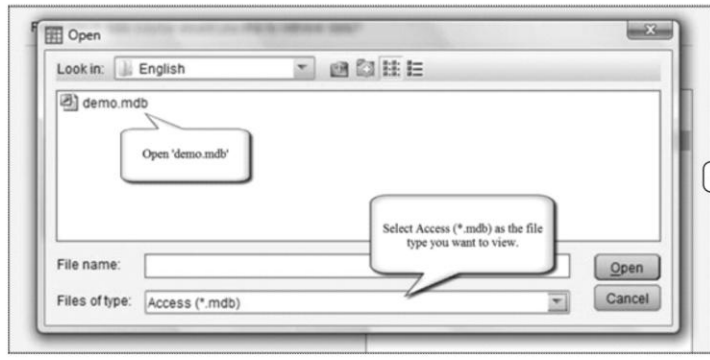
Reading Data from Databases

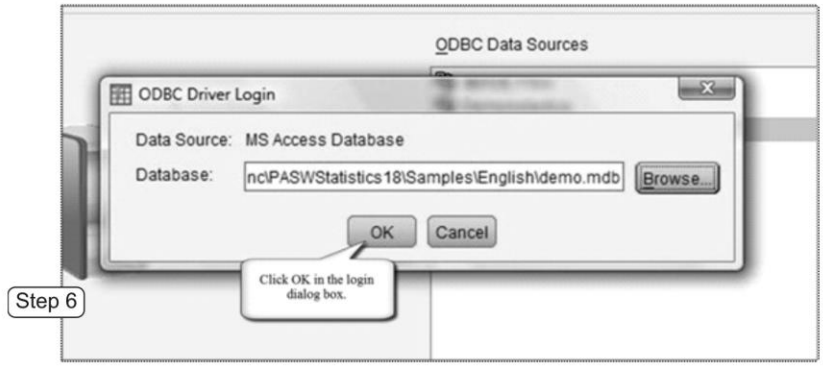


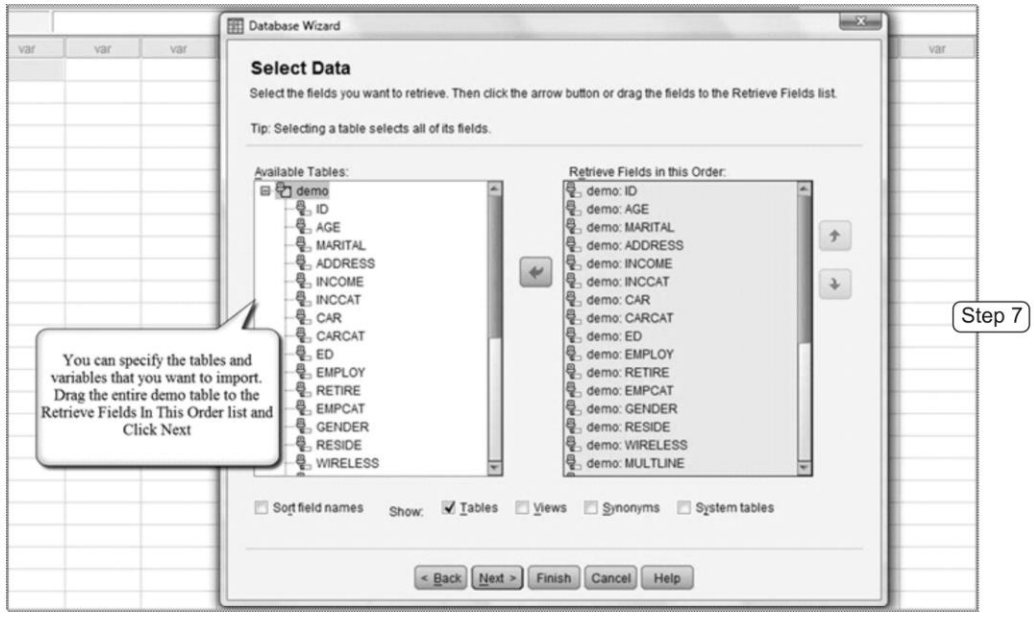

Untitled1 [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons
New
Open
Open Database
Read Text Data...
Close Ctrl+F4
Save Ctrl+S
Save As...
Save All Data
Export to Database...
Mark File Read Only
New Query...
Edit Query...
Run Query...
From the menus choose:
File
Open Database
New Query...
Step 2

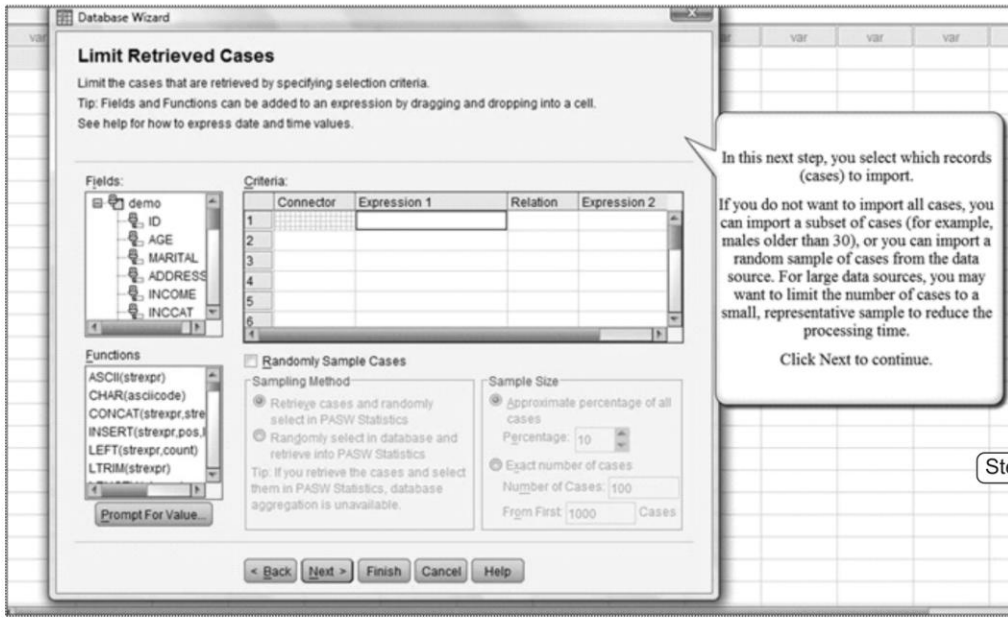

Database Wizard
Welcome to the Database Wizard!
This wizard will help you get data from a non-PASW Statistics data source.
You can choose the data source, specify which cases are retrieved, aggregate and sort the data before retrieval, and specify variable names and properties here. Some features are available only when connected to PASW Statistics Server.
From which data source would you like to retrieve data?
Select MS Access Database from the list of data sources and click Next.
ODBC Data Sources
dBASE Files
Demonstartion
MS Access Database
test
Visual FoxPro Database
Visual FoxPro Tables
Add ODBC Data Source...
< Back
Next >
Finish
Cancel
Help
Step 3


ODBC Data Sources
ODBC Driver Login
Data Source: MS Access Database
Database:
Browse...
OK
Cancel
Click Browse to navigate to the Access database file that you want to open
Step 4

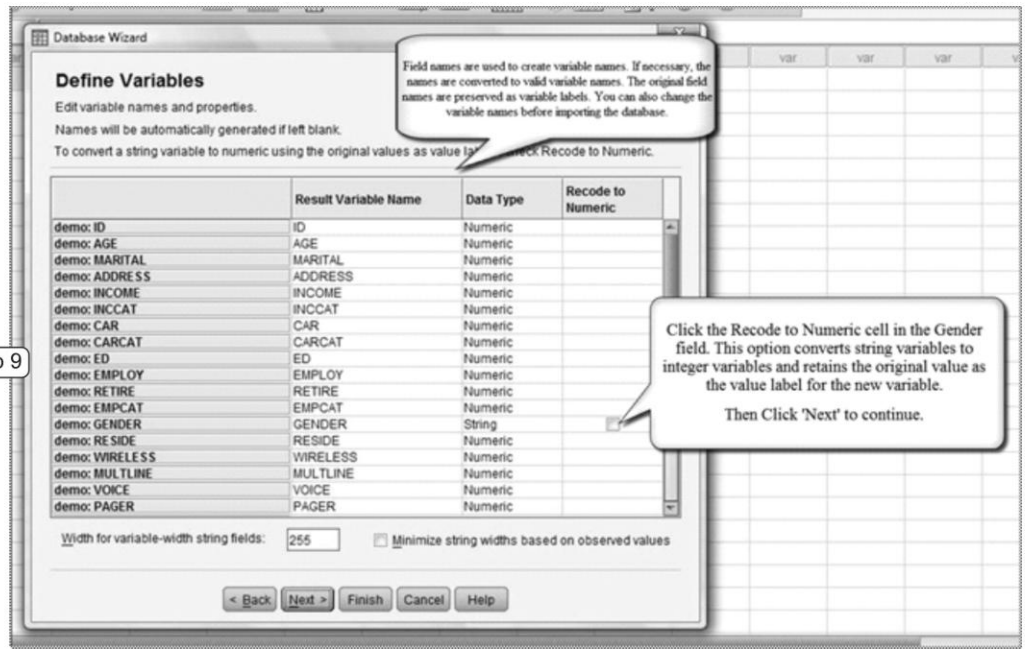

Open
Look in: English
demo.mdb
Open 'demo.mdb'
Select Access (*.mdb) as the file type you want to view.
File name:
Files of type: Access (*.mdb)
Open
Cancel
Step 5


ODBC Data Sources
ODBC Driver Login
Data Source: MS Access Database
Database: nc\PASWStatistics18\Samples\English\demo.mdb
Browse...
OK
Cancel
Click OK in the login dialog box.
Step 6

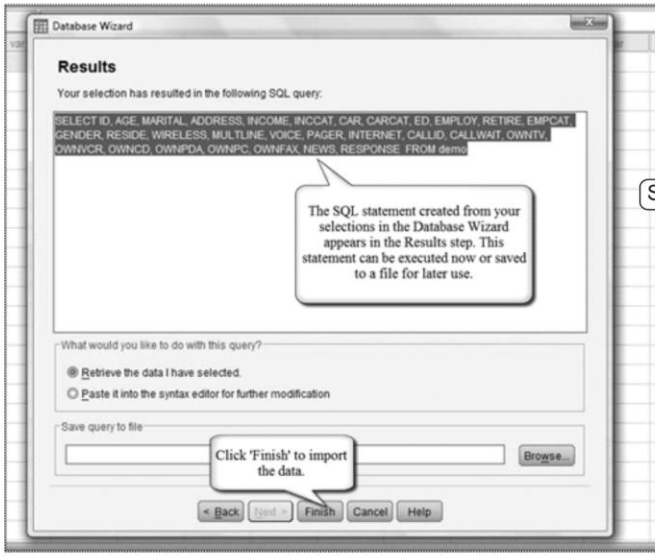

Database Wizard
Select Data
Select the fields you want to retrieve. Then click the arrow button or drag the fields to the Retrieve Fields list.
Tip: Selecting a table selects all of its fields.
Available Tables:
demo
ID
AGE
MARITAL
ADDRESS
INCOME
INCCAT
CAR
CARCAT
ED
EMPLOY
RETIRE
EMPCAT
GENDER
RESIDE
WIRELESS
Retrieve Fields in this Order:
demo: ID
demo: AGE
demo: MARITAL
demo: ADDRESS
demo: INCOME
demo: INCCAT
demo: CAR
demo: CARCAT
demo: ED
demo: EMPLOY
demo: RETIRE
demo: EMPCAT
demo: GENDER
demo: RESIDE
demo: WIRELESS
demo: MULTLINE
Sort field names
Show: Tables Views Synonyms System tables
< Back
Next >
Finish
Cancel
Help
You can specify the tables and variables that you want to import. Drag the entire demo table to the Retrieve Fields In This Order list and Click Next
Step 7

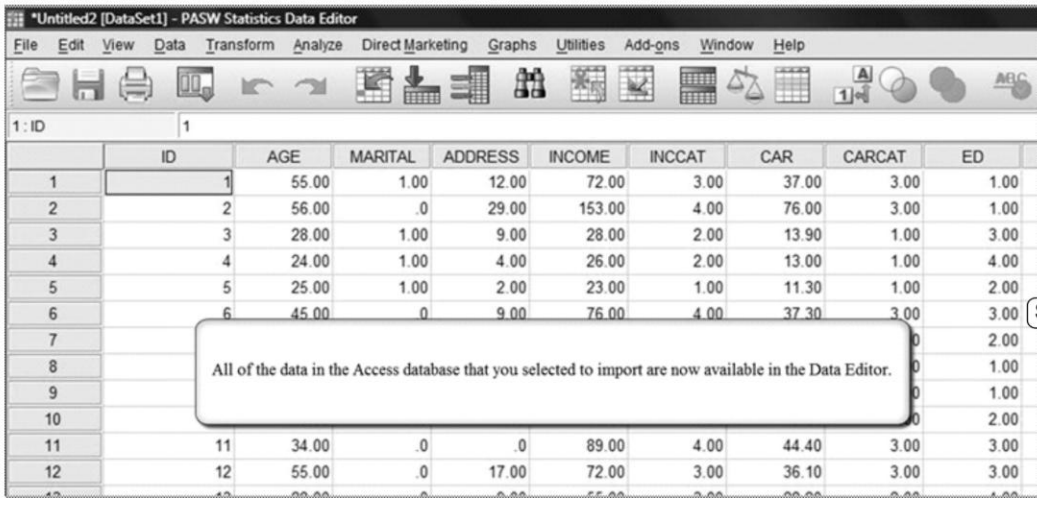

Database Wizard
Limit Retrieved Cases
Limit the cases that are retrieved by specifying selection criteria.
Tip: Fields and Functions can be added to an expression by dragging and dropping into a cell.
See help for how to express date and time values.
Fields:
demo
ID
AGE
MARITAL
ADDRESS
INCOME
INCCAT
Criteria:
Connector
Expression 1
Relation
Expression 2
Functions
ASCII(strexpr)
CHAR(asciicode)
CONCAT(strexpr,stre
INSERT(strexpr,pos,l
LEFT(strexpr,count)
LTRIM(strexpr)
Prompt For Value...
Randomly Sample Cases
Sampling Method
Retrieve cases and randomly select in PASW Statistics
Randomly select in database and retrieve into PASW Statistics
Tip: If you retrieve the cases and select them in PASW Statistics, database aggregation is unavailable.
Sample Size
Approximate percentage of all cases
Percentage: 10
Exact number of cases
Number of Cases: 100
From First 1000 Cases
< Back
Next >
Finish
Cancel
Help
In this next step, you select which records (cases) to import.
If you do not want to import all cases, you can import a subset of cases (for example, males older than 30), or you can import a random sample of cases from the data source. For large data sources, you may want to limit the number of cases to a small, representative sample to reduce the processing time.
Click Next to continue.

