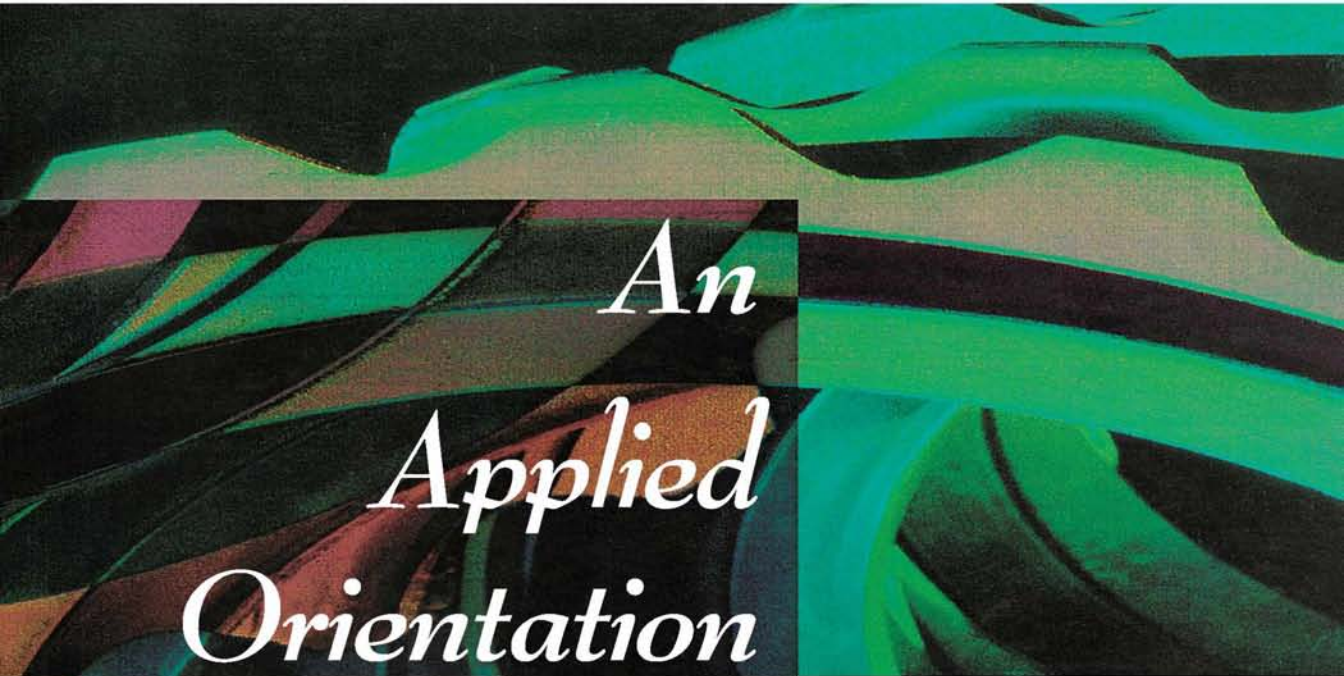


BUSINESS STATISTICS



*An
Applied
Orientation*

P. K. VISWANATHAN

Business Statistics

An Applied Orientation

This page is intentionally left blank.

Business Statistics

An Applied Orientation

P. K. Viswanathan

Visiting Professor

Institute for Financial Management and Research (IFMR)

and

Academy for Management Excellence (ACME)

Chennai, India

PEARSON

Copyright © 2007 Dorling Kindersley (India) Pvt. Ltd

Licenseses of Pearson Education in South Asia

No part of this eBook may be used or reproduced in any manner whatsoever without the publisher's prior written consent.

This eBook may or may not include all assets that were part of the print version. The publisher reserves the right to remove any material present in this eBook at any time.

ISBN 9788131704981

eISBN 9789332506145

Head Office: A-8(A), Sector 62, Knowledge Boulevard, 7th Floor, NOIDA 201 309, India
Registered Office: 11 Local Shopping Centre, Panchsheel Park, New Delhi 110 017, India

to my inspiring Guru

Dr. K. Subramaniam

a teacher par excellence in statistics

This page is intentionally left blank.

About the Author

P. K. Viswanathan is currently a Visiting Professor, Institute for Financial Management and Research (IFMR) and the Academy for Management Excellence (ACME), Chennai, India. He teaches Business Statistics, Operations Research, and TQM for the management students and participates as Faculty in the Executive Development Programs conducted by ACME/IFMR. His areas of expertise also include Marketing Research, Quality Management Tools, SQC, Spreadsheet Modeling, and Customer Satisfaction.

Apart from executing key corporate consultancy assignments, he has designed and conducted training programs for many leading organizations in India.

Mr. Viswanathan has a brilliant academic record-First Rank in MSc Statistics from the University of Madras, Rank holder MBA (FMS, Delhi), and MS degree from the University of Manitoba, Canada on a Fellowship.

His current area of interest is on Web-Based Training Programs suitable to business executives and management students.

In his industrial tenure of more than 15 years, he held senior management positions in Ballarpur Industries (BILT) of the Thapar Group and the J.K. Industries of the J.K. Organization. Mr. Viswanathan is a Visiting Faculty to reputed Management Institutes that include BIM (Tiruchirapalli), TAPMI (Manipal), and Ohio-Manipal School of Business (Bangalore).

As a prolific writer, Mr. Viswanathan has contributed research articles to reputed journals that comprise the article "An Appreciation of the Role of Statistical Hypothesis in Decision Making" hosted in **Hyperstat**, an online site of Rice University under the auspices of Professor David Lane. His article titled "A Step by Step Approach to Marketing Research" that has appeared in **Apeejay Business Review** (July-December 2002) has been widely appreciated. As a course writer, he developed a number of units for the MBA program of Indira Gandhi National Open University (IGNOU) in Research Methodology, Marketing Research and Operations Research.

This page is intentionally left blank.

Foreword

It is a curious fact that despite India's unique role in the history of numbers, the ways of mathematics are still viewed by the typical student of this country with considerable anxiety and even some trepidation or worse, distaste. Generations of school-going children have been attuned to being classified by their peers and family as either math wizards (who aim at the perennial challenge of centum!) or those in the majority who have to struggle against the familiar grain to gain a mastery over this language. This is a great pity, since mathematics is after all an aesthetically pleasing art if viewed in the proper perspective as a delightful new language; and at the same time, a science that propagates the habit of unambiguous thinking that underlies all others. It is in this sense a peculiar amalgam of what CP Snow called the "two cultures".

There is no doubt that but for the natural flair amongst a substantial number of Indians for quantitative methods, we could never have made the impact we have in the wider world of technology particularly software. A less appreciated cousin of mathematics is *statistics*, which one could define as numbers manipulated with a purpose, and therefore an intensely functional, applied, science, as distinct from mathematics and physics where the thinkers at the frontiers for example deal with a form of pure theory almost for its own sake.

Statistics is a modern tool of immense versatility, with which you can both explore and measure the variability and uncertainty inherent in reality in almost any field you can think of. Oil exploration, medical diagnosis, market research, genetics, weather forecasting, analytical finance, and gambling are hardly naturally allied subjects that would be associated in one's mind together, yet none of them would be what they are but for the recent advances in statistics. Some of the seminal concepts of the science such as the "normal" distribution", "probability", "sampling", and the "law of large numbers" have entered into the modern day language to such an extent that we use them unknowingly and imprecisely.

All of us, students and professionals alike, need an intelligent layman's introduction to the subject that is sufficiently precise and yet unthreatening. This is a very difficult challenge that this book addresses – and succeeds. I have known Mr. P.K. Viswanathan, the author of this book as a professor, practitioner, and a trainer for a fairly long while. There are two qualities that mark him as a teacher. The first is a passionate persistence; in fact delight, in teaching even the less able student. The second is a rare combination of great clarity in thinking along with mastery of the subject. Experts on the subject there aplenty, but few are equally effective communicators. The author is clearly among the few who can combine the two – and the book demonstrates it.

Apart from teaching executives and students for many years, the author has had the benefit of testing his ideas as a manager and practitioner. His book is full of evidence of this sensitivity, in the way the chapters are structured, the real life examples, problems, cases, and an interactive style of writing. He has taken pains to produce a number of figures and graphs, which not merely add to the liveliness of the text but create a classroom atmosphere that the student I am sure would find attractive.

Each chapter sets out clearly what its learning objectives are. It is followed by an exposition of concepts, illustrations, and applications from business life. Most importantly, the chapter concludes with a set of self-assessment questions. Thus the structure would be ideal even for a text to accompany a self-paced learning course. The idea of including computer (Excel) based exercises is excellent. With detailed click-by-click guidance, the author walks the student through the material, thus adding to the reader-friendliness of the book. This will also make today's younger generation feel much more at home. The poets and literature students amongst you would be delighted to know that you can approach the book safely, it will not bite you – since no prior mathematical knowledge is presumed beyond that of high school. It is amazing how far one can get with that level of understanding. One will also realize how many people have needlessly made statistics and indeed numerical thinking into a bugbear frightening them out of their commonsense!

Teachers and old fashioned professionals who look first for conceptual clarity and prefer deep foundations to technical wizardry would also feel at home in this book. The coverage is both deep and fairly wide. It goes all the way from simple average to regression, analysis of variance, hypothesis testing, decision analysis, and forecasting. Everywhere the emphasis is on correct problem formulation leaving the number crunching to the PC. It provides everything by way of quantitative analysis that any professional (doctor, accountant, engineer or manager) is likely to need on a day-to-day basis.

I heartily commend this book to what I hope would be an international audience.

S. Ramachander
Director, ACME/IFMR
Chennai, India

Preface

Based on my distilled experience of teaching *Business Statistics* in the last twenty five years both to the students of management and to the corporate executives, I felt a need for a compact and manageable size text book that would help the audience achieve the following objectives :

- Understand and appreciate the most widely used tools of business statistics which form the basis for rational and sound business decisions.
- Focus on problem recognition and test hypothesis/model in the context of managerial decision-making.
- Develop skills in analysis and interpretation of data.
- Handle challenging problems using appropriate analysis tools.

The need identified by me was further reinforced by my students and a large number of business executives who gave me a gentle command to write a text book on business statistics with an applied orientation that would be simple to understand. Thus this book is an outcome of the motivation and encouragement extended by the audience with whom I have been continuously interacting.

This book is designed for the following principal groups :

- MBA students who would like to have a conceptual framework of business statistics as well as to develop skills in applying concepts into decision situations.
- All post graduate students who would like to understand the nitty-gritty of business statistics.
- Corporate executives who would like to comprehend concepts and then use statistical analysis in their functional areas.
- All other professionals who would like to acquire basic knowledge of business statistics that would help them analyze and interpret data.

It may be mentioned here that some of the chapters such as " Decision Analysis", "Forecasting", and "Correlation and Regression" can be used effectively as standalone material for the executive development programs involving a duration of one/two days.

PREREQUISITES FOR THE READERS

- The audience are expected to have knowledge of basic mathematics at the plus two level.
- Statistical thinking is the driving force for this text.
- Knowledge and ability to effectively use Microsoft Excel are desirable.

ABOUT THE BOOK

- The style used is interactive, conversational, direct, and simple. It is designed in such a way that any student after attending the class can treat the book as a teacher in the physical absence of the instructor.
- A number of pictures are interwoven with concepts throughout the book so as to make the reader understand the nuances of some complex problems apart from learning the subject with ease.
- The book is contemporary and uses extensively *Microsoft Excel* to solve a variety of problems. In particular, the learning process is greatly facilitated by the usage of “*Paste Function*” and “*Data Analysis*” of Excel.
- For those who may not always have access to Personal Computer though the probability of this happening is very remote, formula approach is comprehensively delineated for solving problems. Even here, the emphasis is more on electronic *Spreadsheet* software of Excel than on using a calculator. Of course, the student can use the calculator and solve the problems.
- Apart from incorporating the standard pedagogical features of a textbook, every chapter contains a *Discussion Topic* that gives ample scope for the students to debate key issues of the topic that will pave the way for grasping the conceptual philosophy and spirit of the subject. Even though, *Internet* is not exclusively mentioned, knowledgebase available in the Internet can be a great source of help to the students for effectively dealing with the discussion topic.
- Almost all chapters have *Progressive Test Questions* that act as a monitoring and feedback mechanism in the learning process of the subject matter.
- Answers with explanation have been provided to all review questions in every chapter. Also carefully structured *Practice Problems* have been given at the end of each chapter for the students to work out the solution.

ACKNOWLEDGEMENTS

Many persons have been extremely helpful in writing this textbook. First, I would like to acknowledge Dr. K. Subramaniam who was my Professor in the University of Manitoba, Canada for teaching me Statistics in a way I will never forget. He has inspired me tremendously on *Statistical Thinking*.

I would like to acknowledge with reverence, Professor S. Ramachandar, Director IFMR/ACME for his exceptional encouragement and feedback on various aspects concerning the book. In fact, he has graciously agreed to write a Foreword for my book. I am extremely grateful to him for this kind gesture.

I would like to acknowledge Professor Xavier, Dean of IFMR/ACME for his constant support and confidence building measures that gave me the necessary impetus to write the book.

I would like to acknowledge Professor G. Balasubramanian of IFMR/ACME for his valuable feedback on individual chapters as well as on the effective usage of Microsoft Excel.

I would like to place on record my sincere appreciation and gratitude to Dr. T.V. Subramanian, Management Consultant, Chennai, and one of the luminaries in the field of Statistics and Operations Research for going through the chapters minutely. His valuable insights helped me in fine-tuning the get up of the book.

I immensely benefited by my interactions with the two Distinguished Professors-Dr. N. Jayasankaran, Director, BIM (Trichy) and Dr. R. Rajagopalan, Dean Academic Affairs, TAPMI (Manipal), while teaching Statistics in their campuses. I take this opportunity to acknowledge with reverence, their insightfulness into the realm of statistics that helped me write this book with rigor and clarity.

My students at IFMR/ACME have been a source of tremendous strength. Their penetrating questions in the class in all the lecture topics acted as a foundation stone upon which this textbook is built. I would like to wholeheartedly thank them all.

I would like to express my special thanks to Ms. Divya, the first year PGDM student of IFMR for meticulously going through the final proof pages and making corrections.

I would like to acknowledge my management students of BIM, Trichy for their high quality interaction in the classroom in all my visits. This strengthened the conceptual foundation of the subject significantly.

I profusely express my sincere thanks to all the corporate executives with whom I have interacted for many years as a faculty both in in-house and public programs. Their enlightening views helped me write this book with an applied orientation.

I greatly acknowledge Mr. Sanjay K. Singh, Managing Editor, Pearson Education (Singapore) Pte Ltd, and his team for the alacrity with which the editing process of the book has been executed. Mr. Sanjay's follow up action and feedback have been exemplary. My special thanks are due to Mr. G. Loganathan and Kumar Sidharth of Pearson Education for their dedication, involvement, and commitment all through the time horizon of writing the book.

Finally, I would like to express my affection and heartfelt gratitude to my wife Kannamma who stood by me throughout this book project and ensured the completion of all chapters on time by meticulous follow up and gentle reminders.

This page is intentionally left blank.

Contents

<i>About the Author</i>	vii	
<i>Foreword</i>	ix	
<i>Preface</i>	xi	
Chapter 1	An Overview of Statistics	1
	Learning Objectives	1
	Introduction	1
	Chapter Outline	1
	1.1 Why Should I Study Statistics?	2
	1.2 What is Statistics?	2
	1.3 Some Typical Application Areas	3
	1.4 Types of Statistics	4
	1.5 Some Key Terms and Definitions	5
	1.6 Types of Data	6
	1.7 Data Measurement Scales	6
	1.8 Sources of Data	7
	1.9 Step-by-Step Approach to Statistical Investigation	7
	1.10 Chapter Summary	15
	<i>Glossary</i>	15
	<i>Review Questions</i>	16
	<i>Case Study-Savvy Fast Food</i>	17
	<i>Answers to Review Questions</i>	19
Chapter 2	Classifying Data to Convey Meaning	21
	Learning Objectives	21
	Introduction	21
	Chapter Outline	21
	2.1 Meaning and Examples of Raw Data	22
	2.2 Frequency Distribution	23
	2.3 Histogram	24
	2.4 Cumulative Frequency Distribution and Ogive Curve	35
	2.5 Chapter Summary	37
	<i>Glossary</i>	37
	<i>Review Questions</i>	38
	<i>Answers to Review Questions</i>	39
	<i>Practice Problems</i>	40

Chapter 3	Measures of Central Tendency and Dispersion	43
	Learning Objectives	43
	Introduction	43
	Chapter Outline	43
	3.1 Measures of Central Tendency	44
	3.2 Measures of Dispersion (Variation)	50
	3.3 Chapter Summary	57
	<i>Glossary</i>	57
	<i>Review Questions</i>	58
	<i>Answers to Review Questions</i>	59
	<i>Practice Problems</i>	62
Chapter 4	Probability—A Conceptual Framework	65
	Learning Objectives	65
	Introduction	65
	Chapter Outline	65
	4.1 Meaning and Concepts of Probability	66
	4.2 Types of Probability	68
	4.3 Mutually Exclusive Events	70
	4.4 Independent Events	70
	4.5 Rules for Calculating Probability	71
	4.6 Use of Probability Tree	76
	4.7 Chapter Summary	79
	<i>Glossary</i>	79
	<i>Review Questions</i>	80
	<i>Answers to Review Questions</i>	81
	<i>Practice Problems</i>	82
Chapter 5	Probability Distributions	84
	Learning Objectives	84
	Introduction	84
	Chapter Outline	84
	5.1 What is a Probability Distribution?	85
	5.2 The Binomial Distribution	88
	5.3 The Poisson Distribution	93
	5.4 The Normal Distribution	97
	5.5 Chapter Summary	107
	<i>Glossary</i>	108

	<i>Review Questions</i>	109
	<i>Answers to Review Questions</i>	110
	<i>Practice Problems</i>	111
Chapter 6	Basics of Sampling and Sampling Distribution	113
	Learning Objectives	113
	Introduction	113
	Chapter Outline	113
	6.1 What is Sampling and Why Do You Need Sampling?	114
	6.2 Types of Sampling	115
	6.3 Sampling Distribution -A Conceptual Framework	122
	6.4 The Concept of Standard Error	123
	6.5 Sampling Distribution of the Mean from Normal population	124
	6.6 Sampling Distribution of the Mean - Non-Normal Population	128
	6.7 Chapter Summary	129
	<i>Glossary</i>	130
	<i>Review Questions</i>	130
	<i>Answers to Review Questions</i>	131
	<i>Practice Problems</i>	136
Chapter 7	Estimation	137
	Learning Objectives	137
	Introduction	137
	Chapter Outline	137
	7.1 Point Estimation	138
	7.2 Interval Estimation	140
	7.3 Confidence Interval for Population Mean and Proportion- Large Sample	141
	7.4 Confidence Interval for Population Mean - Small Sample (<i>t</i> -Distribution)	146
	7.5 How to Determine Sample Size Using Confidence Interval	149
	7.6 Chapter Summary	151
	<i>Glossary</i>	151
	<i>Review Questions</i>	152
	<i>Answers to Review Questions</i>	153
	<i>Practice Problems</i>	155
Chapter 8	Hypothesis Testing	156
	Learning Objectives	156
	Introduction	156

	Chapter Outline	156
	8.1 Statistical Hypothesis-A Conceptual Framework	157
	8.2 Hypothesis Testing -Univariate Case (One Sample)	160
	8.3 Hypothesis Testing -Bivariate Case (Two Sample)	171
	8.4 Chapter Summary	184
	<i>Glossary</i>	185
	<i>Review Questions</i>	186
	<i>Answers to Review Questions</i>	187
	<i>Practice Problems</i>	188
Chapter 9	Chi-Square Test and Analysis of Variance (ANOVA)	191
	Learning Objectives	191
	Introduction	191
	Chapter Outline	191
	9.1 Chi-Square (χ^2) Analysis-Basics	192
	9.2 Chi-Square Test-Goodness of Fit	192
	9.3 Chi-Square Test of Independence	195
	9.4 ANOVA-Basics	199
	9.5 ANOVA-One-Way Classification	200
	9.6 ANOVA-Two -Way Classification	207
	9.7 Chapter Summary	211
	<i>Glossary</i>	212
	<i>Review Questions</i>	212
	<i>Answers to Review Questions</i>	213
	<i>Practice Problems</i>	215
Chapter 10	Correlation and Regression	217
	Learning Objectives	217
	Introduction	217
	Chapter Outline	217
	10.1 What is Correlation?	218
	10.2 Insights into Correlation	218
	10.3 Basics of Regression	225
	10.4 Regression Model	226
	10.5 Chapter Summary	239
	<i>Glossary</i>	239
	<i>Review Questions</i>	240
	<i>Answers to Review Questions</i>	241
	<i>Practice Problems</i>	242

Chapter 11	Decision Analysis	245
	Learning Objectives	245
	Introduction	245
	Chapter Outline.....	245
	11.1 Steps in Systematic Problem Solving	246
	11.2 How to Structure a Decision Problem.....	249
	11.3 Expected Monetary Value (EMV).....	251
	11.4 Decision Tree	254
	11.5 Value of Sample Information	258
	11.6 Chapter Summary	261
	<i>Glossary</i>	262
	<i>Review Questions</i>	263
	<i>Answers to Review Questions</i>	263
	<i>Practice Problems</i>	265
Chapter 12	Forecasting	268
	Learning Objectives	268
	Introduction	268
	Chapter Outline.....	268
	12.1 Forecasting-Basics	269
	12.2 Qualitative Methods of Forecasting.....	270
	12.3 Quantitative Methods of Forecasting	273
	12.4 Chapter Summary	289
	<i>Glossary</i>	289
	<i>Review Questions</i>	291
	<i>Answers to Review Questions</i>	291
	<i>Practice Problems</i>	291
	<i>References</i>	295
<i>Appendix A</i>	<i>Test Your Knowledge on Business Statistics</i>	296
<i>Appendix B</i>	<i>Binomial Probability Table</i>	301
<i>Appendix C</i>	<i>Poisson Probability Table</i>	307
<i>Appendix D</i>	<i>Normal Distribution Table</i>	312
<i>Appendix E</i>	<i>t Distribution Table</i>	314
<i>Appendix F</i>	<i>Chi-Square Distribution Table</i>	316
<i>Appendix G</i>	<i>F Distribution Table</i>	319

This page is intentionally left blank.

An Overview of Statistics

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Define and appreciate the role of statistics
- Explain descriptive & inferential statistics
- Describe data types
- Describe the various data measurements
- Briefly describe the data sources
- Develop a practical approach towards statistical investigation

CHAPTER OUTLINE

- 1.1 Why should I study Statistics?
 - 1.2 What is Statistics?
 - 1.3 Some Typical Application Areas
 - 1.4 Types of Statistics
 - 1.5 Some Key Terms and Definitions
 - 1.6 Types of Data
 - 1.7 Data Measurement Scales
 - 1.8 Sources of Data
 - 1.9 Step-By-Step Approach to Statistical Investigation
 - 1.10 Chapter Summary
- Glossary
Review Questions
Case Study–Savvy Fast Food
Answers to Review Questions

INTRODUCTION

Managers make sound decisions when they use all relevant information in an effective and meaningful manner. The principal purpose of statistics is to provide decision-makers with a set of techniques for collecting, analyzing, and interpreting data into actionable recommendations. Statistical methods are widely used to aid decision-makers in all functional areas of management. This chapter provides the basic ideas and concepts at a general level.

2 Business Statistics

1.1 WHY SHOULD I STUDY STATISTICS?

Whether you are a student of management or a company executive, you may wonder, "why should I study statistics?" Good Question. Ponder the following decision situations:

Situation 1

A company has to decide whether to introduce a new product into the market or not. The company will introduce the product into the market if 30% of the target audience in the relevant population will accept the product so that the risk of product failure is minimized. Obviously consumer acceptance is paramount in making this decision. To know about the consumer acceptance in a reasonable manner, the company has done a "test marketing" exercise. In the test market, 30% of the sample target audience (based on a sample of 150 consumers) indicate their acceptance of the product. Does the sample result at 95% confidence level suggest that 30% of the target audience in the population (entire market) will accept the product?

Situation 2

A bank which has been steadily losing customers in the light of intense competition wants to investigate the reasons for the loss of customers on account of perceived service quality in critical dimensions like response time, reliability, courtesy of the service staff, and credibility. The bank would like to conduct a comprehensive survey to measure the perceived service quality from the customers' angle on these dimensions with that of competition. This would help the bank develop and implement effective strategies to woo its present customers back as well as to attract new customers.

Can you make the right decision in situation 1 and situation 2 with minimum risk without the help of statistics? The answer is clearly a "No". Information based decision making using statistical analysis is absolutely essential in the present environment characterized by intense competition, onslaught of new products and services, globalization, and revolution of information technology.

1.2 WHAT IS STATISTICS?

By "Statistics" we mean methods specially adapted to the collection, classification, *analysis, and interpretation of data* for making effective decisions in all functional areas of management.

If you carefully go through this definition, you will notice that mere collection and classification of data will not be sufficient. Ponder carefully the expression in italic *-analysis, and interpretation of data* in the definition above. This is crucial for improving your organizational effectiveness and profitability. You will appreciate the profundity of this expression by the following real life example.

Analysis, and Interpretation of Data-Example

American Express Company (AMEX), the pioneer in personal charge cards, during the eighties used to systematically collect customer feedback data from the marketplace on a continuous basis. AMEX is well known for its caring attitude towards customers.

The Analysis and Interpretation of the customer data revealed that the customers wanted the new card to be processed within three weeks where as, AMEX was taking around 5 weeks. AMEX decided to issue new cards within two weeks. Similarly another analysis revealed that the customers wanted the stolen/lost cards to be replaced within two days where as AMEX was taking two or more weeks to issue replacement cards. AMEX decided to replace the lost cards within two days. As a result of these two decisions, AMEX could generate \$1.4 million additional profit per year.

1.3 SOME TYPICAL APPLICATION AREAS

Quality Management

Statistical quality control (SQC techniques) If you are working in a manufacturing organization, SQC techniques help you reduce and stabilize process variation that could cut down your cost significantly. SQC also helps in separating the assignable causes from the chance causes.

Process capability Process Capability enables you to find out whether your company is capable of meeting *Customer Specification* ("Voice of the Customer"). If it is lagging behind the customer requirements, then your management has to take decisions such as modernizing the plant, changing the technology, and buying new equipment so that it achieves process capability. Statistics plays a pivotal role in this area.

Finance

Financial ratio analysis Using the *Balance Sheet* and *Income Statement* data, key financial ratios can be computed to throw light on the financial health of your company. Comparative statistical analysis of the ratios will provide answers to your typical questions that include: How has your company been performing in the last five years? What is the emerging trend? How is your company faring in relation to competition?

Cash forecasting If you are working as a financial analyst in a company, you should forecast cash requirements to meet short-term obligations such as payments to suppliers, payment of salary to the staff, and other working capital needs. Statistical analysis will be helpful to you for arriving at realistic cash forecasting.

4 Business Statistics

Materials Management

Inventory level You can realize tremendous savings that impact directly your profit, if you keep optimum inventory levels for raw materials and finished products. This can be achieved by using appropriate statistical techniques.

Quality assessment for incoming and outgoing items When inputs such as raw material and other items arrive into your plant, they have to be carefully assessed by you for quality standards so that your organization can meet the customer expectations at the minimum cost. Likewise, inspection of finished products before shipment to customers on sample basis helps you in reducing complaints later. *Statistical Sampling Plan* ensures acceptable quality of incoming and outgoing items.

Marketing

Marketing research Marketing Research provides solutions to your problems if you are involved in marketing goods, services, or ideas. Systematic gathering, recording, analyzing, and interpreting marketing data become paramount in this context. Statistical methods play the role of decision support when you as a marketing manager face key decision questions that include: Should I introduce the new product? What type of test marketing is needed? What will be the consumer reaction to price hike? Is the advertisement campaign effective? What is the present level of customer satisfaction my company enjoys in the marketplace?

Demand projections In a highly competitive domestic as well as international market, you need short-term and long-term demand projections for your products and services. Statistical survey methods and statistical analysis of the past data pertaining to the industry and your organization will provide you with a reliable demand and sales projections.

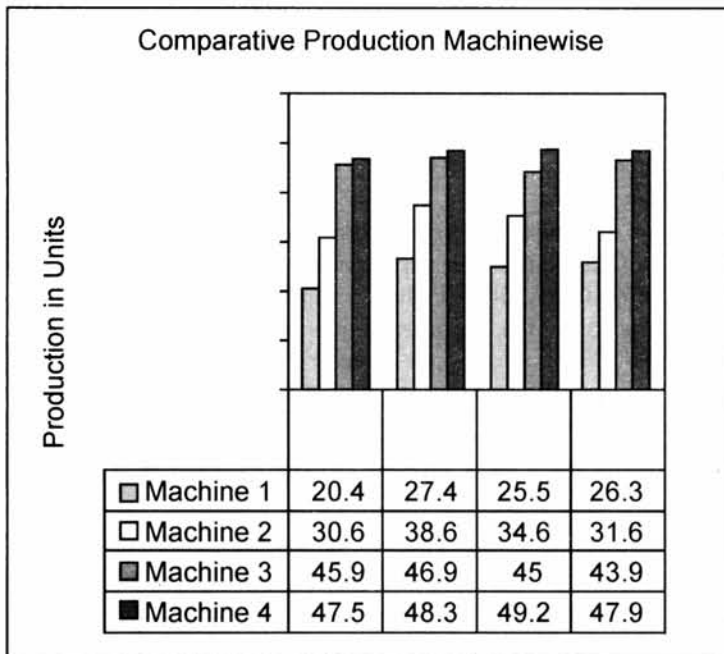
1.4 TYPES OF STATISTICS

Descriptive Statistics is concerned with Data Summarization, Graphs/Charts, and Tables. It processes raw data into information. You know information is key to decision making.

Inferential Statistics is a method used to talk about a Population Parameter from a Sample. It involves Point Estimation, Interval Estimation, and Hypothesis Testing.

Example for Descriptive Statistics

The Quality Control Department of a large manufacturing company would like to compute summary measures such as the mean production per shift for a particular item. The department would also like to get a comparative picture of performance of the mean production of the four machines in the plant by tabulation. Further the company might like to graph the comparative performance of the four machines in the four quarters using a bar chart.



Example for Inferential Statistics

Suppose you, as a marketing manager would like to identify a niche market for your product. You know from your experience that an accurate assessment of the income of a typical family is crucial. The average income of this typical family in the population is estimated by you to be Rs. 320000 based on figures obtained from a sample. In this example, average income based on sample is a *Point Estimate* of the population. The average income that falls within a statistically formed interval of 320000 plus or minus 40000 is called an *Interval Estimate*. The statement that "the average income in the population is more than Rs. 300000 per year" is a *Hypothesis*.

Caution: Inferential Statistics assumes that the sampling methodology is random (i.e. based on probability sampling)!

Discussion Topic

Analyze, criticize, and explain the following statement:

"As a manager, the interval estimate may be much more important to you than the rest (point estimate and hypothesis testing)!"

1.5 SOME KEY TERMS AND DEFINITIONS

Population (Universe)

Population is the collection of all possible observations of a specified characteristic of interest. All students taking the statistics course in a business school is an example of population.

6 Business Statistics

Sample

Sample is a subset of the population. Suppose you want to select a team of 20 students from 200 students in an MBA program for participating in a management quiz. The total number of students 200 is the population. 20 students selected for the quiz is the sample.

Variable

A *variable* is an item of interest that can take on many different numerical values. An example is the number of defective items produced in a factory. If a variable takes different values with associated probability, it is called a *random variable*. For example, the number of times head turns up in a toss of a coin ten times is a random variable.

Parameter

A Population characteristic of interest is called a *parameter*. For example, you want to have some idea about the income level of a particular class of people. The average income of this entire class of people is called a parameter.

Statistic

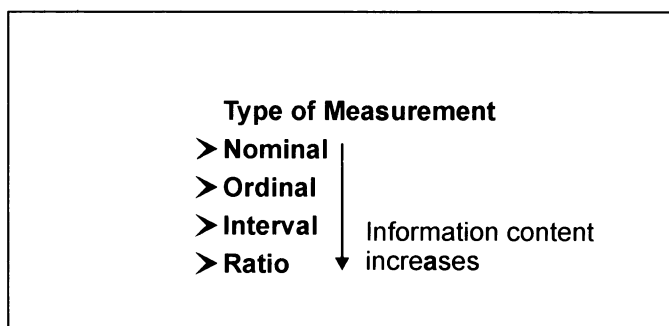
A *statistic* is a type of average that is based on a sample. It is used to make inferences about the population parameter. If you look at the previous example, the average income in the population can be estimated by the average income based on the sample. This sample average is called a statistic.

1.6 TYPES OF DATA

Qualitative Data are nonnumeric in nature and can't be measured. Examples are gender, religion, and place of birth.

Quantitative Data are numerical in nature and can be measured. Examples are balance in your savings bank account, and number of members in your family. Quantitative data can be classified into *discrete type* or *continuous type*. Discrete type can take only certain specific values that are whole numbers such as the number of rooms in a hotel. Discrete numbers cannot take fractional values. Continuous type can take any value within a specific interval such as the production quantity of a particular variety of paper measured in kilograms.

1.7 DATA MEASUREMENT SCALES



Nominal Data They are the weakest of all data measurements. Numbers are used to label an item or characteristic. Categorization is the main purpose of this measurement.

Example A business school may designate subject specialization by numbers such as MBA in Finance = 401, and MBA in Systems = 402.

Various brands of toothpaste, savings bank account numbers are other examples of nominal data. Nominal data are also called categorical data. Note that nominal data cannot be manipulated in a numerical fashion. Thus, they are not amenable to arithmetic operations.

Ordinal or Rank Data Numbers are used to rank objects or attributes. Consider the customer preference for your brand. The least preference is rated at 1, an average preference is rated at 3, and a strong preference is rated at 5 on a scale that ranges from 1 to 5. This is an example of ordinal scale. Simple arithmetic operations are not possible for ordinal data. Ordinal data can also be verbalized on a continuum like excellent, good, fair and poor. In ordinal data, distance between objects or ranks cannot be measured.

Interval Data If you have data with ordinal properties and can also measure the distance between objects, you have an interval measurement. Interval data are superior to ordinal data because, with them, decision makers can measure the distance between two observations, i.e., distances between objects can be measured. For example, frozen-food distributors are concerned with temperature, which is an interval measurement. Interval data have an arbitrary zero point. Basic arithmetic operations are possible with interval data.

Ratio Data It is the highest level of measurement that has the requisite desirable properties and allows you to perform all basic arithmetic operations, including division and multiplication. Data measured on a ratio scale have a fixed zero point. Examples include business data such as cost, revenue, market share, and profit.

1.8 SOURCES OF DATA

Primary Data are collected by the organization itself for a particular purpose of its own. The benefits of primary data are that they fit the needs exactly, and are up to date and reliable. Primary data are generally collected through a statistical sample survey using a questionnaire.

Secondary Data are collected by other organizations or for other purposes. Any data, which are not collected by the organization for the specified purpose, are secondary data. These may be published by other organizations, available from research studies, published by the government, and so on. Secondary data have the advantages of being much cheaper and faster to collect. They also have the benefit of using sources, which are not generally available. For example, companies will respond to a survey by the Government of India, Confederation of Indian Industry, and Central Statistical Organization but they would not answer questions from another company.

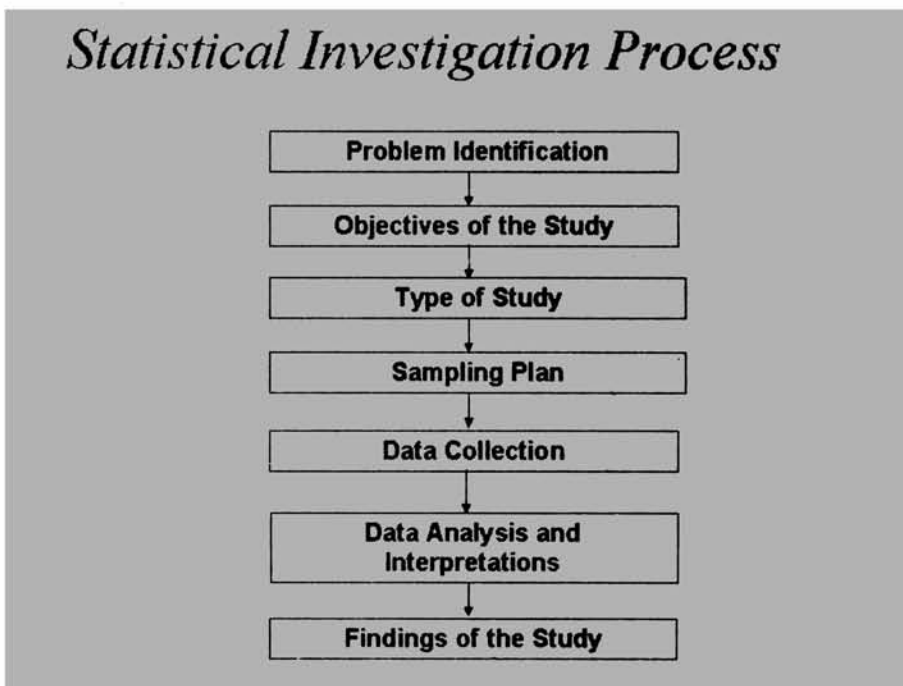
1.9 STEP-BY-STEP APPROACH TO STATISTICAL INVESTIGATION

Having discussed so far the various elements that constitute statistics, it is time for you to appreciate how statistical investigation works in practice.

Statistical investigation is not a fishing expedition or an encyclopedic gathering of assorted facts. It is purposeful and lays the structure for decision-making. It provides answers to various management problems that the decision makers face. It gives a set of actionable recommendations that will help organizations develop and implement superior strategies.

The sequential steps of the statistical investigation process (See visual below) are delineated through an actual case analysis. By no means it is an exhaustive coverage but it would give you a practical approach towards doing statistical investigation. The idea is to motivate you to learn the nitty-gritty of conducting statistical surveys.

Although the findings of this study cannot be generalized at the All India level in view of the fact that it is restricted to BHEL Township in Trichy, it brings out a number of useful hypotheses that could be tested at a macro level through a large-scale survey.



Consumers of Detergent Market and their Attitudes towards Buying in BHEL Township, Trichy - A Case Study

1. Problem Identification

The marketplace is crowded with a number of brands of washing powder and is characterized by intense competition. It is indeed intriguing to study the various aspects of consumer behavior in the context of purchasing a particular brand of detergent. The

idea is to get a clear picture of the varied responses the consumers would give to the different marketing efforts of a player in the market. In particular, we would like to know the profile of the consumers who are buying detergents and their attitudes towards purchase.

2. Objectives of the Study

1. To understand the relationship between income and choice of brand
2. To study the relationship between choice of brand and mode of washing
3. To assess customer loyalty
4. To understand the relative importance of various product attributes perceived by consumers
5. To know who makes the purchase decision

3. Type of Study

This is a descriptive study that deals with determining the various characteristics of the detergent market in the context of consumer attitudes towards purchase. This study also throws light on the frequency with which a marketing phenomenon occurs, or is associated with other relevant variables. The findings of this investigation will help you grasp the buying behavior of consumers to a reasonable extent.

4. Sampling Plan

Sampling frame All past and present users of detergents with income \geq Rs. 5000 per month.

Units of analysis The households in BHEL Township in Trichy.

Sampling method used A representative stratified random sampling procedure was adopted using income as the basis of selection. The categories that were used in classifying income are: up to Rs. 5000 per month, Rs. 5001 to Rs. 12000 per month, Rs. 12001 to Rs. 20000, and above Rs. 20000 per month. A sample of 150 consumers was selected for this study.

5. Data Collection

Primary data collection method was adopted using a questionnaire instrument. Data were collected through personal interview by a group of MBA students of BIM, Trichy who were doing a project under the guidance of the author during 2001.

The questionnaire used was structured and undisguised and just consisted of 7 important questions apart from the classification section. It is given in *Exhibit 1* for a meaningful appreciation of this research study.

6. Data Analysis and Interpretations

(a) What can we say about income versus choice of brand?

Survey Results

	Brand							
	ARIEL		SURF		WHEEL		HENKO	
Income Per Month	Numbers	Percentage	Numbers	Percentage	Numbers	Percentage	Numbers	Percentage
≤ Rs. 5000	8	5.33	6	4.00	12	8.00	15	10.00
Rs. 5001-12000	7	4.67	9	6.00	8	5.33	6	4.00
Rs. 12001-20000	10	6.67	14	9.33	6	4.00	5	3.33
> Rs. 20000	18	12.00	16	10.67	5	3.33	5	3.33

Note Percentage in each cell is worked out to the total respondents (150).

Interpretation

It is interesting to note that consumers in the income slab of Rs.12001 and above prefer Premium Brands (Ariel and Surf) and this is about 39%. Consumers in the income slab of Rs 5000 to Rs.12000 prefer Economy Brands (Wheel and Henko) and this constitutes about 27%. Consumers in the highest income slab of more than Rs. 20000 clearly prefer Premium Brands (23%) and the Economy Brands preference in this category is only about 7%.