Step 8


Database Wizard
Define Variables
Edit variable names and properties.
Names will be automatically generated if left blank.
To convert a string variable to numeric using the original values as value la
Recode to Numeric.
Field names are used to create variable names. If necessary, the names are converted to valid variable names. The original field names are preserved as variable labels. You can also change the variable names before importing the database.
Result Variable Name
Data Type
Recode to Numeric
demo: ID
demo: AGE
demo: MARITAL
demo: ADDRESS
demo: INCOME
demo: INCCAT
demo: CAR
demo: CARCAT
demo: ED
demo: EMPLOY
demo: RETIRE
demo: EMPCAT
demo: GENDER
demo: RESIDE
demo: WIRELESS
demo: MULTLINE
demo: VOICE
demo: PAGER
String
Width for variable-width string fields: 255
Minimize string widths based on observed values
< Back
Next >
Finish
Cancel
Help
Click the Recode to Numeric cell in the Gender field. This option converts string variables to integer variables and retains the original value as the value label for the new variable.
Then Click 'Next' to continue.

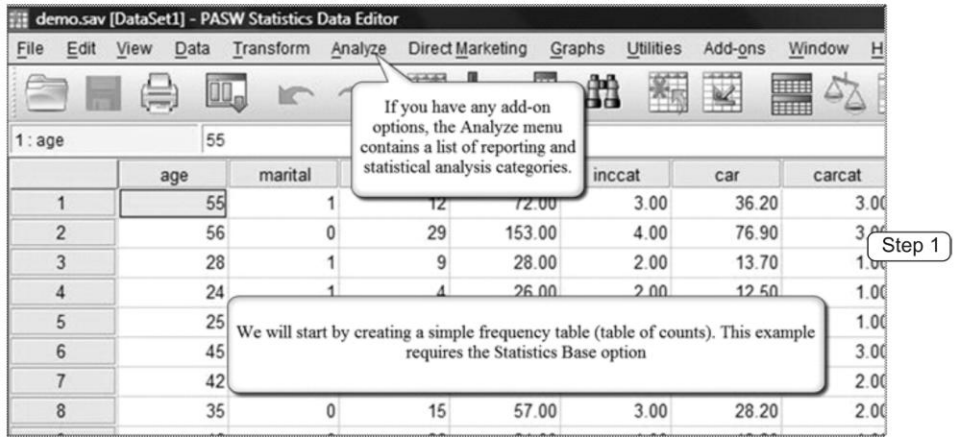
Step 9

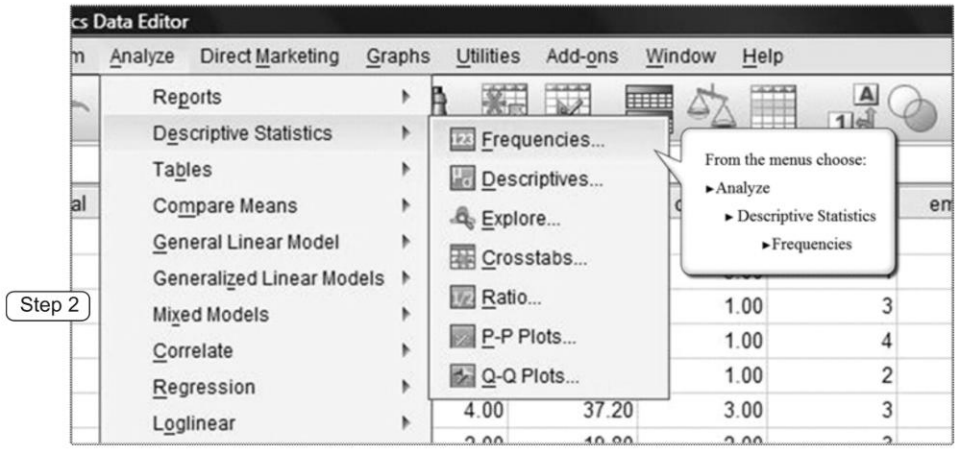


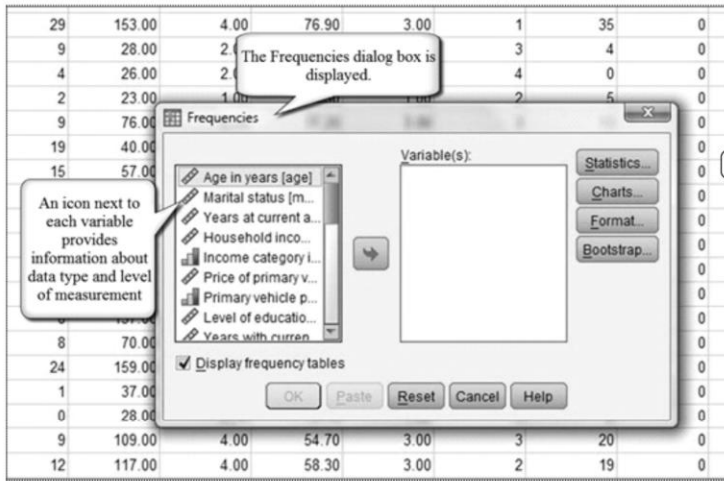


Most analysis can be done using the 'Analyze' option from the menu.

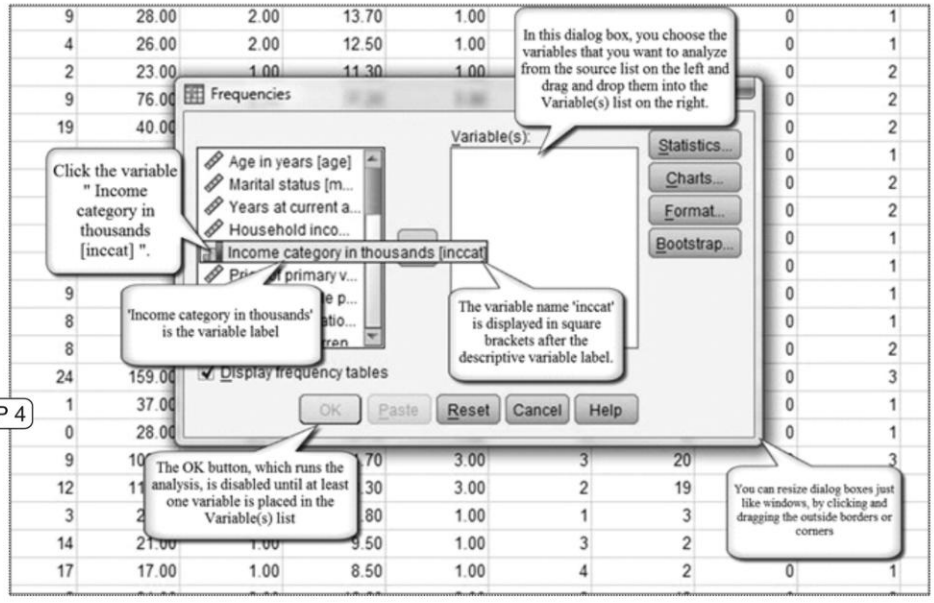
3. RUNNING AN ANALYSIS IN IBM SPSS STATISTICS 18



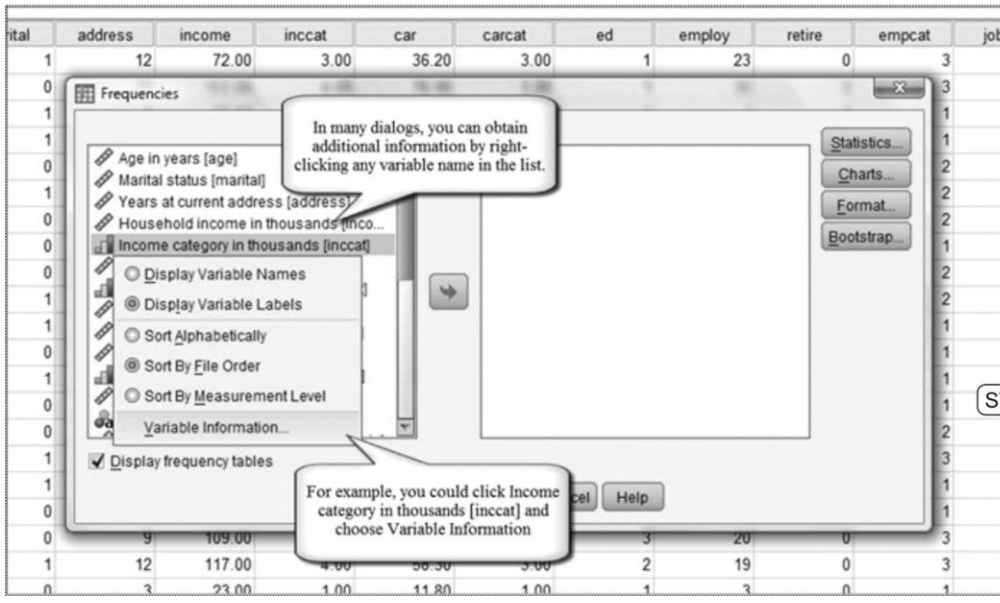


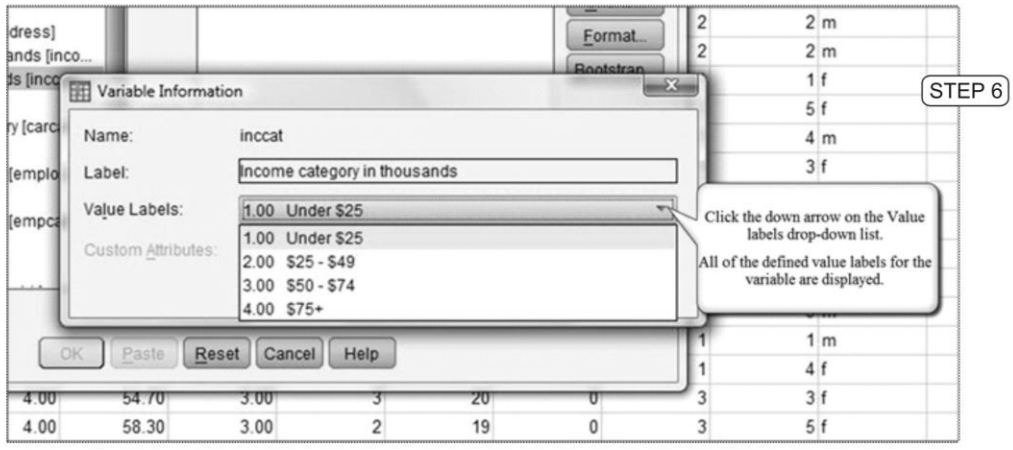


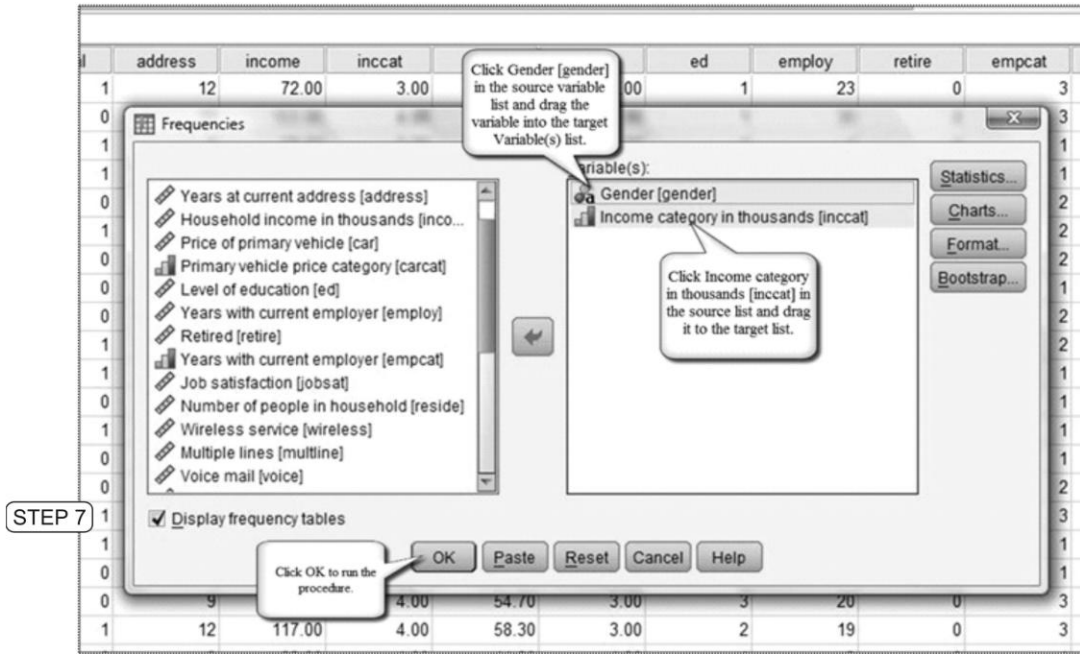
STEP 4

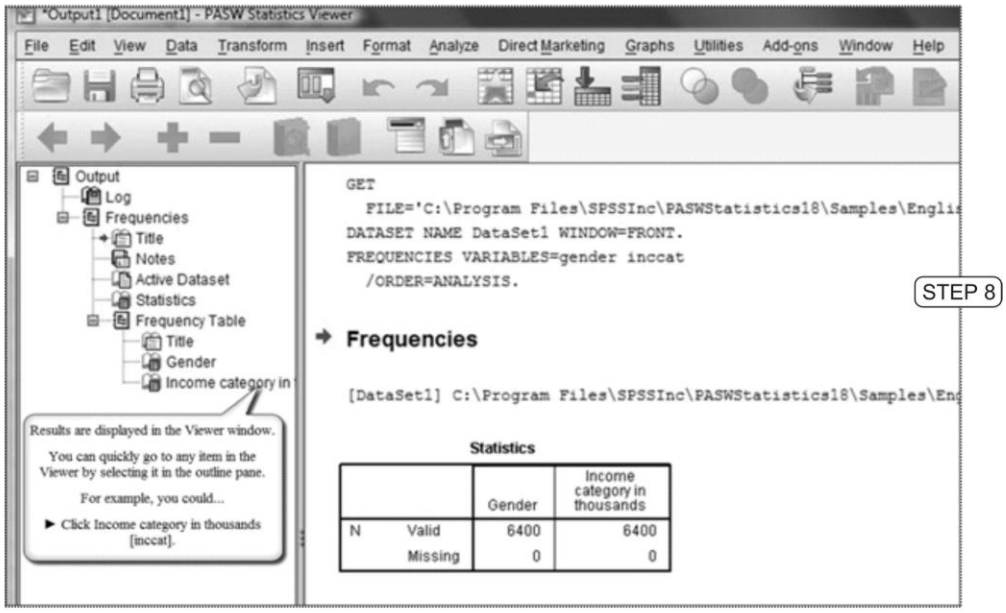

Frequencies
In this dialog box, you choose the variables that you want to analyze from the source list on the left and drag and drop them into the Variable(s) list on the right.
Variable(s):
Statistics...
Charts...
Format...
Bootstrap...
Age in years [age]
Marital status [m...
Years at current a...
Household inco...
Income category in thousands [inccat]
Click the variable " Income category in thousands [inccat] ".
'Income category in thousands' is the variable label
The variable name 'inccat' is displayed in square brackets after the descriptive variable label.
Display frequency tables
OK
Paste
Reset
Cancel
Help
The OK button, which runs the analysis, is disabled until at least one variable is placed in the Variable(s) list
You can resize dialog boxes just like windows, by clicking and dragging the outside borders or corners

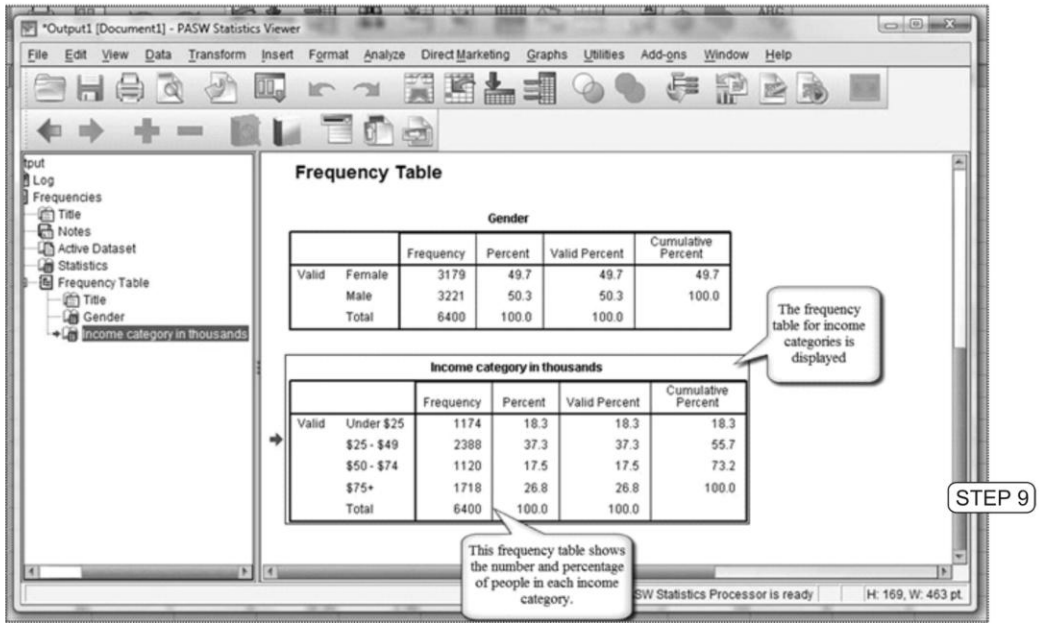
STEP 5


Frequencies
In many dialogs, you can obtain additional information by right-clicking any variable name in the list.
Statistics...
Charts...
Format...
Bootstrap...
Age in years [age]
Marital status [marital]
Years at current address [address]
Household income in thousands [inco...
Income category in thousands [inccat]
Display Variable Names
Display Variable Labels
Sort Alphabetically
Sort By File Order
Sort By Measurement Level
Variable Information...
Display frequency tables
Help
For example, you could click Income category in thousands [inccat] and choose Variable Information

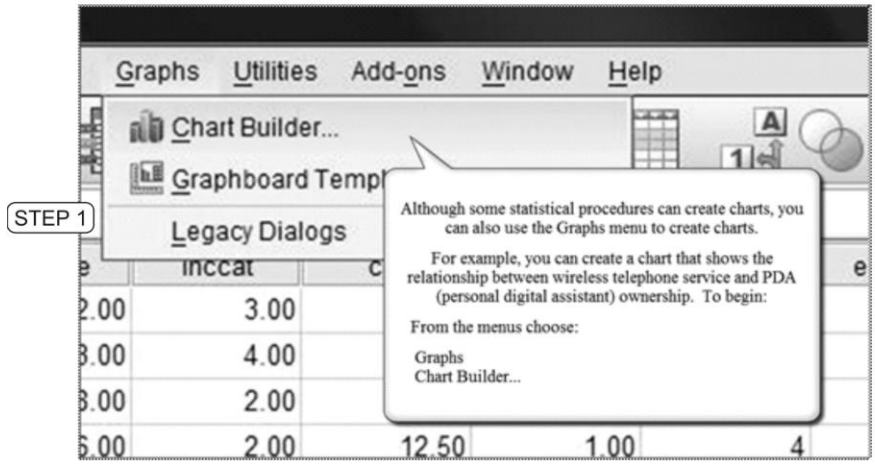

STEP 6
Variable Information
Name: inccat
Label: Income category in thousands
Value Labels: 1.00 Under $25
1.00 Under $25
2.00 $25 - $49
3.00 $50 - $74
4.00 $75+
Custom Attributes:
OK
Paste
Reset
Cancel
Help
Format...
Click the down arrow on the Value labels drop-down list.
All of the defined value labels for the variable are displayed.

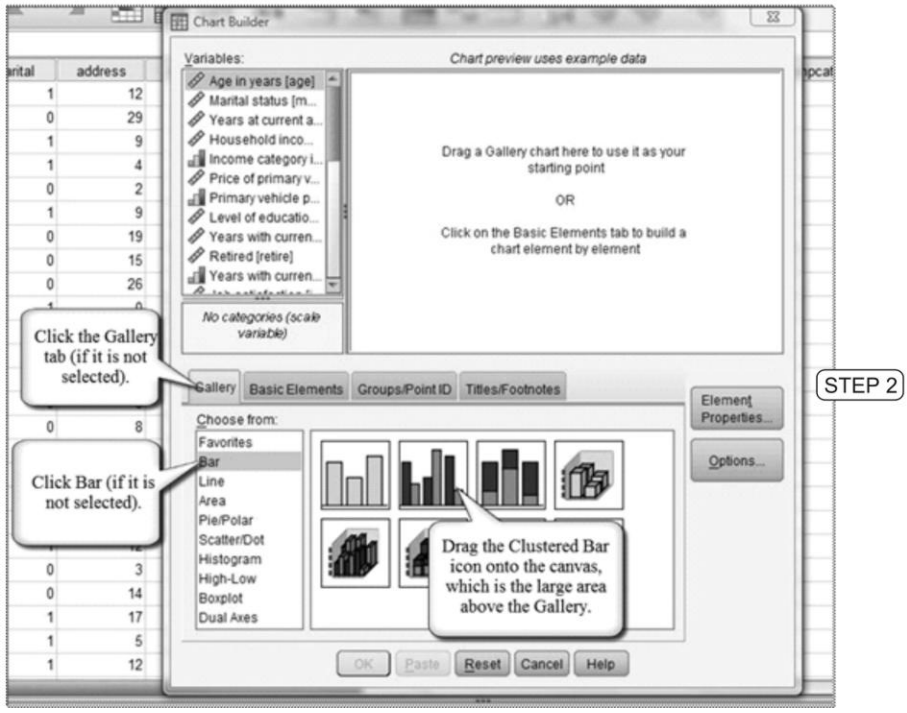

STEP 7
Frequencies
Years at current address [address]
Household income in thousands [inco...
Price of primary vehicle [car]
Primary vehicle price category [carcat]
Level of education [ed]
Years with current employer [employ]
Retired [retire]
Years with current employer [empcat]
Job satisfaction [jobsat]
Number of people in household [reside]
Wireless service [wireless]
Multiple lines [multline]
Voice mail [voice]
Gender [gender]
Income category in thousands [inccat]
Statistics...
Charts...
Format...
Bootstrap...
Display frequency tables
OK
Paste
Reset
Cancel
Help
Click Gender [gender] in the source variable list and drag the variable into the target Variable(s) list.
Click Income category in thousands [inccat] in the source list and drag it to the target list.
Click OK to run the procedure.


*Output1 [Document1] - PASW Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Frequencies
Title
Notes
Active Dataset
Statistics
Frequency Table
Title
Gender
Income category in
Results are displayed in the Viewer window.
You can quickly go to any item in the Viewer by selecting it in the outline pane.
For example, you could...
► Click Income category in thousands [inccat].
GET
FILE='C:\Program Files\SPSSInc\PASWStatistics18\Samples\Englis
DATASET NAME DataSet1 WINDOW=FRONT.
FREQUENCIES VARIABLES=gender inccat
/ORDER=ANALYSIS.
STEP 8
Frequencies
[DataSet1] C:\Program Files\SPSSInc\PASWStatistics18\Samples\En
Statistics
Gender
Income category in thousands
N Valid 6400 6400
Missing 0 0


*Output1 [Document1] - PASW Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Log
Frequencies
Title
Notes
Active Dataset
Statistics
Frequency Table
Title
Gender
Income category in thousands
Frequency Table
Gender
Frequency Percent Valid Percent Cumulative Percent
Valid Female 3179 49.7 49.7 49.7
Male 3221 50.3 50.3 100.0
Total 6400 100.0 100.0
The frequency table for income categories is displayed
Income category in thousands
Frequency Percent Valid Percent Cumulative Percent
Valid Under $25 1174 18.3 18.3 18.3
$25 - $49 2388 37.3 37.3 55.7
$50 - $74 1120 17.5 17.5 73.2
$75+ 1718 26.8 26.8 100.0
Total 6400 100.0 100.0
This frequency table shows the number and percentage of people in each income category.
STEP 9
SW Statistics Processor is ready
H: 169, W: 463 pt

4. PLOTTING GRAPHS IN IBM SPSS STATISTICS 18





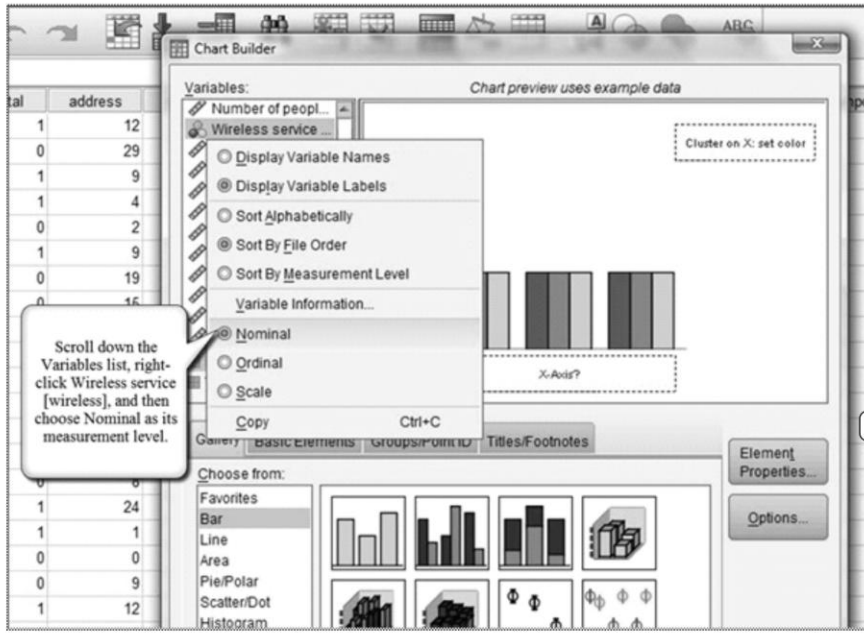


Chart Builder
Variables:
Chart preview uses example data
Number of peopl...
Wireless service ...
Display Variable Names
Display Variable Labels
Sort Alphabetically
Sort By File Order
Sort By Measurement Level
Variable Information...
Nominal
Ordinal
Scale
Copy Ctrl+C
Cluster on X: set color
X-Axis?
Scroll down the Variables list, right-click Wireless service [wireless], and then choose Nominal as its measurement level.
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
Element Properties...
Options...
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
address