Also if we take only the highest income category of more than Rs. 20000, out of the 44 consumers, 34(77%) prefer Premium Brands (Ariel+Surf) and only 10(23%) prefer Economy Brands (Wheel+Henko). Likewise, if we take the lowest income category of up to Rs. 5000, out of the 41 consumers, 27(66%) prefer Economy Brands and 14(34%) prefer Premium Brands.

It seems that income strata impact the brand selection. Brand preference is strongly associated with income level.

(b) What can we say about choice of brand versus mode of washing?

Survey Results

	Brand							
	ARIEL		SURF		WHEEL		HENKO	
Mode of Washing	Numbers	Percentage	Numbers	Percentage	Numbers	Percentage	Numbers	Percentage
Self	18	12.00	12	8.00	26	17.33	9	6.00
Servant	6	4.00	6	4.00	6	4.00	6	4.00
Washing machine	24	16.00	22	14.67	6	4.00	9	6.00

Note Percentage in each cell is worked out to the total respondents (150).

Interpretation

Consumers having washing machine predominantly prefer Premium Brands (31%) and are less enthusiastic about Economy Brands (10%). Consumers doing self-wash use Premium Brands (20%) and Economy Brands (23%).

Also if we take into account the number of consumers using washing machines, out of the 61 consumers, 46(75%) use Premium Brands and only 15(25%) use Economy Brands. Likewise, if we see the self-wash category, out of the 65 consumers, 30(46%) use Premium Brands and 35(54%) use Economy Brands. Out of this 54% using Economy Brands, 40% prefer Wheel and only 14% prefer Henko.

It appears that mode of washing influences choice of brand.

(c) What can we say about customer loyalty?

Survey Results

MODE OF ACTION	PREMIUM BRANDS		ECONOMY BRANDS	
	Numbers	Percentage	Numbers	Percentage
Buy another Brand	21	14.00	15	10.00
Don't Buy any other Brand	67	44.67	47	31.33

Note Percentage in each cell is worked out to the total respondents (150).

Interpretation

Both in Premium and Economy Brands, consumers are showing a strong brand loyalty and are willing to wait till they get their favorite brands. Overall 76% of the consumers are sticking to their brands.

Another interesting observation is that out of 88 consumers of the Premium Brands, 67(76%) are willing to wait for their favorite brands and also out of 62 consumers of Economy Brands, 47(76%) are willing to wait for their favorite brands.

It points to the fact that proportion of customers waiting to buy their favorite brand when it is not available, is much more than proportion of customers switching to another brand.

(d) Which attributes are more important?

Survey Results

Attribute	Numbers	Percentage
Pleasant smell	14	9.33
Price	37	24.67
Lather	21	14.00
Dirt removing efficiency	58	38.67
Cloth quality after wash (Fabric care)	20	13.33
Total	150	100.00

Interpretation

The two most important attributes required for buying, according to consumers of washing powders are, dirt removing efficiency (39%) and price (25%). Lather and cloth quality after wash are in the 3rd and 4th position respectively.

It points to the fact that all attributes are not equally weighted by the consumers when purchasing washing powders.

(e) Are there significant differences in attribute ranking by consumers between Premium and Economy Segments?

Survey Results

<i>Premium Segment</i>		<i>Economy Segment</i>	
<i>Attribute</i>	<i>Sum of ranks of attributes</i>	<i>Attribute</i>	<i>Sum of ranks of attributes</i>
Pleasant smell	393	Pleasant smell	221
Price	249	Price	93
Lather	295	Lather	296
Dirt removing efficiency	161	Dirt removing efficiency	150
Cloth quality after wash	222	Cloth quality after wash	170
Number of Respondents = 88		Number of Respondents = 62	

Note As average rank will pose difficulty in interpretation, we have used sum of the ranks of attribute as the basis for interpretation. Here the attributes are ranked in order of importance by the respondents. Rank 1 for an attribute indicates that it is most important, rank 2 the next most important, and so on. An attribute that has the smallest sum of ranks is the best. Using the sum of ranks as the basis, the interpretation follows:

Interpretation

Under Premium Segment, *dirt-removing efficiency* is the most important attribute followed by *cloth quality after wash*. This seems to weigh more in the minds of the consumers while purchasing a particular detergent. *Price* is ranked only third.

It is also interesting to note that under Economy Segment, price is the most important attribute followed by *dirt removing efficiency*. *Cloth quality after wash* gets the third position.

The results suggest that the ranking of attributes within the premium and economy segments are not the same and there are differences in attribute ranking between Premium and Economy Brands. The expectations of these two groups are not identical.

(f) Who makes the purchase decision?

Survey Results

<i>Decision Maker</i>	<i>Numbers</i>	<i>Percentage</i>
Head of the house	33	22.00
Wife	92	61.33
Elders	19	12.67
Shopkeeper	6	4.00
Total	150	100.00

Interpretation

The results overwhelmingly point to the fact that wives are the main decision makers when it comes to buying washing powder. It constitutes about 61%. Shopkeepers' influence on the purchase decision is only 4%.

It looks pretty obvious that purchase decision of washing powder is predominantly made by wives.

7 Findings of the Study

➤ The impact of Income Levels on the Choice of the Brand is significant

The higher income group prefers Premium Brands like Ariel and Surf Excel while the lower income group prefers Economy Brands like Wheel, and Henko.

➤ The Choice of the Brand also depends on the Mode of Washing

Consumers using washing machines prefer Premium Brands while those doing self-wash prefer Economy Brands.

➤ Customer Loyalty to the Brand is very strong

Proportion of customers willing to wait for their preferred brands when they are not available, is more than the proportion of customers willing to switch to another brand.

➤ Dirt Removing Efficiency is the paramount attribute in choosing a detergent

Dirt removing efficiency is the most important attribute followed by price. Lather and cloth quality after wash are almost tied at third position.

➤ There are significant differences in ranking of attributes by consumers

*It is interesting to note that in the **Premium Segment** Dirt removing efficiency is the most important attribute followed by Cloth quality after wash and then the Price whereas, in the **Economy Segment** Price is the most important attribute followed by Dirt removing efficiency and then the Cloth quality after wash.*

➤ The purchase decision is predominantly made by housewives

The study points out overwhelmingly that wives are the most important people in purchase decision and the shopkeepers are the least important people in purchase decision.

Note All the findings above were statistically validated by standard test procedures that you will be learning in later chapters.

EXHIBIT 1

QUESTIONNAIRE

1. Which brand of washing powder do you use?

2. How long have you been using this powder?

14 *Business Statistics*

3. Which of the following attributes is the most important for you while buying a washing powder?
- (a) Pleasant smell
 - (b) Price
 - (c) Lather
 - (d) Dirt removing efficiency
 - (e) Cloth quality after wash
 - (f) Any other (please specify)

4. Rank the following attributes in the order of importance?

Attribute	Rank
Pleasant smell	
Price	
Lather	
Dirt removing efficiency	
Cloth quality after wash	

5. What is your mode of washing?

- (a) Self
- (b) Servant
- (c) Washing machine

6. If the washing powder of your choice is not available what will you do?

- (a) Try in another shop
- (b) Wait until your brand arrives
- (c) Try a different brand

7. Who makes the purchase decision in your house?

- (a) Head of the Family
- (b) Wife
- (c) Elders
- (d) Shopkeepers

Classification Section:

Name

Number of members in the family

Monthly household income (Rs per month)

- (a) ≤ 5000

- (b) 5001 to 12,000
- (c) 12,001 to 20,000
- (d) Above 20,000

1.10 CHAPTER SUMMARY

In this chapter, you have been exposed to the various facets of statistics. In precise terms, the coverage focused on:

- The Definition and Role of Statistics
- Typical Applications in Management Functions
- Types of Statistics split into Descriptive and Inferential
- Definition and Meaning of Key Terms - Population, Sample, Variable, Parameter, and Statistic
- Types of data split into Qualitative and Quantitative
- Four data measurements - Nominal, Ordinal, Interval, and Ratio
- Data sources categorized into Primary and Secondary
- How Statistical Investigation works in practice through an actual case study using a step-by-step approach

GLOSSARY

Descriptive Statistics *Descriptive Statistics* is concerned with Data Summarization, Graphs/Charts, and Tables that will describe the various facets of a data set.

Inferential Statistics *Inferential Statistics* is a method used to talk about a Population Parameter from a Sample. It involves Point Estimation, Interval Estimation, and Hypothesis Testing.

Interval Data If you have data with ordinal properties and can also measure the distance between objects, you have the *interval measurement of data*. Interval data have an arbitrary zero point.

Nominal Data *Nominal Data* represent the weakest data measurement. Numbers are used to label an item or characteristic. Categorization is the main purpose of this measurement.

Ordinal Data *Ordinal or Rank Data* represent numbers that are used to rank. Ordinal data can also be verbalized on a continuum like excellent, good, fair and poor. In ordinal data, distance between objects or ranks cannot be measured.

Population *Population* is the collection of all possible observations of a specified characteristic of interest.

Parameter A Population characteristic of interest is called a *parameter*.

Primary Data *Primary Data* are collected by the organization itself for a particular purpose of its own.

Qualitative Data *Qualitative Data* are nonnumeric in nature and can't be measured directly.

Quantitative Data *Quantitative Data* are numerical in nature and can be measured.

Ratio Data It is the highest level of measurement that allows you to perform all basic arithmetic operations, including division and multiplication. Data measured on a *ratio scale* have a fixed zero point.

Sample *Sample* is a subset of the population.

Secondary Data *Secondary Data* are collected by other organizations or for other purposes. Any data, which are not collected by the organization for the specified purpose, are secondary data.

Statistics *Statistics* denote methods specially adapted to the collection, classification, analysis, and interpretation of data for managerial decision-making.

Statistic A statistic is computed as a summary measure based on a sample to make inferences about the population parameter.

REVIEW QUESTIONS

- Which one of the following is a true description of the subject "Statistics"?
 - Collection of Data
 - Classification of Data
 - Analysis of Data
 - Interpretation of Data
 - Collection, Classification, Analysis, and Interpretation of Data
- Descriptive Statistics is concerned only with Data Summarization. True or False
- Inferential Statistics is concerned with making decision regarding any population using sample results. True or False.
- Population Characteristic is called a Parameter. True or False
- A sample is a subset of Population. True or False
- Market Share of a company is an example of a
 - Nominal Scale
 - Ordinal Scale
 - Ratio Scale
 - Interval Scale
- Consumers ranking preferences and tastes is an example of
 - Nominal Scale

- (b) Ordinal Scale
 - (c) Ratio Scale
 - (d) Interval Scale
8. A Government Publication is an example of Primary Data. True or False
 9. Primary Data are up to date and are reliable. True or False
 10. Which of the following aptly describes the advantages of Secondary Data?
 - (a) Faster
 - (b) Cheaper
 - (c) Accurate
 - (d) Faster, Cheaper, and Accurate
 - (e) Faster and Cheaper

CASE STUDY-SAVVY FAST FOOD

Savvy Fast Food is located in a metro that is close to a major university in which 30% of the population consists of students. One Mr. Joseph, who holds an MBA degree from a leading business school having graduated seven years ago, owns the restaurant. It is open between 10 a.m and 10 p.m. every day. Well known for its quality and eating facilities, this restaurant usually attracts customers considered to be in the upper income category within the metro. In spite of the above-average prices charged by Savvy Fast Food, its revenue has been growing at a significant rate over the past five years. Currently Mr. Joseph is concerned about possible entry of two new modern fast food restaurants closer to his premises in the next three months.

In order to counteract any potential threat from these two new entrants, Mr. Joseph decided to conduct a customer survey that will throw light on Savvy Fast Food with regard to their demographic and psycho graphic profile as well as their perceptions about Savvy Fast Food. He is very keen to study all these using appropriate scales of measurement.

As a first step, Mr. Joseph designed a brief questionnaire using his knowledge of business statistics course that he underwent in his MBA program. It is given in Exhibit 2. He plans to go for a sample survey of 200 customers out of those who come to Savvy in the next 15 days. On an average, 100 customers visit Savvy every day.

QUESTIONS

1. Keeping in mind the contextual decision situation of Savvy, answer as to what scales of measurement (nominal, ordinal, interval, and ratio) each question in the questionnaire envisages.
2. Will the present questionnaire designed by Mr. Joseph comprehensively meet the objectives of the study?

18 Business Statistics

3. If the answer to question 2 is yes, why? If no, what additional questions will have to be added?

EXHIBIT 2

QUESTIONNAIRE

1. How far do you generally commute to reach Savvy?
_____ 0 to 3 km
_____ 3 to 6 km
_____ 6 to 9 km
_____ 9 and more km
2. How frequently do you visit Savvy in a month?
_____ I have never been before
_____ Once
_____ 2 times
_____ 3 times
_____ more than 3 times
3. How did you first come to know about Savvy?
_____ TV Advertisement
_____ Newspaper advertisement
_____ Friend
_____ Relative
_____ Any Other (Specify)
4. Rank the performance of the following attributes based on your experience of eating in Savvy.
_____ Food quality
_____ Dish variety
_____ Ambience
_____ Prices
_____ Service Quality
5. How would you rate Savvy on the following attributes? (Please tick the appropriate number for each attribute)
- | | | | | | | |
|---------------|---|---|---|---|---|----------------|
| Good Ambience | 1 | 2 | 3 | 4 | 5 | Bad Ambience |
| Expensive | 1 | 2 | 3 | 4 | 5 | Inexpensive |
| Friendly | 1 | 2 | 3 | 4 | 5 | Unfriendly |
| Polite Staff | 1 | 2 | 3 | 4 | 5 | Impolite Staff |

Caring Staff	1	2	3	4	5	Uncaring Staff
Delicious Food	1	2	3	4	5	Tasteless food

6. On an average, how much do you spend in Savvy every time you visit?

7. Sex:

_____ Male

_____ Female

8. Age:

_____ 20 or under

_____ 21 to 25

_____ 26 to 30

_____ 31 and above

9. Your Educational Background :

_____ Plus Two(High school)

_____ Diploma

_____ Bachelor's degree

_____ Master's degree

_____ Any Other(Specify)

10. Marital Status:

_____ Married

_____ Unmarried

ANSWERS TO REVIEW QUESTIONS

- The correct answer is (e) because it describes the entire gamut of the subject "Statistics". Options (a), (b), (c), and (d) are only partial descriptions of "Statistics" and hence, unacceptable.
- False. Descriptive Statistics is concerned not only with Data Summarization but also with Graphs, Charts, and Tables. If the word-only-in the given sentence were not there, then the statement would have been true.
- True. Inferential Statistics is the study of population characteristics using sample results.
- True, because this is how a parameter is defined.
- True. The Statement is true because Sample is always a part of the Population or Universe.
- The correct answer is (c). Market share of a company is expressed in percentage (for example 20.5% is a Ratio Scale). Nominal is not correct because no categorization or labeling is involved here. Ordinal is not correct because no ranking is involved here and interval scale is also incorrect because there is no arbitrary zero point. The zero in a ratio scale is a fixed one. (c) alone is correct.

20 Business Statistics

7. The correct answer is (b). This situation involves ranking of preferences and tastes and therefore, is ordinal. Ordinal scale always has a built-in ranking mechanism. The other choices are obviously, and unambiguously wrong answers.
8. False. This is because they are collected by other organizations such as Central Statistical Organization of the Government of India and are available as published document for use by anybody. Therefore, it is an example of Secondary Data.
9. True. The statement is true because Primary Data are carefully designed and directly collected by the organization for a specific purpose of its own; hence neatly fit into the needs, reliable and up-to-date.
10. The correct answer is (e). (a) is certainly an advantage of using the Secondary Data but not the only advantage. Likewise, (b) is an advantage but not the only advantage. Choice (c) is wrong because Secondary Data are not by and large accurate. (d) is wrong because it includes in the advantages the word "Accurate". (e) alone aptly describes the advantages of using Secondary Data.

Classifying Data to Convey Meaning

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Transform raw data into information
- Construct and use frequency distribution
- Construct histogram and interpret
- Construct cumulative frequency distribution, ogive graph and interpret

CHAPTER OUTLINE

- 2.1 Meaning and Examples of Raw Data
 - 2.2 Frequency Distribution
 - 2.3 Histogram
 - 2.4 Cumulative Frequency Distribution and Ogive Curve
 - 2.5 Chapter Summary
- Glossary
- Review Questions
- Answers to Review Questions
- Practice Problems

INTRODUCTION

When managers are bewildered by plethora of data, which do not convey any sense on the surface of it, they are looking for methods to classify data (data reduction) to convey meaning which will help them draw the right conclusion. This chapter provides the nitty-gritty of classifying data into information.

2.1 MEANING AND EXAMPLES OF RAW DATA

Raw data represent numbers and facts in the original format in which data have been collected. You need to convert the raw data into information for managerial decision-making.

Example of Raw Data

The weekly sales ('000' units) of a product in a region over the past year are:

52	61	59	55	63	70	59	77	81	83	69	91	73
83	90	81	77	77	74	65	33	77	64	49	49	52
50	45	42	46	39	29	38	41	43	23	26	27	22
29	31	29	31	30	30	29	40	44	45	46	47	53

Suppose you present this set of data as it is to the General Manager (Sales). At best, it will be uninspiring to him.

Another example of raw data in a manufacturing unit could be 200 sample measurements taken on the diameter of a shaft in the context of controlling variation. These 200 measurements if you present as they are to the quality assurance manager, he can draw no meaningful conclusion.

Of course, without raw data you cannot get information. Raw data are the *raw materials*, which you will process appropriately to get the finished product- *information*. See the figure 2.1.

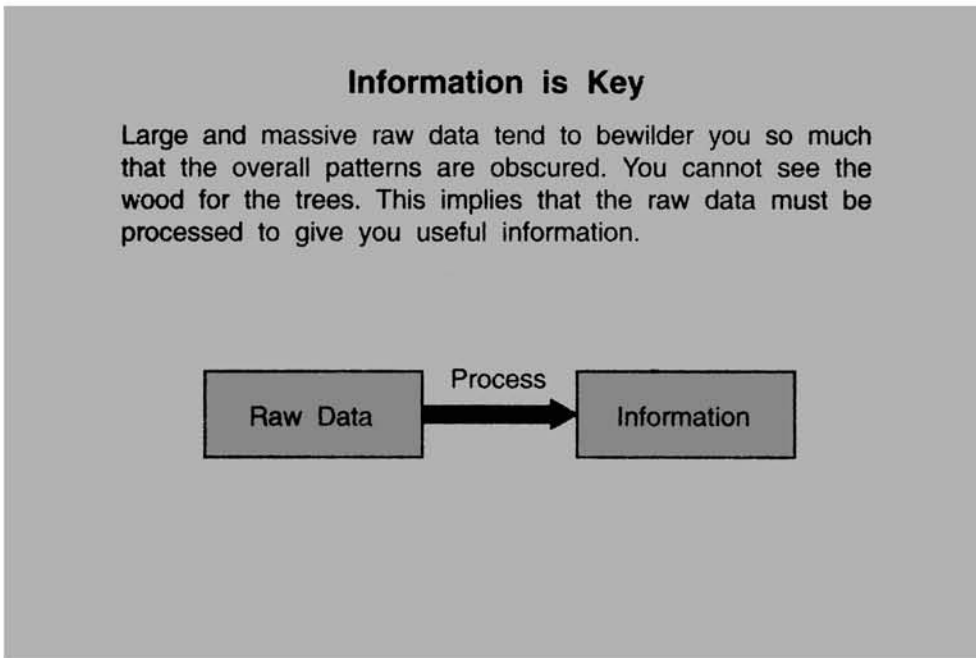


Figure 2.1

2.2 FREQUENCY DISTRIBUTION

Frequency distribution focuses on classifying raw data into information. It is the most widely used data reduction technique in descriptive statistics. When you are looking for pattern that would help you understand the characteristic you measure in a problem situation, frequency distribution is the appropriate tool to use.

In simple terms, *frequency distribution* is a summarized table in which raw data are arranged into classes and frequencies. Classes represent categories or groupings, which contain a lower limit and an upper limit. Classes are formed conveniently following certain guidelines. Against each class, you count and then place the number of observations that fall into it. When you do it for all classes in a given data analysis problem, it becomes a frequency distribution.

Guidelines for Constructing a Frequency Distribution Table

1. Identify the Minimum Value (Min) and Maximum Value (Max) in the given Data Set. Calculate, $Range = Max - Min$
2. Decide on the *Number of Classes* you would like to have. The number of classes can be determined as the square root of the number of observations in the data set. For example, if you have 100 observations, the number of classes can be 10. When perfect square root is not possible, you round off the square root to the nearest whole number. Also for any problem, it is recommended that you have not less than 5 classes and not more than 15 classes.
3. Determine the Width of the Class Interval as $= Range / Number\ of\ Classes$
4. Formulate the Boundaries of the Classes in such a manner that it will include all the observations in the data set. Avoid overlapping of classes. Once class boundary for each class is ready, all you need to do is to tally the number of observations in each class.

Construction of a Frequency Distribution - An Example

You take the example given in the beginning that contains weekly sales of a product in '000' units for the past year in a region. Going by the guidelines given above, you can have 7 classes (Square root of the number of observations, which is 52, is between 7 and 8. You round it off to 7). The range is 69 (Max = 91, Min = 22). The width of the class = range divided by number of classes which is $= 69 / 7 = 10$ approximately. The frequency distribution, which includes all observations, is given below:

<i>Class</i>	<i>Frequency</i>
22-32	12
32-42	5
42-52	11
52-62	7
62-72	5
72-82	8
82-92	4
Total	52

*You will note that in this formation, the classes are not overlapping. 22-32 means 22 and more but less than 32. This continuous way of forming the class interval is key to constructing histogram

2.3 HISTOGRAM

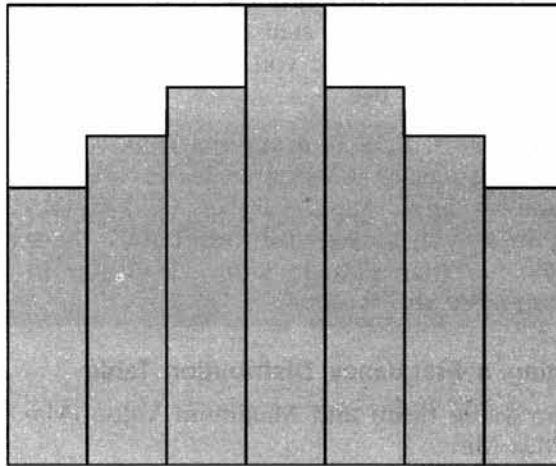


Figure 2.2

Histogram (also known as frequency histogram) is a snap shot of the frequency distribution. Histogram is a graphical representation of the frequency distribution in which the X-axis represents the classes and the Y-axis represents the frequencies. Rectangular bars are constructed at the boundaries of each class with heights proportional to the frequency.

Histogram, like frequency distribution, depicts the pattern of the distribution emerging from the characteristic being measured. If the pattern is symmetrical and bell shaped, then it reflects the normal distribution curve. We will explain the normal distribution later under probability distributions. As you can see, the histogram pictured above reflects a perfect bell shaped normal distribution.

Uses of Histogram

The following visual succinctly brings out the uses of histogram.

Histogram–Uses

- Assessing material strengths
- Estimating process capabilities
- Indicating the necessity for corrective action
- Measuring the effects of corrective action
- Estimating machine capability
- Comparing operators, materials, vendors, and products

Figure 2.3

Construction of Histogram - an illustration An analysis of the records of a hose assembly operation indicated that the "leaks" were a foremost cause of worry. It was decided to examine the hose clamping operation. The hose clamping force (torque) was measured on a sample of twenty five hose assemblies. The data are given below (Figures in foot-pounds). Draw the frequency histogram and comment.

8	13	15	10	16
11	14	11	14	20
15	16	12	15	13
12	13	16	17	17
14	14	14	18	15

Applying the guidelines discussed earlier for constructing histogram, you will notice that the *range* is $20-8 = 12$. You take the *number of classes* as 5 (Note that the square root of the number of observations is $\sqrt{25} = 5$). The width of the class is $\text{range}/\text{number of classes} = 12/5 = 2.4$. Round it to 3. You can now form the boundaries of the classes starting with 8, and then incrementing by 3 successively the lower limit of each class until all the classes are formed. Tally the number of observations under each class. This would give you the following table of frequency distribution.

Class	Frequency
8-11	2
11-14	7
14-17	12
17-20	3
20-23	1

If you graphically depict this frequency distribution by taking classes on the X-axis and frequencies on the Y-axis with rectangular bars at the boundaries of each class, you get the following histogram.

Histogram for the example

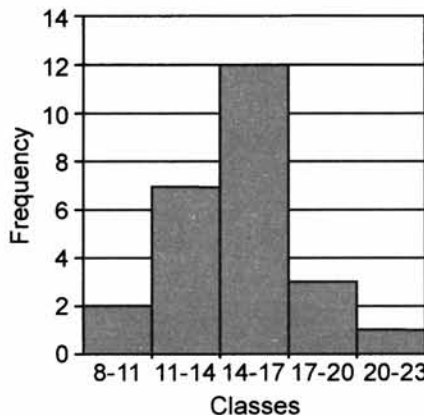


Figure 2.4

Looking at the histogram, it is easy for you to see that the pattern does not show a bell shape curve. Some distortion to normality is caused by the bars adjacent to the class 14-17. It is also evident that the average is in the range 14 to 17. Corrective action may be needed. However, before taking any action, you must be concerned about the fact that the sample size here is only 25 observations. This is a pretty small sample size to warrant an immediate action. Take more measurements and draw the histogram again, before initiating corrective steps.

Computer and Histogram

In the Microsoft Excel, Chart Wizard allows you to create a variety of charts for numerical as well as categorical data. The histogram pictured above, is an output from Chart Wizard.

Also there is a powerful add-in utility supplied by Microsoft Excel called "Data Analysis" in the Tools Menu. This has a variety of analysis tools, which include *Histogram*, *Cumulative Distribution*, *Frequency Distribution*, *Descriptive Statistics*, *Pareto-Chart*, *Correlation*, *Regression*, *Anova*, and many others. Please get familiarized with **Data Analysis** in Excel at the earliest so that you function as a manager capable of taking advantage of computer-based solutions. The power of Excel spread sheet software is amazing. Please ensure that the Data Analysis Pack is properly installed.

Histogram option in the data analysis pack of Excel gives both frequency distribution and histogram together. Excel requires that you enter the upper boundaries of the class intervals. In Excel terminology, upper limit of the class is included in the analysis. In our classification, the class interval is such that it excludes the upper limit. For example, consider the class 8-11. It means 8 and above but below 11. All you need to do while using the histogram template in Excel is, that you enter the upper class limit as 10.99 so that 11 is excluded.

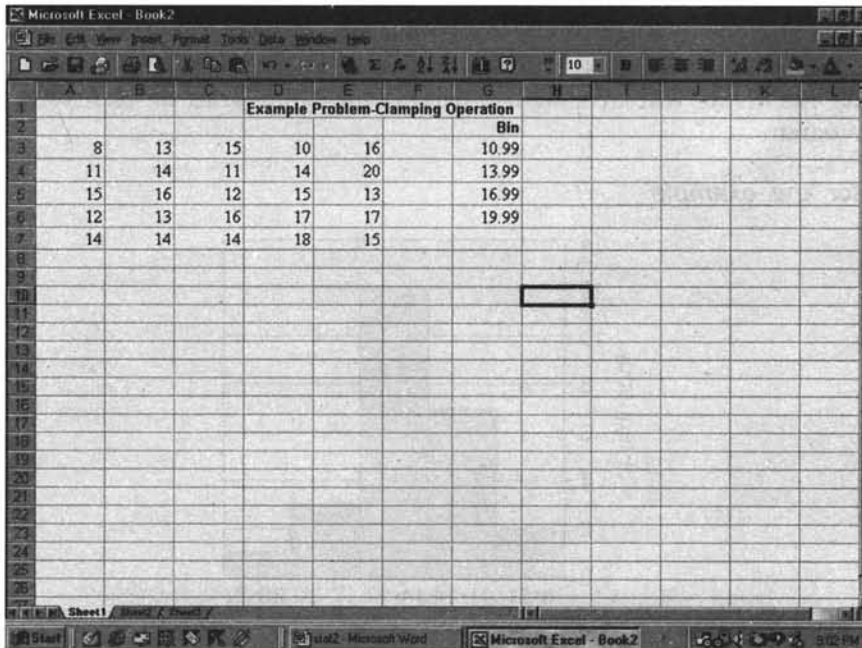


Figure 2.5

The step-by-step approach to getting the histogram from analysis pack is explained below using the same example problem. First ensure that **Data Analysis** is installed, so that it is displayed in the **Tools** menu.

- Step 1** Data entry for this option is simply entering all values of the raw data in a readable format. In a separate column, you enter the **Bin** values, that are nothing but the upper limits of the classes you want. The last upper limit you need not enter as it is automatically taken by default. The initial screen after step1 is shown in figure 2.5.
- Step 2** Click Tools. Click Data Analysis. Click Histogram. As you can see in figure 2.6, Histogram option is highlighted.

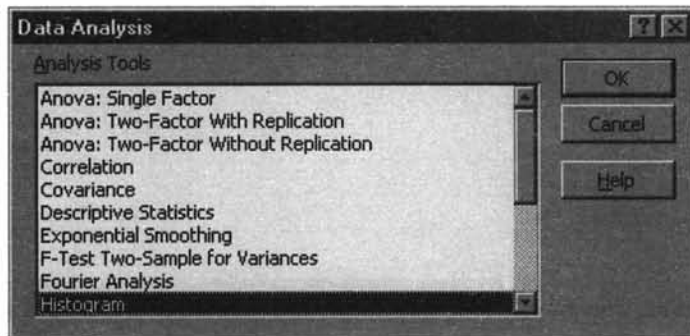


Figure 2.6

- Step 3** Click OK.
- Step 4** Enter values for the **Input Range** by appropriately highlighting the data in the spreadsheet using the mouse. Likewise, enter the **Bin** values by appropriately highlighting the data in the spreadsheet. Bin is the Excel way of understanding the upper limit of the class. Click **Chart Output** in the menu.

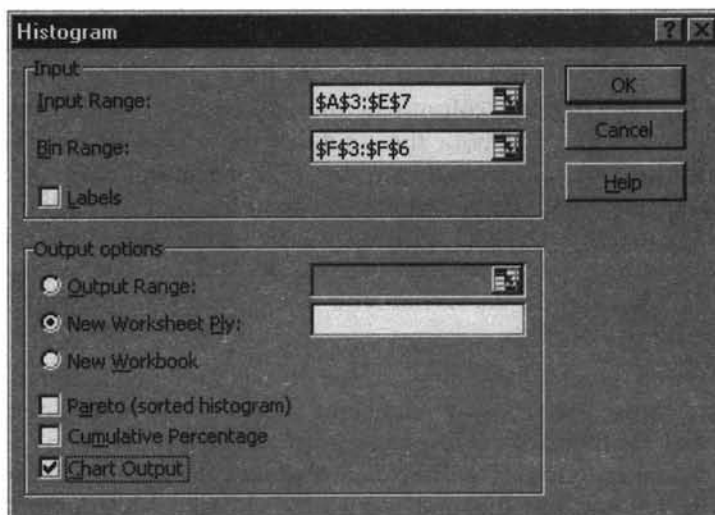


Figure 2.7

Step 5 Click OK. You now get:

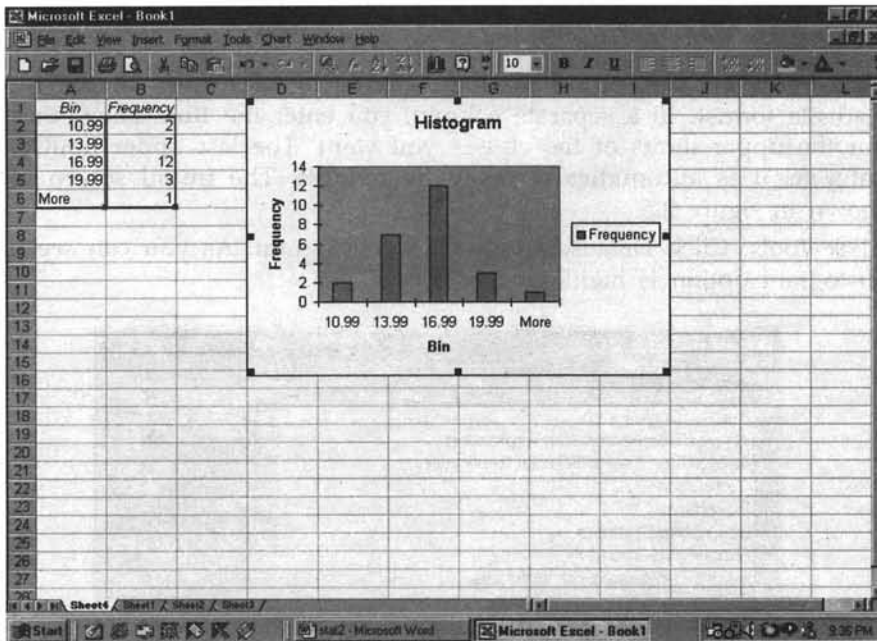


Figure 2.8

Some Fine Tuning

Move the mouse pointer towards the chart area. Right click, and then click *Chart Options* in the Pop up menu. Click *Title*. Change the default title for the X-axis and Y-axis suitably. Please note that the default option for the X-axis is "Bin".

The histogram shown above has gaps between bars that are not permitted. Right click the mouse after bringing the mouse pointer to any one of the bars in the chart. You get in the pop up menu: *Format Data Series*. Click this, you get in the menu *Options*. Click *Options* and then you enter 0(zero) in the *Gap Width* cell. You will now get an orthodox and acceptable histogram. The Bin column in the spreadsheet contains values of the upper limit of the classes. For displaying both the limits of the classes in the X-axis, what you do first is to declare the entire bin column in *Text* format. This is done in order to communicate to Excel that you are not doing any subtraction operation. After highlighting the Bin column with the mouse, click *Format*. Click *Cells*, and click *Text* under *Category*. What you have now is text-formatted cells in column A. Then change each cell by entering 8-11, 11-14, 14-17, 17-20, and 20-23. What you now get is an elegant histogram. This is displayed in figure 2.9.

Note You take the mid point of the class in the X axis and frequency in the Y-axis. Instead of constructing bar against each class with height proportional to the frequency of the class, if you connect all the frequencies corresponding to the respective mid points of the classes by a line graph, you get what is called a "*Frequency Polygon*". To put it succinctly, a frequency polygon is nothing but the mid point of each bar of histogram connected by lines, starting from the top horizontal line of each bar.

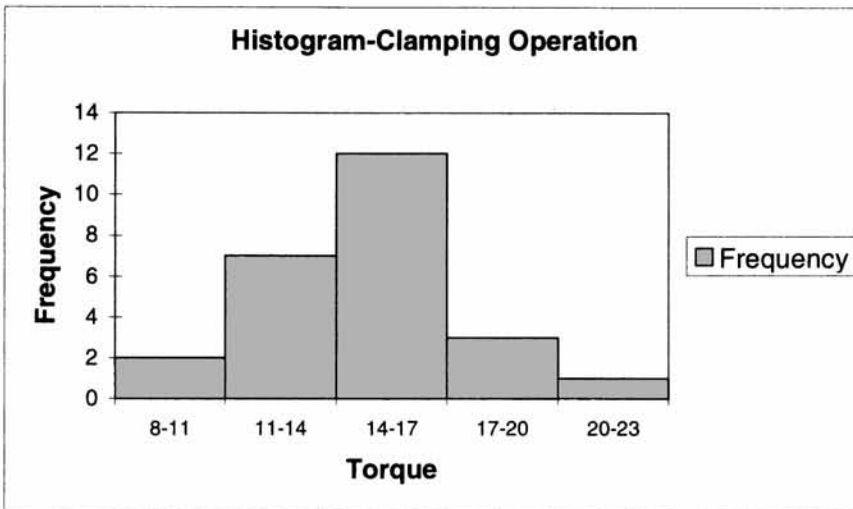


Figure 2.9

Right click the mouse after bringing the mouse pointer inside the histogram chart area. You get in the pop up menu *Chart Type*. Click Chart type, highlight *Line*, and click OK. You get the display of frequency polygon. Right click the mouse after bringing the mouse pointer to Torque(Category X axis). Change the title into **Mid Point of Class**. Likewise, change the Chart title into **Frequency Polygon**. Right click the mouse after bringing the mouse pointer inside the histogram chart area. You get in the pop up menu, *Source Data*. Highlight this and click OK. You see inside the spreadsheet input data highlighted over *Bin*(Class). Change each entry into midpoint. Alternatively, you can click with mouse *sheet 1*, or *sheet 2* etc that appear in the bottom part of the spreadsheet. Click the sheet that contains your data. Change all entries in the *Bin* column into mid point. What you have now is a comprehensive frequency polygon. For this illustration, the frequency polygon is given in figure 2.10.

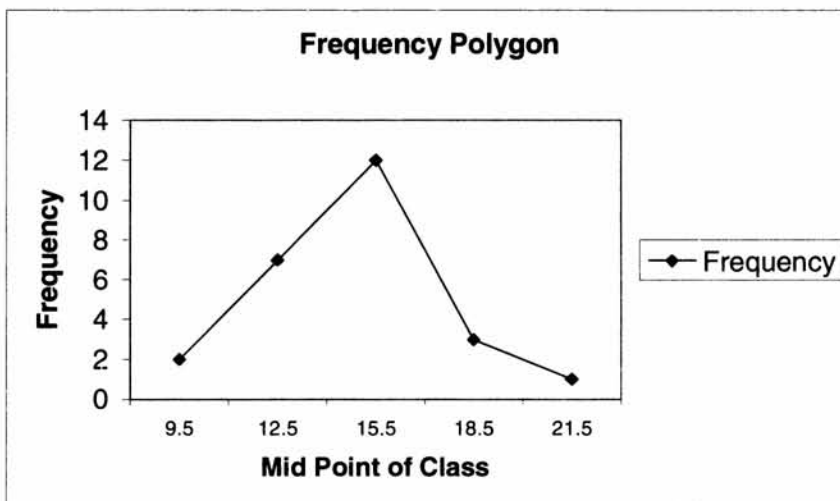


Figure 2.10

Construction of histogram using chart wizard-an example The weekly sales of a product in a region over the past year are (Figures in '000' units):

52	61	59	55	63	70	59	77	81	83	69	91	73
83	90	81	77	77	74	65	33	77	64	49	49	52
50	45	42	46	39	29	38	41	43	23	26	27	22
29	31	29	31	30	30	29	40	44	45	46	47	53

Solution Here, You will be getting the histogram by using the Chart Wizard of Microsoft Excel. For this the input required is the frequency distribution. This is the same example problem we did while constructing the frequency distribution. Let me display to you once again the frequency distribution for this problem.

<i>Class</i>	<i>Frequency</i>
22-32	12
32-42	5
42-52	11
52-62	7
62-72	5
72-82	8
82-92	4
Total	52

Now please follow the step-by step procedure described below:

Step 1 Enter the data in the Excel spreadsheet. It may be pointed out that column B that contains classes will have to be in text format in order to communicate to Excel that you are not doing any subtraction operation. After highlighting the class column with the mouse, click Format. Click Cells and click *Text* under category. What you have now are text-formatted cells in column B. Enter by typing in each cell 22-32, 32-42, etc until the last cell that will have 82-92. The screen will look like this (figure 2.11).

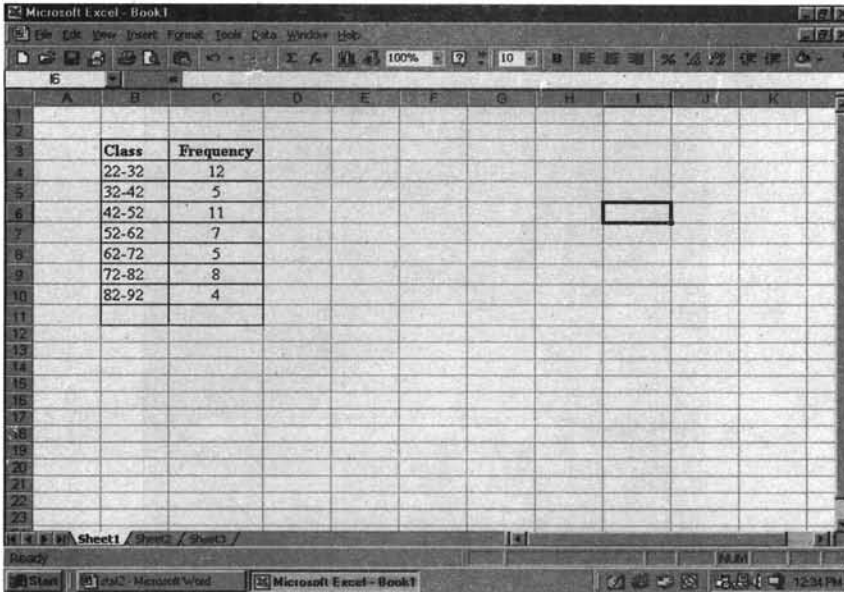


Figure 2.11

Step 2 Click with your mouse the icon depicting Chart Wizard. Highlight with your mouse the choice Column Bar in the menu. You will see the following figure 2.12.