STEP 3

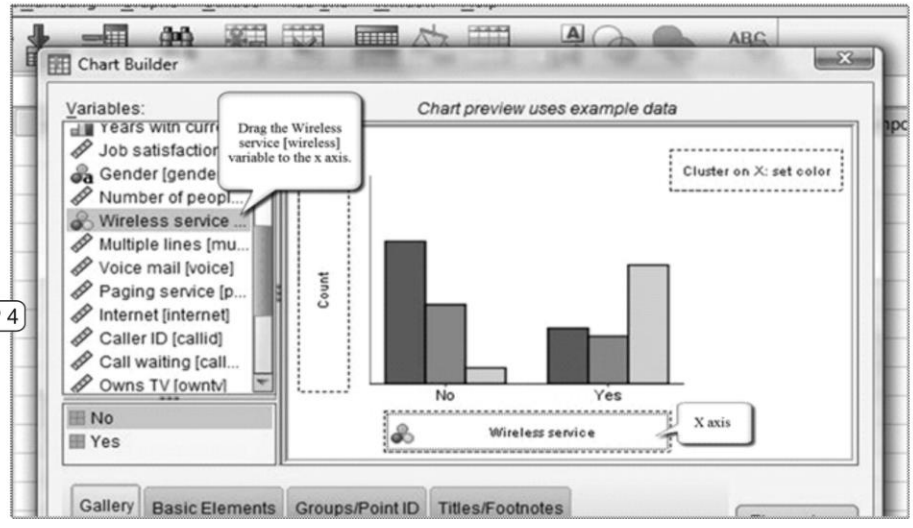


Chart Builder
Variables:
Chart preview uses example data
Drag the Wireless service [wireless] variable to the x axis.
Job satisfactior
Gender [gende
Number of peopl...
Wireless service ...
Multiple lines [mu...
Voice mail [voice]
Paging service [p...
Internet [internet]
Caller ID [callid]
Call waiting [call...
Owns TV [owntv]
No
Yes
Cluster on X: set color
Count
No
Yes
Wireless service
X axis
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes

STEP 4

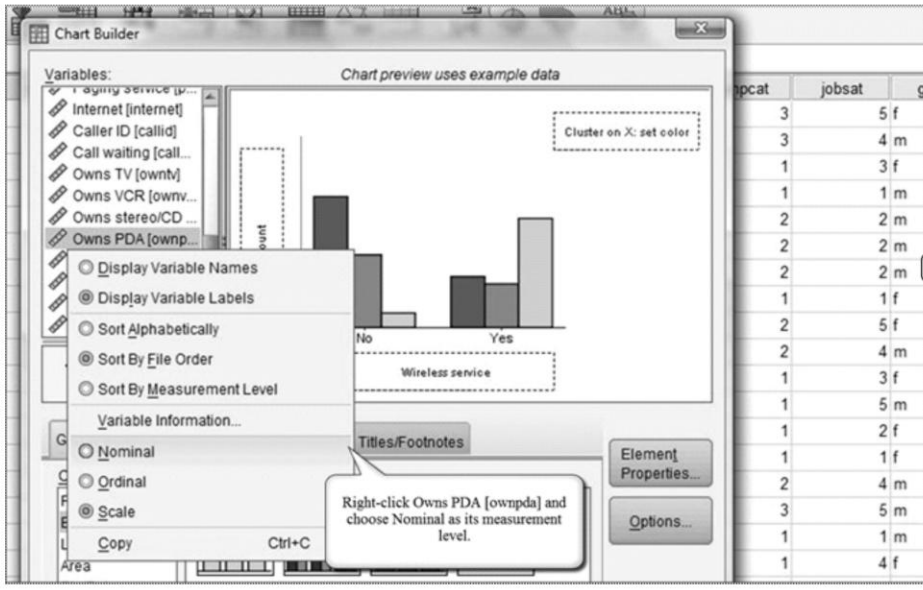


Chart Builder
Variables:
Chart preview uses example data
Internet [internet]
Caller ID [callid]
Call waiting [call...
Owns TV [owntv]
Owns VCR [ownv...
Owns stereo/CD ...
Owns PDA [ownp...
Display Variable Names
Display Variable Labels
Sort Alphabetically
Sort By File Order
Sort By Measurement Level
Variable Information...
Nominal
Ordinal
Scale
Copy Ctrl+C
Cluster on X: set color
No
Yes
Wireless service
Titles/Footnotes
Element Properties...
Options...
Right-click Owns PDA [ownpda] and choose Nominal as its measurement level.
jobsat
STEP 5

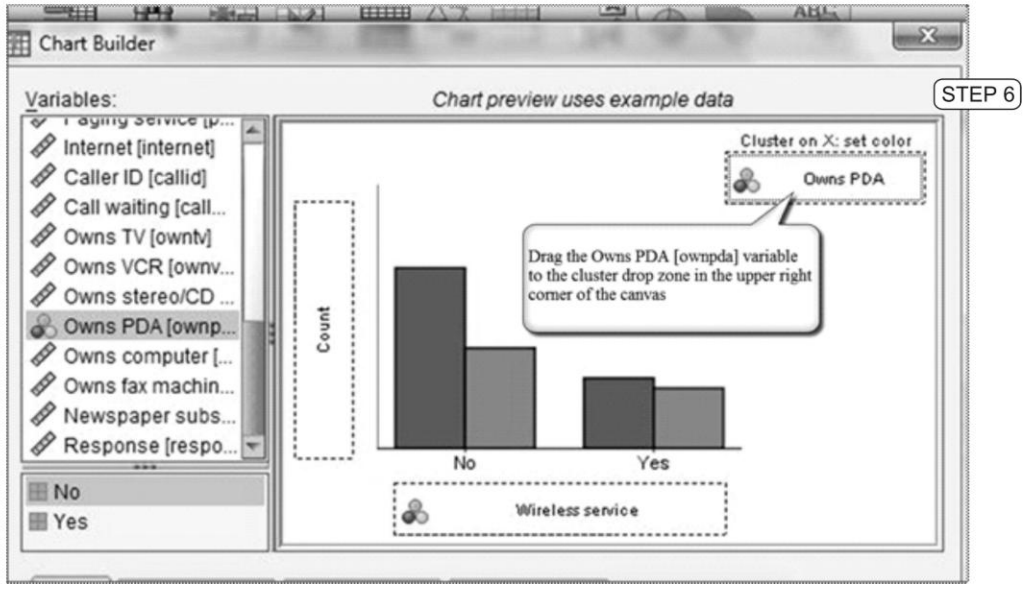


Chart Builder
Variables:
Chart preview uses example data
Internet [internet]
Caller ID [callid]
Call waiting [call...
Owns TV [owntv]
Owns VCR [ownv...
Owns stereo/CD ...
Owns PDA [ownp...
Owns computer [...
Owns fax machin...
Newspaper subs...
Response [respo...
No
Yes
Cluster on X: set color
Owns PDA
Drag the Owns PDA [ownpda] variable to the cluster drop zone in the upper right corner of the canvas
Count
No
Yes
Wireless service
STEP 6

STEP 7

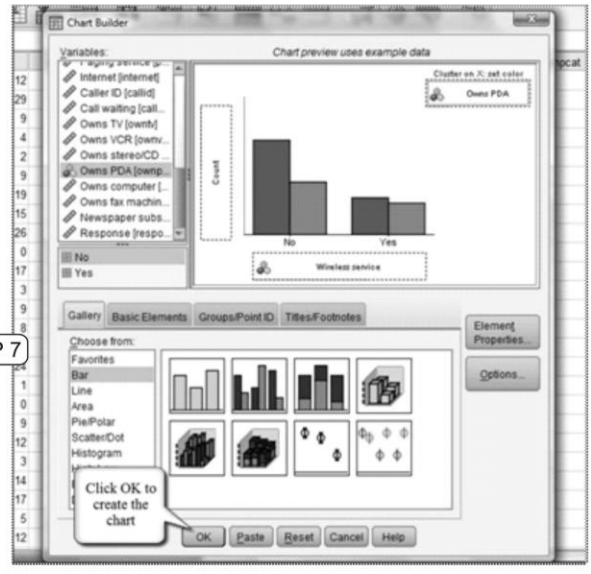
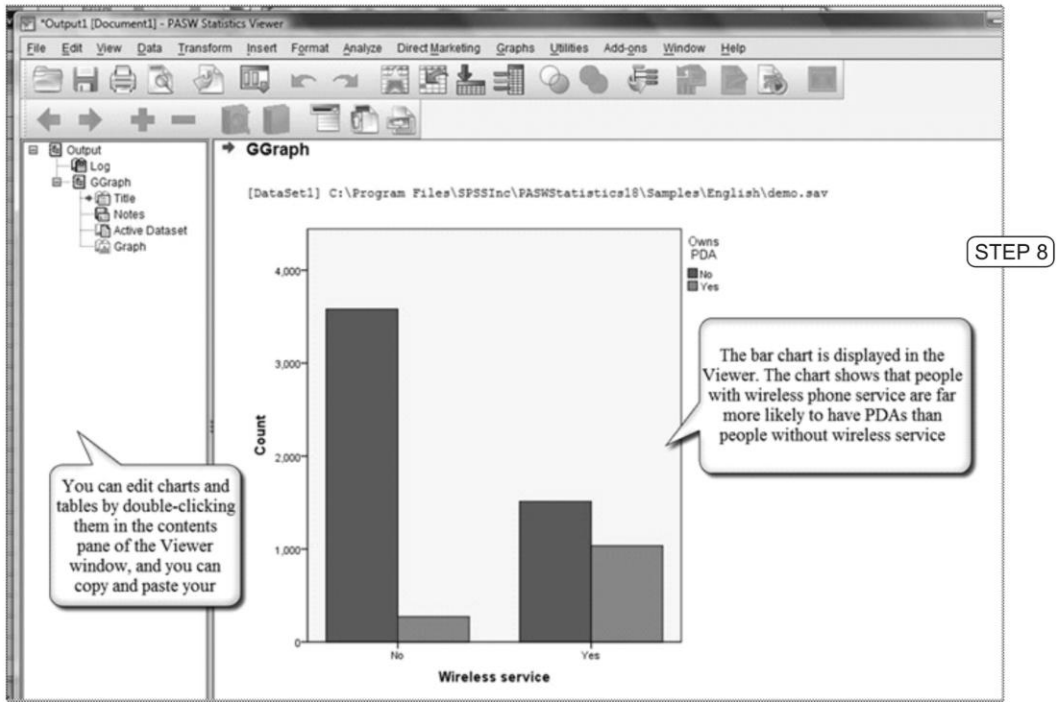


Chart Builder
Variables:
Chart preview uses example data
Internet [internet]
Caller ID [callid]
Call waiting [call...
Owns TV [owntv]
Owns VCR [ownv...
Owns stereo/CD...
Owns PDA [ownp...
Owns computer [...
Owns fax machin...
Newspaper subs...
Response [respo...
No
Yes
Count
No
Yes
Wireless service
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
Element Properties...
Options...
Click OK to create the chart
OK
Paste
Reset
Cancel
Help

STEP 8


*Output1 [Document1] - PASW Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
GGraph
Title
Notes
Active Dataset
Graph
GGraph
[DataSet1] C:\Program Files\SPSSInc\PASWStatistics18\Samples\English\demo.sav
Owns PDA
No
Yes
4,000
3,000
2,000
1,000
0
Count
No
Yes
Wireless service
The bar chart is displayed in the Viewer. The chart shows that people with wireless phone service are far more likely to have PDAs than people without wireless service
You can edit charts and tables by double-clicking them in the contents pane of the Viewer window, and you can copy and paste your

Glossary

Action Research	A research methodology with brainstorming followed by sequential trial-and-error to discover the most effective solution to a problem.
Agglomeration Schedule	Gives information on objects or cases being combined at each stage at hierarchical clustering process.
Alternative Hypothesis	Statement to be accepted if null hypothesis is rejected.
Analysis of Variance: ANOVA	The process for splitting the variation of a group of observations into assignable causes and setting up various significance tests.
Applied Research	Research relating to specific product, service or system or campaign.
Balanced Rating Scale	A scale that has an equal number of categories above and below the midpoint
Basic/Pure/Fundamental Research	Research conducted to extend the existing domain of knowledge about certain subject or topic either in physical form
Beta	Regression coefficients of standardized independent variables
Beta of Stock	A statistical measure which reflects the sensitiveness of a stock to movement in the stock market index like BSE – SENSEX or NSE - NIFTY, as a whole.
Block	A block (analogy with agricultural plots) comprises of same number of experimental units as the number of treatments under experimentation
Blocking	Blocking implies control of factors which are either not of interest or their effect is removed/ filtered/ averaged out.
Brainstorming	is a group creativity technique designed to generate a large number of ideas for the solution of a problem introduced in 1930s by Alex Osborn.
Business Research	It is the process of planning, acquiring, analysing and disseminating relevant data, information and insights to decision makers in an organisation to take appropriate actions that lead to maximising business performance
Canonical Correlation Analysis (CRA)	Used for situations wherein there are several(as compared to one in multiple regression analysis) dependant variables and several independent variables.