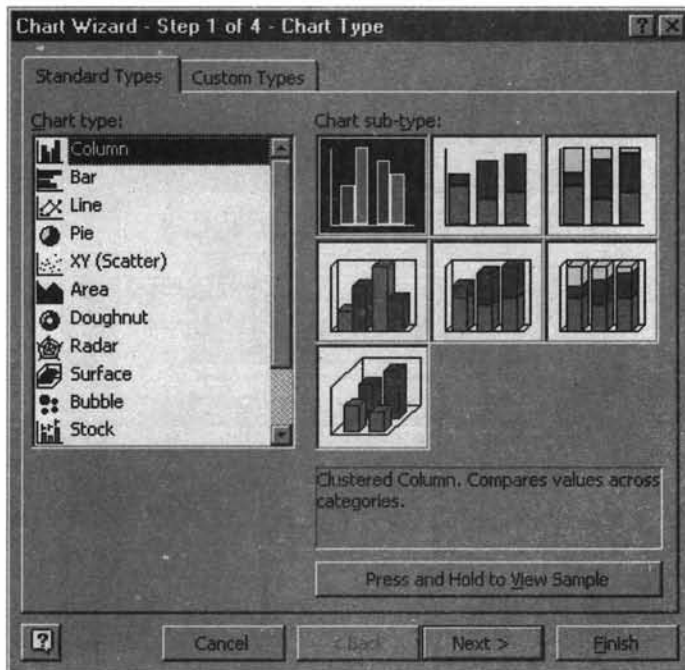


Figure 2.12

32 Business Statistics

Step 3 As you can see in the menu under chart type, *Column* is highlighted. Click Next and you get:

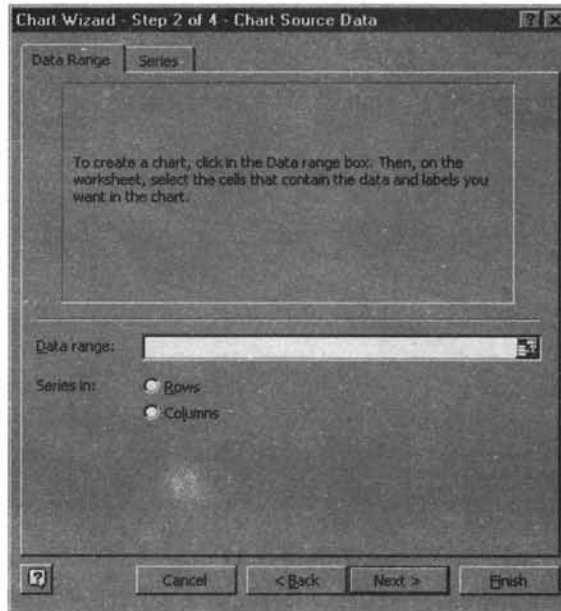


Figure 2.13

Step 4 Highlight with your mouse the input data in the *Data range*. The figure 2.14 will be displayed:

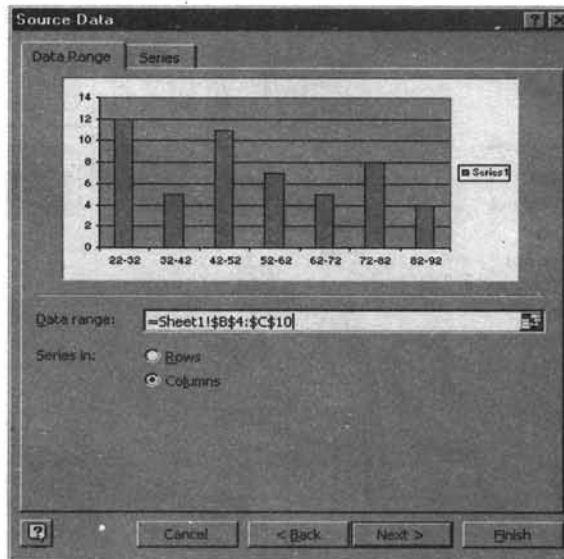


Figure 2.14

- Step 5** Click Next. You get the **Chart Option** menu. Enter a suitable title for the problem as well as titles for X-axis and Y-axis. The figure 2.15 will appear. As you can see, we have given a chart title, "Histogram of Sales Performance", X-axis title as "Class", and Y-axis title as "Frequency".

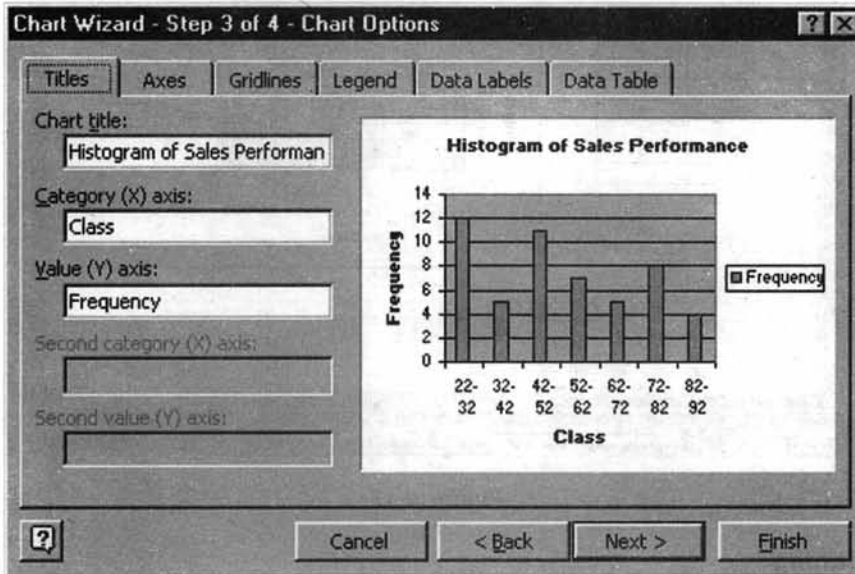


Figure 2.15

- Step 6** Click Next. You get:

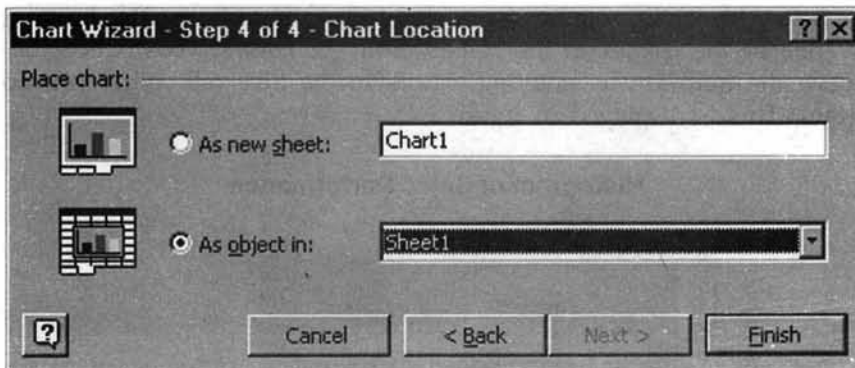


Figure 2.16

- Step 7** You have the option to save the picture either as an object in sheet 1 or in a new sheet. The default option is as object in sheet 1. If you want to exercise the default option Click *Finish*. You get the histogram in the existing sheet. This screen is given in figure 2.17.

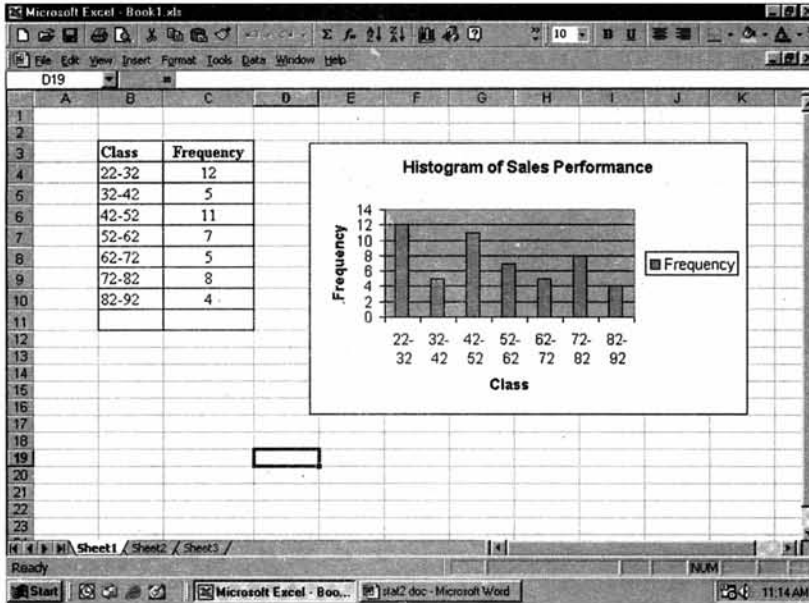


Figure 2.17

Some Fine Tuning

The histogram shown above has gaps between bars that are not permitted. Right click the mouse after bringing the mouse pointer to any one of the bars in the chart. You get in the pop up menu *Format Data Series*. Click this and you get *Options*. Click options and then you enter 0(zero) in the *Gap Width* cell. You will now get an orthodox and acceptable histogram. For better readability of the X-axis, right click the mouse after bringing the mouse pointer to any one of the classes in the X-axis. You get *Format Axis*. Click the *Font* in the menu and reduce the font size by entering, say 8, in the *Size* cell. What you now get is an elegant histogram. This is displayed in figure 2.18.

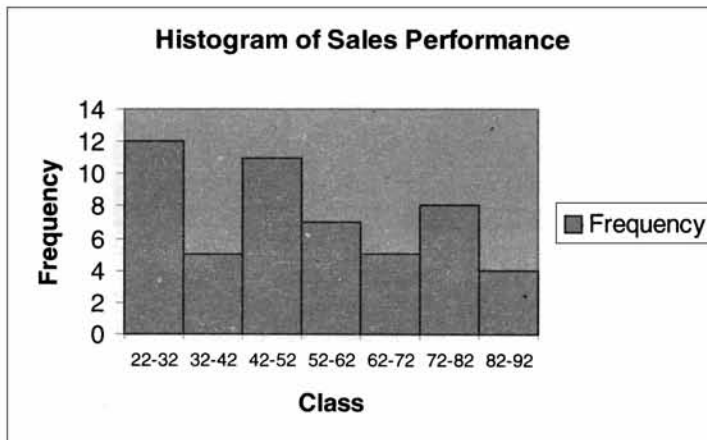


Figure 2.18

2.4 CUMULATIVE FREQUENCY DISTRIBUTION AND OGIVE CURVE

Another extremely useful method of getting the essence of distribution is the *cumulative distribution*. This can be formed from the frequency distribution table. To explain the nuances of frequency distribution along with related terms used, consider the example problem on clamping pressure (torque). Refer page 25.

Summary table for the clamping pressure(torque) example

<i>Class</i>	<i>Frequency Frequency</i>	<i>Relative Frequency as proportion</i>	<i>Relative Frequency as %</i>	<i>Cumulative Frequency in numbers</i>	<i>Cumulative Frequency in percent</i>
8-11	2	0.08	8.00	2	8.00
11-14	7	0.28	28.00	9	36.00
14-17	12	0.48	48.00	21	84.00
17-20	3	0.12	12.00	24	96.00
20-23	1	0.04	4.00	25	100.00
Total	25	1.00	100.00		

Relative frequency as proportion The relative frequency distribution represents the frequency in each class divided by the total frequency (total number of observations). In our illustration, the relative frequency corresponding to the class 8-11 is 2 divided by 25 and = 0.08. Likewise, you can calculate the relative frequencies for the other classes. Relative frequency is a proportion. Relative frequency of any class must be always between 0 and 1, when expressed as a proportion. The relative frequency corresponding to the class 8-11 is .08. This implies that the probability that an observation will fall in this interval is 8%.

Relative frequency as percentage If you multiply the relative frequency in proportions by 100, you get the relative frequency in percentages. In most business applications, percentages and proportions are more useful than mere numbers because they facilitate comparison across groups. For example, if you are interested in comparing the yearly sales performance split into quarters of different regions, relative frequency in proportions or percentages are more meaningful than mere numbers because the sales potential in all regions will not be identical.

Cumulative frequency distribution in numbers For this problem, take the first class 8-11. Now find out the number of observations that are less than 11. It is 8. Take the next class 11-14. How many observations are in this class? It is 7. The number of observations cumulated at this stage will be = 2 + 7 = 9. That is, 9 observations are less than 14. Similarly, the cumulative frequency up to 17 = 2 + 7 + 12 = 21. Likewise, all entries are worked out and filled in the summary table that deals with the clamping pressure example.

Cumulative frequency distribution in percentages Instead of the actual frequency in each class, if you take the percentage (relative frequency) and keep on cumulating exactly in the same manner as you have done while calculating cumulative frequency in numbers, you get the *Cumulative Frequency Distribution* in percentages. These are computed in the table for the clamping pressure example.

The Ogive curve The Ogive curve is a graphical representation of the cumulative frequency distribution using numbers or percentages. In this pictorial representation, less than values are on the X-axis and cumulative frequency in numbers or percentages are on the Y-axis. A line graph in the form of a curve is plotted connecting the cumulative frequencies corresponding to the upper boundaries of the classes. Today, this ogive graph is elegantly and efficiently obtained as output from Chart Wizard of Microsoft Excel. The Ogive graphs for the present torque example, both for numbers and percentages, obtained from Chart Wizard are given below:

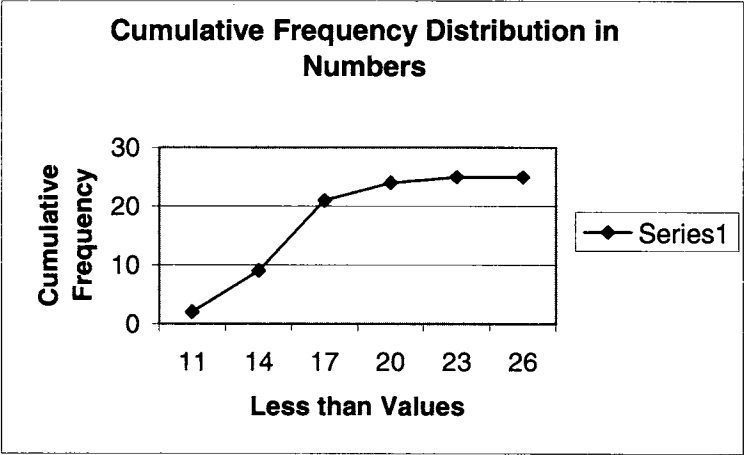


Figure 2.19

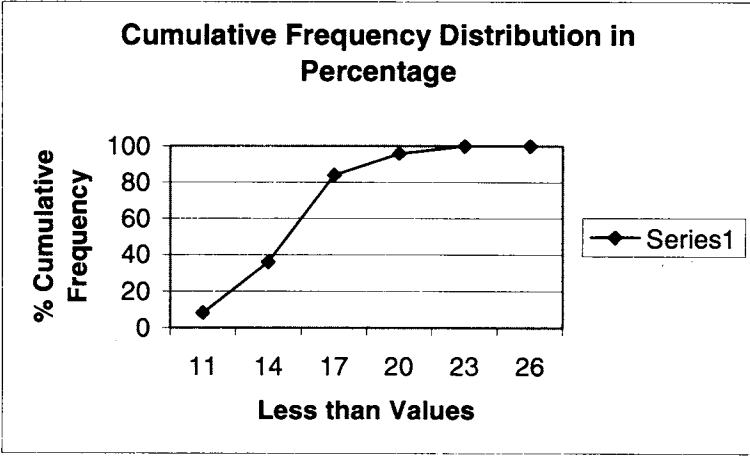


Figure 2.20

Note The procedure for getting the ogive curve from Chart Wizard is identical to getting histogram. The changes envisaged in data entry are to have upper limits of the classes along with the cumulative frequencies in two separate columns. Also in the *Chart Type*, you highlight the "*Line*" graph choice instead of *Bar*.

2.5 CHAPTER SUMMARY

This chapter has given you a conceptual and practical framework of classifying data into information which is key to decision making. In particular, the chapter focused on:

- Explaining the meaning of raw data with examples as well as its inadequacy in the context of decision-making.
- Frequency distribution as a data reduction technique for conveying meaning and recognizing patterns that are essential in analyzing data into information.
- Meaning of histogram and construction of histogram for better and sharper pattern identification as well as understanding the shape of the distribution (Normal or not).
- How to get frequency polygon from histogram
- The features and uses of cumulative frequency distribution along with the related items—relative frequency and ogive curve.
- A step-by-step approach to getting histogram from Microsoft Excel using Data Analysis Pack as well as Chart Wizard.

GLOSSARY

Class A Class represents a category or a group that contains a lower limit and an upper limit.

Cumulative Frequency Distribution A type of frequency distribution that shows how many observations are below the upper boundaries of the classes.

Data Reduction Data reduction involves methods to classify data to convey meaning when managers are bewildered by plethora of data.

Frequency Distribution Frequency distribution is a summarized table in which raw data are arranged into classes and frequencies. Frequencies represent the number of observations that fall into a set of classes.

Frequency Polygon Frequency polygon is nothing but the mid point of each bar of histogram connected by lines starting from the mid point of the top horizontal line of each bar.

Histogram Histogram (also known as frequency histogram) is a snap shot of the frequency distribution. Histogram is a graphical representation of the frequency distribution in which the X-axis represents the classes and the Y-axis represents the frequencies. Rectangular bars are constructed at the boundaries of each class with heights proportional to the frequency.

Information When raw data are processed into suitable formats that convey meaning, they become information.

Ogive Ogive is a graphical representation of the cumulative frequency distribution using numbers or percentages. In this pictorial representation, upper boundaries (less than values) of the classes are on the X-axis and cumulative frequency in numbers or percentages are on the Y-axis. A curve is drawn connecting the cumulative frequencies that correspond to the upper boundaries of the classes.

Raw Data Raw Data represent numbers and facts in the original format in which the data have been collected.

Relative Frequency Relative frequency distribution represents the frequency in each class divided by the total frequency (total number of observations). It is expressed as a percentage or proportion.

REVIEW QUESTIONS

1. Raw Data by and large may not help the manager of an enterprise to take decisions. True or False
2. Frequency Distribution is a data reduction technique to calculate statistical averages. True or False
3. Histogram identifies the pattern as well as the shape of the distribution in any data analysis situation. True or False
4. Histogram is a graphical representation of frequency distribution in the form of column bars in which the width of each bar is equal to:
 - (a) Midpoint of the Class
 - (b) Class Interval
 - (c) Frequency of the Class
 - (d) Height of each Bar
5. The shape of the histogram can be helpful in determining whether the distribution is normal. True or False
6. When you want to compare performance across groups or segments, frequency expressed in numbers is superior to frequency expressed in proportions. True or False

Questions 7-10 refer to the following data:

Daily sales figures (in Rs 000) in a medium restaurant over the last 80 days have been summarized as follows:

Daily Sales	Number of Days
20-30	14
30-40	20
40-50	24
50-60	15
60-70	7

	80

7. The number of days in which the daily sales are between 30-50 is:
 (a) 20 (b) 24 (c) 44 (d) 40
8. The proportion number of days in which the sales will be more than 40 is:
 (a) 0.30 (b) 0.1875 (c) 0.25 (d) 0.575
9. The percentage number of days in which the sales will be less than 30 days is:
 (a) 25% (b) 17.5% (c) 0% (d) 17%
10. The relative frequency corresponding to the class 40-50 is:
 (a) 24% (b) 28% (c) 30% (d) 31%

ANSWERS TO REVIEW QUESTIONS

1. True because the manager will not be in a position to get the essence of information present in the raw data unless they are converted into information by classifying the data.
2. False because Frequency Distribution is a data reduction technique to see any meaningful pattern that may emerge from the characteristic of interest being measured.
3. True because this is the principal purpose of histogram. The histogram is a snap shot of the distribution identifying the pattern and the shape.
4. The correct answer is choice b). The width of the bar can never be the midpoint of the class and hence a) is wrong. Choice c) is not correct because frequency of the class is the height of the bar. Choice d) is wrong because height of the bar has nothing to do with the width of the bar in a histogram.
5. True because histogram plays a pivotal role in understanding the shape of the distribution. It tells at a glance by its pictorial representation, whether the distribution is a bell shaped symmetrical normal distribution.
6. False because frequencies expressed in proportions are percentages and hence, superior to frequency expressed in numbers when it comes to comparison across groups or segments.
7. Choice (c) is correct because the number of days in which the daily sales are between 30 – 50 is = frequency in 30-40 and frequency in 40 – 50 = 20 + 24 = 44. Therefore, the other choices are wrong by sheer logical calculations.
8. The correct answer is choice d). First, let us calculate the number of days in which the sales will be more than 40. This is = adding the number of days in 40 – 50, 50 – 60, 60 – 70. This is = 24 + 15 + 7 = 46 days. The proportion number of days in which the sales will be more than 40 = $46/80 = 0.575$. The other choices are obviously wrong based on actual calculations.
9. Correct answer is b). The number of days in which the sales will be less than 30 days is = number of days corresponding to the class 20-30. This is = 14. Percentage number of days in which the sales will be less than 30 = $(14/80)$ multiplied by 100 = 17.5%. Obviously, other choices are wrong.

10. Correct answer is choice c). The relative frequency corresponding to the class 40-50 is = $24/80 = 0.30$. Since the choices are given in percentages, 0.30 is same as 30%. Obviously, the other choices are rejected.

PRACTICE PROBLEMS

1. Case Study -Waiting Time in ATM Counters

A private bank having a number of ATMs in various locations is contemplating to install another ATM at one of its potentially strong business city. As a first step in assessing commercial feasibility, the bank authorities had collected data on the waiting time customers face at the existing ATM counters. One hundred customers were studied and the waiting time in minutes each one spent at the counter was recorded. The data are given below:

2.9	4.5	3.8	3.0	3.3	4.3	3.0	4.4	3.7	3.0
4.3	2.2	3.6	3.7	6.2	4.1	4.9	6.4	6.2	3.5
3.1	3.8	3.6	5.9	5.8	5.1	3.3	3.1	5.4	3.6
6.0	5.3	4.7	2.2	4.3	6.1	4.7	5.1	4.4	3.1
5.4	2.9	4.2	6.2	5.7	4.1	4.0	4.3	4.3	4.5
2.9	5.5	4.5	3.2	6.7	5.6	5.8	4.6	5.5	5.6
8.3	2.6	6.0	3.1	6.8	8.4	5.7	4.8	5.1	5.9
5.5	3.1	4.8	6.8	7.9	4.2	5.3	4.4	5.7	5.6
3.0	6.5	3.2	4.0	4.0	8.0	4.9	5.5	4.0	7.6
3.5	3.2	3.9	4.9	4.0	5.6	7.6	4.7	6.8	4.0

Construct a frequency distribution and histogram. Interpret the results.

2. Case Study-Shaft Diameter

In the context of studying the behavior of diameter of a shaft, diameter measurements in centimeters were taken. Five samples in succession were taken every half hour. The data are given below:

Time	Sample1	Sample2	Sample3	Sample4	Sample5
10:30	2.52	2.54	2.54	2.51	2.53
11:00	2.54	2.49	2.55	2.51	2.53
11:30	2.53	2.53	2.49	2.53	2.54
12:00	2.52	2.56	2.48	2.57	2.52
12:30	2.53	2.49	2.54	2.52	2.53
13:00	2.53	2.53	2.54	2.54	2.54
13:30	2.50	2.54	2.56	2.53	2.52
14:00	2.53	2.54	2.56	2.52	2.51

Time	Sample1	Sample2	Sample3	Sample4	Sample5
14:30	2.49	2.53	2.51	2.53	2.54
15:00	2.54	2.51	2.52	2.52	2.53
15:30	2.54	2.52	2.52	2.54	2.55
16:00	2.54	2.58	2.51	2.50	2.52
16:30	2.53	2.55	2.55	2.52	2.54
17:00	2.53	2.54	2.57	2.54	2.54
17:30	2.55	2.54	2.51	2.55	2.54
18:00	2.52	2.53	2.52	2.54	2.55
18:30	2.53	2.52	2.53	2.53	2.53
19:00	2.51	2.53	2.52	2.54	2.54
19:30	2.54	2.51	2.53	2.53	2.55
20:00	2.57	2.54	2.57	2.50	2.54

Construct a histogram for the study and interpret the results. If the customer specification is 2.52 cm plus or minus 0.03 cm, how many observations are outside the specification?

3. For problems 1) and 2), construct the cumulative distribution curves and give your comments.

4. Case Study-Electricity Charges

The following data are obtained from a random sample of 50 households with regard to electricity charges in Rs for the month of April 2001. These households belong to the middle class income category.

96	86	167	149	104	93
171	161	195	157	172	104
202	192	100	214	159	168
178	168	126	183	163	136
147	137	182	131	139	99
102	92	121	136	152	120
153	143	158	152	196	176
197	187	223	176	175	122
127	117	140	117	148	114
82	72	175	175	158	143

- (a) Construct a frequency distribution and histogram. Comment on your findings.
- (b) Construct a frequency polygon and interpret the same.

5. Case Study- Money Spent On Fast Food

The following data represent the amount (in Rs) college going students spend on fast food according to a survey involving a random sample of 30 students.

162	162	172	168	182	172
122	132	266	262	198	168
154	222	182	142	178	152
142	112	222	236	192	166
138	162	172	176	216	152

Construct a frequency histogram for this data set. Do you find any pattern with regard to the spending on fast food?

Measures of Central Tendency and Dispersion

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Appreciate the need for summary measures
- Explain and calculate the measures of location
- Explain and calculate measures of spread (dispersion)
- Discuss the strengths and limitations of these measures

CHAPTER OUTLINE

- 3.1 Measures of Central Tendency
- 3.2 Measures of Dispersion (Variation)
- 3.3 Chapter Summary
- Glossary
- Review Questions
- Answers to Review Questions
- Practice Problems

INTRODUCTION

Raw Data are the raw materials that will have to be converted into finished products (Information). From a voluminous database containing raw data, it is impossible to see any pattern unless they are converted into information by data reduction. The reduction can be achieved by summary measures, which are concise and yet, give a reasonably accurate view of the original data. This chapter covers the important summary measures of central tendency and measures of dispersion (variation).

3.1 MEASURES OF CENTRAL TENDENCY

What is Central Tendency?

Whenever you measure things of the same kind, a fairly large number of such measurements will tend to cluster around the middle value. The question that arises is, "is it possible to define one typical representative average in such a manner that the remaining items in the data set will cluster around this value?" This value is called a measure of "Central Tendency". Other synonymous terms include "Measures of Location", and "Statistical Averages".

Measures of Central Tendency

Quantitative Specialists, Statisticians, and Information Analysts rely heavily on summary measures when a large mass of data will have to be analyzed to help decision-makers. You as a manager need these summary measures of central tendency that act as point estimates (sample averages) in your functional area of operation. The most widely used measures of central tendency are *Arithmetic Mean*, *Median*, and *Mode*.

Arithmetic Mean

Arithmetic Mean (called mean) is the most common measure of central tendency used by all managers in their sphere of activities. It is defined as the sum of all observations in a data set divided by the total number of observations. For example, consider a data set containing the following observations:

4, 3, 6, 5, 3, 3. The arithmetic mean = $(4 + 3 + 6 + 5 + 3 + 3)/6 = 4$.

In symbolic form the mean for raw data is given by the following visual.

Arithmetic Mean

$$\bar{X} = \frac{\Sigma X}{n}$$

\bar{X} = Arithmetic Mean

ΣX = Indicates sum all X values in the data set

n = Total number of observations (Sample size)

Arithmetic mean calculation for raw data- An example The inner diameter of a particular grade of tire based on 5 sample measurements are as follows: (figures in millimeters)

565, 570, 572, 568, 585

Calculate the Mean.

Applying the formula,

$$\bar{X} = \frac{\sum X}{n}$$

We get mean = $(565 + 570 + 572 + 568 + 585)/5 = 572$ millimeters.

Caution *Arithmetic Mean is affected by extreme values or fluctuations in sampling. It is not the best average to use when the data set contains extreme values (very high or very low values).*

Median Median is the middle most observation when you arrange data in ascending or descending order of magnitude. That is, the data are ranked and the middle value is picked up. Median is such that 50% of the observations are above the median and 50% of the observations are below the median.

Median is a very useful measure for ranked data in the context of consumer preferences and rating. It is not affected by extreme values but is affected by the number of observations. It may be mentioned here that a large number of market research studies that deal with ordinal data use the median as a summary measure extensively.

The following visual gives the formula for median that uses raw data.

Median

$$\text{Median} = \frac{n+1}{2} \text{th value of ranked data}$$

n = Number of observations in the sample

Note If the sample size is an odd number then median is $(n + 1)/2$ th value in the ranked data. If the sample size is even, then median will be between two middle values. You take the average of these two middle values.

Median calculation for raw data - An example for odd size sample Marks obtained by 7 students in Computer Science Exam are given below:

Compute the median.

45 40 60 80 90 65 55

Arranging the data after ranking gives

90 80 65 60 55 45 40

Median = $(n + 1)/2$ th value in this set = $(7 + 1)/2$ th observation = 4th observation = 60.

Hence, Median = 60 for this problem.

Median calculation for raw data - An example for even sample size Diameter of a shaft, in millimeters, in a manufacturing unit is given below for 10 samples. Calculate the median value.

2.50 2.45 2.55 2.60 2.46 2.43 2.56 2.58 2.66 2.65

Arranging the data in the ascending order, you will get:

2.43 2.45 2.46 2.50 2.55 2.56 2.58 2.60 2.65 2.66

The median falls between 5th and 6th observation. That is between 2.55 and 2.56.

Hence, median = $(2.55 + 2.56)/2 = 2.555$

Mode Mode is that value which occurs most often. It has the maximum frequency of occurrence. Mode is not affected by extreme values.

Mode is a very useful measure when you want to keep in the inventory the most popular shirt in terms of collar size during festival season. Median and mean will not be helpful in this type of situation. Another example where mode is the only answer is in determining the most typical shoe size to be kept in stock in a shop selling shoes.

Caution In a few problems in real life, there will be more than one mode such as bimodal and multi-modal values. In these cases mode cannot be uniquely determined.

Mode calculation for raw data - An example The life in number of hours of 10 flashlight batteries are as follows: Find the mode.

340 350 340 340 320 340 330 330 340 350

340 occurs five times. Hence, mode = 340.

Mean, median, and mode for grouped data (frequency distribution) Information based decision making is the order of the day. When a large volume of raw data will have to be processed into information, you need the measures of central tendency- mean, median, and mode to be calculated based on the frequency distribution (also called grouped data). The formulas in symbolic form and the associated calculations are explained one after the other.

Mean for Grouped Data

$$\bar{X} = \frac{\sum fX}{n}$$

Where

\bar{X} = Mean

$\sum fX$ = Sum of cross products of frequency in each class with midpoint X of each class

n = Total number of observations (Total frequency = $\sum f$)

Example Find the arithmetic mean for the following continuous frequency distribution:

Class	0-1	1-2	2-3	3-4	4-5	5-6
Frequency	1	4	8	7	3	2

Solution for Mean

The midpoint of the first class is $(0 + 1)/2 = 0.5$; the midpoint of the second class, is $(1 + 2)/2 = 1.50$, and so on. Again these calculations can be done on a spreadsheet, as shown below:

	A	B	C	D
1	Class	X	f	fX
2	0-1	0.5	1	0.5
3	1-2	1.5	4	6.0
4	2-3	2.5	8	20.0
5	3-4	3.5	7	24.5
6	4-5	4.5	3	13.5
7	5-6	5.5	2	11.0
8	Total		25	75.5
9	Mean			3.02

$$\text{Mean} = \bar{X} = \frac{\sum fX}{n} = 75.5/25 = 3.02$$

It is easy to see from the spreadsheet that $\sum fX = 75.5$ and $n = \sum f = 25$.

Note The mean calculated for grouped data is only an approximation because you assume in this formula that all the values in any class is equal to the mid point of that class. This assumption leads to the conclusion that the mean of 3.02 computed above is only a good approximation of the actual mean.

Median for Grouped Data

$$\text{Median} = L + \frac{(n/2) - m}{f} \times c$$

Where

L = Lower limit of the median class

n = Total number of observations = $\sum f$

m = Cumulative frequency preceding the median class

f = Frequency of the median class

c = Class interval of the median class

Example Find the median for the following continuous frequency distribution:

Class	0-1	1-2	2-3	3-4	4-5	5-6
Frequency	1	4	8	7	3	2

Intuitive approach for calculating median for this grouped data

$n/2 = 25/2 = 12.5$, which falls in the class 2-3 because the cumulative frequency corresponding to this class is 13. So, the median lies in the class 2-3. Can you say why?

Up to the second class, the cumulative frequency (m) is 5. The median in the ranked data has to be 7.5th observation in the class 2-3. As there are 8 observations in this class, to get median, you add to the lower limit of this class 2-3 a correction factor of $7.5/8$ times the width of this class (which is =1 in this case). Hence, median has to be $=2+7.5/8 = 2.9375$. You get exactly the same value by the formula for median given below. As in the case of mean, median computed for grouped data is only a good approximation of the actual median.

Solution for median in the orthodox way by formula approach

Class	Frequency	Cumulative Frequency
0-1	1	1
1-2	4	5
2-3	8	13
3-4	7	20
4-5	3	23
5-6	2	25

Substituting in the formula the relevant values,

$$\text{Median} = L + \frac{(n/2) - m}{f} \times c$$

$$2 + \frac{(25/2) - 5}{8} \times 1 = 2.9375$$

Mode for Grouped Data

$$\text{Mode} = L + \frac{d_1}{d_1 + d_2} \times c$$

Where L = Lower limit of the modal class

$$d_1 = f_1 - f_0 \quad d_2 = f_1 - f_2$$

f_1 = Frequency of the modal class

f_0 = Frequency preceding the modal class

f_2 = Frequency succeeding the modal class

c = Class Interval of the modal class

Example Find the mode for the following continuous frequency distribution:

Class	0-1	1-2	2-3	3-4	4-5	5-6
Frequency	1	4	8	7	3	2

Intuitive approach for calculating mode for this grouped data

The largest frequency of occurrence is 8, which corresponds to the class 2-3. Hence, mode lies in the class 2-3. The mode has to be 2+something. Can you guess what this something is? . There are 4 observations in the class preceding the modal class (class 1-2 has a frequency of 4) and 7 observations in the class succeeding the modal class (class 3-4 has a frequency of 7). The differences between the frequency of the modal class and the frequencies preceding and succeeding the modal class are respectively, $8 - 4 = 4$ and $8 - 7 = 1$. Intuition suggests that you could use the ratio $4/(4+1)$ or $4/5$ times the width of the modal class (which is =1 in this case) as a correction factor to be added to the lower limit of the modal class. Therefore, mode = $2 + 4/5 = 2.80$. This is exactly same as computed by the formula given below. Just like mean and median, mode computed for grouped data is only a good approximation of the actual mode.

Solution for mode in the orthodox way by formula approach

Class	Frequency
0-1	1
1-2	4
2-3	8
3-4	7
4-5	3
5-6	2

Substituting in the formula the values of the relevant terms,

$$\begin{aligned} \text{Mode} &= L + \frac{d_1}{d_1 + d_2} \times c \\ &= 2 + \frac{4}{5} \times 1 = 2.8 \end{aligned}$$

(Note $d_1 = f_1 - f_0 = 8 - 4 = 4$. $d_2 = f_1 - f_2 = 8 - 7 = 1$)

Comparative Picture of Mean, Median, Mode

Mean	Median	Mode
1. Defined as the arithmetic average of all observations in the data set.	Defined as the middle value in the data set arranged in ascending or descending order.	Defined as the most frequently occurring value in the distribution; it has the largest frequency.

<i>Mean</i>	<i>Median</i>	<i>Mode</i>
2. Requires measurement on all observations.	Does not require measurement on all observations.	Does not require measurement on all observations.
3. Uniquely and comprehensively defined.	Cannot be determined under all conditions	Not uniquely defined for multi-modal situations.
4. Affected by extreme values.	Not affected by extreme values.	Not affected by extreme values.
5. Can be treated algebraically. That is, Means of several groups can be combined.	Cannot be treated algebraically. That is, Medians of several groups cannot be combined.	Cannot be treated algebraically. That is, Modes of several groups cannot be combined.

3.2 MEASURES OF DISPERSION (VARIATION)

Measures of Central Tendency by themselves are not adequate in analyzing business data. Consider the following data of salaries paid to Executives of three departments with each having on its roster 4 executives.

Figures: Salaries per month in Rs

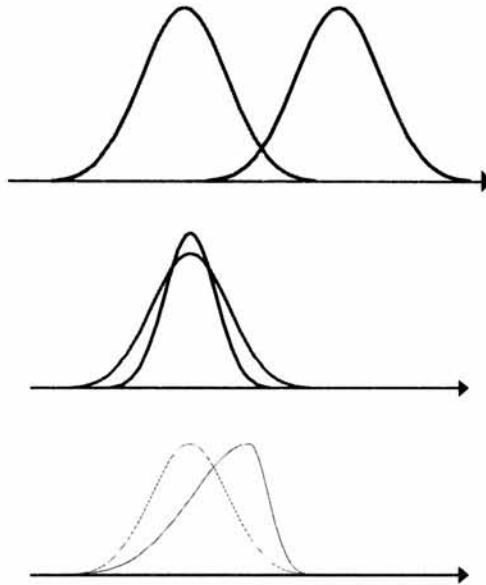
<i>Department A</i>	<i>Department B</i>	<i>Department C</i>
25000	24000	18000
25000	26000	32000
25000	23000	15000
25000	27000	35000

If you carefully study the above table, you will notice that the average salary per executive is Rs.25000 for all the three departments. Department A has no variation and all are getting identical salaries. Department B shows some variation but not too bad. Department C shows a high level of Dispersion and Scattering. Therefore, you have to study the spread or dispersion around the central tendency. In other words, statistical measures of central tendency and measures of dispersion (variation) will have to be studied together for drawing meaningful conclusions.

Measures of Dispersion (Spread)

In simple terms, measures of dispersion indicate how large the spread of the distribution is around the central tendency. It answers unambiguously the question, " what is the magnitude of departure from the average value for different groups having identical averages?" It is important to study the central tendency along with dispersion to throw light on the shape of the curve; to gauge whether there is distortion to the bell shaped symmetrical normal distribution curve that forms the foundation stone upon which the entire statistical inference is built.

Consider the following pictures:



Inferences from the pictures above The first one represents two symmetrical bell shape distributions with different central tendencies. The second one represents two symmetrical bell shape distributions with same mean but different spread or variation. The third one reveals the shape of two distributions that have different central tendencies and spread. It is easy for you to see that the shape of the distribution is influenced both by the central tendency and the dispersion.

You will be exposed to the popular measures of dispersion- *Range, interquartile range Mean Absolute Deviation (MAD), Standard Deviation, and coefficient of variation.*

Range

Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in the data set.

$$\text{Range} = X_{\text{Maximum}} - X_{\text{Minimum}}$$

Example for Computing Range

The following data represent the percentage return on investment for 10 mutual funds per annum. Calculate Range.

12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 9

$$\text{Range} = X_{\text{Maximum}} - X_{\text{Minimum}} = 18 - 9 = 9$$

Range is a good measure of spread in a distribution only when the data set shows a stable pattern of variation without extreme values. If one of the components of range, namely the maximum or minimum value becomes an extreme value, then range should not be used. The other weakness of range is that it is based only on two observations. Suppose you have 1000 observations measured on a particular characteristic, range will ignore 998 observations! Range is a popular measure of variation in quality control applications where you take samples of 4 or 5 observations at regular intervals of time.

Interquartile range Range is entirely dependent on maximum and minimum values in the data set and is highly misleading when one of them is an extreme value. To overcome this deficiency, you can resort to interquartile range. It is computed as the range after eliminating the highest and lowest 25% of observations in a data set that is arranged in ascending order. Thus, this measure is not sensitive to extreme values.

Interquartile range = Range computed on middle 50% of the observations.

Example for interquartile range The following data represent the percentage return on investment for 9 mutual funds per annum. Calculate interquartile range.

12, 14, 11, 18, 10.5, 12, 14, 11, 9

Arranging in ascending order, the data set becomes:

9, 10.5, 11, 11, 12, 12, 14, 14, 18

Ignore the first two (9, 10.5) and last two (14, 18) observations in this data set. The remaining contains 50% of the data. They are 11, 11, 12, 12, and 14. For this, if you calculate range, you get interquartile range.

Interquartile range = $14 - 11 = 3$.

Note The range for this problem is $18 - 10.5 = 7.5$. Interquartile range 3 is much smaller than the range 7.5 thus proving the point that it is less sensitive to extreme values.

Mean Absolute Deviation (MAD)

Mean Absolute Deviation (MAD) is defined as the average based on the deviations measured from arithmetic mean, in which all deviations are treated as positive ignoring the actual sign. Unlike range, MAD is based on all observations. Hence it reflects the dispersion of every item in the distribution. In symbolic form, it is defined by the following formula.

$$\text{MAD} = \frac{\sum |X - \bar{X}|}{n}$$

Where

$\sum |X - \bar{X}|$ represents sum of all deviations from arithmetic mean after ignoring sign

\bar{X} = Arithmetic Mean

n = Number of observations in the sample (sample size)

Caution Mean Absolute Deviation (MAD) has two weaknesses. 1) It cannot be combined for several groups. 2) Ignoring the sign has serious implications to a business manager attempting to measure the spread of the distribution in a scientific manner.