Cases	The objects under study
Case Study	The collection and presentation of detailed information about a particular participant or small group, frequently including the accounts of subjects themselves.
Case Study Method of Research	Case study involves in-depth study of several variables related to a single unit, instance, event, etc
Causal or Explanatory Hypotheses	Such hypotheses encompasses situations where we study the impact or influence of one factor (represented by a variable) on some other factor(represented by another variable).
Causal Research	Research concerned with finding the root cause of a symptom
Centralised Approach of Research	In such a setup, in an organisation,, there is an exclusive Department devoted to research activities. All the problems that are perceived or arrive at any branch, region or corporate level are referred to this Department which is generally equipped with adequate expertise to undertake the assignments.
Centroid	Mean values for discriminant scores for a particular group. The number of centroids equals the number of groups, being one for each group. Means for a group on all the functions are the group centroids.
Classification Matrix or Confusion Matrix	Also called assignment, or prediction matrix or table, is used to assess the performance of DA. This is a table in which the rows are the observed categories of the dependent and the columns are the predicted categories of the dependents. When prediction is perfect, all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications. This percentage is called the **hit ratio**.
Cluster Analysis	It is an analytical technique that is used to develop meaningful subgroups of entities which are homogeneous or compact with respect to certain characteristics.
Cluster Centers	These are initial starting point in non Hierarchical clusters . Cluster are built around these centers these are also termed as seeds.
Cluster Centroid	Cluster centrioids are mean values of variables under consideration (variables given while clustering) for all the cases in a particular cluster. Each cluster will have different centroids for each variables.
Cluster Membership	It is the cluster to which each case belongs. It is important to save cluster membership to analyse cluster and further perform ANOVA on the data
Coding	The process of converting responses into numeric symbols called codes.
Coefficient of Contingency	Extent of association between two variables or factors, one of the two or both of whom may not be measurable i.e. categorical or nominal-scale.
Coefficient of Determination.	The square of the correlation coefficient**.**

Common Factor Analysis (CFA)	It is a statistical approach that is used to analyse inter- relationships among a large number of variables (indicators) and to explain these variables (indicators) in terms of a few unobservable constructs (factors).
Communality	The amount of variance, an original variable shares with all other variables included in the analysis. A relatively high communality indicates that a variable has much in common with the other variables taken as a group.
Comparative Scaling Techniques	The comparative scales involve direct comparison of the different objects.
Completely Randomised Experiment	The experiment concerned with the study of only one factor. Each treatment/factor is assigned or applied to the experimental units without any consideration
Concept	concepts are components of constructs and are concrete, and are therefore measurable
Concurrent Deviations.	Measure of correlation that depends only on the sign (and not magnitude) of the deviations of the two variables, say x and y, recorded at the same point of time from their values at the preceding point of time.
Confidence Coefficient	Complement of level of significance
Confidence Interval or limits	The interval or limits within which the true value of the parameter lies.
Confidence Level	It is expressed in percentage e.g. 95%, and indicates the degree of confidence that the true value of the parameter lies in the specified interval
Confirmatory Factor Analysis (CFA)	This technique is used when the researcher has the prior knowledge(on the basis of some pre-established theory) about the number of factors the variables will be indicating. This makes it easy as there is no decision to be taken about the number of factors and the number is indicated in the computer based tool while conducting analysis.
Conjoint Analysis	Involves determining the contribution of variables (each of several levels) to the choice preference over combinations of variables that represent realistic choice sets (products, concepts, services, companies, etc.)
Consistent Estimator	A property which implies that the estimate tends to the true value of the parameter as the sample size increases
Constant Sum Scaling	In this technique a respondent is asked to allocate certain points, out of a fixed sum of points, for each object according to the importance attached by him/her to that object. If the object is not important, the respondent can allocate zero point, and if an object is most important he/she may allocate maximum points out of the fixed points.
Construct	A construct is an abstract based on concepts, or can be thought of as a conceptual model that has measurable aspects.

Construct Validity	Seeks an agreement between a theoretical concept and a specific measuring device, such as observation.
Content Validity	The extent to which a measurement reflects the specific intended domain of content
Contingency Coefficient	Measure of association between two factors in a contingency table
Continuous Rating Scale/ Graphic Rating scale	In this type of scale, the respondents indicate their rating by marking at appropriate distance on a continuous line. The line is labeled at both ends usually by two opposite words. For simplicity and understanding of the respondent, the line may contain points like 1 to 100.
Control Group	In the experimental design, the group which is not subjected to a treatment. Used to evaluate impact of the treatment given to other group(s).
Controlling	This is the managerial part of overseeing that the project is completed as per the schedule, and if something does not go as per schedule, then to take appropriate corrective action.
Convenience Sampling	Such sampling is dictated by the needs of convenience rather than any other consideration.
Convergent Thinking	The thinking which analyses divergent thinking into a specific product, service or system.
Convergent Validity	The general agreement among ratings, gathered independently of one another, where measures should be theoretically related.
Correlation Coefficient	The quantitative measurement of the degree of linear relationship between two variables, say 'x' and 'y'
Correlation Matrix	This is the matrix showing simple correlations between all possible pairs of variables.
Correlational Hypotheses	Such hypotheses are used when we wish to test whether there is any correlation between two variables.
Covariate	The independent variable that is metric in nature is termed as covariate.
CPM	Critical Path Method i.e. CPM, is activity based method used for planning, implementing and controlling of projects.
Creativity	is the trait of a person to think, create or do something new. Creativity is also defined as the ability to generate new ideas or concepts.
Criterion Related Validity	Used to demonstrate the accuracy of a measuring procedure by comparing it with another procedure which has been demonstrated to be valid; also referred to as instrumental validity.
Cross – Sectional Study	A study that is conducted over a group of companies or organisations at the same point of time
Data Mining	A specialised branch encompassing Information Technology and Statistics. It uses statistical techniques such as Outlier Analysis, Correlation and Regression Analysis, Analysis of Variance, Discriminant Analysis, Cluster analysis, Factor analysis, etc. for extracting valuable information like patterns, associations, etc. from the stored data

Data Warehouse	A repository of an organization's electronically stored data.
Decentralised Approach of Research	In such a system, each department has its own research team to take care of its research requirements that arise from time to time.
Deduction/Deductive Logic/Deductive Approach	Deduction is reasoning from generalizations to specifics that flow logically from the generalizations It may also be termed as 'top-down' approach.
Defining Problem	This is a stage of research process where the research problem is defined, based on the problem faced or posed by the decision maker.
Dendrogram	This is the graphical summary of the cluster solution. This is more used while interpreting results than the Agglomeration Schedule, as it is easier to interpret. The cases are listed along the left vertical axis. The horizontal axis shows the distance between clusters when they are joined. This graph gives an indication of the number of clusters the solution may have. The diagram is read from right to left. Rightmost is the single cluster solution just before right is two cluster solution and so on. The best solution is where the horizontal distance is maximum. This could be a subjective process.
Dependence Techniques	The techniques that define some of the variable/s as independent variable/s and some others as dependent variable/s. Aim at finding the relationship among these variables.
Dependent Jobs/Activities	That can be started only when the earlier job(s) is/are completed.
Dependent Variable	A dependent variable is one which depends on an independent variable defined in a study.
Descriptive Hypotheses	Such hypotheses relate to an assumption or statement relating to a population. Such hypotheses also postulate either difference between a variable of two different populations or of the same populations before and after some factor causes a change
Descriptive Research/ Statistical Research	any research which aims to describe the characteristics of data like mean, median, mode, standard deviation, coefficient of variation, measure of skewness, etc.
Design flexibility	A form of reasoning in which conclusions are formulated about particulars from general or universal premises
Discriminant Analysis(DA)	It is a statistical technique for classification or determining a linear function, called discriminant function, of the variables which helps in discriminating between two groups of entities or individuals.
Discriminant Function	A linear combination of discriminating (independent) variables.
Discriminant Score	The, also called the DA score, is the value resulting from applying a discriminant function formula to the data for a given case. The *Z score* is the discriminant score for standardized data.
Discriminating Variables	The independent variables which are used as criteria for discrimination.

Divergent Thinking	The thinking that is quite different from usual ways of doing and observing. It helps to develop insights and new ideas.
Dummy Activities	The activities which do not consume any resources. These are used just to connect events.
Economy of a Questionnaire	Is defined as the time spent by a respondent to answer the questionnaire.
Editing	To ensure consistency in the responses, and to locate omission(s) of any response(s) as also to detect extreme responses, if any. It also checks legibility of all the responses and seeks clarifications in the responses wherever felt necessary
Efficient Estimator	A property which implies that the variance of the estimator is minimum as compared to any other estimator.
Eigen value for Discriminant Function	Eigen value for a discriminating function is defined as ratio of between groups to within group sum of squares.
Eigenvalue in Factor Analysis	Eigenvalue for each factor is the total variance explained by each factor.
Empirical Research	Research based on observed data without the support of any theory or model.
Estimator	A function of sample values to estimate a parameter of a population.
Ethics (also known as moral philosophy)	is defined as a branch of philosophy which seeks to address questions about morality; that is, about concepts like good and bad, right and wrong, justice, virtue, etc.
Ex Post Facto Design	The design used for studying phenomenon / event which is already occurred.
Experiment	The study conducted by manipulating the independent variables to understand the relationships between the independent and the dependent variables and
Experimental Designs	Involves specification of treatments and the method of assigning experimental units to each treatment
Experimental Unit	The object on which the experiment is be conducted.
Exploratory Factor Analysis (EFA)	Used when there is no prior knowledge about the number of factors that the variables will be indicating. In such cases, computer based techniques are used to indicate appropriate number of factors.
Exploratory Research	Research conducted to explore a problem, at its preliminary stage, to get some basic idea about the solution at preliminary stage of a research study.
External research	Research conducted for an organisation by an outside agency like a consultant, consultancy firm or a professional.
External Validity	Related to generalisability of the findings/results to other situations.
Extraneous variable	It is outside or external to the situation under study, and its impact on dependent variable is beyond the scope of the study.

Factor	It represents the underlying dimensions(constructs) that summarise or account for the original set of observed variables.
Factor Loadings	Or component loadings in PCA, are the correlation coefficients between the variables (given in output as rows) and factors (given in output columns). These loadings are analogous to Pearson's correlation coefficient; the squared factor loading is defined as the percent of variance in the respective variable explained by the factor.
Factor Matrix	This contains factor loadings on all the variables on all the factors extracted.
Factor Plot or Rotated Factor Space	This is a plot where the factors are on different axis and the variables are drawn on these axes. This plot can be interpreted only if the number of factors are 3 or less
Factor Scores	Each individual observation has a score, or value, associated with each of the original variables. Factor analysis procedures derive factor scores that represent each observation's calculated values, or score, on each of the factors. The factor score representd an individual's combined response to the several variables representing the factor. The component scores may be used in subsequent analysis in PCA. When the factors are to represent a new set of variables that they may predict or be dependent on some phenomenon, the new input may be factor scores.
Factorial Designs	A factorial experiment is an experiment whose design consists of two or more factors, each with discrete possible values or 'levels', and whose experimental units take on all possible combinations of these levels across all such factors.
Field notes	These are the notes prepared while observing a phenomenon, event, action, etc.
Focus Group Technique	It involves a moderator facilitating discussion within or interview with a small group of selected individuals who are well informed or concerned with a particular topic.
Focus Groups	These are small and selected group of participants that are interviewed by a trained researcher. The participants are from a target research audience, whose opinion is of interest to the researcher & the client.
Forced-Choice Scale	A forced-choice scale requires that participants select one of the offered alternatives
Formulating Research Problem	The formulation of research problem is the crystallisation of the thinking and deliberations about a research problem that is ultimately taken up for research study
Goodness of a Factor	How well can a factor account for the correlations among the indicators ? One could examine the correlations among the indicators after the effect of the factor is removed. For a good factor solution, the resulting partial correlations should be near zero, because once the effect of the common factor is removed , there is nothing to link the indicators.

Haphazard Sampling	In such sampling, the units from the population are selected without any set criteria. They are selected based on the preference, prejudice or bias of the person(s) selecting the sample.
Historical Research	The process of systematically examining past events to give an account; may involve interpretation to recapture the nuances, personalities, and ideas that influenced these events; to communicate an understanding of past events.
Hypothesis	Educated or informed guess about the answer to a question framed in a particular study. A statement or assumption or a claim about a parameter of a population.
Icicle Diagram	It displays information about how cases are combined into clusters at each iteration of the analysis.
Implementing	This is the physical part of carrying out the project as per the schedule.
Independent /Explanatory/ Causal Variable	is a variable which influences the other variables, under consideration, in the study. The value of this variable can be decided or controlled by the researcher.
Independent Jobs/Activities	A job which can be started on its own is called 'independent' job. Two jobs are said to be independent of each other, if both the jobs can be done independently or simultaneously.
Induction/Inductive Logic/ Inductive Approach	Induction reasoning starts from 'Specific' observations or set of observations to Generalised Theory or Law. It could be termed as 'bottoms-up' approach.
Inductive	A form of reasoning in which a generalized conclusion is formulated from particular instances
Inductive Analysis	A form of analysis based on inductive reasoning; a researcher using inductive analysis starts with answers, but forms questions throughout the research process.
Interaction of Factors	Simultaneous impact of two factors
Interdependence Techniques	Interdependence techniques do not assume any variable as independent /dependent variables.
Internal Consistency	The extent to which all questions or items assess the same characteristic, skill, or quality.
Internal Research	A research conducted by a team of in-house experts of the organisation.
Internal Validity	The rigour with which the study is conducted (e.g., the study's design, the care taken to conduct measurements, and decisions concerning what to be measured and the extent to which alternative explanations for any causal relationships are taken into account.
Internal Validity	Internal validity describes the ability of the research design to unambiguously test the research hypothesis.

Interval Scale	A measurement scale, whose successive values represent equal value or amount of the characteristic that is being measured, and whose base value is not fixed, is called an interval scale.
Intervening Variable	In a study involving independent and dependent variables, there could be a variable / factor which might affect the dependent variable, but it cannot be directly observed or measured then this could be considered as intervening variable.
Interviews	A research tool in which a researcher asks questions of participants; interviews are often audio/video-taped for transcription and analysis, later on.
Inverse Sampling	A method of sampling which requires that drawings of random samples shall be continued until certain specified conditions dependent on the results of the earlier drawings have been fulfilled, e.g. until a given number of units of specified type have been found.
Investigative Questions/ Issues	The next level of the questions' hierarchy is the Investigative questions. These questions disclose specific information that is useful to answer the research question.
Itemised Rating Scale	In an itemised scale, respondents are provided with a scale having numbers and/or brief descriptions associated with each category. The categories are usually ordered in terms of scale position. The respondents are asked to select one of the categories, that best describes the product, brand, company or any other attribute being studied.
Judgement Sampling	In such type of sampling, the selection of units, to be included in the sample, depends on the judgement or assessment of the person(s) collecting the sample.
Kaiser Meyer Olkin (KMO) Measure of Sampling Adequacy	This is an index used to test appropriateness of the factor analysis. High values of this index, generally, more than 0.5 , may indicate that the factor analysis is an appropriate measure, where as the lower values (less than 0.5) indicate that factor analysis may not be appropriate.
Levels of a Factor	Different values of a factor. Values of the factor that are used like 10 gms, 12 gms and 15 gms per sq. meter of land, dosages of medicine, duration of training, Percentages of discounts(5%,10%, 15%)
Likert Scale/Summated Rating Scales	comprise of statements that express either a favourable or an unfavourable attitude toward the object of interest on a 5 point,7 point or on any other numerical value. The respondents are given a list of statements and asked to agree or disagree with each statement by marking against the numerical value that best fits their response. The scores may be summed up to measure the respondent's overall attitude
Longitudinal Studies	Studies which are conducted over a period of time
Management Question/ Issues	The Management dilemma gets translated into management Questions. The management questions convert the dilemma in to question form

Maximum Likelihood Estimation	This method is used in logistic regression to predict the odd ratio for the dependent variable. In least square estimate, the square of error is minimized, but in maximum likelihood estimation, the log likelihood is maximized
Measurement Questions/Issues	These questions allow the researcher to collect specific information required for the research study. For each investigative question, the measurement questions are asked.
Moderating Variable	In a study involving an independent variable and a dependent variable, a relationship could be established through a variable. However we may come across a third variable, which is not an independent variable but forms strong contingent/ contextual effect on the relationship of the independent and dependent variable is moderating variable.
Monte Carlo Simulation	This technique uses modeling of key variables with defined random distributions to cover potential values in solving analytical problems.
Multicollinearity	Refers to the existence of high correlation between independent variables.
Multidimensional Scaling	It is a set of procedures for drawing pictures of data so as to visualise and clarify relationships described by the data more clearly.
Multiple Choice, Multiple-Response Scale/Check List	In this scales the respondent is given a list of multiple choices and can choose more than one choices from the list.
Multiple Choice, Single-Response Scale.	In such scales there are multiple options for the respondent , but only one answer can be chosen.
Multiple Regression Analysis	Deals with the study of relationship between one metric dependent variable and more than one metric independent variables
Multivariate Analysis of Variance (MANOVA)	It explores, simultaneously, the relationship between several non-metric independent variables and **two or more** metric dependant variables.
Nominal Scale	A qualitative scale without any order is called nominal scale.
Non- Participant Observations	This is the qualitative method of data collection where the researcher does not become a part of the group for observation.
Non-Comparative Scaling Techniques	The non-comparative scales involve scaling of each object independently of other objects.
Non-Parametric Tests	Tests of significance used when certain assumptions about the usual tests of significance are not valid or doubtful.
Non-Random/ Non-Probability Sampling	In this type of sampling scheme, the selection of units is subjective and not based on any probability considerations.
Normative Exploratory Research	This is the normative research of exploratory nature
Normative Research	Usually conducted while developing a new product, service or system to assure whether desired objective / standard has been achieved.
Null Hypothesis	An assumption or claim about a specific parameter

Objects	Individuals, Items, Units, Entities, etc.
Observation	is a qualitative method of data collection. In this technique, the information is captured by observing, objects, human behaviour, systems and processes, etc.
One –Way or One – Factor ANOVA	When the source of variation in the observations is primarily due to one factor.
Ordinal scale	is a scale that does not measure values of the characteristic(s) but indicates only the order or rank. In other words a qualitative scale with order
'p' Value	The probability of getting the given sample when the null hypothesis is true.
Paired Comparison	In paired comparison scales, the respondent is asked to select one object from the list of two objects, on the basis of some criteria. This forces the respondent to compulsorily select one of the two.
Participant Observation	This is the qualitative method of data collection where the researcher becomes a part of the group for observation or plays the role of a participant
PERT	Program Evaluation and Review Technique is event based method used for planning implementing and controlling of projects.
Pilot Study/Pretesting	It's a testing method wherein, after the questionnaire is constructed, it is given to a set of people to test various aspects like completeness, time taken to answer, clarity of questions, etc.
Point Estimation	A single value estimate like 20.
Population (or Universe)	The collection of all the units of a specified type at a particular point or period of time is called a population or universe
Post Hoc Test	Test carried out based on the result of the earlier test
Power of a Test (1–β)	The probability that a test will lead to rejection of a statement when it is false. Equal to 1 – Type II error.
Primary Data	The data which is directly collected by the researcher for some specific purpose or study.
Primary Research	Original research reports published by the collectors of data
Principal Component Analysis (PCA)	Technique for forming set of new variables that are linear combinations of the original set of variables, and are uncorrelated. The new variables are called Principal Components.
Principle of Least Squares	Sum of squares of differences between the observed values and the estimated values is *minimum*
Projective Techniques	It is a technique of interviewing. It is an indirect method of questioning in which the interviewer directs the questioning to receive responses on questions not directly related to the respondent, and analysing the response behaviour to matters not directly related to him.
Purposive Sampling	As the name implies, under this type of sampling, units of the population are selected according to the relevance and the nature of representativeness of sampled units.

Qualitative/Categorical/ Non-Metric Variables	The variables which cannot be quantified into some numeric value and the arithmetic calculations like addition, subtraction, etc. cannot be performed.
Qualitative Factor	Not measured numerically like gender, colour, location
Qualitative Research	Researcher immersed in the phenomenon to be studied, gathering data which provide a detailed description of events, situations and interaction between people and things, providing depth and detail. In which the researcher explores relationships using textual, rather than quantitative data. Case study is considered a form of qualitative research. Results are not usually considered generalisable, but are often transferable.
Quantitative/Numeric/ Metric Variables	The variables which can be represented with some numeric value and the arithmetic calculations like addition, subtraction, etc. can be performed.
Quantitative Factor	Measured on a numerical score like discount in price, is based on
Quantitative Research	quantitative or qualitative data relating to measurement, counting and frequency of occurrences and other statistical analysis. In which the researcher explores relationships using numeric data. Survey is generally considered a form of quantitative research. Results can often be generalized, though this is not always the case.
Quasi-experiment	A scientific research method primarily used in the social sciences. "Quasi" means likeness or resembling, and therefore quasi-experiments share characteristics of true experiments which seek interventions or treatments.
Questionnaire	A set of questions asked to individuals to obtain useful information about a given topic of interest.
Quota Sampling	Such sampling is, sometimes considered a type of purposive sampling. It is usually resorted when some quota about the number of units to be included in the sample is fixed.
Random/Probability Sampling	It's a sampling procedure where each and every unit of population has some pre defined probability to be selected in a sample.
Randomised Block Design	Such experimental design involves study of two or more factors. One experimental unit from each of the blocks, say 'n' in number, is assigned to each of the, say, 'm' treatments. Thus, 'n' blocks have 'm' treatments in each block.
Rank Order Scaling	Rank order scaling technique prompts respondents to rank a given list of objects
Rank Sum	Sum of ranks
Ratio scale	is quantitative measure with fixed or true zero.
Refining Problem	This is a stage of research process where the problem is redefined by further investigation in the research study.
Regression Analysis	Study of the relationship among two or more variables

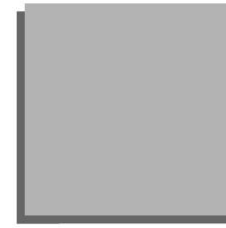
Relational Hypotheses	Such hypotheses are concerned with studying or analysing relationship or correlation between two variables (characteristics) or among more than two variables (characteristics) of a population
Relative Percentage of a Discriminant function	Equals a function's eigenvalue divided by the sum of all eigenvalues of all discriminant functions in the model. Thus it is the percent of discriminating power for the model associated with a given discriminant function.
Reliability	Reliability indicates the confidence one could have in the measurement obtained with a scale. It tests how consistently a measuring instrument measures a given characteristic or concept.
Request for Proposal (RFP)	A document relating to a project that is prepared and issued by the promoter, for competitive bidding by the interested agencies.
Research	It is an organised systematic data-based scientific inquiry, or investigation into a specific problem, undertaken with the purpose of finding answers or solutions to it.
Research Design	A comprehensive plan of the sequence of operations that a researcher intends to carry out to achieve the objectives of a research study. It provides the conceptual structure or blue print for the conduct of the research study.
Research Methodology	"It is the analysis of the principles of methods, rules and postulates used in a field of study." "It encompasses the systematic study of methods that are useful in a field of study."
Research Process	The methodology of conducting a research assignment / project / study in a scientific and systematic manner. The step by step scientific process that is followed to conduct research.
Research Proposal	It encompasses the methodology of conducting the research to solve the formulated research problem.
Research Questions/Issues	The questions that are selected by the researcher for further analysis, Out of the several management questions.
Response	It is the dependent variable of interest
Run	A succession of values with the same sign or type(in case of qualitative data)
Sample	One or more units, selected from a population according to some specified procedure.
Sampling Frame	Sampling frame is assigning unique number to each and every unit of population.
Scatter Diagram	Values of x and y depicted with the help of the rectangular co-ordinate system, by plotting the observed **pairs of values** of x and y.
Scheduling	It implies indicating the starting and completion times for all the activities.

Scientific Research	The research conducted in Science subjects such as Physics, Chemistry, Biology, etc., or research study conducted using scientific process.
Scree Plot	A plot of Eigen values against the factors in the order of their extraction.
Secondary Data	The data which is disseminated through some media
Secondary Research	Reports/Publications analyzing and evaluating data collected by others
Semantic Differential Scale	This scale provides a measure to the psychological meaning of an attitude or an object, using bipolar adjectives. The respondents mark in the blank spaces provided between the two objects, indicating how they would best rate the object. Commonly this is rated on 7-point scale.
Semi-Structured Interviews	This method is used when the researcher asks the respondent some basic questions, and then lets the respondent answer, interfering whenever necessary. In this method, the interviewer sets some guidelines for the questions to be asked. The succeeding questions are generally on the basis of the preceding questions.
Signed Rank	The ranks assigned to observations are attached a sign viz. + or –
Significant	Difference between the sample value and population value is said to be significant if the calculated statistic falls in the rejection or critical region.
Similarity/ Distance Coefficient Matrix	It is a matrix containing the pair wise distances between the cases.
Simple Category Scales	This scale is also termed as a dichotomous scale. It offers two mutually exclusive response choices, typically a 'Yes' or 'No' type of response.
Simple Random Sampling	It's a sampling procedure where each and every unit of population has equal probability to be selected in a sample.
Simple Regression Analysis	Study of the relationship between two variables
Snowball Sampling/ Chain Referral Sampling	In such sampling, the sampling units are not fixed in advance but are decided as the sampling proceeds.
Social/Behavioural Research	Refers to research conducted by social and behavioural researchers into sociology, political science, behavioural science, education, etc.
Spurious Correlation	When two variables change due to different factors and not due to interdependence on each other.
Rank Correlation	Association between two variables where data is given in the form of the ranks of two variables based on some criterion.
Standard Error	Standard Deviation of an Estimate from a Sample
Standardised Discriminant Coefficients	Also termed as *standardised canonical discriminant function coefficients*. Used to compare the relative importance of the independent variables, like beta weights in regression.

Stapel Scales	It is a unipolar rating scale with 10 categories numbered from -5 to +5 without a neutral point or zero.
Statistic	A function of sample values
Stepwise Method	Helps in finding out the independent variables, out of a set of independent variables, that contribute most significantly in the regression model.
Stratified Sampling	Involves classifying the population into a certain number of non-overlapping homogeneous groups, called strata, and then selecting samples independently from each stratum.
Structure Correlations	Also known as discriminant loadings, are defined as simple correlations between the independent variables and the discriminant functions
Structured Interviews	The set of designed questions are personally asked by the researcher or interviewer. In this method, either the questionnaire is handed over personally, and taken back after completion by the respondent, or he is asked the questions orally and the responses are noted down by the interviewer
Sufficient Estimator	A property which implies that after the calculation of such estimator, the sample does not contain any worthwhile information about the estimated parameter.
Survey	A method for collecting quantitative/qualitative information about units in a population, usually through a questionnaire.
Systematic Sampling	A method of sampling in which, only the first unit of the sample is selected randomly, and the rest are selected automatically according to a pre- determined pattern.
Conventional Scales	Mostly used in the questionnaire format.
Non-conventional Scales	Used for unconventional collection of data through games, quizzes etc.
Think Tank	An embodiment of persons engaged in strategic or policy level thinking over some issues with the objective of arriving at suitable action plan or course of action to resolve the issue.
Trace	The sum of squares of the values on the diagonal of the correlation matrix used in the factor analysis. It represents the total amount of variance on which the factor solution is based.
Treatment or Factor	Those independent variables whose effect on the response is of interest to the experimenter
Two – Way or Two – Factor ANOVA	When there are two factors as sources of variation in the observations.
Type-I Error (α)	Probability that a test will result in the rejection of a hypothesis when it is true. Denoted by Greek letter 'α'.
Type-II Error (β)	Probability that a test will result in the rejection of a hypothesis when it is false. Denoted by Greek letter 'β'
Unbalanced Rating Scale	A scale that has unequal number of favourable and unfavorable response choices.

Unbiased Estimator	A property which implies that its expected value is equal to the population value.
Unforced-Choice Rating Scale	This allows participants to express no opinion when they are unable to make a choice among the alternatives offered.
Unstructured Interviews	Such interviews allow the interviewer to get opinions and get a feel of general attitudes of the respondents. In this type of interviewing method, the questions asked are not structured, and different questions may be asked to different participants.
Validation	Process of checking and correcting data for data entry errors. Also to ensure that the data has been collected as per the specifications in the prescribed format or questionnaire.
Validity	The degree to which a study accurately reflects or assesses the specific concept that the researcher is attempting to measure.
Validity of a Measuring Instrument	Indicates the extent to which an instrument/scale tests or measures what it is intended to measure.
Variable	Observable characteristic that varies among individuals/items/ units/ entities, etc.
Wilks' lambda	Used to test the significance of the discriminant function as a whole. The "Sig." level for this function is the significance level of the discriminant function as a whole

Some Other Useful Books on Business Research Methodology



Aczel Amir D. and Jayavel Sounderpandian: Complete Business Statistics, Tata McGraw-Hill Publishing Company Ltd (2006) Sixth Edition.