An example for calculating MAD The following data represent the percentage return on investment for 10 mutual funds per annum. Calculate MAD (Please note that this is the same example used for computing Range).

12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 9

Solution

$$\bar{X} = \frac{\sum X}{n} = (12 + 14 + 11 + 18 + 10.5 + 11.3 + 12 + 14 + 11 + 9)/10 = 12.28$$

$$\sum |X - \bar{X}| = |12 - 12.28| + |14 - 12.28| + |11 - 12.28| + |18 - 12.28| + |10.5 - 12.28| + |11.3 - 12.28| + |12 - 12.28| + |14 - 12.28| + |11 - 12.28| + |9 - 12.28| = 18.32$$

$$MAD = \frac{\sum |X - \bar{X}|}{n} = 18.32/10 = 1.832$$

Standard deviation Standard deviation forms the basis for the discussion on *Inferential Statistics*. It is a classic measure of dispersion. It has many advantages over the rest of the measures of variations. It is based on all observations. It is capable of being algebraically treated which implies that you can combine standard deviations of many groups. It plays a very vital role in testing hypotheses and forming confidence interval. You will appreciate the use of standard deviation in later chapters that deal with statistical inference.

To define standard deviation, you need to define another term called variance. In simple terms, standard deviation is the square root of variance. In order to get a complete picture, look at the following table of terms and notations.

Important Terms with notations	Key Remarks
Sample Variance $S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$	1. $S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$ is an unbiased estimator of $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$
Sample Standard Deviation $S = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$	
Population Variance = $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$	2. $\bar{X} = \frac{\sum X}{n}$ is an unbiased estimator of $\mu = \frac{\sum X}{N}$
Population Standard Deviation $\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$	

Important Terms with notations	Key Remarks
<p>Where $\bar{X} = \frac{\sum X}{n}$ (Sample Mean) and $\mu = \frac{\sum X}{N}$ (Population Mean)</p> <p>n = Number of observations in the sample (Sample size)</p> <p>N = Number of observations in the Population (Population Size)</p>	<p>3. The divisor $n-1$ is always used while calculating sample variance for ensuring property of being unbiased.</p> <p>4. Standard deviation is always the square root of variance.</p>

An example to calculate standard deviation The following data represent the percentage return on investment for 10 mutual funds per annum. Calculate the sample standard deviation (Please note that this is the same example used for computing Range and MAD).

12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 9

It is easy to compute the variance and standard deviation using Microsoft Excel spreadsheet. The Calculations are shown below:

A	B	C	D
1			
2	X	$X - \bar{X}$	$(X - \bar{X})^2$
3	12	-0.28	0.08
4	14	1.72	2.96
5	11	-1.28	1.64
6	18	5.72	32.72
7	10.5	-1.78	3.17
8	11.3	-0.98	0.96
9	12	-0.28	0.08
10	14	1.72	2.96
11	11	-1.28	1.64
12	9	-3.28	10.76
13	Mean =		56.96
14	12.28	Variance =	6.33
15		Standard Deviation =	2.52

- From the spreadsheet of Microsoft Excel, it is easy to see that Mean = $\bar{X} = \frac{\sum X}{n}$ =12.28 (In column B and row14, 12.28 is seen).

- Sample Variance = $S^2 = \frac{\sum (X - \bar{X})^2}{n-1} = 6.33$ (In column D and row 14, 6.33 is seen)
- Sample Standard Deviation = $S = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}} = 2.52$ (In column D and row 15, 2.52 is seen)

A comparative picture of the spread calculated by the three Measures of Dispersion For this illustration:

Dispersion Used	Calculated Value	Remarks
Range	9	Range is not reflecting the reality in this example. This is because one of its components bears the maximum value of 18, which is an extreme value.
Mean Absolute Deviation (MAD)	1.832	Mean Absolute Deviation (MAD) seems better than Range but ignoring the sign of deviation has serious implications to a business manager.
Standard Deviation	2.52	Standard Deviation is an elegant measure here. It brings out profoundly the risk factor in the return on investment of the mutual funds as 2.52. This means that the departure from an average return of 12.28 could be 2.52 on either side (+2.52 or -2.52).

All factors considered, Standard Deviation is the Measure of Dispersion you will be using as a manager while analyzing the spread or Scattering around the central tendency in any business analysis and problem solving exercise.

Standard deviation calculation for grouped data (Frequency Distribution)

The standard deviation for sample data, based on frequency distribution is given by

$$S = \sqrt{\frac{\sum f(X - \bar{X})^2}{n-1}}$$

which is used to estimate the

Population Standard Deviation.

Here $\bar{X} = \frac{\sum fX}{n}$

n is the Sample Size = $\sum f$, X = Mid point of each class

An example: Frequency distribution of return on investment of mutual funds

Percentage Return	Number of Mutual Funds
5-10	10
10-15	12
15-20	16
20-25	14
25-30	8
Total	60

You can do this calculation of Sample Standard Deviation using the spreadsheet of Microsoft Excel.

A	B	C	D	E	F	G	H
1	Return on	Investment		No. of			
2			MidPoint	Funds			
3	Lower limit	Upper Limit	X	f	fx	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
4	5	10	7.50	10	75	96.69	966.94
5	10	15	12.50	12	150	23.36	280.33
6	15	20	17.50	16	280	0.03	0.44
7	20	25	22.50	14	315	26.69	373.72
8	25	30	27.50	8	220	103.36	826.89
9				60	1040		2448.33
10				Mean =	17.333		
11				Sample Variance =			41.50
12				Sample Standard Deviation =			6.44

Note Mean = $\bar{X} = \frac{\sum fX}{n} = 1040/60 = 17.333$ (Cell F10),

$$\text{Standard Deviation} = S = \sqrt{\frac{\sum f(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{2448.33}{59}} = 6.44 \text{ (Cell H12)}$$

Coefficient of variation (Relative Dispersion) Coefficient of Variation (CV) is defined as

the ratio of Standard Deviation to Mean. In symbolic form, $CV = \frac{S}{\bar{X}}$ for sample data and

$\frac{\sigma}{\mu}$ for the population data. CV is the measure to use when you want to see the relative spread across groups or segments. It also measures the extent of spread in a distribution as a percentage to the mean. Larger the CV, greater is the percentage spread. As a manager, you would like to have a small CV so that your assessment in a situation is robust and the percentage risk is minimized.

An example for CV Consider two Sales Persons working in the same territory. The sales performance of these two in the context of selling PCs are given below. Comment on the results.

<i>Sales Person 1</i>	<i>Sales Person 2</i>
Mean Sales (One year average) 50 units	Mean Sales (One year average) 75 units
Standard Deviation 5 units	Standard Deviation 25 units

The CV is $5/50 = 0.10$ or 10% for the Sales Person1 and $25/75 = 0.33$ or 33% for Sales Person2. It seems Sales Person1 performs better than Sales Person2 with less relative dispersion or scattering. Sales Person2 has a very high departure or standard deviation from his average sales achievement. The moral of the story is "don't get carried away by absolute number". Look at the scatter. Even though, Sales Person2 has achieved a higher average, his performance is not consistent and is erratic.

3.3 CHAPTER SUMMARY

In this chapter, you have been exposed to the concepts and applications of the summary measures of central tendency and measures of dispersion along with their strengths and limitations. To be specific, this chapter focused on the following:

- The meaning, definition, and role of the three measures of central tendency- arithmetic mean, median, and mode along with the calculations of these three for raw data and grouped data.
- The meaning, definition, and role of the measures of dispersion - range, interquartile range, mean absolute deviation(MAD), and standard deviation along with the calculation of standard deviation for raw data and grouped data.
- The definition and use of coefficient of variation(CV) with an example.
- Spreadsheet methodology for calculations of summary measures wherever required.

GLOSSARY

Central Tendency Whenever we measure things of the same kind, a fairly large number of such measurements will tend to cluster around the middle value. This is known as central tendency.

Coefficient of Variation (Relative Dispersion) Coefficient of Variation (CV) is defined as the ratio of Standard Deviation to Mean. CV measures the relative spread across groups or segments.

Dispersion (Variation) Dispersion indicates how large the spread of the distribution is around the central tendency.

Interquartile Range It is the range of numbers after eliminating the top 25% and bottom 25% of the observations in a given data set that is arranged in ascending order.

Mean (Arithmetic Mean) Mean (Arithmetic Mean) is the most common measure of central tendency used by all managers in their sphere of activities. It is defined as the sum of all observations in a data set divided by the total number of observations.

Mean Absolute Deviation Mean Absolute Deviation (MAD) is a measure of dispersion that is computed as the average based on the deviations measured from arithmetic mean in which all deviations are treated as positive, ignoring the actual sign.

Measure of Central Tendency If one typical representative average is defined in such a manner that the remaining items in the data set will cluster around this value, such a value is called a measure of "Central Tendency".

Measure of Dispersion A measure that assesses the magnitude of departure from the average value (central tendency) is called measure of dispersion.

Median Median is the middle most observation when we arrange data in ascending or descending order of magnitude. Median is such that 50% of the observations are above the median and 50% of the observations are below the median.

Mode Mode is that value which occurs most often. It has the maximum frequency of occurrence.

Range It is calculated as the difference between maximum and minimum value in a data set.

Symmetrical It is a property of a distribution in which the area covered in one half portion is the mirror image of the other half portion.

Standard Deviation Standard deviation is the positive square root of variance. Its unit of measurement is same as that of the original data.

Variance It is a measure of dispersion that is computed as the average sum of square of each item from the mean in a data set.

REVIEW QUESTIONS

Questions 1 to 3 refer to the following data:

2, 6, 5, 3, 8, 2, 1, 9, 7, 7, 6, 4

1. The arithmetic mean is:
(a) 4 (b) 5.5 (c) 4.25 (d) 5
2. The median is:
(a) 4 (b) 5.5 (c) 4.25 (d) 6
3. Which of the following is always true? As a measure of central tendency, the arithmetic mean:
(a) is generally a misleading measure of central tendency
(b) is always superior to median and mode
(c) is preferred to median and mode except when all the three measures are equal
(d) should not be used when the data set contains extreme values

Questions 4 to 6 should be answered based on the following situation:

Sales Target (Rs Lacs)	Number of Times Achieved
10-20	6
20-30	8
30-40	12
40-50	9
50-60	5

4. The arithmetic mean is:
 (a) 36 (b) 35 (c) 34.75 (d) 37
5. The median is:
 (a) 35 (b) 36 (c) 34.5 (d) 35.5
6. The mode is:
 (a) 35 (b) 36 (c) 35.71 (d) 36.5
7. While calculating mean absolute deviation (MAD), the actual signs are ignored. This will never have any significant adverse effect in a real business situation. True or False.

Questions 8-10 refer to the following situation:

Bulb life in hours of two suppliers A and B are given below based on a sample of 9 bulbs for each supplier.

Supplier	Bulb Life (hours)
A	40, 45, 50, 45, 40, 35, 55, 40, 60
B	35, 55, 65, 75, 45, 50, 70, 65, 30

8. The range of supplier A is more than supplier B. True or False.
9. The standard deviation of supplier B is more than the standard deviation of supplier A. True or False.
10. If you take coefficient of variation (CV) as the basis of comparison, which one of the following is true?
 (a) Supplier B is better than Supplier A
 (b) Supplier A has the same CV as Supplier A
 (c) Supplier A is superior to Supplier B

ANSWERS TO REVIEW QUESTIONS

1. Correct answer is d). Let us calculate arithmetic mean to justify this choice. Mean

$$= \bar{X} = \frac{\sum X}{n} = (2 + 6 + 5 + 3 + 8 + 2 + 1 + 9 + 7 + 7 + 6 + 4)/12 = 60/12 = 5.$$

Obviously, other choices are wrong.

2. Choice (b) is the correct answer. Arranging the data in the ascending order, we have the following array.

1, 2, 2, 3, 4, 5, 6, 6, 7, 7, 8, 9. The middle value is the median. Since, this is an even sample size case, the average of the 6th and 7th value = median. That is, median = $(5 + 6)/2 = 5.5$. The other choices are, therefore, rejected.

3. (d) is the right choice because arithmetic mean is affected by extreme values present in the data set. (a) is not correct because arithmetic mean is generally a good measure of central tendency possessing outstanding properties. Its singular weakness is that it is affected by extreme values (very high or very low). (b) is not correct because you can not say it is always superior to median and mode. When ranking is involved median is superior to arithmetic mean. Whenever a particular value has the largest frequency of occurrence, mode is the preferred measure. By this same logic choice (c) is rejected. Choice (d) is always true and hence, correct.

4. (c) is the right answer. Calculations are shown below:

<i>Class</i>	<i>MidPoint</i>	<i>Frequency</i>	
	<i>X</i>	<i>f</i>	<i>fx</i>
10-20	15	6	90
20-30	25	8	200
30-40	35	12	420
40-50	45	9	405
50-60	55	5	275
		40	1390
		Mean =	34.75

$$\text{Mean} = \bar{X} = \frac{\sum fX}{n} = 1390/40 = 34.75 \text{ (See calculations done on the spreadsheet above)}$$

5. (a) is the right choice.

Please note here that median lies in the class 30-40 because $n/2 = 40/2 = 20$ implies that median is in this class (cumulative frequency is 26 for class 30-40).

<i>Class</i>	<i>Frequency</i>	<i>Cumulative</i>
	<i>f</i>	<i>Frequency</i>
10-20	6	6
20-30	8	14
30-40	12	26
40-50	9	35
50-60	5	40

Applying the formula for median using grouped data in the usual notation, we have

$$\text{Median} = L + \frac{(n/2 - m)}{f} \times c. \quad L = 30, n = 40, m = 14, f = 12 \text{ and } c = 10. \text{ Substituting, we}$$

have median = $30 + \frac{(40/2 - 14)}{12} \times 10 = 35$. Hence (a) is the answer. The other choices are rejected automatically.

6. (c) is the right choice. Let us see the calculation to justify the answer.

You see that the largest number falls in class 30-40 (frequency =12). Hence, mode lies in this class.

Class	Frequency <i>f</i>
10-20	6
20-30	8
30-40	12
40-50	9
50-60	5

Applying the formula for mode using grouped data in the usual notation, we have

$$\text{Mode} = L + \frac{d_1}{d_1 + d_2} \times c = 30 + \frac{4}{4 + 3} \times 10 = 35.71 \text{ on simplification. (Please note that}$$

$d_1 = f_1 - f_0 = 12 - 8 = 4$. $d_2 = f_1 - f_2 = 12 - 9 = 3$). Hence, (c) is the correct answer. Other choices are, obviously, wrong.

7. The given statement is false because ignoring the sign while computing the mean absolute deviation (MAD) has serious implications to managers. The risk factor increases with regard to decision making, particularly when the deviation is large. Hence, ignoring the sign does impact the decision at times seriously.
8. The given statement is false. Let us see the calculation to justify our choice. Range = Maximum Value - Minimum Value in any data set. Range for Supplier A = 60-35 = 25. Range for Supplier B = 75-30 =45. We see that Supplier B has a larger range than Supplier A.
9. True. Let us see how this statement is true.

Sample Standard Deviation in the usual notation is given by $S = \sqrt{\frac{(X - \bar{X})^2}{n - 1}}$.

		<i>Supplier A</i>		<i>Supplier B</i>	
		<i>X</i>	$(X - \bar{X})^2$	<i>X</i>	$(X - \bar{X})^2$
		40	30.86	35	378.09
		45	0.31	55	0.31
		50	19.75	65	111.42
		45	0.31	75	422.53
		40	30.86	45	89.20
		35	111.42	50	19.75
		55	89.20	70	241.98
		40	30.86	65	111.42
		60	208.64	30	597.53
Sum =		410.00	522.22	490	1972.22
Mean = \bar{X} =		45.56		54.44	
Standard Deviation $S = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$ =			8.08		15.70

From the calculations above, it is easy to see that the standard deviation of supplier B(15.70) is more than the standard deviation of supplier A (8.08). Hence the statement is true.

10. (c) is the right choice. Coefficient of Variation (CV) = Standard Deviation/Mean.

For Supplier A, $CV = 8.08/45.56 = 0.1773$ or 17.73%. For Supplier B, $CV = 15.70/54.44=0.2883$ or 28.83%. Supplier A is superior to Supplier B based on CV. Hence, (c) is the right choice.

PRACTICE PROBLEMS

1. A factory has two units producing and packing a particular type of Ice Cream in suitable boxes. These boxes are specially designed to maintain cool temperature so that they comfortably reach the destinations, namely, ice cream parlors. The process time to produce, pack, and ship the ice cream for these two units are given below:

<i>Unit 1</i>	<i>Unit 2</i>
11	19
11	23
33	33
22	25
23	52

Unit 1	Unit 2
43	31
17	29
17	26
11	27
23	20
23	12
15	25
15	18
16	26
9	12
21	5
15	29
19	11
15	13
18	19

- (a) Compute the measures of Central Tendency (Arithmetic Mean, Median, and Mode) for the two units and interpret the results.
- (b) Compute the Standard Deviation and Coefficient of Variation for the two units. Which shows better performance- Unit 1 or Unit 2?

2. Mr. Chopra is a salesperson working in a company. In the context of sales, the amount spent by him in his five trips in the past two months throws the following figures:

Trip No	Days	Amount Spent (Rs)	Amount per day (Rs)
1	3	810	270
2	9	1620	180
3	1	700	700
4	10	1700	170
5	5	900	180
Totals	28	5730	1500

Mr. Chopra's boss felt that his expenses were very high compared to the normal spending. He said, the average spending per day of Mr. Chopra was Rs. 300, while other salespersons spend on an average Rs 250 a day. Explain the fallacy in this problem.

3. The following data represent the monthly income of workers in two factories.

Income	1800-1900	1900-2000	2000-2100	2100-2200	2200-2300	Total
Factory A	24	39	64	36	22	185
Factory B	34	48	75	45	30	232

Compute mean, median, and mode for the two factories and interpret the results.

64 Business Statistics

4. Samples of light bulbs were bought from two suppliers and were subjected to destruction test in the lab. The following data were collected on the life:

<i>Life in hours</i>	<i>700-800</i>	<i>800-900</i>	<i>900-1000</i>	<i>1000-1100</i>	<i>Total</i>
Supplier A	14	74	29	13	130
Supplier B	12	58	32	18	120

- (a) Which supplier provides greater average life? Supplier A or Supplier B
 (b) Which supplier provides uniform quality? Supplier A or Supplier B
 (c) Which supplier would you prefer? Supplier A or Supplier B
5. The distance between residence and office for company executives in a city according to a market survey is found to be as follows:

<i>Distance in km</i>	<i>% Executives</i>
0	1
1	8
2	12
3	15
4-5	18
5-7	10
7-10	10
10-14	6
14-20	6
20 and above	14

The mean distance from residence to office is estimated to be 8.5 km. Calculate the mode, the median and standard deviation.

Probability—A Conceptual Framework

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Appreciate the use of probability in decision making
- Explain the types of probability
- Define and use the various rules of probability depending on the problem situation
- Appreciate the use of Probability Tree

CHAPTER OUTLINE

- 4.1 Meaning and Concepts of Probability
- 4.2 Types of Probability
- 4.3 Mutually Exclusive Events
- 4.4 Independent Events
- 4.5 Rules for Calculating Probability
- 4.6 Use of Probability Tree
- 4.7 Chapter Summary
- Glossary
- Review Questions
- Answers to Review Questions
- Practice Problems

INTRODUCTION

Managers need to cope with uncertainty in many decision situations. For example, you as a manager may assume that the volume of sales in the next year is known to you exactly. This is not true because you know roughly what the next year sales will be, but you cannot give the exact number. There is some uncertainty. Concepts of probability will help you measure uncertainty and perform associated analyses. This chapter provides the conceptual framework of probability and the various probability rules that are essential in business decisions.

4.1 MEANING AND CONCEPTS OF PROBABILITY

- In simple terms, *probability* refers to chance or likelihood of a particular *event*-taking place. For example, you may like to ask, "what is the chance that the company will achieve a sale of more than Rs. 50 lakhs in the coming quarter?"
- An *event* is an outcome of an *experiment*.
- An *experiment* is a process that is performed to understand and observe possible outcomes. Examples include tossing a coin, drawing a card from a well-shuffled pack of cards, or rolling a pair of dice.
- Set of all outcomes of an experiment is called the *sample space*.

An Example

In a manufacturing unit three parts from the assembly are selected. You observe whether they are defective or non-defective. Determine:

- (a) The sample space.
- (b) The event of getting at least two defective parts.

Solution

Let defective be designated as D and Non-defective be designated G.

- (a) Let S = Sample Space. It is pictured as under:



In symbols, this is given by the following set:

$$S = \{GGG, GGD, GDG, DGG, GDD, DGD, DDG, DDD\}$$

Here, the sample space is the set of all outcomes of the experiment of observing all the three parts for defective or non-defective. GGG represents all the three are non-defective, GGD represents first is non-defective, second is non-defective, and third is defective, and so on. Thus, there are eight possibilities.

- (b) Let E denote the event of getting at least two defective parts. This implies that E will contain two defectives, and three defectives. Looking at the sample space diagram above, $E = \{GDD, DGD, DDG, DDD\}$. It is easy to see that E is a part of S and commonly called as a subset of S. Hence, an event is always a subset of the sample space.

Progressive Test Question

What is the sample space when you toss the coin twice? Use the symbol T =Tail and H = Head.

- (a) $S = \{TH\}$
 (b) $S = \{HH\}$
 (c) $S = \{TT\}$
 (d) $S = \{HT, HH, TH, TT\}$

You try to do it first and then check with the correct answer given below.

Answer The correct answer is (d). There are four possibilities in this experiment of tossing the coin twice. It would be head and tail, head and head, tail and head, tail and tail. This is shown in choice (d). Choice (a), choice (b) and choice (c) are only partially correct and hence, rejected.

Definition of Probability Probability of an event A is defined as the ratio between two numbers m and n . In symbols,

$$P(A) = \frac{m}{n}$$

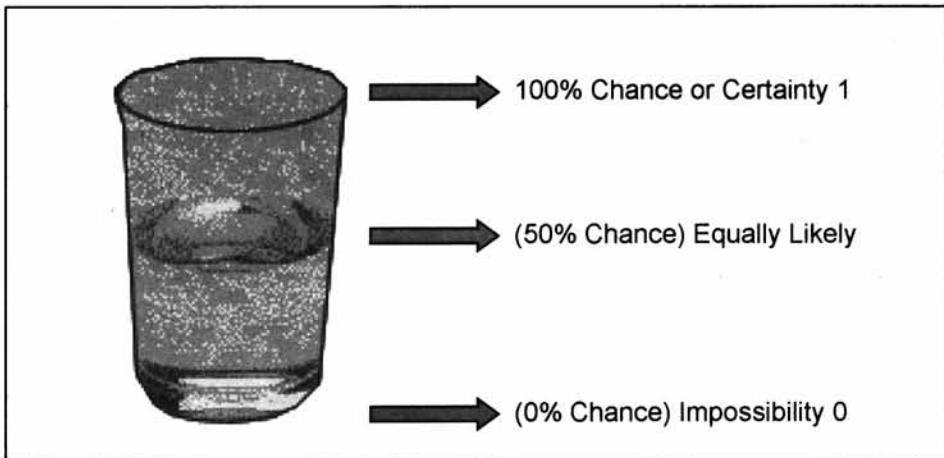
Where

m = number of ways that are favorable to the occurrence of A and

n = the total number of outcomes of the experiment (all possible outcomes)

Please note that $P(A)$ is always ≥ 0 and always ≤ 1 . $P(A)$ is a pure number.

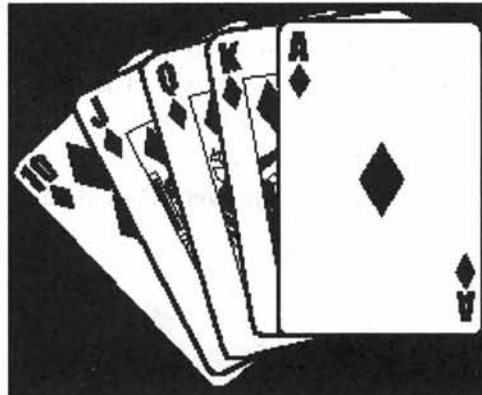
The range within which probability of an event lies can be best understood by the following diagram. The glass shows three stages (for the purpose of illustration)- Empty, half-full, and full to explain the properties of probability.



Man giving birth to a child is an impossibility. Sun rises in the east is a certainty. Getting a head in one toss of a coin is 50%. Likewise, getting tail in one toss of a coin is

50%. Of course, probability of an event can take any value between 0 and 1. The above picture shows three typical and distinct values of probability of an event.

Example From a well-shuffled pack of cards, if you pick up one card at random, what is the probability that the selected card is a King?



Let K = Getting a King

m = number of ways in which King can occur = 4 (There 4 Kings in a pack of cards)

n = Total number Possibilities (Sample Space) = 52 (There are 52 cards)

$$P(K) = \frac{4}{52} = \frac{1}{13}.$$

Note Random means that every unit in the population has the same chance of being selected. Here, it implies that while picking up a card, every one of the 52 cards in the pack has the same probability of being selected. Another term that is used in this context is "equally likely".

Progressive Test Question A quality assurance manager says "I am 200 percent certain that our product will meet the specification of the customer". Is this statement true or false?

Answer The statement is false because probability cannot be more than 100 percent. When the event is a certainty, probability equals 1 (100%). Probability is always between 0 and 1.

4.2 TYPES OF PROBABILITY

There are three types of probability, which you should be aware of. They are:

1) Classical Probability

Classical Probability is same as the one we have used while defining probability in the beginning. This is called *a priori probability* because you can compute the answer in

advance without actually having to perform any experiment. Consider for example probability of getting a King from a pack of cards. You can compute the answer in advance as $4/52$. Even though classical probability comprehensively defines probability, its use in business analysis for assessing uncertainty is seriously handicapped because you cannot compute probability in advance with out performing trials (experiments), or observing the behavior of the random variable. Here, *random variable* means a variable that takes many values with corresponding probabilities.

2) Relative Frequency Probability

Relative Frequency Probability

"What is the chance that our company will achieve a sale of more than 15000 tons of specialty paper?". "What is the probability that you will get a score of at least 75% in the course Quantitative Methods in your MBA program?". "What is the chance that there will be no stock out problem this year?". The answers to these questions cannot be obtained in advance with out performing experiments or using the past data. The tool organizations deploy is a typical frequency distribution of the past data expressed as a relative frequency that was discussed in Chapter 2. For example, your company can use the past data on sales for assessing the probability of achieving a particular target. All you need to do is to convert the past sales figures into a well-structured frequency distribution containing target range in every class and the number of times that was achieved in every class. If you now express relative frequency either in proportion or in percentages to the total frequency, you get the probability of achieving particular target sales.

Please note that relative frequency is a pragmatic way of assessing probability in real life business situations.

3) Subjective Probability

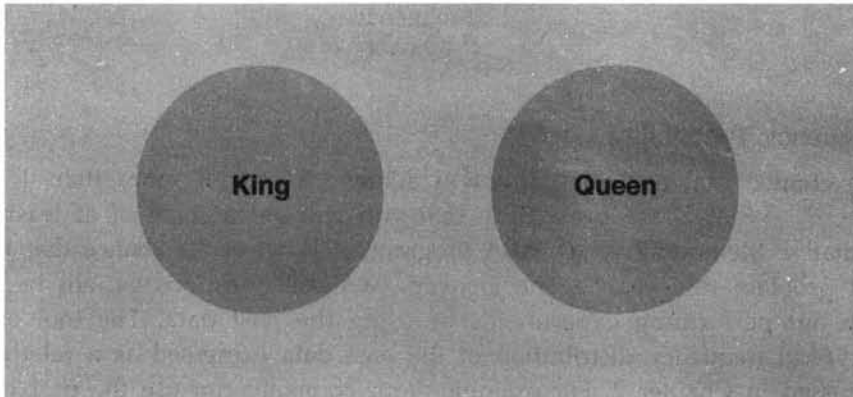
Subjective Probability

When you cannot get probability worked out either based on classical approach or based on relative frequency of past data, the only option available to you is to use subjective probability. Assigning subjective probabilities to various events is based on some past experience, personal opinion, and an analysis of the business situation. Survey of Experts' Opinion for knowing the chances of certain key events affecting business is an example of subjective probability.

The major drawback of the subjective probability is that it can be tremendously influenced by personal prejudices, likes, and dislikes.

4.3 MUTUALLY EXCLUSIVE EVENTS

Two events A and B are said to be mutually exclusive if the occurrence of A precludes the occurrence of B . For example, from a well shuffled pack of cards, you pick up one card at random and would like to know whether it is a King or a Queen. The selected card will be either a King or a Queen. It cannot be both a King and a Queen. If King occurs, Queen will not occur and Queen occurs, King will not occur. This is shown in the following diagram known by the name, Venn diagram. You will appreciate that there are no common elements shared between the events King and Queen.



The principle of “mutually exclusive” can be extended to more than two events. The occurrence of one event precludes the occurrence of the remaining events.

4.4 INDEPENDENT EVENTS

Two events A and B are said to be independent if the occurrence of A is in no way influenced by the occurrence of B . Likewise, occurrence of B is in no way influenced by the occurrence of A .

For example, when you toss a coin twice, getting head in the second trial is in no way influenced by what happened in the first toss. In the first toss you might have got a head or a tail. In other words, the outcome of the second toss has nothing to do with what happened in the first toss.



The concept of independence can be extended to more than two events. In such situations, the occurrence of a particular event is in no way influenced by the occurrence of the remaining events.

Progressive Test Question In a market survey, the respondents were asked, "Do you live in a flat or in an independent house? This is an example of mutually exclusive events. Is it True or False?

Answer True because the occurrence of one event precludes the occurrence of the other. The respondent can either live in a flat or in a house. You cannot live at the same time in both the places.

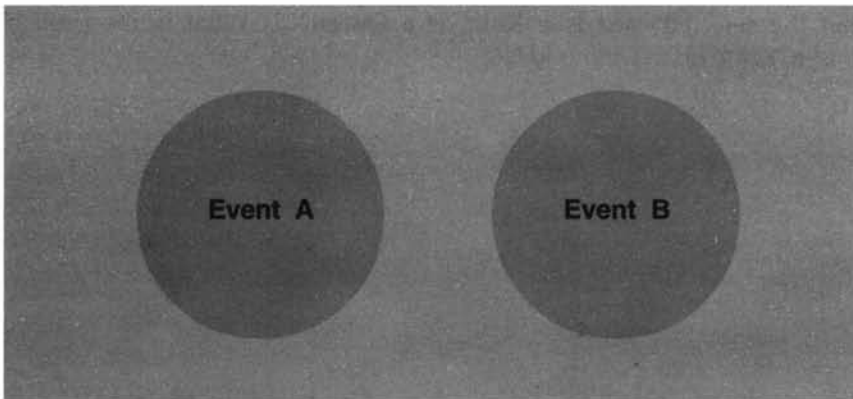
4.5 RULES FOR CALCULATING PROBABILITY

1. Addition Rule – Mutually Exclusive Events

$$P(A \cup B) = P(A) + P(B)$$

This rule says that the probability of the union of A and B is determined by adding the probability of the events A and B .

Here, the symbol $A \cup B$ is called A union B meaning A occurs, or B occurs or both A and B simultaneously occur. When A and B are mutually exclusive, A and B cannot simultaneously occur. The Venn diagram depicting this is as under. This is similar to the example given for mutually exclusive events discussed earlier. There are no common elements for A and B when they are mutually exclusive.

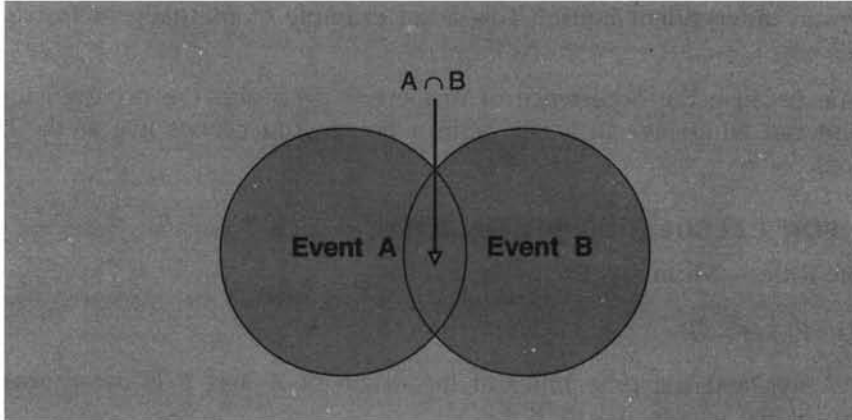


2. Addition Rule – Events are not Mutually Exclusive

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Here, the symbol $A \cup B$ is called A union B meaning A occurs, or B occurs or both A and B simultaneously occur. $A \cap B$ is called A intersection B meaning both A and B simultaneously occur.

The above addition rule says the probability of the union of A and B is determined by adding the probability of the events A and B and then subtracting the probability of the intersection of the events A and B . The Venn diagram depicting $A \cup B$ is given below:

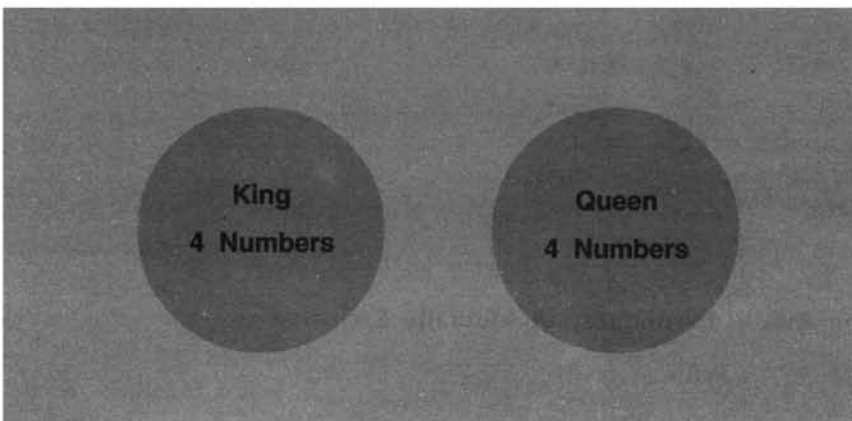


You will appreciate from the diagram that there are common elements between A and B designated as $A \cap B$. The subtraction of the term $A \cap B$ is because of double counting. This is included in event A as well as in event B .

Example Problem for Addition Rule

From a pack of well-shuffled cards, a card is picked up at random. 1) What is the probability that the selected card is a King or a Queen? 2) What is the probability that the selected card is a King or a Diamond?

Solution to 1.



Let A = getting a King

Let B = getting a Queen

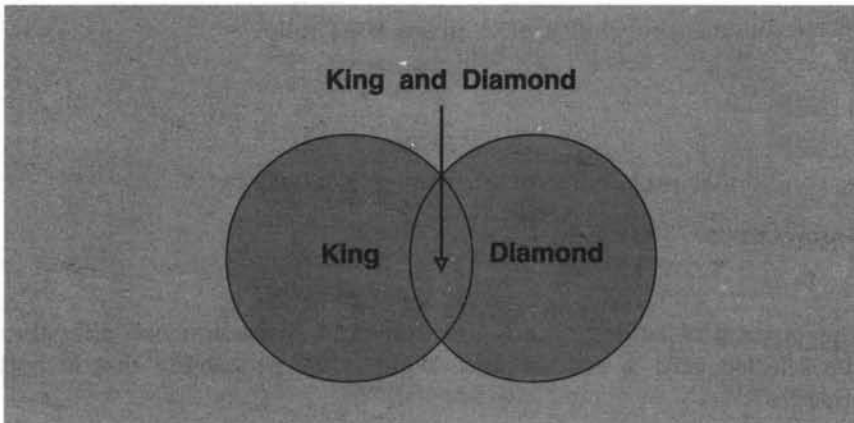
There are 4 kings and there are 4 Queens. The events are clearly mutually exclusive.

Applying the formula $P(A \cup B) = P(A) + P(B) = 4/52 + 4/52 = 8/52 = 2/13$

Hence the answer is $2/13$.

Solution to 2. There are totally 52 cards in a pack out of which 4 are Kings and 13 are Diamonds. Let A = getting a King and B = getting a Diamond. The two events here are not mutually exclusive because you can have a card, which is both a King and a Diamond called King Diamond. Please see diagram below: Applying the formula for non-mutually exclusive case, namely, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, we have $P(A) = 4/52$, $P(B) = 13/52$, and $P(A \cap B) = 1/52$ (chance of getting both a King and a Diamond). Substituting and simplifying, we have:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 4/52 + 13/52 - 1/52 = 16/52 = 4/13$$



3. Multiplication Rule -Independent Events

$$P(A \cap B) = P(A) \cdot P(B)$$

This rule says when the two events A and B are independent, the probability of the simultaneous occurrence of A and B (also known as probability of intersection of A and B) equals the product of the probability of A and the probability of B . Of course, this rule can be extended to more than two events.

Example The probability that you will get an A grade in Quantitative Methods is 0.7. The probability that you will get an A grade in Marketing is 0.5. Assuming these two courses are independent, compute the probability that you will get an A grade in both these subjects.

Solution Let A = getting A grade in Quantitative Methods

Let B = getting A grade in Marketing

Answer The correct choice is (c). Let us see how this is the right choice. This problem involves application of conditional probability because late and missing the bus are not independent.

Let A = Miss the bus

$$P(A) = 1 - \text{probability of catching the bus} = 1 - 0.30 = 0.70$$

Let B = Late

$$P(B/A) = 0.60 \text{ (60\% chance of being late/miss the bus)}$$

$$\text{Probability that you will be late} = P(A \cap B) = P(A) \cdot P(B/A) = 0.70 \cdot 0.60 = 0.42 = 42\%$$

Choice (a) is wrong because it only shows that there is 70% chance that you will miss the bus. Choice (b) is wrong because it only shows that there is a 60% chance of being late given the fact that you miss the bus. Hence (c) alone is the correct answer.

5. Marginal, Joint and Conditional Probability using a Contingency Table

Contingency table consists of rows and columns of two attributes at different levels with frequencies or numbers in each of the cells. It is a matrix of frequencies assigned to rows and columns.

The term marginal is used to indicate that the probabilities are calculated using a contingency table (also called joint probability table). *Marginal probability* means the probability of a simple event like $P(A)$ or $P(B)$.

Joint probability is the simultaneous occurrence of events. Probability that both A and B occur ($P(A \cap B)$) is an example of joint probability.

Conditional probability like $P(A/B)$ and $P(B/A)$ have already been covered.

Let us take an example to calculate these probabilities from a contingency table.

A survey involving 200 families was conducted. Information regarding family income and whether the family buys a car are given in the following table.

<i>Family</i>	<i>Income below Rs 4 lakhs/year</i>	<i>Income of Rs. 4 lakhs and above</i>	<i>Total</i>
Buyer of Car	38	42	80
Non-Buyer	82	38	120
Total	120	80	200

- What is the probability that a randomly selected family is a buyer of the car?
- What is the probability that a randomly selected family is both a buyer of car and belongs to income of Rs. 4 lakhs and above?
- A family selected at random is found to be belonging to income of Rs 4 lakhs and above. What is the probability that this family is buyer of car?

Solution to the Problem

<i>Family</i>	<i>Income below Rs 4 lakhs/year</i>	<i>Income of Rs. 4 lakhs and above</i>	<i>Total</i>
Buyer of Car	38	42	80
Non-Buyer of Car	82	38	120
Total	120	80	200

- (a) What is the probability that a randomly selected family is a buyer of the Car?

Look at the contingency table above. There are 80 families in total are buyers of a car. There are 200 families in the survey. Hence, the probability that a randomly chosen family is a buyer of the car $=80/200 = 0.40$. **Note** This is a case of *marginal probability* of a family buying a car.

- (b) What is the probability that a randomly selected family is both a buyer of car and belongs to income of Rs. 4 lakhs and above?

Again look at the table above. Both a buyer of car and an income of Rs 4 lakhs and above = 42 families. Total families in the survey =200. Hence the probability that a randomly selected family is both a buyer of car and belongs to income of Rs. 4 lakhs and above $=42/200 = 0.21$. **Note** This is a case of *joint probability (buyer and Rs. 4 lakhs and above)*

- (c) A family selected at random is found to be belonging to income of Rs 4 lakhs and above. What is the probability that this family is buyer of car?

A family is selected and found to be having an income of Rs 4 lakhs and above. In this category there are totally 80 families out of which 42 are buyers. Hence the probability $= 42/80 = 0.525$. **Note** This is a case of *conditional probability of buyer given income is Rs. 4 lakhs and above.*

4.6 USE OF PROBABILITY TREE

Probability concepts are more effectively understood using a probability tree as the basis. The branches of the trees represent the probabilities of associated events. The tree can solve intriguing problems including Bayes' Theorem of posterior analysis. Posterior analysis is a process by which the probabilities are revised based on new or additional information. Bayes' Theorem is essentially a case of conditional probability. Let us look at an example to comprehend the power of the probability tree including Bayes' Method.

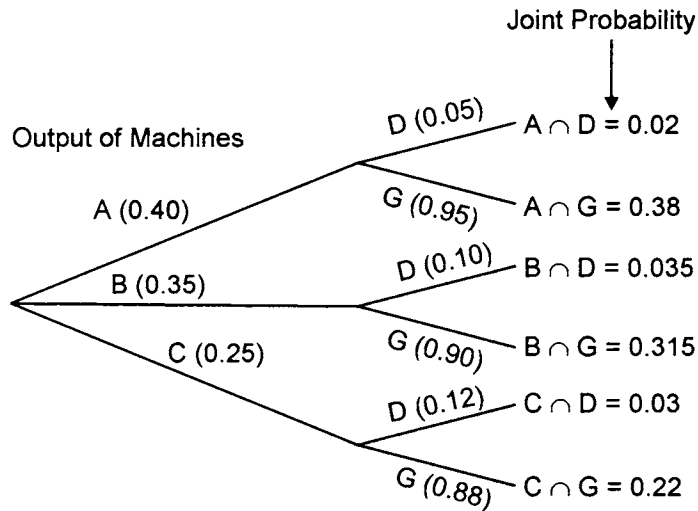
There are three Machines designated as *A*, *B*, and *C* producing the same item. The output of *A*, *B*, and *C* are 40%, 35% , and 25% respectively. *A* produces 5 % defectives, *B* produces 10% defectives and *C* produces 12% defectives. Draw the probability tree and then answer the following;

- (a) If an item is selected from the total output at random, what is the probability that it is defective?
- (b) If an item is selected from the total output at random, what is the probability that it is non-defective (good piece)?
- (c) An item is selected and found to be defective. What is the chance that it is produced from *A*? Produced from *B*? Produced from *C*?

Solution to the Problem Let *D* = Defective. Let *G* = Non-Defective (Good).

The following tree is drawn. In the last part towards the end of the tree, each branch indicates the joint probability. For example, $A \cap D$ means Machine *A* and Defective and the probability of this happening = $0.02(0.4 \text{ multiplied by } 0.05 = 0.02)$. Likewise, all probabilities are computed for all branches.

Probability Tree-Example Problem

**Solution details**

- (a) If an item is selected from the total output at random, what is the probability that it is defective?

Look at the tree. Defective occurs in three places. *A* and Defective, *B* and Defective, *C* and Defective. Adding these three joint probabilities, namely $0.02 + 0.035 + 0.03$ you get the answer. The answer is 0.085 or 8.5%. That is $P(D) = 0.085$ (Marginal Probability of *D*)

- (b) If an item is selected from the total output at random, what is the probability that it is non-defective (good piece)?

Again look at the tree. There are three branches in which you find good pieces. *A* and Good, *B* and Good, *C* and Good. Adding these three joint probabilities, you get the answer. That is adding $0.38 + 0.315 + 0.22 = 0.915$ or 91.5%. This is the answer. (Note: $P(G) = 0.915$ is the marginal probability of *G*)

- (c) An item is selected and found to be defective. What is the chance that it is produced from *A*? Produced from *B*? Produced from *C*?

This part involves the application of Bayes' Theorem of computing conditional probability. From the tree, computing these probabilities is very simple.

Let us first take Machine *A*. You want $P(A/D) = \frac{P(A \cap D)}{P(D)} = 0.02/0.085 = 0.23529$.

Likewise, for Machine *B*, $P(B/D) = \frac{P(B \cap D)}{P(D)} = 0.035/0.085 = 0.41177$.

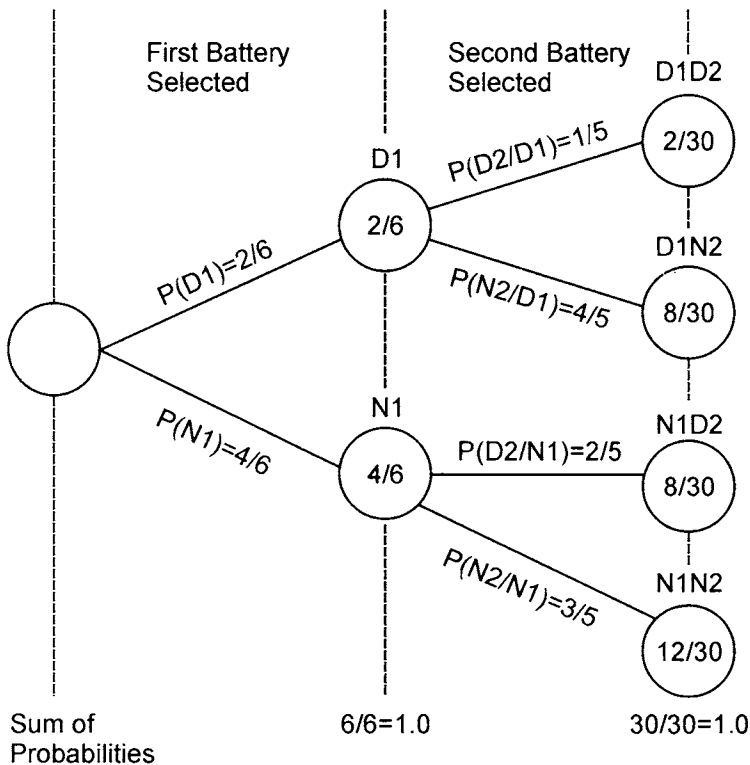
For Machine C, $P(C/D) = \frac{P(C \cap D)}{P(D)} = 0.03/0.085 = 0.35294$.

Comprehensive Example

There are 6 batteries in an equipment. Two of them are defectives but you do not know which of the two are defectives. The only way to know is to remove the batteries and test them one by one using a testing device. Assume that after each battery is tested, it is not replaced. Draw the probability tree and then answer the following:

1. What is the probability that one battery is defective?
2. What is the probability that at least one battery is defective?
3. Suppose the first battery selected is defective. What is the probability that the second battery is defective?

Solution Look at the tree drawn below. All questions can be answered.



In the picture above, the symbols are designated as under:

$D1$ = First battery defective, $N1$ = first battery non-defective

$D2$ = Second battery defective and $N2$ = Second battery non-defective

1. Probability that one battery is found to be defective corresponds to the outcome D1N2 and N1D2. This = $(8/30 + 8/30) = 16/30$.
2. Probability of at least one battery is defective is = sum of probabilities of one defective battery and two defective batteries. This corresponds to D1N2, N1D2, and D1D2. This = $(8/30 + 8/30 + 2/30) = 18/30$.
3. This requires the conditional probability the second battery is defective given that the first battery is defective. That is, we need $P(D2/D1)$. This is = $1/5$ which can be directly read from the diagram.

Discussion Topic

A company is currently interested in launching a new product. This product, if introduced, will be the first of its kind in the marketplace. The problem facing the company is a reasonable assessment of the demand potential for this item. This is really a challenging problem to the company in view of the fact that no past data are available. The company is interested in estimating the probability associated with various demand levels. Discuss how would you proceed with this task. Make sure you highlight the strengths and limitations of the methods you propose.

4.7 CHAPTER SUMMARY

This chapter provided a conceptual framework on probability concepts with examples. Specifically this chapter focused on:

- The meaning and definition of the term Probability interwoven with other associated terms-Event, Experiment, and Sample Space
- The three types of Probability - Classical Probability (A Priori Probability), Relative Frequency Probability, and Subjective Probability
- The concept of Mutually Exclusive Events and Independent Events
- The Rules for Calculating Probability which include Addition Rule for Mutually Exclusive Events and Non-Mutually Exclusive Events, Multiplication Rule for Independent Events and Non-Independent (Dependent) Events, and Conditional Probability
- Contingency Table to explain and calculate Marginal Probability, Joint Probability, and Conditional Probability
- The use of Probability Tree to solve intriguing problems including the application of Bayes' Theorem in a pragmatic manner

GLOSSARY

A Priori Probability This is a probability that you can compute in advance without actually having to perform any experiment.

Bayes' Theorem Bayes' Theorem is essentially a case of conditional probability when the events are dependent.

Classical Probability The number of cases that are favorable to the occurrence of an event divided by the total number of cases in an experiment.

Conditional Probability It is the probability of an event occurring given the fact that another event has taken place.

Contingency Table Contingency table consists of rows and columns of two attributes at different levels with frequencies or numbers in each of the cells. It is a matrix of frequencies assigned to rows and columns.

Equally Likely Equally likely means every event in an experiment has the same chance of being selected.

Experiment An experiment is a process that is performed to understand and observe possible outcomes.

Independent Events Two events are said to be independent if the occurrence of one event is in no way influenced by the occurrence of another event.

Joint Probability Joint probability is the simultaneous occurrence of events. Probability of two events taking place in succession is an example of joint probability.

Marginal Probability *Marginal probability* is the probability of a simple event. That is, it represents the probability of a single event.

Mutually Exclusive Events The occurrence of one event precludes the occurrence of another event.

Posterior Probability Posterior probability is the probability that is revised based on new or additional information.

Probability Probability refers to chance or likelihood of a particular event-taking place.

Probability Tree A visual display of events shown as a tree. The branches of the tree represent the probabilities of associated events.

Relative Frequency Probability Relative frequency when expressed either in proportion or in percentages to the total frequency, it becomes relative frequency probability.

Random Variable A variable that takes many values with corresponding probabilities.

Sample Space Set of all outcomes of an experiment is called the sample space.

Subjective Probability Subjective probability is based on some past experience, personal opinion, and an analysis of the business situation as perceived by a person who is assessing the probability.

Venn Diagram A visual display in which the sample space is represented by a rectangle and the events in the sample space represented by circles. Its purpose is to facilitate understanding of probability in intriguing problems.

REVIEW QUESTIONS

1. A box contains 12 soft drink bottles, 4 of which are defectives. A bottle is selected at random and is found to be defective; it is not replaced. What is the probability that the next bottle selected will also be defective?
(a) 0.19 (b) 0.20 (c) 0.30 (d) 0.27

2. A single die is rolled. What is the probability of getting 4?
 (a) $2/6$ (b) 0 (c) $4/6$ (d) $1/6$

Questions 3 to 6 will have to be answered with reference to the following contingency table:

A survey of 300 families was conducted to study income level versus brand preference.

<i>Income \ Brand Preferred</i>	<i>Brand1</i>	<i>Brand2</i>	<i>Brand3</i>	<i>Total</i>
High	55	45	20	120
Medium	45	25	25	95
Low	25	35	25	85
Total	125	105	70	300

3. If a family is selected at random, what is the probability that it belongs to High Income?
 (a) 0.35 (b) 0.40 (c) 0.45
4. The probability that a randomly selected family prefers Brand 3 is:
 (a) 0.26 (b) 0.30 (c) $7/30$
5. A family is selected at random and found to prefer Brand 2. What is the chance that it belongs to Medium Income?
 (a) $1/12$ (b) $5/19$ (c) $5/21$
6. What is the probability that a randomly chosen family prefers Brand 3 or belongs to the High Income Group?
 (a) $19/30$ (b) $17/30$ (c) $7/30$
7. If S represents the sample space, the probability of S designated as $P(S)$ equals 1. True or False.
8. If A and B are mutually Exclusive, $P(A \cap B) = 0$. True or False.

ANSWERS TO REVIEW QUESTIONS

1. The correct choice is (d)
 This problem uses the concept of conditional probability. The probability that the first bottle selected is defective = $4/12$ (There are 4 defective bottles and there are 12 bottles in total). The first bottle is not replaced. The probability that the second bottle is defective given that the first bottle is defective = $3/11$ (There are now 3 defective bottles out of a total of 11 bottles). Simplifying, $3/11 = 0.27$. Obviously, the other choices are omitted.
2. The correct choice is (d). When you roll the die once, there are 6 possibilities in total namely you can get 1, 2, 3, 4, 5, 6. Getting 4 can happen only in one way (4 is one of the faces of a die. Hence, the probability of getting 4 = $1/6$. The other options are wrong.

82 Business Statistics

Answers to questions 3-6 uses the contingency table given in the problem. For clarity, it is given below:

A survey of 300 families was conducted to study income level versus brand preference.

<i>Income \ Brand Preferred</i>	<i>Brand1</i>	<i>Brand2</i>	<i>Brand3</i>	<i>Total</i>
High	55	45	20	120
Medium	45	25	25	95
Low	25	35	25	85
Total	125	105	70	300

- The correct choice is (b). There are 120 families in the High Income Group. Total families surveyed = 300. Therefore, the probability that the chosen family belongs to High Income = $120/300 = 0.40$. The other options are rejected.
- The correct choice is (c). The number of families preferring Brand 3 is 70. Total families = 300. Hence the probability that a randomly chosen family prefers Brand 3 is $70/300 = 7/30$. Obviously, the rest of the options are wrong.
- The correct choice is (c). This is a case of conditional probability. We want the probability of Medium Income given that the family prefers Brand 2. This is equal to Probability of Medium Income and Brand 2 divided by Probability of Brand 2. From the table, we get this = $25/105 = 5/21$. The other choices are rejected.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 70/300 + 120/300 - 20/300 = 170/300 = 17/30$. The correct choice is (b). Hence, other choices are wrong.
- True because this amounts to adding the individual probability of all the events in the sample space S. It is always 1.
- True because when A and B are mutually exclusive, both A and B cannot occur. Hence, $P(A \cap B) = 0$ (See definition of mutually exclusive events in the chapter which says that the occurrence of A precludes the occurrence of B).

PRACTICE PROBLEMS

- A machine has three components A, B, and C. Even if one of the components fails, the machine will not work. The failure occurrence of the three components are independent. Component A has a failure rate of 8%, component B has a failure rate of 6% and component C has a failure rate of 9%. What is the probability that the machine will fail?
- If the probability of an income tax return having a single error is 0.06, what is the probability that an income tax return is error free?
- The inspection of 215 parts with regard to conformance to specification from three suppliers throws up the following data. All these 215 parts are contained in a box.

<i>Supplier</i>	<i>Number of Conformance</i>	<i>Number of Nonconformance</i>	<i>Total</i>
A	100	8	108
B	55	4	59
C	45	3	48
Total	200	15	215

- (a) What is the probability that a part randomly selected will belong to supplier *B* or supplier *C*?
- (b) What is the probability that a randomly selected part belongs to nonconformance category?
- (c) A part is randomly selected and found to belong to supplier *A*. What is the probability that it belongs to conformance category?
- (d) A part is randomly selected and found to be defective. What is the probability that it is supplied by supplier *C*?
- (e) What is the probability that a part selected at random will belong to supplier *B* or nonconformance category?
4. A box contains 44 pieces of cabbage of which 6 are spoiled. If a sample of two pieces is drawn in succession and is not replaced, what is the probability that both are spoiled pieces? What is the probability that both pieces are spoiled if the piece selected is replaced after the draw?
5. There are three dispatch assistants *A*, *B*, and *C* in an office. Assistant *A* processes 45% of letters, *B* processes 30% of letters, and *C* processes the remaining. The error rate of *A* is 6%, *B* is 9%, and *C* is 4%. The manager picked up one letter at random on a particular day and found it to be defective.
- (a) What is the probability that it was processed by assistant *A*?
- (b) What is the probability that it was processed by assistant *B*?
- (c) What is the probability that it was processed by assistant *C*?

Probability Distributions

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Appreciate what is a Probability Distribution
- Explain and use The Binomial Distribution
- Explain and use The Poisson Distribution
- Explain and use The Normal Distribution

CHAPTER OUTLINE

- 5.1 What is a Probability Distribution?
 - 5.2 The Binomial Distribution
 - 5.3 The Poisson distribution
 - 5.4 The Normal distribution
 - 5.5 Chapter Summary
- Glossary
Review Questions
Answers to Review Questions
Practice Problems

INTRODUCTION

Very often we want to describe and analyze numbers in the form of a distribution. You have seen this sharply in chapter 2 and chapter 3. Distributions can be formed by two methods:

1) From data collected and 2) with the help of standard distributions that have stood the test of time. This chapter provides the conceptual framework and applications of the three widely used probability distributions namely, The Binomial Distribution, The Poisson Distribution, and The Normal Distribution. While the first two are discrete in nature, the last one is a continuous distribution.

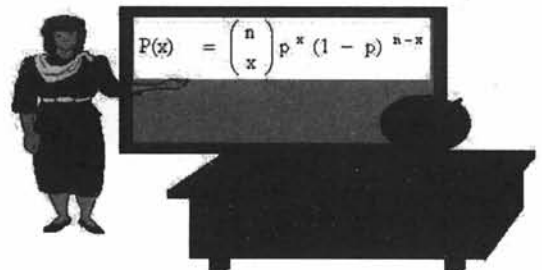


Figure 5.1

5.1 WHAT IS A PROBABILITY DISTRIBUTION?

To grasp the essence of a Probability Distribution, let us look at an example. This example will bring into sharp focus the underlying concepts of a probability distribution.

You rotate a pair of dice and observe the sum by totaling the numbers that turn up on both the dice. Let us designate this sum by a random variable X . When the first die shows up number 1, the second die could be 1, 2, 3, 4, 5, 6. Likewise, when the first die shows up number 2, the second die could be 1, 2, 3, 4, 5, 6. If you continue this experiment this way until all the possibilities are exhausted, the sample space will contain 36 elements. The distribution that can be observed will contain a sum = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. Please note that the starting sum will be = 2 because when both the dice show the number 1, the sum = $1 + 1 = 2$. Likewise, the maximum sum in the distribution will be 12 when both the dice show up a number 6. If we write the values of the random variable (sum) with associated probabilities, it will become what is called a Probability Distribution.



Figure 5.2

Sample Space for the Example

(D1 D2)	(D1 D2)	(D1 D2)	(D1 D2)	(D1 D2)	(D1 D2)
(1 1)	(2 1)	(3 1)	(4 1)	(5 1)	(6 1)
(1 2)	(2 2)	(3 2)	(4 2)	(5 2)	(6 2)
(1 3)	(2 3)	(3 3)	(4 3)	(5 3)	(6 3)
(1 4)	(2 4)	(3 4)	(4 4)	(5 4)	(6 4)
(1 5)	(2 5)	(3 5)	(4 5)	(5 5)	(6 5)
(1 6)	(2 6)	(3 6)	(4 6)	(5 6)	(6 6)

Note D1 represents Die1 and D2 represents Die2

D1 and D2 in the sample space above are used with a view to facilitating the sum of numbers in both the dice. For example, you take the case of getting eight. Adding $(2\ 6) = 8$. Adding $(3\ 5) = 8$. Adding $(4\ 4) = 8$. Adding $(5\ 3) = 8$. Adding $(6\ 2) = 8$. You see, there are 5 possibilities in which the sum 8 can occur. If you exhaust all the values that the random variable sum can assume, there are 36 elements in the sample space. Hence, the probability of getting a sum of 8 = $5/36$. Likewise, the Probabilities are computed for all other values of the sums associated with this experiment.

In precise terms, a *probability distribution* is a total listing of the various values the random variable can take along with the corresponding probability of each value. A real life example would be the pattern of distribution of the machine breakdowns in a manufacturing unit. The random variable in this example would be the various values the machine breakdowns could assume. The probability corresponding to each value of the breakdown is the relative frequency of occurrence of the breakdown. The probability distribution for this

example is constructed by the actual breakdown pattern observed over a period of time. Statisticians use the term "observed distribution" of breakdowns.

Probability Distribution for the Example

X = Sum	P(X)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Please note that the probability of the sample space = 1. (Adding all the probabilities above you get 1)

Progressive Test Question The relative frequency distribution is an example of a probability distribution. True or False.

Answer True because relative frequency of each class is a proportion or percentage of the frequency of that class to the total frequency. This indicates, therefore, the probability of the particular class occurring.

Another Example for a Probability Distribution

A multinational bank is concerned about the waiting time of its customers before they could use the ATM for their transactions. A study of a random sample of 500 customers reveals the following probability distribution.

X (Waiting Time per Customer in Minutes)	P(X)
0	0.20
1	0.18
2	0.16
3	0.12
4	0.10
5	0.09
6	0.08
7	0.04
8	0.03
Total	1.00

- (a) What is the probability that a customer will wait more than 5 minutes?
 (b) What is the probability that a customer need not wait?
 (c) What is the probability that a customer will wait less than 4 minutes?

Solution

(a) $P(X > 5) = P(6) + P(7) + P(8) = 0.08 + 0.04 + 0.03 = 0.15$

(b) $P(X = 0) = 0.20$

(c) $P(X < 4) = P(0) + P(1) + P(2) + P(3) = 0.20 + 0.18 + 0.16 + 0.12 = 0.66$

Expected Value of a Random Variable is another fundamental term, which you must explain in the context of decision making under uncertainty. The expected value of a random variable is computed by multiplying each value the random variable can assume by the corresponding probability of occurrence of that value and then taking the summation of these cross product terms. The probabilities are the weights that are attached to the respective values the random variable can take. In other words, expected value of a random variable is a weighted average of the all the possible outcomes of an experiment.

To understand how the expected value is calculated, let us take the multi national bank example discussed earlier.

<i>X (Waiting Time per Customer in Minutes)</i>	<i>P(X)</i>	<i>X × P(X)</i>
0	0.20	0.00
1	0.18	0.18
2	0.16	0.32
3	0.12	0.36
4	0.10	0.40
5	0.09	0.45
6	0.08	0.48
7	0.04	0.28
8	0.03	0.24
Total	1.00	2.71

From the table in the last row in column 3 against Total you see the value 2.71. This is the Expected value of X designated as E(X) and it is found to be equal to 2.71 Minutes. Thus, the average waiting time of a customer before getting access to ATM is 2.71 minutes.

Discrete probability distribution The two examples we have taken to explain the concept of a probability distribution are called discrete probability distributions. The probability distribution that uses a discrete random variable is called a **discrete probability distribution**. A **discrete random variable** implies that the random variable can assume only a restricted number of distinct values that are whole numbers. The number of units demanded per day of a product is an example of a discrete random variable. The number of cars passing through a street per hour during peak traffic is another example of a discrete random variable. In this chapter, we will deal with two types of discrete probability distribution namely, 1) The Binomial Distribution, and 2) The Poisson Distribution.

Continuous probability distribution The probability distribution that uses a continuous random variable is called a *continuous probability distribution*. A *continuous random variable* is a random variable, which can take any value within some interval of real numbers. Measurement of the height and weight of the respondents is an example of a continuous random variable. Similarly, voltage, pressure, and temperature, are examples of continuous random variables. In this chapter, we will deal with the important continuous distribution namely, the Normal Distribution.

5.2 THE BINOMIAL DISTRIBUTION

The Binomial Distribution is a widely used probability distribution of a discrete random variable. It plays a major role in **quality control** and **quality assurance** function. Manufacturing units use the binomial distribution for **defective** analysis. An item is considered defective if it has one or more types of defects. Reducing the number of defectives using the proportion defective control chart (p chart) is an accepted practice in manufacturing organizations.

Binomial distribution is also being used in **service organizations** like banks, and insurance corporations to get an idea of the proportion of customers who are satisfied with the service quality.

Another application area of binomial distribution is in the context of deciding whether to accept or reject a lot, containing components/finished product based on statistically designed sampling plan. As you know, in quality control function, companies do inspect the incoming quality of inputs as well as quality of outputs (finished products) before being shipped to customers. Binomial distribution plays a pivotal role in this inspection.

The conditions for applying Binomial Distribution are enunciated in the *Bernoulli Process*. They are given below:

Conditions for Applying Binomial Distribution (Bernoulli Process)

- Trials are independent and random.
- There are fixed number of trials (n trials).
- There are only two outcomes of the trial designated as **success** or **failure**.
- The probability of success is uniform through out the n trials.

Binomial Probability Function

Under the conditions of a Bernoulli process,

The probability of getting x successes out of n trials is indeed the definition of a Binomial Distribution. The Binomial Probability Function is given by the following expression:

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Where $P(x)$ is the probability of getting x successes in n trials.

$\binom{n}{x}$ is the number of ways in which x successes can take place out of n trials = $\frac{n!}{x!(n-x)!}$.

p is the probability of success, which is the same through out the n trials.

p is the parameter of the Binomial distribution.

x can take values $0, 1, 2, \dots, n$.

Note In the usual notation $n! = 1.2.3.4.5 \dots n$ (It is the product of the first n terms). For example, $5! = 1.2.3.4.5 = 120$.

Example A bank issues credit cards to customers under the scheme of Master Card. Based on the past data, the bank has found out that 60% of all accounts pay on time following the bill. If a sample of 7 accounts is selected at random from the current database, construct the Binomial Probability Distribution of accounts paying on time.

Solution This problem can be structured as a Bernoulli Process, where an account paying on time is taken as success, and an account not paying on time is taken as failure. The random variable x represents here an account paying on time, which can take values $0, 1, 2, 3, 4, 5, 6, 7$. You need to prepare a table containing x and $P(x)$ for all the values of x .

Performing calculations using Binomial Probability Function, $P(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, 3, 4, 5, 6, 7$ is very tedious manually and tedious with a calculator. One option available to you is to use the Binomial Probability Table that gives the probability value for a given value of n, p , and x . The best option is to use the Electronic Spread sheet of *Microsoft Excel* to calculate the Binomial Probabilities both for individual terms and for the cumulative probabilities. This facility is available under the option "Paste Function". Click *Statistical*, and then click *Binomial*. The form of the function is:

BINOMDIST

Number_s = number

Trials = number

Probability_s = number

Cumulative = logical

Returns the individual term binomial distribution probability.

Number_s is the number of successes in trials.

Formula result =

? OK Cancel

Figure 5.3

= *BINOMDIST*($x, n, p, 0$ or 1) where x is the number of successes, n is the number of trials, and p is the probability of success in each trial. The last term 0 or 1 performs a logical operation. If you enter 0, the computer returns the individual probability value; if 1 is entered, the computer gives the cumulative probability. Making use of this facility, the probability distribution for this example is worked out and displayed next.

If some of you want to use the binomial probability table, by all means do so. It is given in Appendix B.

Spread Sheet Output -The Binomial Probability Distribution for the Example given in the previous page.

A	B	C	D
1		Example Problem-Master Card	
2			
3	x	P(x)	Cumulative
4			Probability
5	0	0.0016384	0.0016384
6	1	0.0172032	0.0188416
7	2	0.0774144	0.0962560
8	3	0.1935360	0.2897920
9	4	0.2903040	0.5800960
10	5	0.2612736	0.8413696
11	6	0.1306368	0.9720064
12	7	0.0279936	1.0000000

Please note the following from the spreadsheet above

Column B has the random variable x = an account paying on time, column C has $P(x)$ which gives the probability corresponding to the value of x . Column D gives the Cumulative Probability of x .

Progressive Test Questions From the Binomial Probability Distribution example in the previous page, answer the following fill in the blank questions.

1. What is the probability of getting not more than 3 accounts paying on time?
The answer is -----.
2. What is the probability of at least 4 accounts paying on time?
The answer is -----.
3. What is the probability of no account paying on time?
The answer is -----.

Solution Look at the spreadsheet that is given below again for answering the questions.

A	B	C	D
1		Example Problem-Master Card	
2			
3	x	P(x)	Cumulative
4			Probability
5	0	0.0016384	0.0016384
6	1	0.0172032	0.0188416
7	2	0.0774144	0.0962560
8	3	0.1935360	0.2897920
9	4	0.2903040	0.5800960
10	5	0.2612736	0.8413696
11	6	0.1306368	0.9720064
12	7	0.0279936	1.0000000

1. You want $P(x \leq 3)$. The answer is 0.2897920. Why? Can you justify? Column D in the spreadsheet of Microsoft Excel gives the cumulative probability corresponding to the x value. Look at Column B, which gives the values that can be assumed by x . When $x = 3$, the cumulative probability in column D says the required answer is 0.2897920.
2. You want $P(x > 4) = P(4) + P(5) + P(6) + P(7) = 1 - P(x \leq 3) = 1 - 0.2897920 = 0.710208$.
3. You want $P(x = 0) = P(0) = .0016384$.

You will appreciate by now that you can tackle through Microsoft Excel intriguing and complicated problems on the Binomial Distribution.

The **Mean** and **Standard Deviation** of the Binomial Distribution is given in the following visual.

Mean and Standard Deviation of the Binomial Distribution

The mean μ of the Binomial Distribution is given by

$$\mu = E(x) = np$$

The Standard Deviation σ is given by

$$\sigma = \sqrt{np(1-p)}$$

Mathematically, we can derive both these measures. The derivations are not really important for managers taking decisions that involve the use of the Binomial Distribution.

We therefore, omit the derivations here. Intuitively speaking, it makes sense that the mean = np . For example, in a production process, if the proportion defective (p) is 0.10 for a particular component, then over a period of time, it is reasonable to expect that the defective components in the output would be 0.10 times the total output. The standard deviation gives an idea about the spread and scattering around the mean and in this case, it measures the departure from the average defective output.

Example for Mean and Standard Deviation

A bank issues credit cards to customers under the scheme of Master Card. Based on the past data, it has found out that 60% of all accounts pay on time following the bill. If a sample of 7 accounts is selected at random from the current database, calculate the mean and standard deviation of the accounts paying on time?

This problem can be structured as a Bernoulli Process where an account paying on time is taken as success and an account not paying on time is taken as failure. The random variable x represents here an account paying on time, which can take values 0, 1, 2, 3, 4, 5, 6, 7. Applying the formula for mean, $\mu = np$, you get mean = $7 \times 0.6 = 4.2$. That is, you can expect on an average 4.2 of the accounts will pay on time.

Applying the formula for standard deviation, $\sigma = \sqrt{np(1-p)}$, you get standard deviation = $\sqrt{4.2(1-0.60)} = 1.30$.

Example from Quality Control Function

Using the Microsoft Excel provide solution to the following Problem involving the Binomial Distribution.

From the shop floor of a manufacturing company, the quality control department selects a sample of 15 items. According to the requirement, if 3 or more of the items in the sample are found to be defective, the entire production lot will be rejected and then the lot will be subjected to 100% inspection. From the past data, it is known that the probability of an item being defective is 0.04.

QUESTIONS

1. What is the probability that the production lot will be rejected?
2. Find the mean and standard deviation of the binomial random variable (x = number of defectives).
3. Plot a curve by taking on the x-axis a set of values for percentage defectives, and on the y-axis, the probability of acceptance corresponding to the percentage defective values on the x-axis. Interpret your results.

SOLUTION

1. Probability of rejecting the lot = 1-probability of accepting the lot. Using paste function = BINOMDIST of Microsoft excel, we have the probability of acceptance for $n = 15$,

$x = 2$, $p = .04$, read as 0.9797. So, probability of rejecting the lot = $1 - 0.9797 = 0.0203$. (Please go through this chapter for paste function example on Binomial. Keep practicing).

- Mean = $np = (15)(.04) = 0.60$, Standard Deviation = $\sqrt{np(1-p)} = \sqrt{0.60(0.96)} = 0.76$.
- For this part see diagram given below:

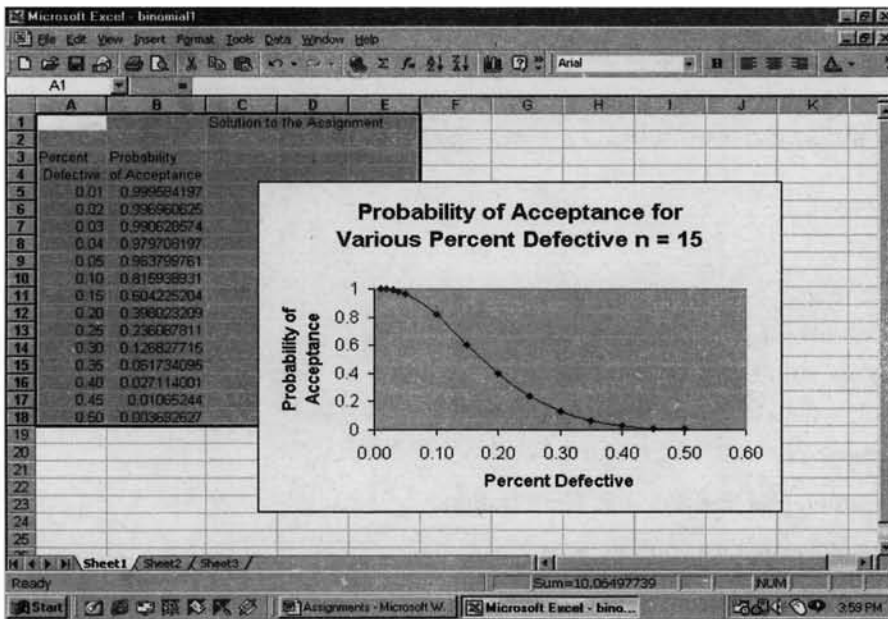


Figure 5.4

The shape of the curve clearly shows that as the proportion defective keeps on increasing, the probability of acceptance keeps on decreasing steeply. By the way, in quality control parlance, this curve is called the operating characteristic curve.

5.3 THE POISSON DISTRIBUTION

Poisson Distribution is another discrete distribution which also plays a major role in **quality control** in the context of reducing the number of defects per standard unit such as number of defects per item, number of defects per transformer produced, number of defects per 100 m² of cloth, etc. Other real life examples would include:

- The number of cars arriving at a highway check post per hour.
- The number of customers visiting a bank per hour during peak business period.

Poisson Probability Distribution Function is worked out based on the Poisson Process characterized by the following:

Poisson Process

The probability of getting exactly one success in a continuous interval such as length, area, time and the like is constant.

The probability of a success in any one interval is independent of the probability of success occurring in any other interval.

The probability of getting more than one success in an interval is 0.

Poisson Distribution Formula

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

where

$P(x)$ = Probability of x successes given an idea of λ .

λ = Average number of successes.

$e = 2.71828$ (based on natural logarithm).

x = successes per unit which can take values 0, 1, 2, 3,..... ∞ .

λ is the Parameter of the Poisson Distribution.

Mean of the Poisson Distribution is = λ .

Standard Deviation of the Poisson Distribution is = $\sqrt{\lambda}$.

Example If on an average, 6 customers arrive every two minutes at a bank during the busy hours of working, a) what is the probability that exactly four customers arrive in a given minute? b) What is the probability that more than three customers will arrive in a given minute?

Solution $\lambda = 3$ (6 customers every two minutes implies 3 customers per minute) (a) You are required to find out $P(x = 4)$. Substituting in the formula the value of

$x = 4$, you can get the answer after simplifying the expression $P(x) = \frac{e^{-\lambda} \lambda^x}{x!}$. However, you

are strongly encouraged to use the Microsoft Excel. Standard Poisson table or manual calculations cannot match "Excel".

The Poisson probability table is given in Appendix C. You can use it if you feel the necessity.

Use the Paste Function option and then click *Poisson* under *Statistical*. Follow the prompts, you get the answer. Just like the Binomial Distribution, here also, the logical operator 0 gives the individual probabilities, and 1 gives the cumulative probabilities. The output from Excel is given below for the values of x ranging from 0 to 10.

POISSON

X = number

Mean = number

Cumulative = logical

=

Returns the Poisson distribution.

X is the number of events.

Formula result =

OK Cancel

Figure 5.5

A	B	C	D	E	F	G	H
1		Poisson Distribution for the Example (x ranges from 0 to 10)					
2					$\lambda = 3$		
3			Cumulative				
4	x	P(x)	Probability				
5	0	0.049787	0.049787				
6	1	0.149361	0.199148				
7	2	0.224042	0.423190				
8	3	0.224042	0.647232				
9	4	0.168031	0.815263				
10	5	0.100819	0.916082				
11	6	0.050409	0.966491				
12	7	0.021604	0.988095				
13	8	0.008102	0.996197				
14	9	0.002701	0.998898				
15	10	0.000810	0.999708				

- (a) Look at the table and you will find that corresponding to $x = 4$ in column B, the probability $P(x = 4)$ is 0.168031 found in column C.
- (b) You want $P(x > 3) = 1 - P(x \leq 3) = 1 - 0.647232 = 0.352768$ (Please note that $P(x \leq 3)$ is found in Column D corresponding to the value of $x = 3$ in column B. You could see 0.647232 in row 8 and column D.

When can poisson distribution be used as an approximation to the binomial distribution?

There is a general consensus among the statisticians that when n is greater than or equal to 20, and p is less than or equal to 0.05, the Poisson Approximation to the Binomial is a valid assumption. In this approximation, $np = \lambda$. Now, use the Poisson Probabilities as an approximation to the Binomial Probabilities by substituting $np = \lambda$ in the expression

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$
. When you have the power of Microsoft Excel, this approximation loses its

significance. At any rate conceptually, you should know as to when the Poisson Distribution can be used in the place of the Binomial Distribution.

Progressive Test Question The number of arrivals of trucks per hour at a loading booth is an example of the:

- (a) Binomial distribution
- (b) Normal distribution
- (c) Poisson Distribution

Answer The right choice is (c) First Poisson distribution is a discrete distribution. Choice (c) is wrong because Normal distribution is a continuous distribution. Trucks arrival per hour time interval fits into the Poisson process. Number of arrivals per hour indicates the rate of arrival (which we term as λ) and hence, Binomial distribution is not appropriate here.

Application in Airlines

An airlines in a large city, gets on an average 5 claims per day for missing baggage. Using Microsoft Excel, construct the Poisson probability distribution (both individual and cumulative) for this problem. Based on this output, answer the following:

- (a) What is the probability that at least there are 7 claims for missing baggage?
- (b) What is the probability of no claims for the missing baggage?
- (c) What is the probability that there are exactly 4 claims for missing baggage?

Using paste function = Poisson from Microsoft Excel, we have the following table giving individual and cumulative probabilities. All parts can be answered from this.

x	P(x)	Cumulative
0	0.006738	0.006738
1	0.03369	0.040428
2	0.084224	0.124652
3	0.146274	0.270926
4	0.175467	0.446393
5	0.175467	0.62186
6	0.146274	0.768127
7	0.104445	0.872572
8	0.065276	0.937848
9	0.03369	0.971538
10	0.016845	0.988383
11	0.006738	0.995121
12	0.002775	0.997896
13	0.001071	0.998967
14	0.000412	0.999379

Figure 5.6

Using the above table, we have,

- (a) Probability of at least 7 missing baggage claims = $1 - P(x \leq 6) = 1 - 0.7621 = 0.2379$
- (b) Probability of no claim of missing baggage = $P(0) = 0.0067$
- (c) Probability of exactly 4 claims of missing baggage = $P(4) = 0.1755$

5.4 THE NORMAL DISTRIBUTION

The Normal Distribution is the most widely used continuous distribution. It occupies a unique place in the field of statistics. In fact, the entire quality control function that employs the statistical techniques heavily will come to a grinding halt without the use of the normal distribution. The control charts for reducing and stabilizing variation rely on the normal distribution. Process capability studies to meet the customer specifications need the normal distribution. The whole subject matter- inferential statistics is based on the normal distribution. In all management functions including the human side, the observed frequency distributions encountered are all fairly close to the normal distribution when the sample size is reasonably large.

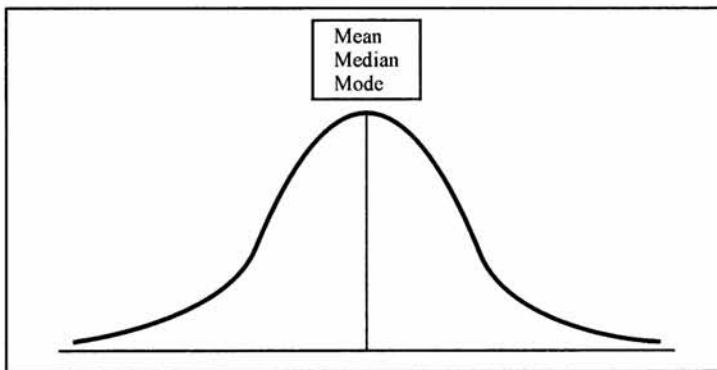


Figure 5.7

Properties of the Normal Distribution

1. The normal distribution is a continuous distribution looking like a bell. Statisticians use the expression "Bell Shaped Distribution".
2. It is a beautiful distribution in which the mean, the median, and the mode are all equal to one another.
3. It is symmetrical about its mean.
4. If the tails of the normal distribution are extended, they will run parallel to the horizontal axis without actually touching it. (asymptotic to the x-axis).
5. The normal distribution has two parameters namely the mean and the standard deviation.

Normal Probability Density Function

In the usual notation, the probability density function of the normal distribution is given below:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]}$$

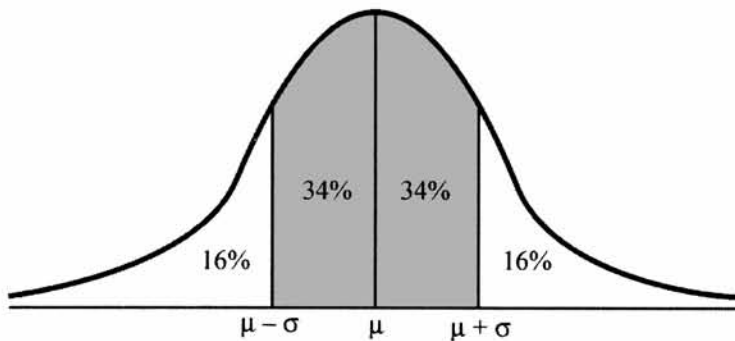
x is a continuous normal random variable with the property $-\infty < x < \infty$ meaning x can take all real numbers in the interval $-\infty < x < \infty$. You need not worry about this complicated function because all calculations can be done without the help of this expression.