Anderson, David R., Sweeney Dennis J and Williams, Thomas A : Statistics for Business and Economics. Thomson Asia Pte. Ltd (2002) Eighth Edition.

Anderson, T.W.: An introduction to Multivariate Statistical Analysis, John Wiley & Sons Inc. (2003) Third Edition.

Beri, C.G.: Business Statistics, Tata McGraw-Hill Publishing Company Ltd., (2005), Second Edition.

Cochran, William G and Cox, G.M.: Experimental Designs, John Wiley & Sons, Inc. (1977) Third Edition

C.R. Kothari: Research Methodology—Methods and Techniques, New Age Publications (2002) Second Revised Edition.

Donald R. Cooper and Pamela S. Schindler: Business Research Methods, Tata McGraw-Hill Publishing Company Ltd. (2009), Ninth Edition.

Fred N. Kerlinger and Howard B. Lee: Foundations of Behavioural Research, Harcourt College Publishers (2000), Fourth Edition.

Frye, Microsoft® Excel Version 2002 Step By Step, Microsoft Press (2002)

Guy Hart Davis : How to do Everything with Microsoft Excel 2007, Tata McGraw-Hill Ltd. (2007)

Josef F. Hair, Jr., William C. Black, Barry J. Babin, Ralf E. Anderson and Ronald L. Tatham: Multivariate Data Analysis, Pearson Prentice Hall (2006), Sixth Edition.

Hastings NAJ and Peacock JB: Statistical Distributions, London Butterworth (1975)

Johnson Richard A. and Wichern Dean W.: Business Statistics – Decision Making with Data, John Wiley & Sons, Inc. (2003)

Kanji, Gopal K.: 100 Statistical Tests, SAGE Publications Ltd., (1999), New Edition

Lee Cheng F., Lee John and Lee, Alice C L: Statistics for Business and Financial Economics, World Scientific Publishing Company Pte. Ltd (2000) Second Edition

Levin Richard I. and Rubin David S.: Statistics For Management, Prentice Hall of India Pvt. Ltd., (2002) Seventh Edition.

Malhotra, Naresh K.: Marketing Research—An Applied Orientation, Pearson Prentice Hall (2006) Fifth Edition

Murray R. Spiegel: Schaum's Outline Series—Theory and Problems of Statistics, McGraw-Hill Book Company (1972)

Pande Peter S, Neuman Robert P and Cavanaugh Ronald R: The Six Sigma Way, Tata McGraw-Hill (2003)

Paul McFedries: Formulas and Functions with Microsoft Office Excel 2007, Pearson Education Asia (2007)

Rajendra Nargudkar: Marketing Research – Text and Cases, Tata Mc Graw-Hill Publishing Company Ltd. (2009) Third Edition.

Richard A. Johnson and Dean W. Wichern: Applied Multivariate Analysis, Pearson Prentice Hall (2007) Sixth Edition.

Sharma Subhash: Applied Multivariate Techniques, John Wiley & Sons (1996) First Edition

Siegel Sidney and Castellan, Jr. N. John : (Nonparametric Statistics for the Behavioral Sciences) McGraw-Hill International Edition (1988),

Subhash Sharma: Applied Multivariate Techniques, John Wiley and Sons, Inc. (1996).

Uma Sekaran: Research Methods for Business–A Skill Building Approach, John Wiley and Sons (2003) Fourth Edition.

Vohra N.D: Quantitative Techniques in Management, Tata McGraw-Hill (2007) Third Edition

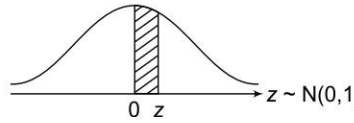
William G. Zikmund: Business Research Methods, Thomson South Western (2006) Seventh Edition.

Yule G U and Kendall M.G: An Introduction to the Theory of Statistics, Charles Griffin & Co. Ltd., (1973) Eleventh Edition.

Zikmund, William: Business Research Methods, Thomson Asia Pte. Ltd (2004) 7th Edition.

Statistical Tables

Table T1 Areas under the Standard Distribution from 0 to z



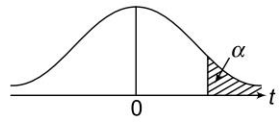
z	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952

(*Contd*)

(*Contd*)

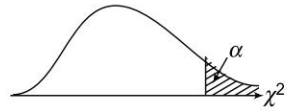
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

Table T2 Student's 't' Distribution-Values of 't'



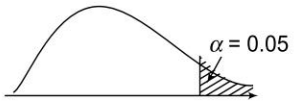
df \ α	*0.1*	*0.05*	*0.025*	*0.01*	*0.005*
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∝	1.282	1.645	1.960	2.326	2.576

Table T3 Chi-square Distribution-Values of χ^2 for Various Levels of Significance (α)



df	α	*.10*	*.05*	*.02*	*.01*
1		2.706	3.841	5.412	6.635
2		4.605	5.991	7.824	9.210
3		6.251	7.815	9.837	11.345
4		7.779	9.488	11.668	13.277
5		9.236	11.070	13.388	15.086
6		10.645	12.592	15.033	16.812
7		12.017	14.067	16.622	18.475
8		13.362	15.507	18.168	20.090
9		14.684	16.919	19.679	21.666
10		15.987	18.307	21.161	23.209
11		17.275	19.675	22.618	24.725
12		18.549	21.026	24.054	26.217
13		19.812	22.362	25.472	27.688
14		21.064	23.685	26.873	29.141
15		22.307	24.996	28.259	30.578
16		23.542	26.296	29.633	32.000
17		24.769	27.587	30.995	33.409
18		25.989	28.869	32.346	34.805
19		27.204	30.144	33.687	36.191
20		28.412	31.410	35.020	36.566
21		29.615	32.671	36.343	38.932
22		30.813	33.924	37.659	40.289
23		32.007	35.172	38.968	41.638
24		33.196	36.415	40.270	42.980
25		34.382	37.652	41.566	44.314
26		35.563	38.885	42.856	45.642
27		36.741	40.113	44.140	46.963
28		37.916	41.337	45.419	48.278
29		39.087	42.557	46.693	49.588
30		40.256	43.773	47.962	50.892
40		51.805	55.759	60.436	63.692
50		63.167	67.505	72.613	76.154
60		74.397	79.082	84.580	88.379

Table T4 'F'-Distribution-Values of 'F' Level of Significance $\alpha = 0.05$



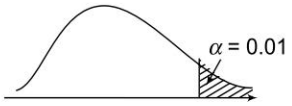
n_1 = degrees of freedom for numerator

n_2 = degrees of freedom for denominator

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∝
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	2.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	21.5	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	3.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.43	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∝	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

(Contd.)

Table T4 (Contd.) 'F'-Distribution-Values of 'F' Level of Significance $\alpha = 0.01$



n_1 = degrees of freedom for numerator

n_2 = degrees of freedom for denominator

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Table T5 Random Number Table

2315	7548	5901	8372	5993	7624	9708	8695	2303	6744
0554	5550	4310	5374	3508	9061	1837	4410	9622	1343
1487	1603	5032	4043	6223	5005	1003	2211	5438	0834
3897	6749	5194	0517	5853	7880	5901	9432	4287	1695
9731	2617	1899	7553	0870	9425	1258	4154	8821	0513
1174	2693	8144	3393	0872	3279	7331	1822	6470	6850
4336	1288	5911	0164	5623	9300	9004	9943	6407	4036
9380	6204	7838	2680	4491	5575	1189	3258	4755	2571
4954	0131	8108	4298	4187	6953	8296	6177	7380	9527
3676	8726	3337	9482	1569	4195	9686	7045	2748	3880
0709	2523	9224	6271	2607	0655	8453	4467	3384	5320
4331	0010	8144	8638	0307	5255	5161	4889	7429	4647
6157	0063	6006	1736	3775	6314	8951	2335	0174	6993
3135	2837	9910	7791	8941	3157	9764	4862	5848	6919
5704	8865	2627	7959	3682	9052	9565	4635	0653	2254
0924	3442	0068	7210	7137	3072	9757	5609	2982	7650
9795	5350	1840	8948	8329	5223	0825	2122	5326	1587
9373	2595	7043	7819	8885	5667	1668	2695	9964	4569
7262	1112	2500	9226	8264	3566	6594	3471	6875	1867
6102	0744	1845	3712	0794	9591	7378	6699	5361	9378
9783	9854	7433	0559	1718	4547	3541	4422	0342	3000
8916	0971	9222	2329	0637	3505	5454	8988	4381	6361
2596	6882	2062	8717	9265	0282	3528	6248	9195	4883
8144	3317	1905	0495	4806	7469	0075	6765	0171	6545
1132	2549	3142	3623	4386	0862	4976	6742	2452	3245
6475	5838	8584	1222	5920	1769	6156	5595	0459	5947
1030	2522	8977	4363	4430	3811	2490	6707	3482	3328
7101	7984	9551	3085	0374	6659	1028	8753	7656	9149
6001	2556	0588	4103	4879	7965	5901	6978	8000	3666
3733	0946	5649	1614	2802	4827	4547	5544	5536	5090
4786	9870	0131	5911	2273	6062	6128	2234	6916	1212
3804	0427	3764	1678	9578	3932	3493	2488	4343	8706
7350	8309	0883	0548	0078	3666	9302	9556	4604	5336
3262	3464	7484	0610	4324	2062	8373	1932	3564	3969
9759	1995	4936	6303	5106	6206	9929	7595	3205	7734
7401	2319	5559	7909	6982	6622	4240	1596	7490	7589
5675	4264	5713	3510	5014	9096	6336	7469	0963	3488
4980	0499	0854	8312	1998	0852	8263	7292	9236	5026
4358	4896	4724	8785	6670	0022	1501	9399	5916	2377
1665	3796	6460	3257	1301	3574	2836	3673	0588	7229
4850	2690	5565	3225	8748	3144	6802	3731	2529	6367
9676	5546	9236	3168	6230	4829	6383	5223	8166	4094
3892	3615	5080	3578	1784	2344	4124	6333	9922	8128
7795	8816	9425	2250	5587	5107	3010	7060	2186	1961
1792	8280	6525	5860	8771	0264	1850	6465	7964	8170
9403	6859	7802	3180	4499	4105	4105	3187	4312	1596
4746	0604	7956	2304	8417	1437	2851	6727	5580	0368
4785	6560	8851	9928	2439	4064	4171	7013	4631	8288
5761	6346	5392	2986	2018	1037	5765	1562	9869	0756
0830	0927	0466	7526	6610	5718	8791	0754	2222	2013

Table T6 Minimum Value of Pearson's Correlation Coefficient 'r' Considered Significant

Sample Size	*Level of Significance (α)*			
n	*0.1*	*0.05*	*0.02*	*0.01*
4	0.9000	0.9500	0.9800	0.9900
5	0.8054	0.8783	0.9343	0.9587
6	0.7293	0.8114	0.8822	0.9172
7	0.6694	0.7545	0.8329	0.8745
8	0.6215	0.7067	0.7887	0.8343
9	0.5822	0.6664	0.7498	0.7977
10	0.5494	0.6319	0.7155	0.7646
11	0.5214	0.6021	0.6851	0.7348
12	0.4973	0.576	0.6581	0.7079
13	0.4762	0.5529	0.6339	0.6835
14	0.4575	0.5324	0.612	0.6614
15	0.4409	0.514	0.5923	0.6411
16	0.4259	0.4973	0.5742	0.6226
17	0.4124	0.4821	0.5577	0.6055
18	0.4000	0.4683	0.5425	0.5897
19	0.3887	0.4555	0.5285	0.5751
20	0.3783	0.4438	0.5155	0.5614
21	0.3687	0.4329	0.5034	0.5487
22	0.3598	0.4227	0.4921	0.5368
23	0.3515	0.4132	0.4815	0.5256
24	0.3438	0.4044	0.4716	0.5151
25	0.3365	0.3961	0.4622	0.5052
26	0.3297	0.3882	0.4534	0.4958
27	0.3233	0.3809	0.4451	0.4869
28	0.3172	0.3739	0.4372	0.4785
29	0.3115	0.3673	0.4297	0.4705
30	0.3061	0.361	0.4226	0.4629
35	0.2826	0.3338	0.3916	0.4296
40	0.2638	0.312	0.3665	0.4026
45	0.2483	0.294	0.3457	0.3801
50	0.2353	0.2787	0.3281	0.361

Table T7 Minimum Value of Rank Correlation Coefficient Considered Significant

	Level of Significance			
n	*0.10*	*0.05*	*0.025*	*0.01*
4	0.8000			
5	0.8000	0.9000	0.9000	
6	0.7714	0.8286	0.8857	0.9429
7	0.6786	0.7450	0.8571	0.8929
8	0.6190	0.7143	0.8095	0.8571
9	0.5833	0.6833	0.7667	0.8167
10	0.5515	0.6364	0.7333	0.7818
11	0.5273	0.6091	0.7000	0.7455
12	0.4965	0.5804	0.6713	0.7273
13	0.4780	0.5549	0.6429	0.6978
14	0.4593	0.5341	0.6220	0.6747
15	0.4429	0.5179	0.6000	0.6536
16	0.4265	0.5000	0.5824	0.6324
17	0.4118	0.4853	0.5637	0.6152
18	0.3994	0.4716	0.5480	0.5975
19	0.3895	0.4579	0.5333	0.5825
20	0.3789	0.4451	0.5203	0.5684
21	0.3688	0.4351	0.5078	0.5545
22	0.3597	0.4241	0.4963	0.5426
23	0.3518	0.4150	0.4852	0.5306
24	0.3435	0.4061	0.4748	0.5200
25	0.3362	0.3977	0.4654	0.5100
26	0.3299	0.3894	0.4564	0.5002
27	0.3236	0.3822	0.4481	0.4915
28	0.3175	0.3749	0.4401	0.4828
29	0.3113	0.3685	0.4320	0.4744
30	0.3059	0.3620	0.4251	0.4665