What is important is that you should understand the normal distribution and you should be able to apply it to practical problems. For this purpose, the knowledge of the actual probability density function is not required.

If x follows a normal distribution with mean μ and standard deviation σ , it is a general practice among statisticians to symbolize this expression as $x \sim N(\mu, \sigma)$.

The use of symmetry property of the normal distribution The symmetry property of the normal distributing is a beauty. The area under the normal curve has got three distinct positions. They are pictured as follows: Please see comments given in the pictures. Please also note that the values of the area covered under each picture are very close to the actual values that one could compute based on the probability density function of the normal distribution.

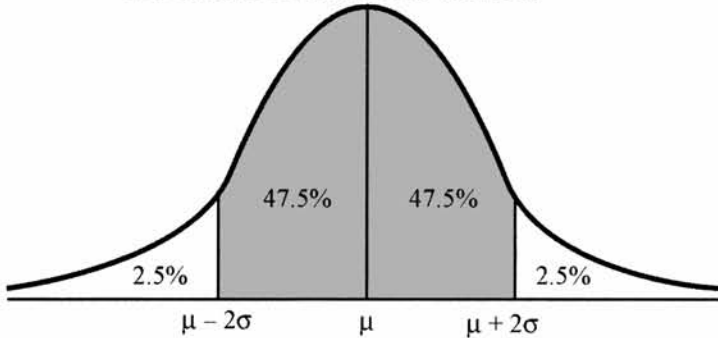
Picture 1: Area Under the Normal Curve within One Standard Deviation from the Mean



The picture shows that the total area covered within one Standard Deviation from the Mean is 68%. Because of the symmetry, 68% gets divided into 34% on either side of the Mean

Figure 5.8

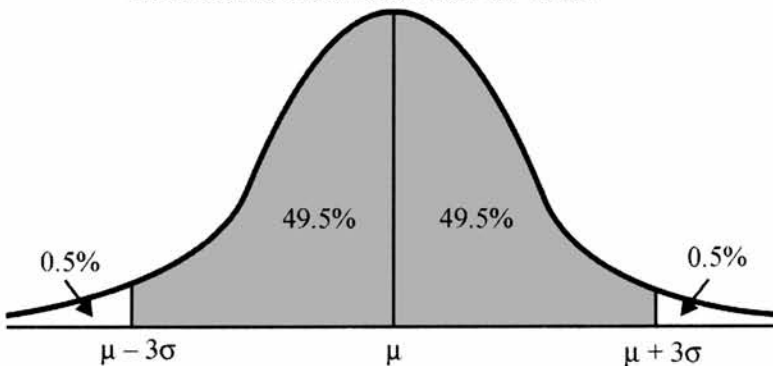
Picture 2: Area Under the Normal Curve within Two Standard Deviation from the Mean



The picture shows that the total area covered within Two Standard Deviation from the Mean is 95%. Because of the symmetry, 95% gets divided into 47.5% on either side of the Mean

Figure 5.9

Picture 3: Area Under the Normal Curve within Three Standard Deviation from the Mean



The picture shows that the total area covered within Three Standard Deviation from the Mean is 99%. Because of the symmetry, 99% gets divided into 49.5% on either side of the Mean

Figure 5.10

The standard normal distribution For practical problems, which require the help of the normal distribution for solutions, it is not possible to compute the probability in a straightforward manner because each of those normal variables may have different units of

measurements. One problem may involve the unit meter, another may involve kilograms, and so on. Therefore you need a probability distribution that can tackle any unit of measurement. Fortunately, you have the Standard Normal Distribution defined as follows:

$$z = \frac{x - \mu}{\sigma}$$

z is called the standard normal variable. Please note that z is a pure number

independent of the unit of measurement. The random variable z follows a normal distribution with mean = 0 and standard deviation = 1. You first convert the original variable in a given problem into z . It is only a transformation of scale. The probability table for z is available for computing the necessary probabilities for a given situation. The elegant way of getting the probability value based on z is to take advantage of the Microsoft Excel.

Example Problem for the Normal Distribution

The mean weight of a morning breakfast cereal pack is 0.295 kg with a standard deviation of 0.025 kg. The random variable weight of the pack follows a normal distribution.

- What is the probability that the pack weighs less than 0.280 kg?
- What is the probability that the pack weighs more than 0.350 kg?
- What is the probability that the pack weighs between 0.260 kg to 0.340 kg?

Solution Draw the correct diagram to the original problem. This step is crucial for getting the answer. Let us draw the diagram for each part and then provide the solution using Microsoft Excel.

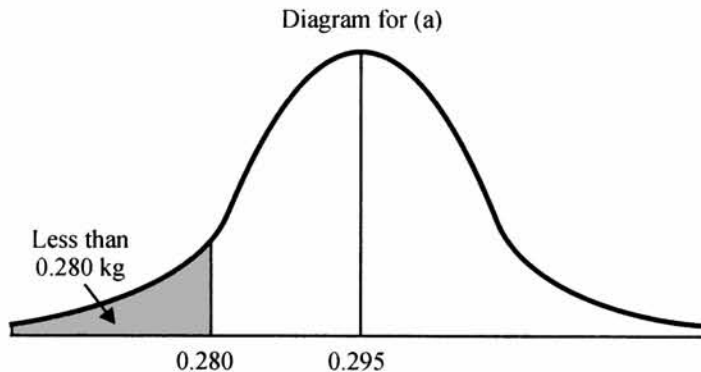


Figure 5.11

- See diagram above in which the probability of the shaded area is wanted.

$z = \frac{x - \mu}{\sigma} = (0.280 - 0.295)/0.025 = -0.6$. Click "Paste Function" of Microsoft Excel, then click the *Statistical* option, and then click *NORMSDIST* (the standard normal distribution option). You get the following:

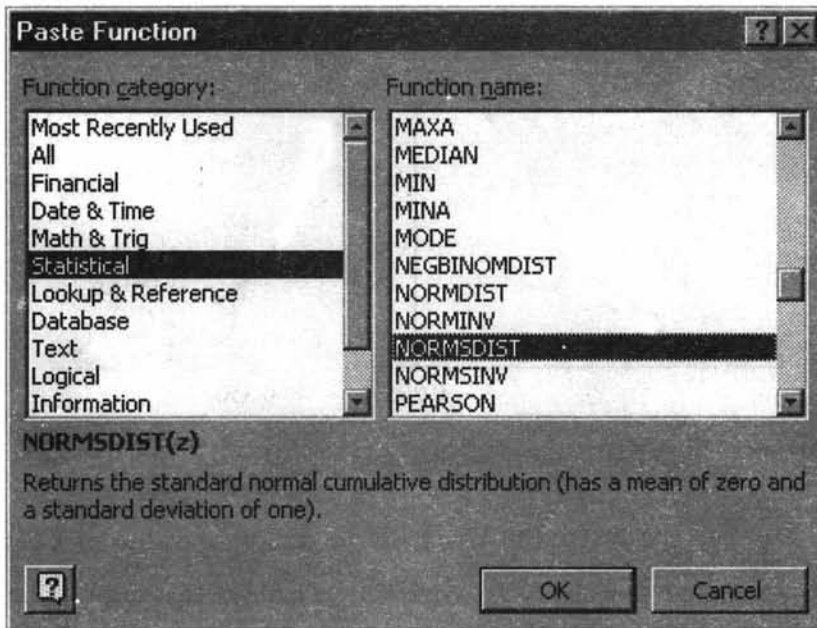


Figure 5.12

Now click OK and enter the z value in the reference cell provided for Z . You get the answer. Excel accepts directly both the negative and positive values of z . Excel always returns the cumulative probability value. When z is negative, the answer is direct. When z is positive, the answer is $= 1 -$ the probability value returned by Excel.

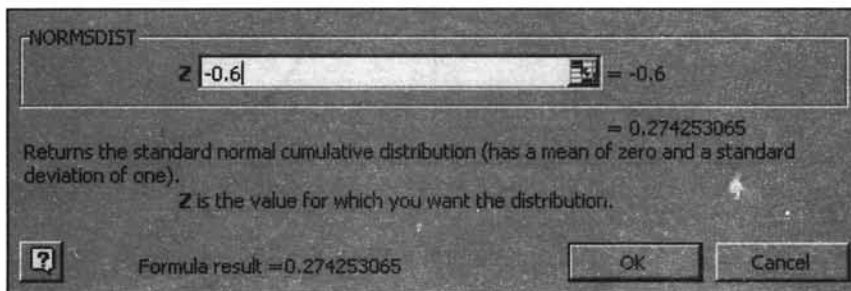
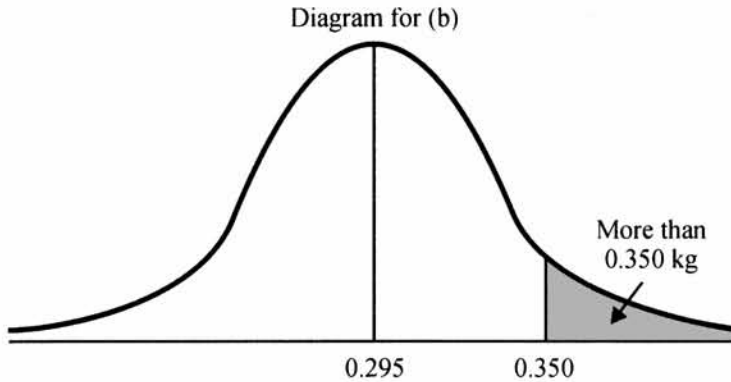


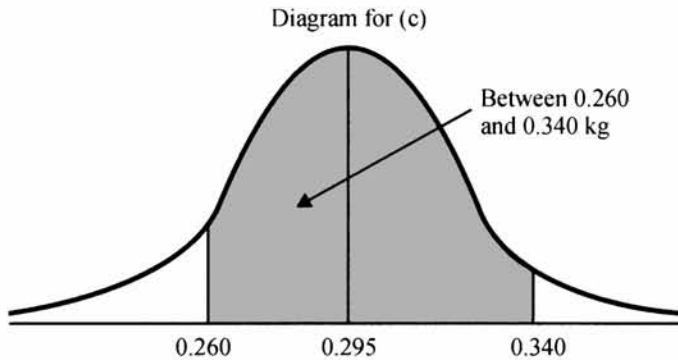
Figure 5.13

The answer for part a) of the question $= 0.2743$ (Direct from Excel since z is negative). Please see the answer displayed in NORMSDIST. We have rounded off to 4 decimal places. This says that 27.43 % of the packs weigh less than 0.280 kg.

Proceeding in the same manner, part b) and part c) are answered below: To repeat again, click paste function, click *Statistical*, click *NORMSDIST*, click OK, and then enter the value for z in the reference cell. You get the answer.

**Figure 5.14**

- (b) $z = \frac{x - \mu}{\sigma} = (0.350 - 0.295)/0.025 = 2.2$. Excel returns a value of 0.9861. Since z is positive, the required probability is $= 1 - 0.9861 = 0.0139$. This means that 1.39% of the packs weigh more than 0.350 kg.

**Figure 5.15**

- (c) For this part, you have to first get the cumulative probability up to 0.340 kg and then subtract the cumulative probability up to 0.260. $z = \frac{x - \mu}{\sigma} = (0.340 - 0.295)/0.025 = 1.8$ (up to 0.340). $z = \frac{x - \mu}{\sigma} = (0.260 - 0.295)/0.025 = -1.4$ (up to 0.260). These two probabilities from Excel are 0.9641 and 0.0808 respectively. The answer is $= 0.9641 - 0.0808 = 0.8833$. This means that 88.33% of the packs weigh between 0.260 kg and 0.340 kg.

The standard normal table is given in Appendix D. This table displays the probability between 0 and positive value of z . Using the table in Appendix D, let us answer all the parts of the above example.

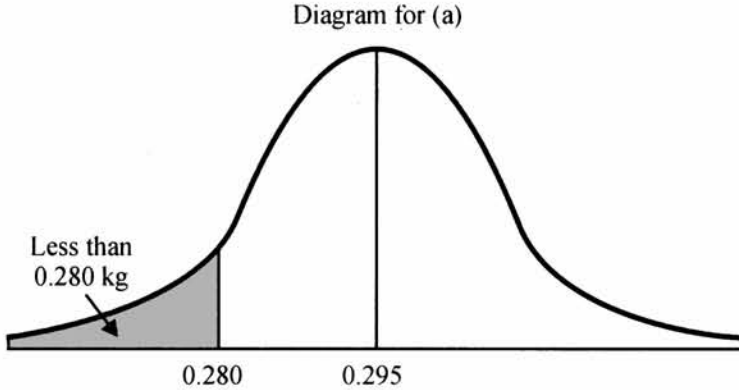


Figure 5.16

- (a) See diagram above, in which, the probability of the shaded area is wanted.

$z = \frac{x - \mu}{\sigma} = (0.280 - 0.295)/0.025 = -0.6$. In the format of the table in Appendix D, this is = 0.5- area up to $z = -0.6$. By property of symmetry, this is same as 0.5 - area from 0 to 0.6. This is = $0.5 - 0.2257 = 0.2743$. This is same as what you have got earlier, using paste function of Microsoft Excel.

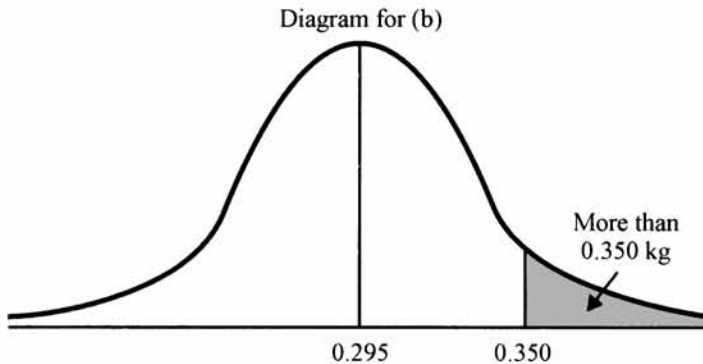


Figure 5.17

- (b) $z = \frac{x - \mu}{\sigma} = (0.350 - 0.295)/0.025 = 2.2$. What you want is the area that is greater than 2.2. This is same as 0.5 - area from 0 to 2.2. This is = $0.5 - 0.4861 = 0.0139$.

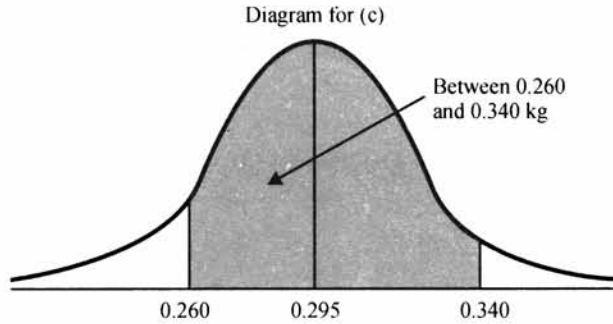


Figure 5.18

(c) What you want is the area shaded above. This is the sum of area between $z = \frac{x - \mu}{\sigma}$
 $= (0.260 - 0.295)/0.025 = -1.4$ and $z = \frac{x - \mu}{\sigma} = (0.340 - 0.295)/0.025 = 1.8$. This is same as -1.4 to 0 plus 0 to 1.8 . But area -1.4 to 0 is same as 0 to 1.4 . Therefore, the answer is = area of 0 to 1.4 plus area of 0 to 1.8 . This is = $0.4192 + 0.4641 = 0.8833$.

Critical Thinking Skills

There is another fantastic feature in Microsoft Excel by which you can get the cumulative probabilities for any normal distribution problem directly. You need not use the z distribution. Find out how is this possible from the paste function. Have a detailed discussion on this feature

Progressive Test Question A large number of line voltage tests of residences reveal a mean voltage of 200 V and a standard deviation of 30 V. The distribution pattern in this case is found to be normal. Determine the percentage of voltage data between 180 V and 210 V.

Solution First, draw the appropriate diagram. In this case it is:

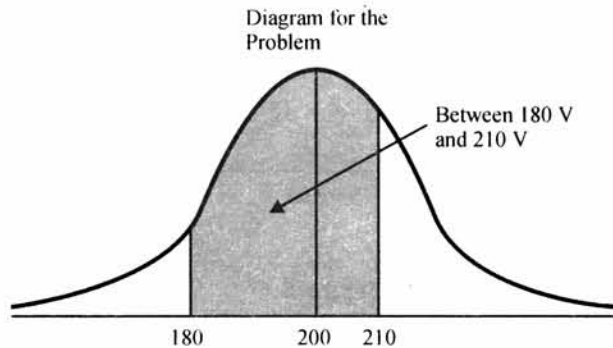


Figure 5.19

Using Excel first compute the cumulative probability up to 210 V, and then subtract the cumulative probability up to 180 V to get the answer.

Let us use the direct feature from Excel that computes the cumulative probability by using *NORMDIST* (x, mean, standard deviation, cumulative). The last parameter is a logical one. If you want the cumulative, enter 1. Otherwise enter 0 (in this case you will get the probability mass function).

Click *paste function*, click *Statistical*, and click *NORMDIST*. You get the following:

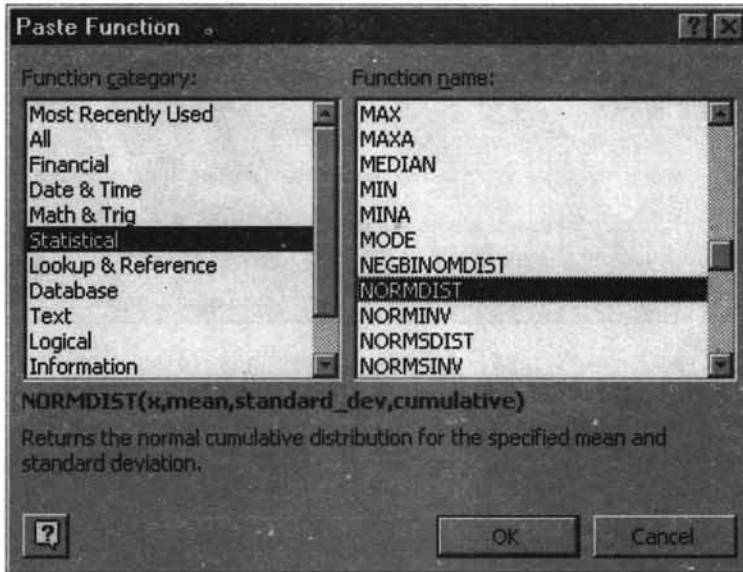


Figure 5.20

Click OK and you get:

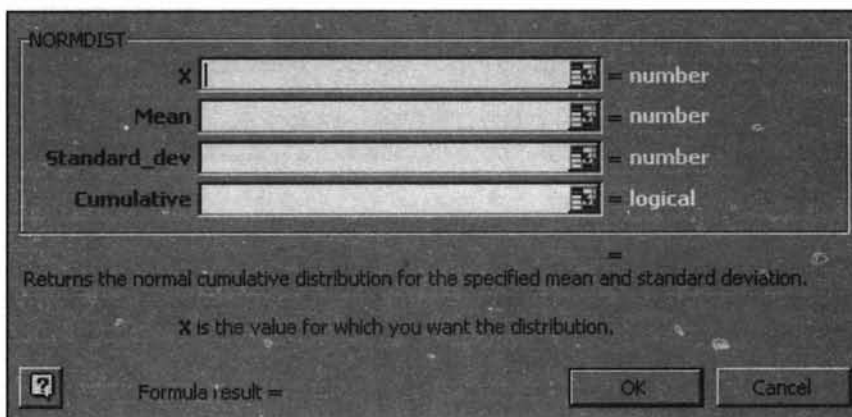


Figure 5.21

Enter in the reference cell for *X*, the value 210, reference cell for *Mean*, the value 200, reference cell for *Standard_dev*, the value 30, and in the *Cumulative* cell, the value 1. Click OK. You now get the cumulative probability up to 210. Instead of 210 for *X*, if you enter 180, you get the cumulative probability up to 180.

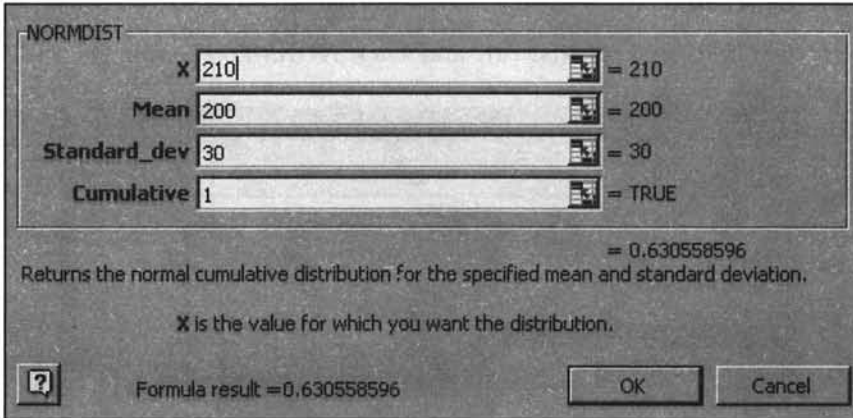


Figure 5.22

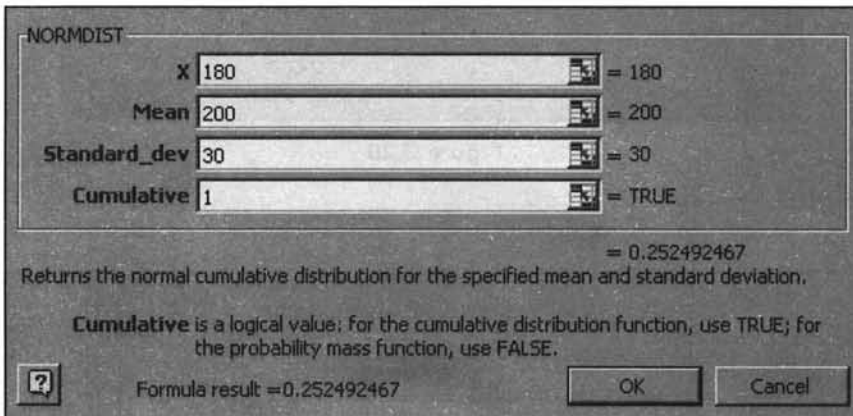


Figure 5.23

From Excel output display in the above two NORMDIST screens, up to 210 the cumulative probability is 0.6306. Similarly up to 180, the cumulative probability is 0.2525. Hence the answer is $0.6306 - 0.2525 = 0.3781$. That is about 37.81% of the voltage data fall in the range of 180 V to 210 V.

Progressive Test Question The probability that the line voltage is more than 210 V is 36.94 %. True or False.

Solution True because the cumulative probability up to 210 V is 0.6306. Hence the probability that the line voltage is more than 210 V = $1 - 0.6306 = .3694 = 36.94\%$

THE NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

The diagram given below shows a pattern of a symmetrical bell shaped normal distribution. In the binomial distribution by taking $n = 10$, and $p = 0.5$, a histogram was first drawn in which the horizontal axis represents the success (x) and the vertical axis represents the probability of success $P(x)$. A normal distribution is superimposed on the histogram. The shape of the histogram coincides with the normal curve almost perfectly. Hence, it is agreed by many practitioners that the Normal approximation for the Binomial holds true when the value of np is at least = 5.

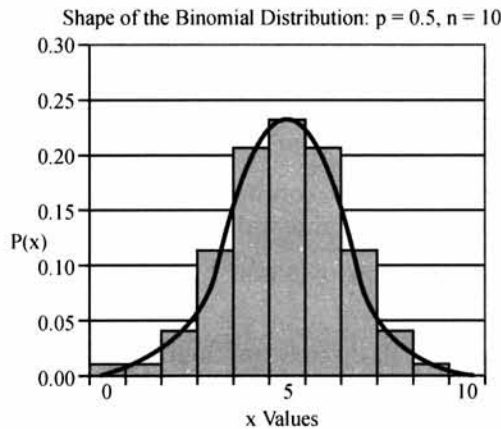


Figure 5.24

5.5 CHAPTER SUMMARY

This chapter is extremely important from the point of view of many fascinating aspects of statistical inference that would follow in the subsequent chapters. Certainly, it is expected from you that you master the nitty-gritty of this chapter. This chapter specifically focused on:

- The definition, meaning, and concepts of a probability distribution
- The related terms-discrete random variable, continuous random variable, and observed distribution
- Discrete probability distribution, and continuous probability distribution
- The Bernoulli process with its elements
- The Binomial distribution and its role in business problems
- The Poisson process
- The Poisson distribution and its uses

- The Poisson approximation to the Binomial distribution
- The Normal distribution and its role in statistical inference
- The concept of the Standard Normal Distribution and its role
- The Normal approximation to the Binomial distribution
- All through this chapter the pivotal role of the Microsoft Excel for probability calculations for all the distributions.

GLOSSARY

Bernoulli Process It is a process in which trials are independent and random, there are fixed number of trials (n trials), there are only two outcomes of the trial usually designated as success or failure, and the probability of success is uniform through out the n trials.

Binomial Distribution It is a probability distribution of a discrete random variable under the conditions of Bernoulli process.

Continuous Probability Distribution It is a probability distribution that uses a continuous random variable.

Continuous Random Variable A continuous random variable is a random variable, which can take any value within some interval of real numbers.

Discrete Probability Distribution It is a probability distribution that uses a discrete random variable.

Discrete Random Variable A discrete random variable is a random variable that can assume only a restricted number of distinct values.

Expected Value of a Random Variable Expected value of a random variable is a weighted average of the all the possible outcomes of an experiment with probabilities of outcomes as weights.

Normal Distribution It is a continuous probability distribution which looks like a bell. Statisticians use the expression "Bell Shaped Distribution". It is a beautiful distribution in which the mean, the median, and the mode are all equal to one another. It is symmetrical about its mean. If the tails of the normal distribution are extended, they will run parallel to the horizontal axis without actually touching it (asymptotic to the x -axis). The normal distribution has two parameters-mean, and standard deviation.

Poisson Distribution Poisson distribution is probability distribution of a discrete random variable based on the Poisson Process.

Poisson Process It is a process characterized by the following:

The probability of getting exactly one success in a continuous interval such as length, area, time and the like is constant. The probability of a success in any one interval is independent of the probability of success occurring in any other interval. The probability of getting more than one success in an interval is 0.

Probability Distribution A probability distribution is a total listing of the various values the random variable can take along with the corresponding probability of each value.

Standard Normal Distribution It is normal distribution with mean = 0 and standard deviation = 1

REVIEW QUESTIONS

1. In the Binomial Distribution, the outcome of one trial is independent of the outcome of any other trial. True or False
2. The expected value of a random variable is a weighted average of all possible outcome of an experiment. True or False
3. What probability method is used to compute the probabilities in observed distributions?
 - (a) Subjective
 - (b) A Priori
 - (c) Relative
4. A random sample of 5 gearboxes is selected from the shop floor of an automobile company. The proportion defective gearbox based on a comprehensive pilot study is 0.08. What is the probability of 2 or more defectives?
 - (a) 0.0633
 - (b) 0.0496
 - (c) 0.0544
5. For problem 4 above, the mean and standard deviation are,
 - (a) 0.04, 0.06
 - (b) 0.40, 0.6066
 - (c) 0.08, 0.60
6. The mean billing error encountered in a bank per day is 2.0. What is the probability of getting no billing error?
 - (a) 0
 - (b) 0.25
 - (c) 0.12
 - (d) 0.1353
7. In problem 6) above, what is the probability of getting 3 or less billing errors?
 - (a) 0.1804
 - (b) 0.1675
 - (c) 0.8571
 - (d) 0.1967
8. Which one of the following is not true of the normal distribution?
 - (a) A bell shaped distribution
 - (b) A symmetrical distribution
 - (c) The mean, median, and mode are all equal to one another
 - (d) A discrete probability distribution
9. A machine produces steel rods. The lengths of the rods are normally distributed with a mean of 26 cm and a standard deviation 1 cm. Rods that are longer than 27 cm or shorter than 24 cm have to be discarded. The machine produces 500 rods per shift. How many rods per shift have to be discarded?
 - (a) 70
 - (b) 83
 - (c) 113
 - (d) 91

ANSWERS TO REVIEW QUESTIONS

1. True because one of the crucial assumptions of the Binomial Distribution is that the trials are independent
2. True. Expected value of a random variable is indeed a weighted average of all possible outcome of an experiment (Sample Space) with probabilities as weights.
3. The right choice is (c). In observed distributions, the probability is computed using the relative frequency of the class. That is, for each class, frequency of the class is divided by the total frequency to get the probability of that class. Therefore, the correct answer is c)
4. The right choice is (c). This is a case of Binomial Distribution with $p = 0.08$ and $n = 5$. First $P(x >= 2) = 1 - P(x <= 1)$. $P(x <= 1)$ is the cumulative probability up to 1 which you can pick up from Excel using = BINOMDIST ($x, n, p, 0$ or 1). The answer for this is = 0.9456. So, the probability of getting 2 or more defectives = $1 - 0.9456 = 0.0544$.
5. The right choice is (b). For the Binomial distribution, mean = $np = 5 \times 0.08 = 0.40$. Standard deviation = $\sqrt{np(1-p)} = \sqrt{0.40(1-0.08)} = 0.6066$. Hence, (b) is the correct answer
6. The right choice is (d). This problem requires the use of the Poisson distribution. Here, $\lambda = 2.0$. You want $P(0) = 0.1353$ from Excel (using the Poisson Probability in paste function). The other choices are wrong and are automatically rejected.
7. The correct answer is (c). You want $P(x <= 3)$. From Microsoft Excel, the cumulative probability up to 3 is directly read as 0.8571. The other options are wrong.
8. The correct choice is (d). The normal distribution is a continuous probability distribution. (a) is true-normal distribution is bell shaped; (b) is true-normal distribution is symmetrical; (c) is true- in normal distribution, the mean, the median, and the mode are all equal to one another. (d) says the normal distribution is a discrete probability distribution which is not true.
9. **Solution:** (d) is the right choice. First draw the diagram for the problem.

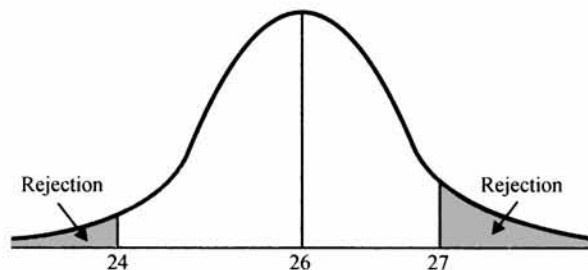


Figure 5.25

If you add the probability of the shaded areas, you get the answer. Using the Microsoft Excel Paste Function = NORMDIST (x , mean, standard deviation, cumulative). The first

one is the direct cumulative up to 24. The value for this = 0.0228. The second probability is 1- cumulative up to 27. = $1 - 0.8413 = 0.1587$. So the required probability = $0.0228 + 0.1587 = 0.1815$. If the machine produces 500 rods, the rods that are to be discarded or rejected = $0.1815 \times 500 = 90.75$ or approximately 91 rods.

PRACTICE PROBLEMS

1. CASE STUDY: BUSINESS STATISTICS COURSE

At the conclusion of a course in business statistics, a group of management students sat for a written examination. The results throw up the following information.

Marks obtained have a mean of 60 and a standard deviation of 12. There were 300 students who wrote the exam. The pattern of marks follows a normal distribution. Answer the following questions:

- (a) The percentage of students who score more than 80
- (b) The percentage of students who score less than 50
- (c) What should be the distinction mark if the highest ten percent of students are to be awarded distinction?

2. CASE STUDY: AUTOMOBILE COMPONENTS

A company manufacturers automobile components of various kinds. The quality control policy for a particular type of component stipulates taking random samples of six components at regular intervals. The number of defective components are kept separately in counting process of every sample. Out of 200 such samples, in 108 cases no defectives, in 64 cases one defective, in twenty cases two defectives and in the remaining cases three defectives were recorded.

Do these results suggest that the process is operating with an average of 11 per cent defectives under the phenomenon of the binomial distribution? Justify your answer.

3. CASE STUDY: CREDIT CARDS

A multinational bank issuing Master Card is monitoring the use of credit card account holders in the context of their spending habits. A market survey shows that the average monthly spending of its regular card users is normally distributed with mean Rs.2800 and standard deviation Rs.900. The customers are classified into four categories according to pattern of spending:

- (a) Category 1 spends less than Rs.2000
- (b) Category 2 spends Rs.2000 or more but less than Rs.3000
- (c) Category 3 spends Rs.3000 or more but less than Rs.4000
- (d) Category 4 spends Rs.4000 or more

What proportion of customers would you expect to fall into each category?

4. CASE STUDY: MOTORCAR ACCIDENTS

Those who drive motorcars are vulnerable to accidents. The reasons attributed for accidents are bad weather, driver error, bad maintenance, and violation of traffic rules while driving. The frequency of accidents is also to a large extent correlated with the age and brand of the car apart from the other factors mentioned above.

For all cars, insurance companies have to decide on the quantum of premium to be charged from the customers to protect against risks. In a dynamic manner, premiums are revised from time to time depending on the circumstances. The following data refer to the number of accidents in a time horizon of 300 days along with the number of vehicles associated with the number of accidents. The pattern and frequency of accidents would pave the way for deciding the quantum of insurance premium.

Number of accidents in 300 days	0	1	2	3	4	5	6	7
Number of cars with involved	46	66	48	20	10	6	2	2

Does this pattern of accidents depict a Poisson distribution? Justify your answer.

Basics of Sampling and Sampling Distribution

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Explain the need for Sampling
- Explain the Types of Sampling
- Define and explain the concept of Sampling Distribution
- Define and explain the concept of Standard Error
- Explain and use the Sampling Distribution of Mean

Before you proceed further in this chapter, please refresh your concepts on the following terms by meticulously going through the first chapter (Chapter1). This is key to understanding this chapter.

- Population (Universe)
- Sample
- Parameter
- Statistic

INTRODUCTION

The aim of sampling is to throw light on the population (universe) parameter that is of interest to the investigator. A well thought out representative random sample most of the times gives meaningful insights into the properties of the population parameters. This is the very foundation of statistical inference. This chapter covers the sampling methodology and the associated sampling distribution.

CHAPTER OUTLINE

- 6.1 What is Sampling and why do you need Sampling?
 - 6.2 Types of Sampling
 - 6.3 Sampling Distribution-A Conceptual Framework
 - 6.4 The Concept of Standard Error
 - 6.5 Sampling Distribution of the Mean from Normal Population
 - 6.6 Sampling Distribution of the Mean-Non-Normal Population
 - 6.7 Chapter Summary
- Glossary
Review Questions
Answers to Review Questions
Practice Problems



Figure 6.1

6.1 WHAT IS SAMPLING AND WHY DO YOU NEED SAMPLING?

Sampling is a method of selecting units of analysis such as households, people, consumers, companies etc from a population (universe) of interest to a manager. By analyzing the data collected from the sample, you draw inferences about the population parameters. In other words, sampling is employed to throw light on the population parameter. In chapter 1, you have been already exposed to the definition and meaning of the terms, "parameter" and "statistic". A statistic is an estimate based on sample data to draw inferences about a population characteristic of interest called the parameter.

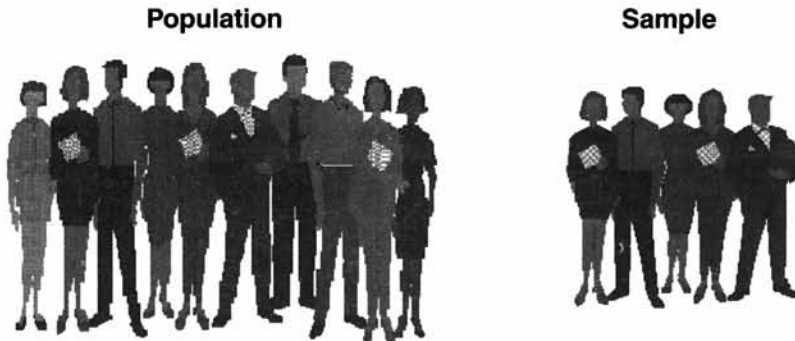


Figure 6.2

Why do you need Sampling?

Suppose a company is interested in launching a new product and wants to get some ideas about the demand potential. There are two ways of doing this:

1. It could ask all potential buyers in the country whether they will actually buy it, and if so how much would they buy.
2. It could take a sample of the potential buyers, ask them how many units of the product would they buy, and then estimate the likely demand for the product in the market as a whole.

The first approach is called a **Census** (also known as complete enumeration). It has two major disadvantages. 1) It is time-consuming 2) It is very expensive.

The second approach that uses **Sampling** procedure has two major advantages. 1) It is significantly less expensive 2) It takes the least possible time.

Also there are situations that involve destruction procedure where sampling is the only answer. A well-designed statistical sampling methodology would give accurate results and at the same time will result in cost reduction, and least time. You know how important it is for a manager to take decisions within reasonable time frame and therefore, sampling is the best option available.

Discussion Analyze, criticize, and explain the following statement.

"Suppose time and cost are not the consideration, 100 percent enumeration (Census) method gives 100 percent accuracy and is superior to Sampling".

6.2 TYPES OF SAMPLING

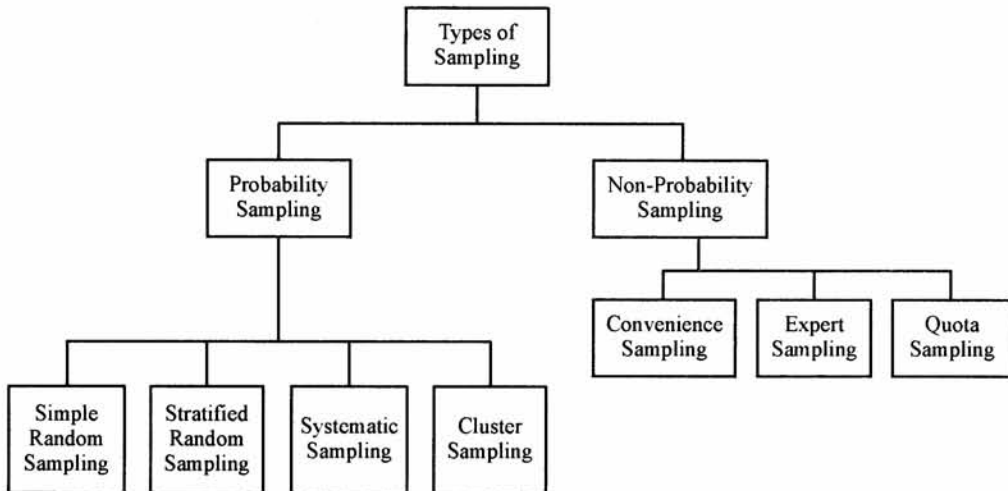


Figure 6.3

Probability Sampling (Random Sampling)

A probability sampling is a method of sampling that ensures that every unit in the population has a known non-zero chance of being selected. Please note that every potential sample need not have the same chance of selection. Practitioners have been using various forms of random selection, the most popular being a random number table. Today, computers have replaced the random number table and the software generates the random numbers in a scientific manner very fast.

Some Key Terms in Sampling

N = Number of units in the Population

n = Number of units in the Sample (Sample Size)

$\binom{N}{n}$ = Number of possible selection of n units from N units

$$\frac{N \times (N-1) \times (N-2) \dots \times (N-n+1)}{1 \times 2 \times 3 \dots \times n} = \text{(without replacement scheme)}$$

$f = n/N$ = Sampling fraction

Sampling Frame is a complete list of the units of analysis of interest from which the samples are selected.

Simple random sampling Simple Random Sampling is the foundation of Probability Sampling. It is a special case of probability sampling in which every unit in the population has the same chance of being selected. If you have to select n units out of N units, every possible selection of n units must have the same probability. Can you say how many ways

are possible to pick up n units out of N units? Of course, you can. Is it equal to $\binom{N}{n}$?

Yes. Refer the table giving some key terms in sampling in the previous page. Simple random sampling guarantees that every possible selection of n units from N units has the

same probability $\frac{1}{\binom{N}{n}}$. We are assuming here that the units are selected without replacement.

Example A bank wants to do a study on the customers' perception of its service quality in the last 12 months with regard to the savings bank account holders. First, you have to prepare the sampling frame for this study. You can go through the bank's records and get a complete list of savings bank account holders. This is your sampling frame. Suppose your sampling frame contains 500 account holders and you have to select 50 out of this 500 and interview them. How to actually draw a sample of 50 account holders out of the 500 account holders?

One method is to prepare 500 small paper slips, each giving the account holder's name and account number. Put these slips in a container, shuffle the container thoroughly and then select 50 slips one after the other from the container.

The efficient way is to use the Microsoft Excel to accomplish this task. First you enter the list of account holders' names into a column in the EXCEL spreadsheet. Then, in the next column paste the function =RAND() which is Excel's way of generating a random number between 0 and 1 into the cells. Now sort both the columns namely the list of account holders' names and the random numbers using the random numbers as the key for sorting. This will arrange the data set from the lowest to the highest random number. All you need to do is to pick up the first 50 names in this sorted list. Very simple. Is it not?

Progressive Test Questions

1. Probability Sampling is another name for Random Sampling. True or False

Answer True because in sampling techniques probability sampling and random sampling are synonymous.

2. You have to select 2 customers out of 10 customers for a depth interview. The number of ways in which you can select 2 customers out of 10 customers is:

(a) 90 (b) 60 (c) 45

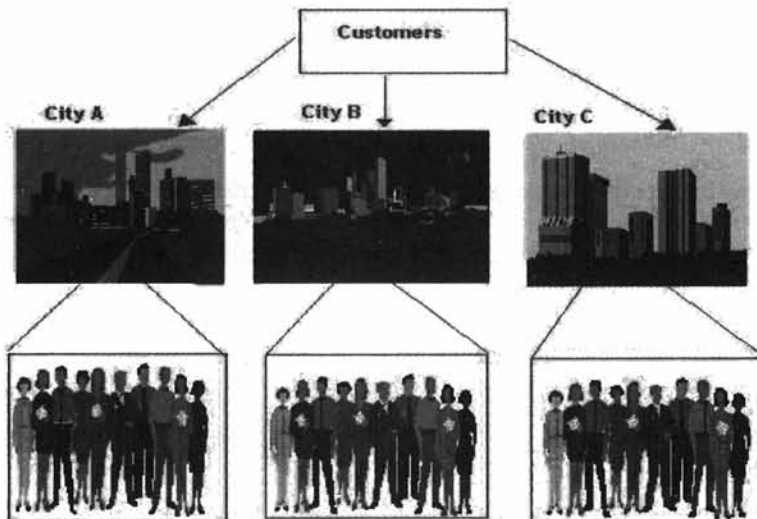
Answer The correct choice is (c). There are $\binom{10}{2}$ ways in which we can select 2 customers out of 10 customers = $(10 \times 9)/(1 \times 2) = 45$. Obviously, the other choices are wrong.

3. Generate 10 random numbers in the interval 0 to 1 using the paste function RAND () in the Microsoft Excel.

Answer One set of random numbers using the Excel is given below:

	A
1	0.846051
2	0.849978
3	0.731355
4	0.460597
5	0.590597
6	0.713204
7	0.751197
8	0.037322
9	0.790024
10	0.988960

Stratified random sampling



Stratified Random Samples from City A, City B, and City C

Figure 6.4

Imagine that you are working as a Marketing Manager in a Consumer Product company. Suppose you are studying the customer attitudes towards your product in order to improve your sales. Suppose there are three typical cities that will influence your sales. Suppose the customers within each city are similar and between cities are vastly different. Selection of the customers for the study has to be a random sample of customers chosen from each city so that meaningful and reliable inferences can be drawn, which in turn will enable the marketing manager to develop suitable strategies. This is an example of stratified random sampling.

Stratified Random Sampling involves dividing the population into a number of groups called strata in such a manner that the units within a stratum are homogenous and the units between the strata are heterogeneous. Having divided the population into a number of strata, now select a simple random sample of appropriate size from each stratum. The sample size in each stratum is equal to the overall sample size multiplied by the proportion number of units of that stratum to the total number of units in the population. This is called *proportionate stratified random sampling*. Suppose the overall sample size in a study is set at 200 units. You are interested in working out the sample size for stratum 1. Suppose stratum 1 has a total number of 2000 units and the population has 20000 units. The sample size for stratum 1 is $= 200 (2000/20000) = 20$.

You can also have a *disproportionate stratified random sampling* that requires some idea of the standard deviation of the distribution of the parameter of interest, within the strata. As this information is not easy to obtain managers may have to rely on intuition and logic to determine the sample size for each stratum. For example, bigger retail outlets may have greater variation in sales of certain products compared to small retail outlets. Therefore, it is appropriate to have a disproportionately large sample size for bigger retail outlets. One common strategy managers follow is that, first select equal sample size from each stratum and then give weights according to the stratum's proportion of total population.

Stratified random sampling is appropriate when the population is heterogeneous and you are keen to get a realistic picture of the overall population. An ordinary simple random sampling performed on the entire population that is heterogeneous will be highly misleading.

You would prefer stratified sampling to simple random sampling in the case of a heterogeneous population because it assures that you will be able to represent not only the overall population, but also the key strata of the population.

Systematic random sampling In systematic random sampling, the units are drawn from the population at regular intervals clearly defined. It is one of the easiest procedures to follow. The steps involved in constructing a systematic sampling scheme are as below:

- Compute $K = (N/n)$ and take the integer value. K is called the sampling interval.
- Select a random number between 1 and K .
- Starting with this number select every K th number until all the n units are selected.

Let us take an example to explain systematic sampling procedure: Suppose in a market survey, you have to select 5 households out of 50 households in a Block. The selection

procedure is pictured below. The table containing all the households is serially numbered from 1 to 50.

Systematic Sampling Procedure-An Example		
Number of units in the Population $N = 50$		
Number of units in the Sample $n = 5$		
Sampling Interval $K = (N/n) = (50/5) = 10$		
Select a random number between 1 and 10		
Suppose the selected random number is =5.		
Starting with 5, select every 10th unit. The selected units are highlighted		

1	21	41
2	22	42
3	23	43
4	24	44
5	25	45
6	26	46
7	27	47
8	28	48
9	29	49
10	30	50
11	31	
12	32	
13	33	
14	34	
15	35	
16	36	
17	37	
18	38	
19	39	
20	40	

In the systematic sampling procedure, it is necessary that the units in the population are randomly arranged on the basis of the characteristics you are measuring. Why should you use systematic random sampling? First, it is very easy to implement. You just have to select the first number at random. The rest of the units are determined automatically after the first random selection. Secondly, systematic sampling is more pragmatic than simple random sampling. Finally, there are situations where systematic sampling is the only way out especially when you have to strike a balance between precision on the one hand and cost on the other hand.

Cluster sampling (Area Random Sampling) One of the problems encountered with probability sampling methods is that you have to apply sampling procedure to a population that is scattered across a number of wide geographic regions. In these cases, you will have to cover a lot of distance in order to have access to the units you propose to sample.

Suppose you want to do a simple random sample survey of all the residents in India who belong to the highest income category. Your interviewers will have to do a tremendous amount of traveling. It is for this reason cluster-sampling method is followed. The steps involved in cluster sampling are:

- > Divide the population into a number of clusters based on geographic boundaries
- > Select a random sample of clusters from this population of clusters

- Either measure all units within the randomly chosen clusters or do further random sampling in each cluster.



Figure 6.5

Strictly speaking, when you measure all the units in the selected clusters, the procedure is called cluster sampling. Suppose you do further sampling within each cluster by adopting a simple random sampling or stratified random sampling, the procedure becomes a *multi-stage sampling*.

Progressive Test Questions

1. In which of the following sampling procedures are the N units in the population divided into separate groups each with a common characteristic?
 - (a) Systematic Sampling
 - (b) Cluster Sampling
 - (c) Stratified Sampling

Answer (c) is the right choice. In stratified sampling, you divide the population into a number of groups called strata in such a manner that the units within a stratum are similar (common characteristic). Homogeneity within a stratum and heterogeneity among strata are the distinguishing features of stratified sampling. Obviously, the other options are incorrect.

2. A dairy corporation that home delivers milk in a city would like to estimate the average milk consumption per household per month. If it picks up at random three of its fourteen delivery routes and obtains the relevant information for each household within these three routes, which type of sampling technique it has adopted?
- (a) Stratified Random Sampling
 - (b) Systematic sampling
 - (c) Simple random sampling
 - (d) Cluster Sampling

Answer (d) is the right choice. There are 14 delivery routes, which can be taken as 14 clusters. Three clusters are picked up at random and measurements taken on each household within the three routes (three clusters). This is the very definition of cluster sampling. Other choices are incorrect.

Non-probability sampling The fundamental difference between non-probability sampling and probability sampling is that in non-probability sampling procedure, the selection of the sample units does not ensure a known chance to the units being selected. In other words, the units are selected without using the principle of probability.

Even though the non-probability sampling has advantages such as reduced cost, speed, and convenience in implementation, it lacks accuracy in view of the selection bias. Another negative point of the non-probability sampling is its inability to generalize results from the sample to the population. It is mandatory in inferential statistics to use only probability sampling for valid conclusions. Non-probability sampling is suitable for pilot studies and exploratory research.

Convenience sampling Using college and university students in studies involving attitudes towards co-education is basically a matter of convenience. In consumer panel studies you may use clients who are available to you as your respondents for giving their opinion on products and services. In many research projects, you simply look for volunteers to participate. This is how the convenience sampling is done. For heaven's sake, don't generalize results based on convenience sampling.

Expert opinion sampling Expert Opinion Sampling involves gathering a set of people who have the knowledge and expertise in certain key areas that are crucial to decision making. In qualitative methods of demand projection for a new product, you use the expert opinion method to arrive at a reasonable forecast. The advantage of this sampling is that it acts as a support mechanism for some of your decisions in situations where virtually no data are available. The major disadvantage is that even the experts can have prejudices, likes, and dislikes that might distort the results.

Quota sampling In simple terms, quota sampling is stratified random sampling without probability principle being applied to the selection of the sample units. Suppose in an opinion study, you want both men and women to participate. You know that in the population category of interest, 65% are men and 35 % are women. If your sample size is fixed at 200, you will have a quota of 130 men and 70 women. It doesn't matter how you

get them as long as you have met the quota. There are some socio-economic studies where quota sampling is the only way out because of practical considerations. You can do the descriptive statistics, graphs, charts, and summary table and stop there. That is it. Drawing any possible conclusions from a quota sampling will be highly tentative. None of the statistical inference techniques should be applied when you have followed quota sampling or for that matter any non-probability sampling procedure

6.3 SAMPLING DISTRIBUTION -A CONCEPTUAL FRAMEWORK

How do you go about using a sample statistic to estimate a population parameter? To answer this question, you will have to understand the concept of the sampling distribution. The entire inferential statistics is built on the foundation of the sampling distribution.

Sampling Distribution -Definition

The probability distribution of all the possible values a sample statistic can take is called the **sampling distribution** of the statistic. The key word here is "sample statistic". Sample mean and sample proportion based on a random sample are examples of sample statistic(s). Please note that we are not interested in the probability distribution of a set of numbers. We are interested in the probability distribution of a statistic which can assume different values in an experiment that involves taking a large number of times random samples of same sample size from a population, and computing the statistic afresh every time.

Cautionary Note There is a feeling among many students that the expression "sampling distribution" automatically implies the sampling distribution of the mean. This is not correct. You can have a sampling distribution of the median as well. Please notice that the sampling distribution of the mean is not the same as the sampling distribution of the median. They are entirely different.

In order to understand the sampling distribution better, let us look at an example. Suppose that you take a sample of four youth from a population of youth numbering 10000. The mean weight of the four youth is worked out. Again take another fresh sample of four youth from the same population and work out the mean weight. If you repeat this process an infinite number of times, the probability distribution of these infinite number of sample means would become the sampling distribution of the mean. Let us assume that the mean weight of the first sample is 65 kg. The mean weight of the second sample is 70 kg, and so on. Suppose you keep on taking a sample of 4 youth from the same population and keep on working out the sample mean every time, then the mean of the means of the infinite samples will approach the true mean of the population. Let us assume that the true mean is 70 kg. The histogram and the distribution of the sample mean for this example will look like the following bell shaped normal distribution. (figure 6.6).

This is the concept of the sampling distribution. Likewise, you can also have a sampling distribution of the median, if you compute the median of each sample instead of the mean. From the point of view of statistical inference, sampling distribution of the mean is very important. Shortly, we will focus on the sampling distribution of the statistic (mean) when samples are drawn from a normal population. Suppose the original distribution is not normal, what happens? Relax! You have the central limit theorem to help you out. Even in the example above, the original population is not normal!

Sampling Distribution of the Mean weight of the Youth

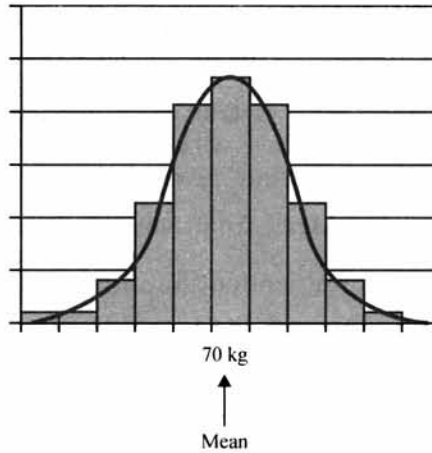


Figure 6.6

6.4 THE CONCEPT OF STANDARD ERROR

What is the standard deviation of the sample statistic called? Can you guess? It is called the Standard Error of the Statistic. In other words, the standard deviation of the Sample Statistic is called the Standard Error of the Estimate. Please note that any sample statistic is an estimate that is used to estimate the population parameter. The standard deviation of the distribution of the sample means is called the *standard error of the mean*. Likewise, the standard deviation of the distribution of the sample proportions is called the *standard error of the proportion*.

Let's explain once more. A standard deviation is the spread or departure of the values from the mean in a single sample. The standard error is the spread or departure of the values from the mean of means in a sampling distribution. You got it now?

In the context of sampling, the standard error is popularly known as the *sampling error*. We told earlier that a sample statistic is an estimate. Sampling error throws light on the precision and accuracy of our estimate. The logical question is how to compute the sampling error? Can you say how? Of course you can, if you think intuitively. The larger the sample standard deviation, the larger is the standard error. The larger the standard error, the larger is the sampling error.

Standard error and the sample size The standard error is inversely proportional to the sample size. The larger the sample size, the smaller is the standard error. Can you guess why? Because, as the sample size increases, the standard error goes on declining and approaches the value zero, theoretically, when n tends to ∞ . In practice, this implies that for a large sample size, the sample mean is almost equal to the population mean barring some minor deviations. In the next chapter (chapter 7) you will learn how to determine the sample size for a given situation by using the concept of confidence interval and standard error. At this stage, it is enough if you know how important the standard error is in the context of the sampling distribution.

6.5 SAMPLING DISTRIBUTION OF THE MEAN FROM NORMAL POPULATION

If you have gone through the initial discussion on the concepts of the sampling distribution of the mean along with the example given, the sampling distribution of the mean from the normal population is a logical extension of the same principle. The samples are randomly drawn from a normal population with certain mean and standard deviation. The original population is distributed normally. You have a clear structure for the sampling distribution of the mean with terms and notations that are given below:

Sampling Distribution of the Mean from Normal Population - Statistical Description

- If $X_1, X_2, X_3, \dots, X_n$ are n independent random samples drawn from a Normal Population with Mean = μ , and Standard Deviation = σ , then the sampling distribution of \bar{X} follows a Normal Distribution with Mean = μ , and Standard

$$\text{Deviation} = \frac{\sigma}{\sqrt{n}}.$$

- $\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n}$ = the Sample Mean. Please note that \bar{X} is a random variable and will be different every time when a fresh sample of n observations are taken.

- \bar{X} is called an Unbiased Estimator of the Population Mean, μ . This implies that $E(\bar{X}) = \mu$. Symbolically, this is written as $\mu_{\bar{x}} = \mu$.

- The Standard Deviation of the Sample Mean $\bar{X} = \frac{\sigma}{\sqrt{n}}$. This is written as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

- $\sigma_{\bar{x}}$ is called the Standard Error of the Sample Mean \bar{X} . You will appreciate that the standard error keeps on decreasing as n keeps on increasing.

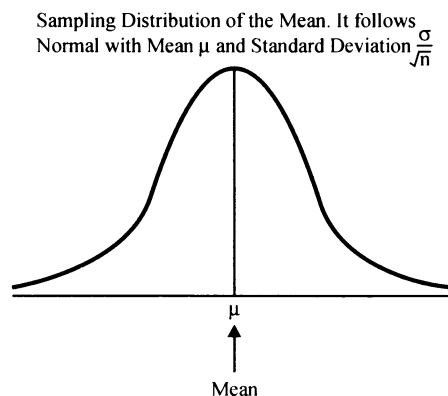


Figure 6.7

Are the mathematical symbols and expressions in the previous page confusing and mystifying? Don't worry. To remove the mystic feeling, let us take the very same example that was discussed earlier in the context of the concept of a sampling distribution. The only modification we make in that example is that the original youth population of interest follows a normal distribution with mean weight 70 kg and standard deviation 5 kg.

Suppose that you take a sample of four youth from this normal population of youth numbering 10000. The mean weight of the four youth is worked out. Again take another fresh sample of four youth from the same population and work out the mean weight. If you repeat this process an infinite number of times, the probability distribution of these infinite number of sample means would become the sampling distribution of the mean. Let us assume that the mean weight of the first sample is 65 kg. The mean weight of the second sample is 70 kg, and so on. Suppose you keep on taking a sample of 4 youth from the same population and keep on working out the sample mean every time, then the mean of the means of the infinite samples will approach the true mean of the population. So what can you say about this sampling distribution?

1. The mean weight of the youth follows a Normal Distribution with Mean = $\mu = 70$ kg and Standard deviation = $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{4}} = 2.5$ kg. In other words $\mu_{\bar{x}} = 70$ kg and $\sigma_{\bar{x}} = 2.5$ kg.
2. \bar{X} is an Unbiased Estimator of the Population Mean μ which says that the $E(\bar{X}) = \mu$. That is, the mean of $\bar{X} = \mu_{\bar{x}} = 70$ kg.
3. What is the Standard Deviation of the Sample Mean \bar{X} ? Of course, it is $\sigma_{\bar{x}}$ and is = 2.5 kg. Is not the standard deviation of \bar{X} called the standard error? Yes.
4. One interesting feature you notice here is that the standard deviation of the original distribution is 5 kg where as the standard deviation of the sampling distribution of \bar{X} is only 2.5 kg. So the standard deviation of the sampling distribution of the mean is always less than the standard deviation of the individual items in the population.

What impact does the Standard Error have on the Sampling Distribution when you increase the sample size? Look at the following diagram.

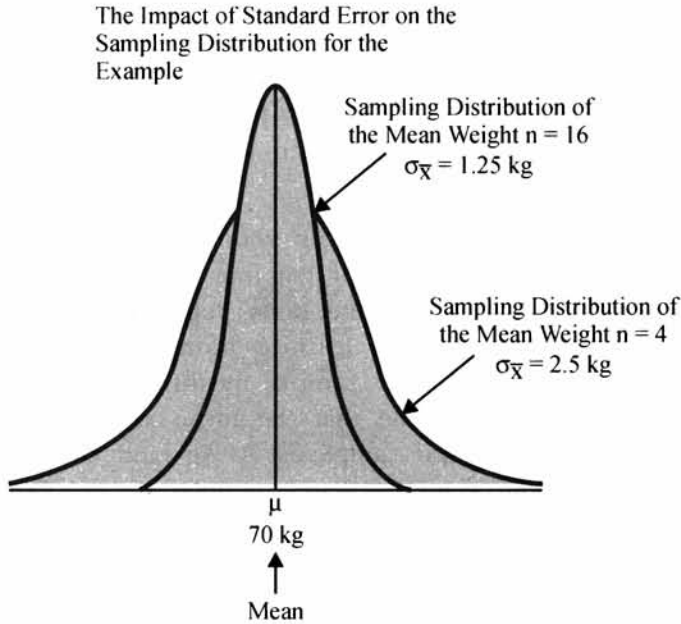


Figure 6.8

In our example, we have taken that the youth population is normally distributed with mean = 70 kg and a standard deviation = 5 kg. The first experiment involves drawing a random sample of 4 observations from this population and the second experiment involves drawing a random sampling of 16 observations. The two normal distributions have the same mean but the spread is not the same. As you can see the standard error of the sample mean is 2.5 kg when $n = 4$ and 1.25 kg when $n = 16$. How far can you go on increasing the sample size? Even though the aim is to reduce the sampling error, you have to strike a balance between the cost on the one side and the accuracy on the other side while working out the optimal sample size. The standard error plays a major role in deciding the right kind of sample size.

Example The marks obtained by students in an aptitude test are normally distributed with a mean of 60 and a standard deviation of 20. A random sample of 16 students is drawn from this population.

- (a) What is the standard error of the sampling mean?
- (b) What is the probability that the mean of a sample of 16 students will be either less than 50 or greater than 80?

Solution

(a) The standard error of the sample mean is given by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{16}} = 5$.

(b) Let us draw the picture of the sampling distribution of the mean.

Sampling Distribution of the Mean for the Example

$$n = 16 \quad \mu_{\bar{x}} = 60 \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 20 / 4 = 5$$

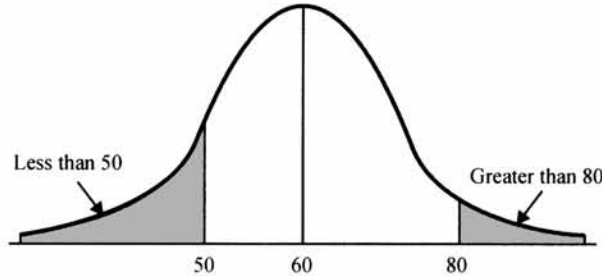


Figure 6.9

If you add the probability of the shaded area of the two portions, you get the answer. Using the paste function in the Microsoft Excel, you can work out these two probabilities. You don't even need to convert the original variable into z. Let me show you step by step how to calculate the cumulative probability for portion 1 (up to 50) using Excel. Click the paste function and you see the paste function lay out given below:

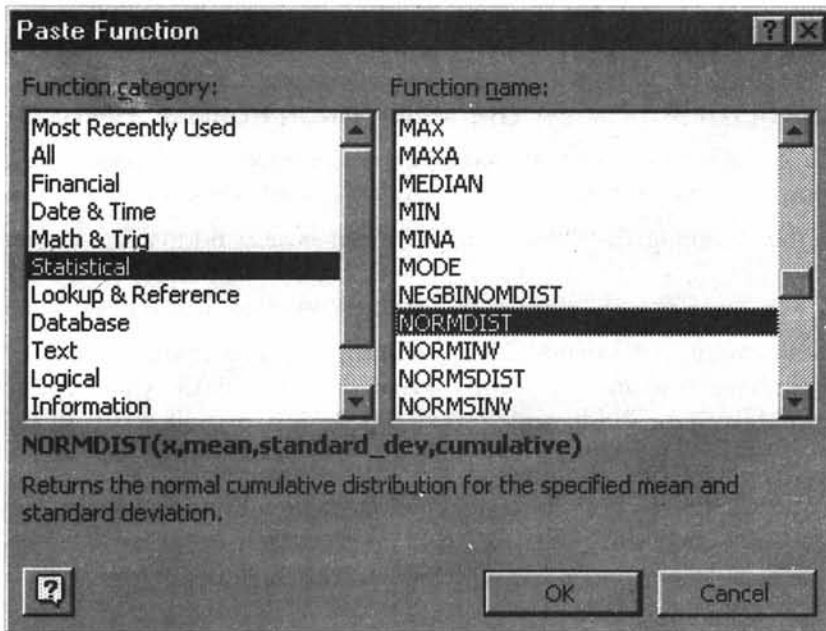


Figure 6.10

The function NORMDIST is highlighted in the paste function above. Click OK and you get the function NORMDIST. Enter the values in the cells. Now you get the NORMDIST function with values entered given below:

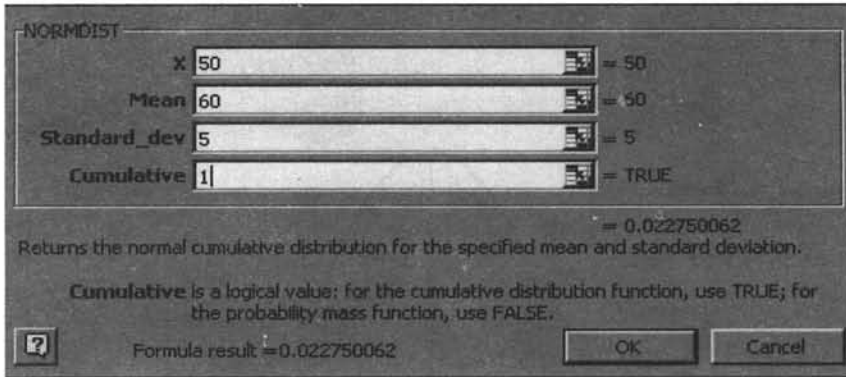


Figure 6.11

Now either you click OK or alternatively, you see the answer corresponding to the "Formula result". The answer is 0.022750062 or = 0.0228(4 decimal places).

The cumulative probability greater than 80 is = 1 - cumulative probability up to 80. Repeat the same procedure you get cumulative probability up to 80 = 0.99997. The required probability for portion 2 = 1 - 0.99997 = 0.00003.

The answer to the example question is = .0228 + .00003 = .0228 (rounded to 4 places of decimal). This means that there is about 2.28% chance that the mean score will be either less than 50 or greater than 80.

6.6 SAMPLING DISTRIBUTION OF THE MEAN - NON-NORMAL POPULATION

So far, we have discussed in detail the sampling distribution of the mean when the samples are drawn from a normal population with mean μ and standard deviation σ . We have concluded that the sampling distribution of the mean is also normal with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. What happens to the sampling distribution of mean when the original distribution is not normal? Managers in real life situation have to deal with population distributions that are not normal. What do you do? Do you have a solution? Yes. You have fortunately a brilliant and amazing solution provided by the *Central Limit Theorem*.

In simple terms, the *central limit theorem* says that from any given population with mean μ and standard deviation σ , if we pick up a random sample of n observations, the sampling distribution of the mean will approach a normal distribution with a mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size increases and becomes large.

The distinguishing and unique feature of the central limit theorem is that irrespective of the shape of the distribution of the original population, the sampling distribution of the mean will approach a normal distribution as the size of the sample increases and becomes

large. How large is large? A thumb rule based on experience says a sample size of 30 and above is considered large. It works reasonably well in a large number of problems. Please remember that n is the sample size for each mean computed and not the number of samples. Also note that theoretically speaking you compute the mean for infinite number of samples. Of course, in practice this could mean a sufficiently large number of samples.

Diagram depicting the Central Limit Theorem

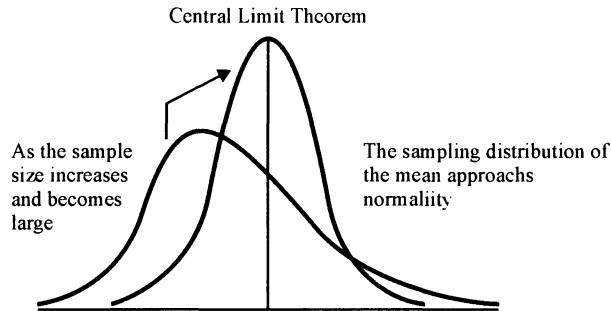


Figure 6.12

Central Limit theorem is indeed a hallmark of statistical inference. It permits a manager to make inference about the population parameter based on random samples drawn from populations that are not normally distributed. Now you read again attentively the first example we have used to explain the concepts of the sampling distribution. The shape of the curve there is a bell shaped normal because of the central limit theorem.

6.7 CHAPTER SUMMARY

This chapter has exposed you to the basics of sampling techniques and the concepts of sampling distribution. Remember that the entire inferential statistics is built on the foundation of the sampling distribution. In particular, this chapter focused on:

- The meaning and the need for sampling.
- The types of sampling categorized into probability sampling and non-probability sampling.
- The most widely used sampling methods that include simple random sampling, stratified random sampling, systematic sampling, and cluster sampling under probability sampling and convenience sampling, expert opinion sampling, and quota sampling under non-probability sampling.
- The conceptual framework of the sampling distribution.
- The meaning and role of standard error.
- The sampling distribution of the mean from normal population.
- The sampling distribution of the mean from non-normal population interwoven with the central limit theorem.

➤ How to use Microsoft Excel for solution to problems in this chapter.

GLOSSARY

Census Complete enumeration of every unit in the population.

Central limit Theorem Central limit theorem assures that irrespective of the shape of the distribution of the original population, the sampling distribution of the mean will approach a normal distribution as the size of the sample increases and becomes large.

Cluster sampling It is one of the probability sampling methods in which the population is divided into a number of clusters. A random sample of clusters is chosen and then all the units in the selected clusters are measured.

Convenience Sampling It is a non-probability sampling procedure in which you simply look for volunteers to participate in your study. Selection of sample units is just a matter of convenience.

Probability Sampling It is a method of sampling that ensures that every unit in the population has a known non-zero chance of being selected.

Random Sampling Same as probability sampling

Sampling Distribution The probability distribution of all the possible values a sample statistic can take is called the sampling distribution.

Sampling Error It is the difference between the sample statistic and the population parameter based on a random sample of observations.

Sampling Fraction It is the ratio between the sample size and population size.

Sampling Frame It is a complete list of the units of analysis of interest from which the samples are selected.

Simple Random Sampling It is a special case of probability sampling in which every unit in the population has the same chance of being selected.

Standard Error The standard deviation of the Sample Statistic is called the Standard Error.

Stratified Sampling Stratified Random Sampling involves dividing the population into a number of groups called strata in such a manner that the units within a stratum are homogenous and the units between the strata are heterogeneous. Having divided the population into a number of strata, now select a simple random sample of appropriate size from each stratum.

Systematic Sampling It is a probability sampling in which the units are drawn from the population at regular intervals clearly defined. The regular interval is computed by taking the integral part of the reciprocal of the sampling fraction.

REVIEW QUESTIONS

1. The standard deviation of the distribution of the sample means is called:
 - (a) The standard deviation of the sample
 - (b) The standard error of the sample
 - (c) The standard error of the mean

2. The central limit theorem states that irrespective of the shape of the original population distribution, the sampling distribution of the mean will approach the normal distribution:
 - (a) As the size of the population standard deviation increases
 - (b) As the sample size increases and becomes larger
 - (c) As the number of samples gets larger
3. When the original population distribution is not normal, the central limit theorem assures the managers that the sampling distribution of the mean will be normal and helps them in making inferences about the population parameter regardless of the scheme of sampling followed, as long as the sample size is large. True or False.
4. If a random sample of size 64 is taken from a population whose standard deviation is equal to 32, then the standard error of the mean is:
 - (a) 0.50
 - (b) 2.0
 - (c) 4.0
 - (d) 32
5. Data collected by employing the non-probability sampling can be very effectively used to make inferences about the population parameter. True or False.

Questions 6 to 8 should be answered based on the following mini case.

MINI CASE

A company is seriously considering buying one type of machine that can save significant labor hours. The labor hour saved by the machine follows a normal distribution with mean = 2200 and a certain standard deviation. It is known that there is a fifty- fifty chance that the labor hours saved by the machine is either greater than 2400 or less than 2000. The price of the machine is Rs. 860000. The incremental cost of a labor hour incurred in the company is currently Rs. 400. The company has also performed 36 trial runs to find out how the machine is fairing. The company would like to make some preliminary assessment before buying the machine. It requires your help to answer the following fill in the blank questions:

6. The standard deviation of the population distribution of the labor hours saved by the machine is -----.
7. The standard error of the sample mean of labor hours saved is -----.
8. The probability that the sample mean of labor hours saved will exceed the break-even labor hours is -----.

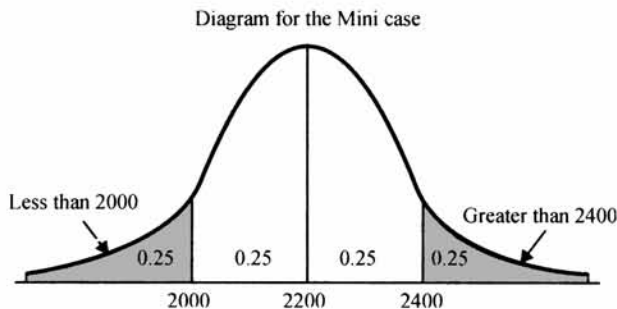
ANSWERS TO REVIEW QUESTIONS

1. (c) is the correct answer. Please read the definition and meaning of the standard error of the mean. The standard error of the mean is nothing but the standard deviation of the distribution of the sample means.
2. (b) is the right answer. As the sample size increases and becomes larger, the sampling distribution of the mean approaches normality regardless of the shape of the original distribution of the population. The other options are obviously wrong.

3. The statement is false because even though the original distribution of the population need not be normal for the central limit theorem to operate, the samples drawn from the population must be random samples (following the probability sampling scheme). If the samples are drawn based on non-probability sampling, then the sampling distribution of the mean need not be normal.
4. (c) is the right answer. The standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{32}{\sqrt{64}} = 4.0$. Other options are wrong.
5. The statement is false. The very weakness of the non-probability sampling procedure is that the sample results cannot be used to make inferences about the population parameter because the selection of the units from the population does not involve the probability principle. At best, the data of the non-probability sampling can be used for descriptive statistics that include summarization, tabulation, and charts.

Answers to Mini Case It is an interesting problem giving ample scope to the students to think and act. The standard deviation of the original distribution is not given. Is it 200? If you say yes, then you are wrong.

Some basic facts The cost of the machine is Rs. 860000. The incremental cost of a labor hour = Rs. 400. Therefore the break-even point labor hours = $860000/400 = 2150$. If the machine can save more than 2150 labor hours, it can be considered for buying. Then, the case says that there is a 50-50 chance that the labor hours saved by the machine is either greater than 2400 or less than 2000. What do you mean by this? It means that $P(X > 2400) + P(X < 2000) = 0.50$ where X is the random variable denoting the mean labor hours saved. Can you draw the correct diagram to the problem? Then you can easily answer the questions of the case.



Please note that the property of symmetry divides the area into 4 equal parts of 0.25. Can you say why? If the shaded areas are added, it must equal 0.50 because the case says that there is a 50-50 chance that the labor hours saved is either less than 2000 or greater than 2400. Naturally, the remaining area between 2000 and 2400 must be equal to 0.50. Applying the property of symmetry of the normal curve, the area between 2000 and 2200 must be the same as the area between 2200 and 2400. Hence, each one of the areas must be = 0.25. Therefore, it follows that all the four portions of the curve = 0.25.

6. The answer is 297 hours. The cumulative probability up to 2000 = 0.25 (given). It is easy to solve this problem using z because you can get the value of z corresponding to the cumulative probability using NORMSINV of Excel. Here $z = (x - \mu)/\sigma = (2000 - 2200)/\sigma = -200/\sigma$. For what value of $-200/\sigma$, is the cumulative probability = 0.25. The steps are given below one by one. First, click Paste Function in Excel.

The following paste function will appear:

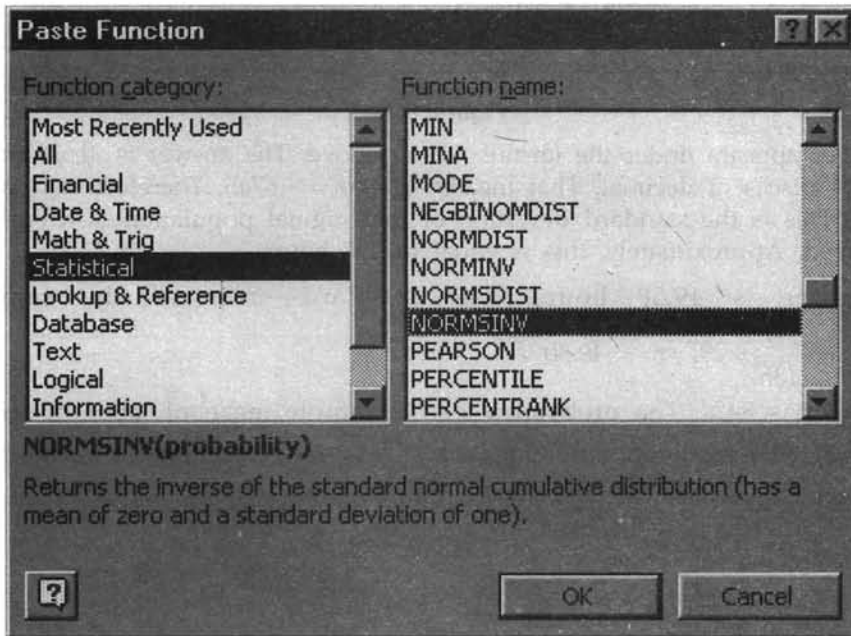


Figure 6.14

You see in the above function NORMSINV is highlighted corresponding to "Statistical" on the left side. Click OK and you will get:

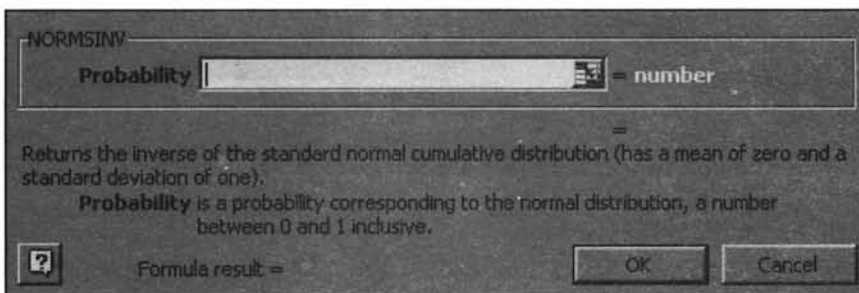


Figure 6.15

Enter the probability value 0.25 in the cell above and then click OK. You will get:

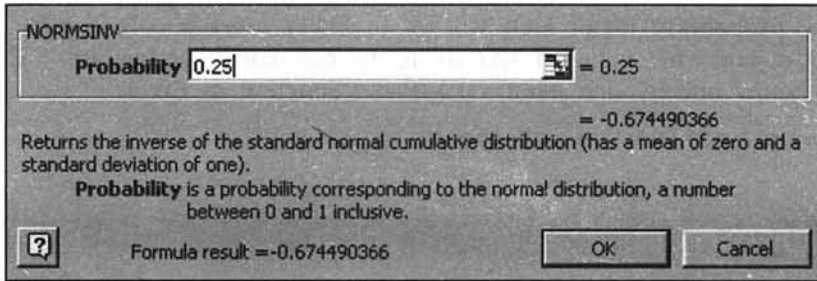


Figure 6.16

The answer appears under the formula result above. The answer is -0.6745 if we round it off to 4 places of decimal. That means $-200/\sigma = -0.6745$. Therefore $\sigma = 200/0.6745 = 296.5159$. This is the standard deviation of the original population distribution of labor hours saved. Approximately, this is equal to 297 hours.

7. The answer is 49.50 hours. The standard error of the sample mean $= \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{297}{\sqrt{36}} = 297/6 = 49.50$.
8. The answer is 84%. The probability of the sample mean of labor hours exceeding break-even point labor hours is same as $P(\bar{X} > 2150)$ is wanted. Using the NORMDIST we can get the answer directly. Please note that the standard deviation of the sample mean that is 49.50 is to be used for answering this question. Click Paste Function in Excel. You get:

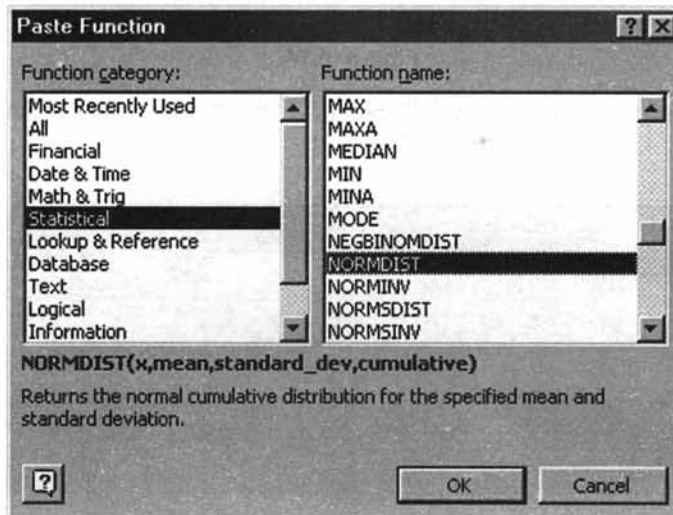


Figure 6.17

You see NORMDIST is highlighted against Statistical. Click OK and you get:

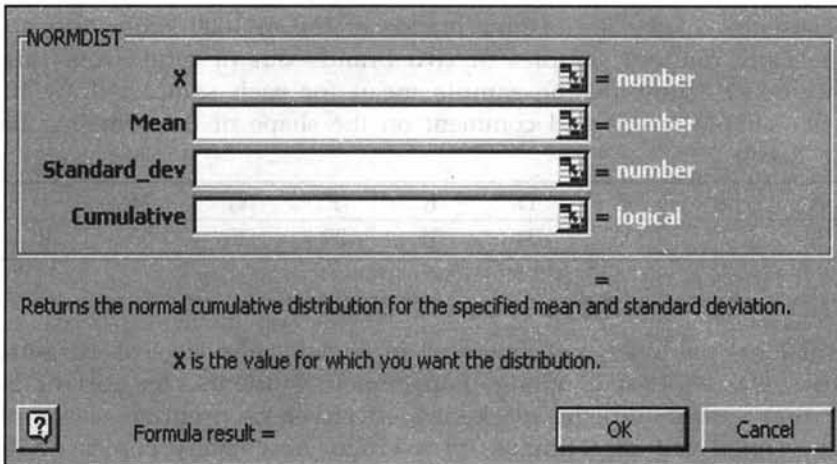


Figure 6.18

Enter the value 2150 for X, mean =2200, standard deviation = 49.5, and 1 in the cell of cumulative. Click OK. You get:

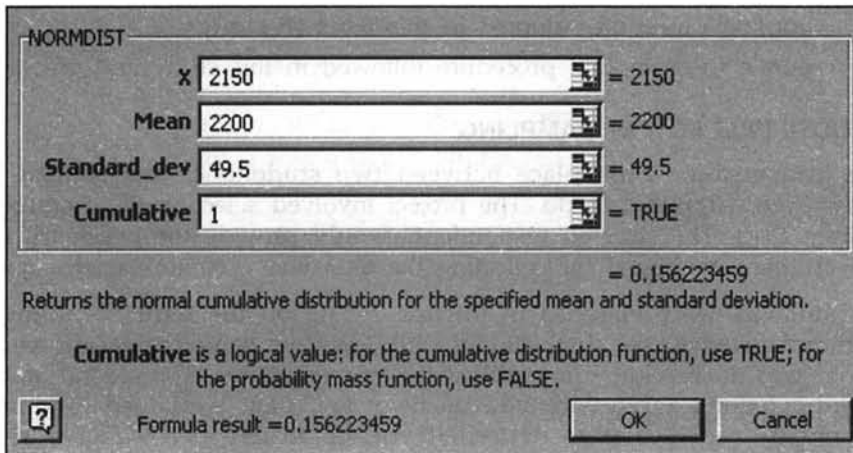


Figure 6.19

In the above table of Excel you see the cumulative probability up to 2150 = 0.1562. Therefore, the probability of getting more than 2150 = $1 - 0.1562 = 0.8438$. That is, there is about 84 % chance that the distribution of the mean labor hours saved by the machine will be more than the break-even labor hours when a sample of 36 trials is taken.