Table T8 Acceptance Region for Values of 'R' in the RUN Test for Randomness

Sample Size	*Minimum Value*	*Maximum Value*
20	6	16
24	7	19
30	10	22
40	14	28

Table T9 Minimum (Critical) Value of 'T' for Signed Rank Test for Mean

Sample Size	*Level of Significance (α)*	
n	*0.05*	*0.01*
10	8	10
14	9	11
20	10	14
25	11	15
30	12	16

Table T10 Critical Values of D in the Kolmogorov-Smirnov One Sample Test

Sample Size	*Level of Significance for D = Maximum [$F_0(X) - F_e(X)$]*				
n	*.20*	*.15*	*.10*	*.05*	*.01*
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	.356
25	.21	.22	.24	.27	.32
30	.19	.20	.22	.24	.29
35	.18	.19	.21	.23	.27
Over 35	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Table T11 Critical Values of U in the Mann-Whitney Test

Critical Values for One-Tail Test at $\alpha = .025$ or a Two-Tail Test at $\alpha = .05$

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	66	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

Critical Values for One-Tail Test at $\alpha = .05$ or a Two-Tail Test at $\alpha = .10$

n_2 \ n_1	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

Table T12 Wilcoxon-Wilcox Test for Comparison of Multiple Treatments (Level of Significance $\alpha = 0.05$)

	k		
n	*3*	*4*	*5*
4	6.6	9.4	15.1
5	7.4	10.5	13.6
6	8.1	11.5	14.9
8	9.4	13.3	17.3
10	10.5	14.8	19.3

ANSWERS TO EXERCISES

Chapter 9

1. CV = 109.1143. It is a measure of disparities in GDP among various countries and remains unchanged as 109.1143.
2. Total sales (2005) = 80.4 Crores. Total sales (2006) = 111.2 Crores. Growth: 38.3%.
3. Median – 76,075, IQR - 13,366, Sample Std Dev - 13,582.67, Coefficient of Variation - 20%
4. (i) 68.58 (ii) 67.1% (iii) 7855.82 (iv) 30.86%
5. Jyoti: Mean = 80% CV = 9.048481
Anuj : Mean = 79% CV = 3.10062
Their average is about the same but Anuj has more consistent rating than Jyoti.
6. 18.92%
7. While mean has reduced by 1.14 days, s.d. slightly increased by 0.48. However, CV has increased. i.e. variability is increased by 1%
8. Mean - 10.25, Median - 9.00, Coefficient of Variation - 57%
ROI 12%, Rating: 66.33%
9. Average Rate of Interest or Cost of Funds = 5.366%

Chapter 10

1. (a) Net profit = 983.4 + 0.055 × Net Sales, $r = 0.3663$, $r^2 = 0.1342$
(b) Net sales = 7818.07 + 3.02 × P/E Ratio, $r = 0.0231$, $r^2 = 0.0005$.
2. (i)0.8909 (ii) 0.7576 (iii) 0.8667
3. (a) (i) 0.104378 (ii) 0.146146 (iii) 0.073593
(b) (i) Beta of ICICI stock = 1.0273
This implies that the ICICI Stock is 2.73% more aggressive than BSE
(ii) Beta of Reliance Industries stock = 0.8523
This implies that the Reliance Industries Stock gives 85.23 % returns if the return on BSE is 100%.
(iii) Beta of L & T stock = 0.981
This implies that the L & T Stock gives 98.1 % returns if the return on BSE is 100%.
4. Sales Revenue = 60.94737 + 4.302632 × Advertising Expenses
Sales Revenue = 190.0263 Crores for Advertising expenses = 30 Crores.

5. r = 0.685315.
It is desirable that the correlation is high for consistency in rankings by the two executives. They may not be able to recruit the right candidates.

6. (a) Profit = 6.5625 + 121.875 × Expenditure on R&D
(b) 128.4375 (c) 0.97745 (d) 95.54%

7. 0.818182 Only 67% of variation in deposits is explained by customer satisfaction. The bank might be getting deposits due to other reasons also such as higher rate of deposits, confidence reposed by customers, flexibility in deposit schemes, etc.

8. (i)Demand = 188.522 – 7.33 × Price (ii) 78,560 units (iii) 90.86%

9. (a) Marks = 1 + 0.654 × I.Q. (b) 79.515 (c) 0.9014
(d) 81.25%

10. $r = 0.6232$ $r^2 = 0.3884$. Thus, only about 39% of variation in 'Score on Job' is explained by 'Academic Score'

Chapter 11

1. 0.7 ± 0.0421 = (0.6579, 0.7421)

2. Sample size = 42

3. $H_0 : m = 1000$ ml.
$H_1 : m \neq 1000$ ml.
Test Statistic $z = -4$. z tabulated = – 2.576. Reject *H*o. The machine does not fill 1000 ml.

4. $H_0 : m = 15.6$
$H_1 : m \neq 15.6$
Test Statistic $z = -1.93$, z tabulated = 2.575. Do not reject *H*o. The mean breaking strength of the lot could be 15.6

5. $H_0 : m = 2000$
$H_1 : m < 2000$
Test Statistic $z = -2.5$. z tabulated = –2.05 Reject *H*o. The population mean is less than 2000 hrs.

6. $H_0 : p = 0.90$
$H_1 : p < 0.90$
Test Statistic $z = -4.714$. z tabulated = – 1.645. Reject *H*o. The claim that the medicine is effective for 90% of people is not justified.

7. $H_0 : m_1 = m_2$
$H_1 : m_1 \neq m_2$
Test Statistic = –3.95. Reject *H*o. There is significant difference in the two brands of gloves.

8. $H_0 : m_1 = m_2$
$H_1 : m_1 > m_2$
Test Statistic = 6.3246. Reject *H*o. The process is effective in reducing time.

9. $H_0 : m_1 = m_2$
$H_1 : m_1 < m_2$
Test Statistic = 2.1921. Reject. The additive has increased the mileage of cars.

10. 90% : (29.013,30.987), 95% : (28.824,31.176), 99% : (28.4545, 31.5455)

11. $H_0 : p_1 = p_2$
$H_1 : p_1 < p_2$

Test Statistic = 1.572.. z tabulated = 1.645. Do not reject *H*o. There is no significant improvement by the sponsorship of the movie.

12. $H_0 : d = 0$
$H_1 : d \neq 0$
Test Statistic = 30.98 Reject *H*o. There is significant change in perception of the product's effectiveness.

13. Test statistic $\chi^2 = 2.9946$. Do not reject the hypothesis that they are independent.

14. Test statistic $\chi^2 = 10.302$. Reject hypothesis that approval/rejection is not independent of the officer.

15. $H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$
$F = 1.01$ Do not reject *H*o. There is no significant difference in the variances of the service times of the two operators.

Chapter 12

1. ANOVA Table

Source	*SS*	*df*	*MS*	*F*	$F_{critical}(\alpha = 5\%)$	*Result*
Between Card Types	1203.33	2	601.67	18.513	3.8853	Reject
Within Card Types	390	12	32.5			
Total	1593.33	14				

There is difference in the billing amounts of different cards. This implies that the amount spent depends on the type of the card.

2. ANOVA Table

Source	*SS*	*df*	*MS*	*F*	$F_{critical}(\alpha = 5\%)$	*Result*
Between Institutes	33.617	2	16.808	1.2865	3.8853	Do not reject
Within Institutes	156.783	12	13.065			
Total	190.4	14				

There is no difference in the effectiveness of the salesmen of these different Institutes.

3. ANOVA Table

Source	*SS*	*df*	*MS*	*F*	$F_{critical}(\alpha = 5\%)$	*Result*
Between Shares of Companies	6.9744	2	3.4872	2.0085	4.1028	Do not reject
Within Shares of Companies	17.3625	10	1.7363			
Total	24.3369	12				

There is no difference in the rates of returns among shares of the Companies.

4. ANOVA Table

Source	*SS*	*df*	*MS*	*F*	$F_{critical}(\alpha = 5\%)$	*Result*
Between Offices	4735.6	2	2367.8	2.4516	3.8853	Do not reject
Within Offices	11590	12	965.83			
Total	16325.6	14				

There is no difference among the three Offices.

5. ANOVA Table

Source	*SS*	*df*	*MS*	*F*	$F_{critical}(\alpha = 5\%)$	*Result*
Between Cities	1845.5	3	615.17	95.252	3.4903	Reject
Within Cities	77.5	12	6.4583			
Total	1923	15				

There is difference in the prices among the cities.

6. ANOVA Table

Source	*SS*	*df*	*MS*	*F*	$F_{critical}(\alpha = 5\%)$	*Result*
Between Locations	6333.3	2	3166.7	2.9208	3.8853	Do not reject
Within Locations	13010	12	1084.2			
Total	19343.3	14				

There is no difference among the locations.

7. ANOVA Table

Source	*SS*	*df*	*MS*	*F*	$F_{critical}(\alpha = 5\%)$	
Between Car Brands	115.58	3	38.5278	5.9021	4.7571	Reject
Between Tyre Brands	68.167	2	34.0833	5.2213	5.1433	Reject
Error	39.167	6	6.52778			
Total	222.92	11				

There is difference in the sales among the Car Brands.
There is difference in the sales among the Tyres.

8. ANOVA Table

Source	*SS*	*df*	*MS*	*F*	$F_{critical}(\alpha = 5\%)$	*Result*
Between Sales Persons	10443.3	2	5221.7	1.6577	3.8853	Do not reject
Within Sales Persons	37800	12	3150			
Total	48243.3	14				

There is no difference in the sale prices among the Sales Persons.

Chapter 13

1. Critical Values of R for number of '+' signs = 17 and number of '–' signs = 13, as per Table T11, are ≤ 10 and ≥ 22. Since R = 16 lies in acceptance region, accept the hypothesis that the pattern of returns is random.

2. H_0 : m = 12, H_1: m ≠ 12
T = Min.(11, 55) = 11
Minimum (critical) value of 'T' at α =0.05 from Table T12 is 8. Since T(= 11) > 8, accept H_O that the mean is equal to 12.

3. Tabulated Value of 'D' = 0.0813. Since, Max 'D'(= 0.0107) < Tabulated value of 'D', therefore, H_O is accepted, and data follows Poisson distribution.

4. Tabulated value of 'D' = 0.136. Since, Max 'D'(= 0.02) < Tabulated value of 'D', therefore, H_O is accepted, i.e. percentage of MBA girl students in all the five Institutes are the same.

5. 'U' (for Company Recruited Officers) = 10 × 10 + (10 × 11)/2 – 130.5 = 34.5
'U' (for Market Recruited Officers) = 10 × 10 + (10 × 11)/2 – 79.5 = 75.5

Tabulated value of 'U' with $n_1 = n_2 = 10$ and $\alpha = 0.05$, is = 23. Since Minimum (34.5, 75.5) = 34.5 > 23, reject H_O. Thus, scores of company recruited and market recruited officers are not equal.

6. 'U' = $10 \times 10 + (10 \times 11)/2 - 82.5 = 72.5$
'U' = $10 \times 10 + (10 \times 11)/2 - 127.5 = 27.5$
Tabulated value of 'U' with $n_1 = n_2 = 10$ and $\alpha = 0.05$ is = 23
Since Minimum (72.5, 27.5) = 27.5 > 23, reject H_O. Thus, lives of two types of batteries is not equal.

7. Data about daily rates of returns and their ranks among the three companies

Date	*BSE*	*ICICI Bank*	*Reliance Industries*	*L & T*
6/3/2006	—	—	—	—
7/3/2006	–0.093	–2.047	–0.007	3.331
8/3/2006	–2.014	–1.682	–1.694	–2.181
9/3/2006	0.619	1.897	1.008	0.121
10/3/2006	1.806	1.853	0.75	2.878
13/3/2006	0.362	–1.599	0.027	0.939
16/3/2006	0.685	0.73	4.936	–1.139
17/3/2006	–0.165	–0.37	0.826	–1.619
20/3/2006	0.746	0.025	0.174	–0.215
21/3/2006	–0.329	–1.255	0.528	0.187

Assigning rank 1 to lowest rate of return (–2.181), and 27 to highest rate (4.936)

Sum of ranks for ICICI = 105
Sum of ranks for Reliance Industries = 147
Sum of ranks for L&T = 126
Calculated Value of '*H*' = 8.14 > 5.99 (Tabulated Value of χ^2 at $\alpha = 0.05$)

Therefore, reject the null hypothesis that the daily rates of returns for the three companies are equal.

8. Since tabulated value of 'F' at 27,9 d.f. (2.88) is less than calculated value of 'F'(4.09), therefore, reject equality of ranks of companies on the three parameters.

9. From the table of net differences in rank sums of the pairs of four indices, it is observed that none of the differences is significant i.e. greater than the critical value for the difference in 'Rank Sums' for number of indices as 4, number of observations for each index = 11, and level of significance = 5 %. Thus, there is no significant difference in the returns on BSE 30, BSE 100, BSE 200 and BSE 500.

10. Since the calculated value of 'F'(2.9) < 7.815 (Tabulated value of χ^2 at 5 % level of significance and, 3 d.f.), there is no significant difference among daily rates of return on the four BSE indices.

Answers

ANSWERS TO OBJECTIVE TYPE QUESTIONS

Chapter No.	*Answers*
8	1 (a), 2 (d), 3 (d), 4 (d), 5 (d)
9	1 (c), 2 (c), 3 (c), 4 (b), 5 (c), 6 (b), 7 (a), 8 (b), 9 (c), 10 (a), 11(d), 12 (d), 13 (d), 14 (b), 15 (d)
10	1 (b), 2 (a), 3 (c), 4(b), 5 (a), 6 (b), 7 (a), 8 (d), 9(b), 10 (d), 11(a), 12 (a), 13 (c), 14 (c), 15 (d), 16 (a), 17 (b), 18 (c), 19 (a), 20 (c)
11	1 (d), 2 (c), 3 (c), 4 (a), 5 (b), 6 (d), 7 (b), 8 (b), 9 (c), 10 (b)
12	No Questions
13	1 (c), 2 (b), 3 (d), 4 (b), 5 (c), 6 (a), 7 (b), 8 (c), 9 (a), 10 (b)

Index