PRACTICE PROBLEMS**1. CASE STUDY - TIRE LIFE**

A population consisting of the life of nine brands of tire in 1000 kilometers is given below. Generate all possible random samples of two brands out of nine, from this population (without replacement). Compute the sample mean for each sample; draw the frequency histogram for the sample means and comment on the shape of the sampling distribution of the mean. (Use Excel)

Tire Brand	A	B	C	D	E	F	G	H	I
Life	30	29	33	34	31	29	33	33	30

2. CASE STUDY - BOOK EXPOSURE ON STUDENTS

A leading international publisher was interested in knowing the level of exposure, a book on job opportunities was receiving among management students. As part of an important research study that looked into the marketing effectiveness program, a sample survey of bookshops was carried out to estimate, on average, how many copies of this book are displayed on shelves in various bookstores. The publisher selected at random a sample of twenty universities and a further sample of five colleges offering business management programs from each of these 20 universities. Every bookstore in the vicinity of every sample college selected was personally visited and the number of books on shelves counted.

QUESTIONS

1. Label the sampling procedure adopted in this research study?
2. Critically examine the sampling procedure followed in this study and give your views.

3. CASE STUDY- FALLACY IN SAMPLING

The following conversation took place between two students in the context of a group project that they were supposed to do. The project involved selection of respondents using a suitable sampling plan. The selected respondents would provide the necessary data for the project. The instrument envisaged for collecting the data was a comprehensive questionnaire.

The two students were engaged in intense debate on the term "Random Sampling". There was commotion when the first student said, "random sampling means you just go to any body whom you do not know and fill in the questionnaire". The second student said to the first student " I don't agree with you the way you have explained random sampling. How do you ensure representative pattern if you go to any person?" For this, the first student replied "Very simple. Select using area map of the city. Cover all typical areas. Then in each area select any unknown person for the interview. Keep on doing this until you have got the requisite sample size. Fill in the questionnaire from these chosen respondents and your sample survey using random sampling is done".

QUESTIONS

1. Do you agree with the meaning of random sampling explained by these two students? If your answer is "Yes" why do you say so? If your answer is "No" why do you say so?
2. How would you select the respondents in a random fashion for this study?

Estimation

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Define and Compute Point Estimation
- Define and Compute Interval Estimation
- Determine Sample Size based on Confidence Interval

CHAPTER OUTLINE

- 7.1 Point Estimation
 - 7.2 Interval Estimation
 - 7.3 Confidence Interval for Population Mean and Proportion- Large Sample
 - 7.4 Confidence Interval for Population Mean-Small Sample(t -Distribution)
 - 7.5 How to Determine Sample Size Using Confidence Interval
 - 7.6 Chapter Summary
- Glossary
Review Questions
Answers to Review Questions
Practice Problems

INTRODUCTION

Marketing Manager in an organization needs to estimate the likely market share his company can achieve in the market place. Quality Assurance Manager may be interested in estimating the proportion defective of the finished product before shipment to the customer. Manager of the credit department needs to estimate the average collection period for collecting dues from the customers. How confident are they in their estimates? This chapter provides some insights into point estimation and interval estimation that are essential in business planning. Please remember that the three components-*point estimation*, *interval estimation*, and *hypothesis testing* together constitute the all-important *inferential statistics*.



Figure 7.1

7.1 POINT ESTIMATION

Marketing manager of an enterprise is interested in the population (universe) value, not in a sample value! A sample is of interest, as long as it throws light on the population.

Point Estimation deals with the task of selecting a specific sample value as an estimate for a population parameter. The picture given below along with the description is an example of a point estimate.

The percentage of housewives who prefer model X refrigerator in a random sample of 250 housewives is another example of a point estimate.

Average Income of Purchasers of Brand X Car based on a random sample of 150 purchasers is \$ 50000 per Annum. This is an Example of a Point Estimate.

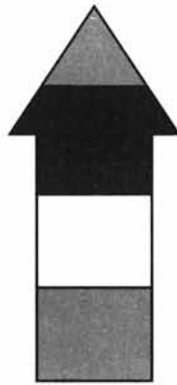


Figure 7.2

A *point estimate* is a specific value of a sample statistic that is used to estimate a population parameter.

Point Estimation - Population Mean

The sample mean \bar{X} is an unbiased estimator of the population mean, μ . An unbiased estimator is one whose expected value is equal to the population parameter. That is, $E(\bar{X})$ is equal to μ . Of course the samples drawn must be independent random samples from the population of interest. For better clarity, read the following example.

Example 1 The marketing manager of a company selling carpet cleaners is interested in estimating the average value (in \$) of sales per call. The marketing manager is interested in this estimate for the future and not for the past. The company follows a policy of feeding its sales persons with leads obtained from enquiries and referrals. The results of calls made from these leads are maintained in a comprehensive customer database. The marketing manager takes a simple random sample of size 40 from these records. The arithmetic mean sales per call is equal to \$ 59.50. What is the best point estimate for the corresponding

population mean? It is \$ 59.50. Why? Because the sample mean is an unbiased estimator of the population mean. The criterion for the "best" is based on the property of unbiasedness. Here, $\bar{X} = 59.50$ and we say that it is an unbiased estimator of the population mean, μ . The meaning in this context is that the arithmetic mean sales value per call based on the sample salespersons is a true reflection of the population average sales value per call based on the total calls made by all the salespersons in the company.



Figure 7.3

Point Estimation-Population Proportion

First you must try to explain as to what do you mean by the word proportion? A *proportion* is a special type of arithmetic mean, special in the sense that the individual elements can take the value either 0 or 1. In a survey if you select a random sample of respondents and ask them whether they have a deluxe car, you may get a reply- yes or no. You will assign "1" to each person who says yes and "0" to each person who says no. The values are summed and the sum is divided by the total number of respondents. The interpretation of a proportion then becomes similar to the interpretation of an arithmetic mean. In other words, proportion is a particular case of arithmetic mean. Let us designate the population proportion as P and sample proportion as p . Sample proportion is an unbiased estimator of the population proportion. That is p is an unbiased estimator of P . So $E(p) = P$.

A particular value of p based on a sample survey becomes a point estimate.

In symbols $p = \frac{\sum X_i}{n}$ where $X_i = 1$, if the i th unit in the sample possesses an attribute of interest such as having a deluxe car. $X_i = 0$, if the i th unit in the sample does not possess the attribute.

Applying the central limit theorem, if the sample size is large, regardless of the shape of the population distribution, the distribution of the sample proportion follows a normal

distribution with mean = P and standard deviation = $\sqrt{\frac{P(1-P)}{n}}$. The standard error of the proportion $\sigma_p = \sqrt{\frac{P(1-P)}{n}}$. We avoid the derivations here because they are not necessary.

Suppose the marketing manager in the previous example (example1) is also interested in estimating the percentage of calls resulting in sales. A random sample of 150 records from the customer database reveals that 20% of the leads results in sales. The sample proportion of leads resulting in sales (p) is an unbiased estimator of the population proportion of leads resulting in sales. The practical implication is that in the population of leads, 20% will result in sales. Suppose the total leads in the population are 1500, 20% of these leads will result in sales. That is 300 out of 1500 leads will result in sales.

These two statistic(s) namely the sample mean and the sample proportion are encountered very frequently in research studies. What percentage of last month's buyers were first time buyers? What is the per capita household consumption of gas in a middle class family? How much time a week does the average youth watch TV? What proportion of major consumer durable purchases are joint decisions by husband and wife? All these examples refer to a proportion or a mean. In each case the sample statistic is an unbiased estimator of the true population parameter, provided random sampling procedure is employed.

Progressive Test Questions A supermarket is concerned about the presence of bacterial infection in a food package. A random sample of 64 packages is selected and sent to a laboratory for test. A score between 0 and 100 is possible in the test result. 0 means totally free from infection and 100 means very highly infected. Tests on the sample of 64 result in a mean score of 20. It is known from pilot studies conducted on such food products that the infection score is normally distributed with a standard deviation of 4.

1. What is the point estimate for the population mean score? (The mean score of all the food packages in the supermarket)

Answer The sample mean is an unbiased estimator of the population mean. The sample mean based on a random sample of 64 is a score of 20. Hence, the point estimate here is 20.

2. The standard error of the sample mean is -----.

Answer The standard error of the sample mean is 0.5.

The standard error of the sample mean = $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{64}} = 4/8 = 0.5$

7.2 INTERVAL ESTIMATION

A point estimate cannot be expected to coincide exactly with the population parameter. Suppose in a survey you find that the average income of a household is Rs. 300000 per year. Is it that the income of every household is Rs. 300000 per year? Certainly not. Some households may have more than Rs. 300000 and some may have less than this amount. In

other words point estimate will not coincide with the population parameter. How do we cope with this problem? Interval Estimation comes to our help in this regard. Interval Estimation establishes an interval consisting of a lower limit and an upper limit in which the true value of the population parameter is expected to fall. This interval is called "**Confidence Interval**" in the parlance of inferential statistics. The meaning of the expression confidence interval is that if you keep on taking repeated samples, the probability that the true value of the population parameter will fall in this interval is a certain percentage. The convention is to use a 95 % confidence level, and some times 99% confidence level. Suppose you choose 95% as the confidence level. Let us interpret the meaning once more. If you keep on taking repeated samples, the probability that the true value will fall in this interval is 95%. In other words, you are 95% confident that the true value of the population parameter will fall in this interval. The actual establishment of the confidence interval is built on the sampling distribution principle. The following diagram captures the meaning of interval estimation.

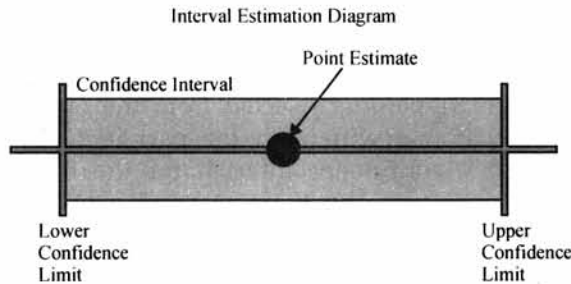


Figure 7.4

The methods of constructing the confidence interval for the population parameters - mean and proportion are discussed next.

7.3 CONFIDENCE INTERVAL FOR POPULATION MEAN AND PROPORTION - LARGE SAMPLE

Confidence Interval for Population Mean (σ Known)

The $(1- \alpha)$ Confidence Interval for the Population Mean μ is given by:

$$\bar{X} - Z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z \frac{\sigma}{\sqrt{n}}$$

Where \bar{X} is the sample mean based on a random sample of size n and would vary for repeated random samples of the same size n .

$\frac{\sigma}{\sqrt{n}}$ is the standard error of the distribution of sample mean \bar{X} .

μ is the Population Mean.

Z is the value corresponding to the area of $\left(\frac{1-\alpha}{2}\right)$ from the mean of the standard normal distribution.

α is the proportion in the tails of the standard normal distribution that is outside the range of the confidence interval.

The meaning of the confidence interval given in this section is described below:

If you take repeated independent random samples of size n from a population with an unknown mean but known standard deviation, the probability that the true population mean

μ will fall in the interval $\bar{X} - Z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z \frac{\sigma}{\sqrt{n}}$ is $(1 - \alpha)$. α is a measure of risk indicating the percentage times the true value of the population mean will fall outside this interval. Please note that the original population need not be normal. When the sample size is sufficiently large, the sampling distribution of the mean follows a normal distribution with

mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ (Central Limit Theorem). An example that follows will help you grasp the meaning of confidence interval.

Example 2 A machine produces components, which have a standard deviation of 1.6 cm in length. A random sample of 64 parts is selected from the output and this sample has a mean length of 90 cm. The customer will reject the part if it is either less than 88cm or more than 92cm. Does the 95% confidence interval for the true mean length of all the components produced ensure acceptance by the customer?

Solution To answer the question of acceptance by the customer, you should first work out the 95% confidence interval for the population mean μ (Here μ is the mean length of the components in the population). The formula for the confidence interval is

$\bar{X} - Z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z \frac{\sigma}{\sqrt{n}}$. When you want 95% confidence level, the Z value is 1.96 from the standard normal distribution. See diagram below:

Standard Normal Distribution $Z = \left(\frac{X - \mu}{\frac{\sigma}{\sqrt{n}}} \right)$

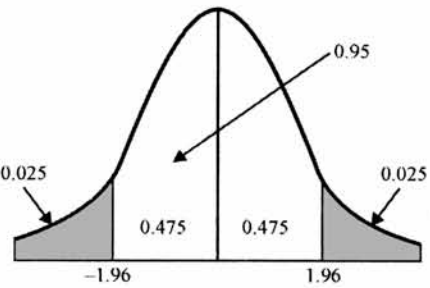


Figure 7.5

Hence 95% confidence interval for the population mean is given by:

$$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$

Now $\bar{X} = 90$, $\frac{\sigma}{\sqrt{n}} = \frac{1.6}{\sqrt{64}} = 0.2$. Substituting these values in the interval

$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$, we have $90 - 1.96(0.2) \leq \mu \leq 90 + 1.96(0.2) = 89.61 \leq \mu \leq 90.39$. This implies that the probability that the true value of the population mean length of the components will fall in this interval of $89.61 \leq \mu \leq 90.39$ is 95%. Hence, we infer that the 95% confidence interval ensures acceptance by the customer. (Please see picture below).

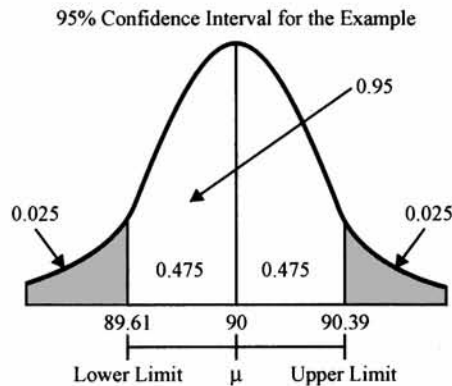


Figure 7.6

Progressive Test Question For the same example discussed, if you want to establish the 99% confidence interval for the population mean μ ,

1. The Z value to be used is _____.
2. The 99% confidence interval for μ is_____.

Solution

1. Using the inverse probability facility of the standard normal distribution from Microsoft Excel, we have $Z = 2.58$ for 99% confidence interval for μ . The answer is 2.58. See diagram below:

Standard Normal Distribution $Z = \left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right)$

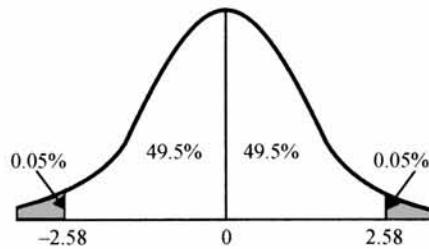


Figure 7.7

2. The 99% confidence interval for μ is given by the interval $\bar{X} - 2.58 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.58 \frac{\sigma}{\sqrt{n}}$

Now $\bar{X} = 90$. $\frac{\sigma}{\sqrt{n}} = \frac{1.6}{\sqrt{64}} = 0.2$. Substituting these values in the interval

$\bar{X} - 2.58 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.58 \frac{\sigma}{\sqrt{n}}$, we have $90 - 2.58(0.2) \leq \mu \leq 90 + 2.58(0.2) = 89.48 \leq \mu \leq 90.52$. This means that the probability that the true value of the population mean length will fall in this interval of $89.48 \leq \mu \leq 90.52$ is 99%.

Confidence interval for the population proportion-large sample Please go through the discussion in this chapter on the concept of proportion before knowing how to set up the confidence interval for the population proportion. Proportion is frequently used in survey research particularly in marketing research. The $1-\alpha$ confidence interval for the population proportion is given below:

$$p - Z \sqrt{\frac{P(1-P)}{n}} \leq P \leq p + Z \sqrt{\frac{P(1-P)}{n}}$$

Where P is the population proportion,

p is the sample proportion,

Z is the value corresponding to the area of $\left(\frac{1-\alpha}{2}\right)$ from the mean of the standard normal distribution,

α is the proportion in the tails of the standard normal distribution that is outside the range of the confidence interval.

Unfortunately, this interval contains the term P , which is the population proportion that we are trying to estimate. It is a practice among the statisticians to use the sample proportion p in the place of P . Please note that p is an unbiased estimator of P . Hence, the $(1-\alpha)$ confidence interval for the population proportion P becomes:

$p - Z\sqrt{\frac{p(1-p)}{n}} \leq P \leq p + Z\sqrt{\frac{p(1-p)}{n}}$. This confidence interval in particular for 95% and 99%

confidence levels are given below:

Confidence Level	Confidence Interval for the Population Proportion P
95%	$p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq P \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}$
99%	$p - 2.58\sqrt{\frac{p(1-p)}{n}} \leq P \leq p + 2.58\sqrt{\frac{p(1-p)}{n}}$

Example 3 In a health survey involving a random sample of 75 patients who developed a particular illness, 70% of them are cured of this illness by a new drug. Establish the 95% confidence interval for the population proportion of all the patients who will be cured by the new drug. This would help assess the market potential for this new drug by a pharmaceutical company.

Solution The 95% confidence Interval for the population proportion is given by:

$$p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq P \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}$$

$$0.70 - 1.96\sqrt{\frac{0.70(1-0.70)}{75}} \leq P \leq 0.70 + 1.96\sqrt{\frac{0.70(1-.70)}{75}}. \text{ Upon simplification, this interval}$$

becomes $0.5963 \leq P \leq 0.8037$. That is, the probability that the population proportion will fall in this interval of $0.5963 \leq P \leq 0.8037$ is 95%.

Progressive Test Question For the example above, the 99% confidence interval for the population proportion is:

(a) $0.5365 \leq P \leq 0.8635$

(b) $0.5635 \leq P \leq 0.8365$

(c) $0.5563 \leq P \leq 0.8635$

Answer The correct choice is (b). Applying the formula for the 99% confidence interval for the population proportion, we have the interval:

$$p - 2.58\sqrt{\frac{p(1-p)}{n}} \leq P \leq p + 2.58\sqrt{\frac{p(1-p)}{n}}$$

$$0.70 - 2.58\sqrt{\frac{0.70(1-0.70)}{75}} \leq P \leq 0.70 + 2.58\sqrt{\frac{0.70(1-0.70)}{75}}. \text{ Simplifying, we get } 0.5635 \leq P \leq 0.8365.$$

7.4 CONFIDENCE INTERVAL FOR POPULATION MEAN - SMALL SAMPLE (t-DISTRIBUTION)

How will you establish a 95% confidence interval for the population mean when the standard deviation is not known? How will you establish a 95% confidence interval for the population mean when the sample size is small? Both these questions can be answered by the t-distribution. William Gosset, under the nickname student, discovered the t-distribution. Hence, it is called "Student's t distribution". For better clarity, see the diagram below:

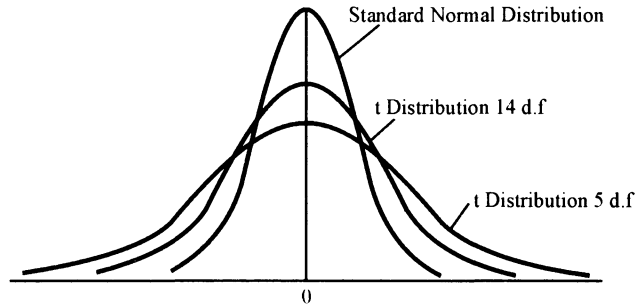


Figure 7.8

Characteristics of the t Distribution

As you can see from the diagram, t distribution is also symmetrical like the normal distribution. However, the t distribution is flatter than the normal distribution. The t distribution depends upon one more factor called the degrees of freedom. You can define the degrees of freedom as the number of unrestricted (independent) moments you can have out of the sample size n . In other words how many of the sample values are free to vary? In the case of t distribution, there are $n-1$ **degrees of freedom**. Intuitively, the degrees of freedom suggest that if we know the values for $n-1$ terms, the n th term can be calculated. According to Ronald Fisher, the father of Statistics, degrees of freedom is a term borrowed from mathematics. In fact he explains the concept of degrees of freedom beautifully while deriving the Chi-Square distribution, which we will cover in a later chapter. When n is large (greater than or equal to 30), the t distribution loses its flatness and becomes a normal distribution. Hence, we can use the normal approximation to t when n is greater than or equal to 30. For estimating population mean involving *small sample*, t distribution is the best choice.

Confidence Interval for Mean using t Distribution

The $1 - \alpha$ confidence interval for the population mean is given by:

$$\bar{X} - t_{n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1} \frac{S}{\sqrt{n}}$$

Where \bar{X} is the sample mean based on independent random samples from a normal population and the sample size is small.

$$S \text{ is the sample standard deviation} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

t_{n-1} is the value of the t distribution with $n-1$ degrees of freedom for an area of $\alpha/2$ in both the tails of the distribution.

Please note that the sample statistic $t = \frac{(\bar{X} - \mu)}{\frac{S}{\sqrt{n}}}$ follows a t distribution with $n-1$ degrees of freedom (d.f.).

An example will help you explain the confidence interval using the t distribution.

Now, you will see this example with solution.

Example 4 The average travel time taken based on a random sample of 10 people working in a company to reach the office is 40 minutes with a standard deviation of 10 minutes.

Establish the 95% confidence interval for the mean travel time of everyone in the company? This will help the company redesign the working hours.

Solution The $1-\alpha$ confidence interval for the population mean is given by

$$\bar{X} - t_{n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1} \frac{S}{\sqrt{n}}. \quad \bar{X} = 40 \quad \frac{S}{\sqrt{n}} = \frac{10}{\sqrt{10}} = 3.16$$

For a 95% confidence level, $\alpha/2 = .05/2 = .025$. Each tail will have probability of .025. For what value of t is this value = .025? Of course, the degrees of freedom is = 9. Remember t distribution is symmetrical about its mean like the normal. It is enough if you look at the upper tail to answer the question. An even easier way out is to avoid looking at t -tables and get it from the Microsoft Excel. The steps are as follows:

Step 1 In the menu bar click the icon of paste function. You get:

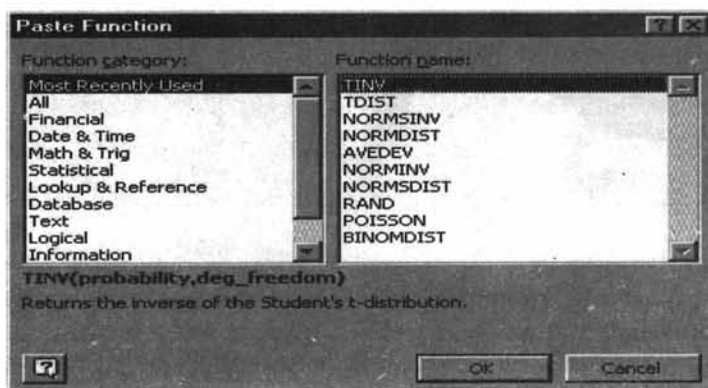


Figure 7.9

Step 2 Click Statistical on the left side and then click TINV on the right side. You get:

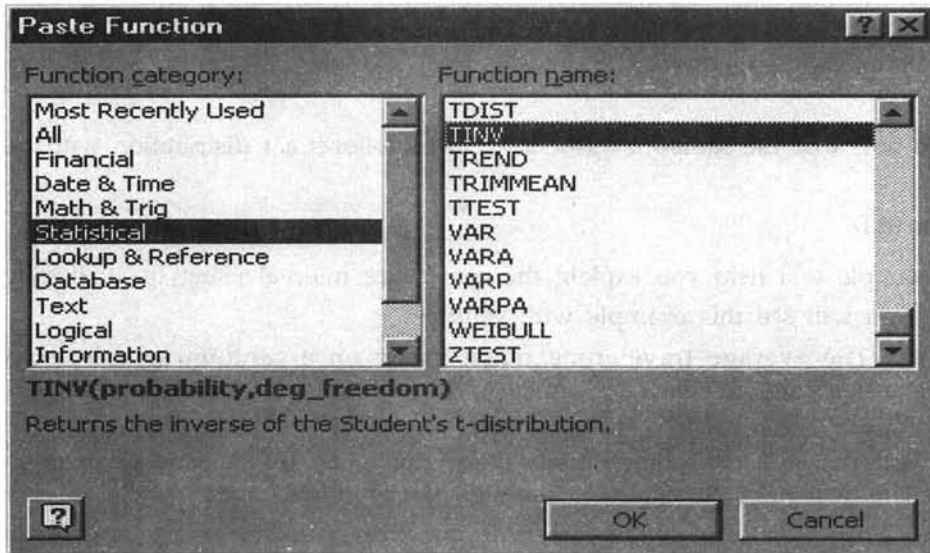


Figure 7.10

Step 3 Now click OK, you get:

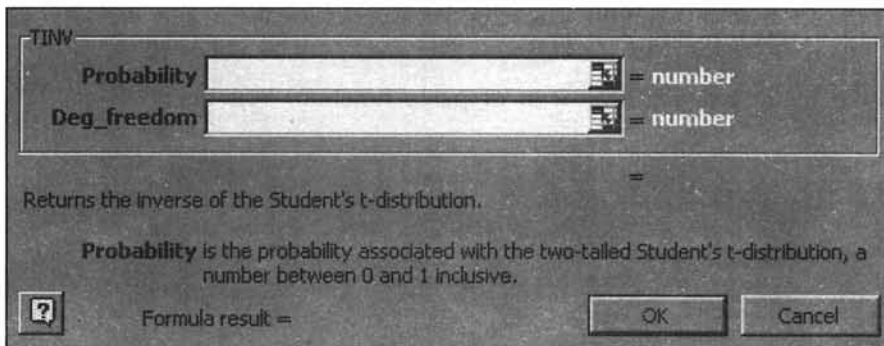


Figure 7.11

You enter the value of α straightaway because Excel returns the t value for a given α associated with the two-tail student's t . No need for giving $\alpha/2$. In our example, you enter in probability cell 0.05 because $\alpha = 0.05$ (you want 95% confidence level). In Deg _freedom cell, enter 9 as there are 10 sample values. Click OK. You get:

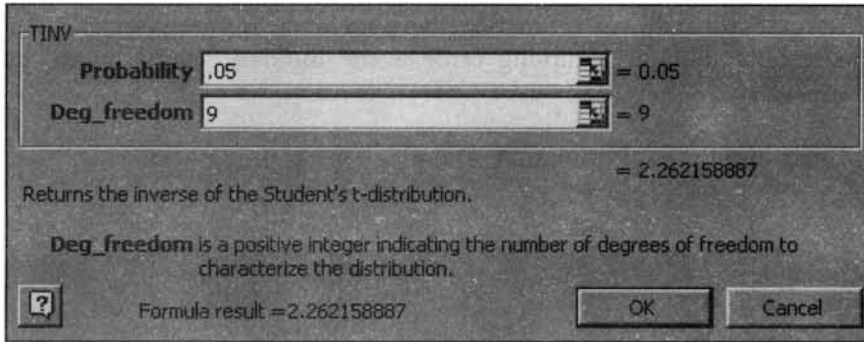


Figure 7.12

As you can see, the t value for 95% confidence level is 2.26 (rounded to two decimal places) appearing under "Formula result =". This is the t value for 9 degrees of freedom (t_{n-1}). Substituting in the expression for the confidence interval $\bar{X} - t_{n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1} \frac{S}{\sqrt{n}}$ all the relevant values, you get $40 - 2.26(3.16) \leq \mu \leq 40 + 2.26(3.16)$ or

$32.86 \leq \mu \leq 47.14$. The probability that the true mean travel time of everyone in the office will fall in this interval of $32.86 \leq \mu \leq 47.14$ is 95%.

If you want to use t -distribution table to answer this example, by all means do so. It is given in Appendix E. Please note that the value of t depends on the degrees of freedom apart from the confidence level.

Progressive Test Question For the previous example, the 90% confidence interval for the population mean is -----.

Solution You know the values of all the factors except the t value for $\alpha = 10\%$ (90% confidence) for 9 d.f. From Excel do exactly what was done for 95% confidence except enter 0.10 in the probability cell. You get the t value. It is = 1.83. As before substituting the values of all the unknown terms in the interval $\bar{X} - t_{n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1} \frac{S}{\sqrt{n}}$, you get $40 - 1.83(3.16) \leq \mu \leq 40 + 1.83(3.16)$ or $34.22 \leq \mu \leq 45.78$.

Discussion Analyze, criticize, and explain the following statement:
 "If a confidence interval is constructed by adding and subtracting the same quantity to the sample statistic, you are assuming that the distribution of the original population is symmetric".

7.5 HOW TO DETERMINE SAMPLE SIZE USING CONFIDENCE INTERVAL

If you specify the sampling error (precision), the confidence level desired and the standard deviation of the original population, you can compute the optimal sample size. You can

determine the sample size both for estimating the population mean as well as the population proportion. Please note that the sampling error is the difference between the estimate and the actual parameter.

Sample Size Determination - Population Mean

Please recall that
$$Z = \left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right)$$

Sampling error $E = \bar{X} - \mu$. Squaring both sides and simplifying we have $n = \frac{Z^2 \sigma^2}{E^2}$. Z is

the value corresponding to the area of $\left(\frac{1-\alpha}{2}\right)$ from the mean of the standard normal distribution. In particular, if you want 95% confidence level, $Z = 1.96$. If you want 99% confidence level $Z = 2.58$. You can get these values from Excel very easily.

Example 5 A marketing manager of a fast food restaurant in a city wishes to estimate the average yearly amount that families spend on fast food restaurants. He wants the estimate to be within \pm Rs 100 with a confidence level of 99%. It is known from an earlier pilot study that the standard deviation of the family expenditure on fast food restaurant is Rs 500. How many families must be chosen for this problem?

Applying the formula $n = \frac{Z^2 \sigma^2}{E^2}$, we have $n = \frac{2.58^2 (500^2)}{100^2} = 166.41 = 166$ rounded to the nearest integer.

Sample Size Determination - Population Proportion

Please recall
$$Z = \left(\frac{p - P}{\sqrt{\frac{p(1-p)}{n}}} \right)$$

Sampling error $E = (p - P)$ Squaring both sides and simplifying, we have,

$n = \frac{Z^2 p(1-p)}{E^2}$. Z is the value corresponding to the area of $\left(\frac{1-\alpha}{2}\right)$ from the mean of the standard normal distribution. In particular, if you want 95% confidence level, $Z = 1.96$.

If you want 99% confidence level $Z = 2.58$.

Example 6 A company manufacturing sports goods wants to estimate the proportion of cricket players among high school students in India. The company wants the estimate to be within ± 0.03 with a confidence level of 99%. A pilot study done earlier reveals that out of 80 high school students, 36 students play cricket. What should be the sample size for this study?

Solution $p = 36/80 = 0.45$

Applying the formula for calculating the sample size $n = \frac{Z^2 p(1-p)}{E^2}$, we have,

$$n = \frac{2.58^2(0.45)(1-0.45)}{.03^2}. \text{ Simplifying, you get } n = 1831.$$

In summary, you can determine the sample size both for estimating the population mean and the population proportion if you know the precision, the confidence level, and the standard deviation of the original population.

7.6 CHAPTER SUMMARY

In this chapter, you have been given a conceptual framework of statistical estimation, which is one of the three components of statistical inference. In particular, this chapter focused on the following:

- The definition and meaning of point estimation for the population mean and population proportion.
- The role of sample mean and sample proportion in estimating the population mean and population proportion with their property of unbiasedness.
- The conceptual framework of interval estimation with its key elements.
- The methodology for establishing the confidence interval for the population mean and the population proportion based on the sample mean and the sample proportion.
- Examples giving the 95% and 99% confidence interval for the population mean and the population proportion for large samples.
- Establishing confidence interval for small samples using the t distribution after explaining the role of degrees of freedom in computing the value of t .
- Determining the optimal sample size based on precision, confidence level, and a knowledge about the population standard deviation.
- As usual, this chapter has emphasized the role of Microsoft Excel in calculating the value of t based on the degrees of freedom by using the paste function.

GLOSSARY

Confidence Interval It is an interval in which the true value of the population parameter will fall with a certain probability.

Confidence Level This is the probability that is associated with an interval estimation of a population parameter. This indicates how confident are we, that the population parameter will fall in this interval.

Degrees of Freedom Degrees of freedom is the number of unrestricted (independent) moments one can have out of the given sample size. In other words, degrees of freedom indicate how many of the sample values are free to vary.

Estimate A particular value of an estimator.

Estimator A statistic that is used to estimate a population parameter.

Interval Estimation Interval Estimation establishes an interval consisting of a lower limit and an upper limit in which the true value of the population parameter is expected to fall.

Point Estimation It deals with the task of selecting a specific sample value as an estimate for a population parameter.

Point Estimate A specific value of a sample statistic that is used to estimate a population parameter.

Student's t Distribution A probability distribution discovered by William Gosset under the nickname, Student. It is similar in structure to the normal distribution except that it depends on degrees of freedom. It is used when the sample size is small (<30) and the population standard deviation is unknown.

Unbiased Estimator An unbiased estimator is one whose expected value is equal to the population parameter.

REVIEW QUESTIONS

- A company sells components in length (cm) with a standard deviation of a 3.5 cm. A random sample of 64 components sold reveals that the mean length of the components in the sample is 105 cm. The 95 per cent confidence interval for the population mean length sold of all the parts is:
 - $103 \leq \mu \leq \bar{X} + 105$
 - $103.5 \leq \mu \leq 105.7$
 - $104.14 \leq \mu \leq 105.86$
 - None of the above.
- It is a general practice to use the normal distribution to establish the confidence interval for the population proportion if the sample size is large. True or False.
- The area of the tails of the t distribution is less than the area of the tails of the normal distribution. True or False.
- The Z-value that is used to establish a 95% confidence interval for the population mean is:

- (a) 1.28 (b) 1.65
(c) 1.96 (d) 2.58.
5. Which of the following is not true of the t distribution?
(a) As the sample size keeps on increasing, it approaches normality
(b) Its value depends on the degrees of freedom
(c) It is a symmetrical distribution
(d) All of the above are true for the t distribution
6. While determining the sample size required for estimating the population mean, given the level of confidence, and standard deviation, if the sampling error increases, then the sample size:
(a) Will remain the same (b) Cannot be estimated
(c) Will decrease (d) Will increase
7. The director of a market research agency wishes to study the reach of a particular advertising campaign. He is concerned with the percentage of the target market that has seen at least a portion of the campaign. The director does not think that the figure will exceed 25 %. What should be the sample size for this study if the director wishes the estimate to be within three percentage points of the true value and 95 per cent confidence level is specified?. The sample size needed for this study rounded to whole number is:
(a) 550 (b) 600 (c) 785 (d) 800
- Questions 8 to 10 refer to the following random sample of nine observations collected from a population that is normally distributed.
6, 5, 8, 3, 9, 5, 8, 2, 8
8. What is the standard deviation of the distribution of the sample means of size 9?
(a) 1 (b) 0.8165
(c) 2.4495 (d) 6
9. The unbiased point estimator of the population mean is:
(a) 6 (b) 3 (c) 2 (d) 4
10. The 95 per cent confidence limits for the population mean is:
(a) 4.11 to 7.89 (b) 3.6 to 7.6
(c) 7.9 to 10.9

ANSWERS TO REVIEW QUESTIONS

1. The correct choice is ((c). Applying the formula for the confidence interval for the population mean $\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$, we have $\bar{X} = 105$, $\frac{\sigma}{\sqrt{n}} = \frac{3.5}{\sqrt{64}} = 0.4375$.

Upon making the substitution in this interval, we have $104.14 \leq \mu \leq 105.86$. Therefore, the right answer is choice (c).

2. The statement is true because the confidence interval for the population proportion is based on the crucial assumption of the sample size being large. This enables us to apply the central limit theorem for the sample proportion (which is a binomial distribution to start with). Regardless of the shape of the original distribution, the sample statistic approaches normality when the sample size is large ($n \geq 30$).
3. The statement is false because t distribution is flatter than the normal distribution and therefore the area of the tails are more than the area of the tails of the normal distribution. Read the characteristics of the t distribution in this chapter along with the picture, in which, the t distribution is compared with the normal distribution.
4. The right choice is (c). Read the explanation and diagram showing the area of the standard normal distribution in this chapter. Therefore, choice(c) is the right answer.
5. The right choice is (d). You will be able to recall from this chapter that (a), (b), and (c) are the distinguishing features of the t distribution
6. (c) is the right answer. When the sampling error increases, naturally the sample size will decrease because you are willing to have more error. Remember that the sampling error $E = \bar{X} - \mu$. Read how to determine the sample size in this chapter. The term E^2 is in the denominator.
7. The right answer is (d). Applying the formula for calculating the sample size in the case of proportion, we have $n = \frac{Z^2 p(1-p)}{E^2}$. Here $p = 0.25$, $E = 0.03$, $Z = 1.96$. Upon substitution in the formula and simplification, the answer is $n = \frac{1.96^2(0.25)(1-.25)}{0.03^2} = 800$ (rounded).
8. The right choice is (b). The standard deviation of the distribution of the sample means is given by $\frac{S}{\sqrt{n}}$, where $S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$. Using Excel we can calculate $\bar{X} = 6$, and $S = 2.4495$. So $\frac{S}{\sqrt{n}} = \frac{2.4495}{\sqrt{9}} = 0.8165$.
9. The correct choice is (a). The sample mean is an unbiased estimator of the population mean. Here the sample mean $\bar{X} = 6$.
10. (a) is the right choice. This problem needs the help of t distribution. The formula for the confidence interval for the population mean based on t distribution is: $\bar{X} - t_{n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1} \frac{S}{\sqrt{n}}$. All the factors are known except the t value with 8

degrees of freedom ($n - 1$). From Excel using the = TINV, we have the t value for 8 degrees of freedom = 2.31 (two places of decimal). Substituting, we have $6 - 2.31(0.8165) \leq \mu \leq 6 + 2.31(0.8165)$. On simplification, this becomes $4.11 \leq \mu \leq 7.89$.

PRACTICE PROBLEMS

- In a corporate headquarters of a company, the accounting division receives documents from the branches. Junior assistants in the head office are assigned the task of collecting, entering the information into computer and file the documents. A pilot study done earlier reveals that the number of documents cleared per junior assistant per day follows a normal distribution with mean 380 and standard deviation 40. There are 8 junior assistants working in the head office.
 - What is the point estimate of the total number of documents cleared by all junior assistants every day?
 - What is the probability that the number of documents processed per assistant per day will exceed 470?
 - Establish a 95% confidence interval for the population mean number of documents that will be cleared by any assistant on a day.
- A random sample of 20 children aged 8 years is taken and their height measured. This sample data set is used to estimate the average height of these children in the population with an accuracy of ± 6 cm at 95% confidence level. What will be the accuracy if the sample size chosen is 500 instead of 20?
- A company sells a particular component to the customers in length (cm). It has a standard deviation of 4 cm. A random sample of 81 components sold reveals that the mean length of the components in the sample is 110 cm. Establish a 99% confidence interval for the population mean length sold of all the parts.
- The average travel time taken based on a random sample of 15 people working in a company to reach the office is 45 minutes with a standard deviation of 9 minutes. Establish the 95% confidence interval for the mean travel time of everyone in the company?
- A private bank offering Internet Banking Services wants to estimate the proportion of customers who are satisfied with its service quality. The bank wants this estimate to be within 0.04 with a confidence level of 95%. A pilot study done earlier reveals that out of 120 customers, 90 are satisfied customers. What should be the sample size for a new comprehensive survey to ascertain satisfaction level?

Hypothesis Testing

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Define and Explain Hypothesis
- Formulate Hypotheses Appropriately
- Test Hypotheses for Large Sample and Small Sample

CHAPTER OUTLINE

- 8.1 Statistical Hypothesis-A Conceptual Framework
 - 8.2 Hypothesis Testing -Univariate Case (One Sample)
 - 8.3 Hypothesis Testing -Bivariate Case (Two Sample)
 - 8.4 Chapter Summary
- Glossary
Review Questions
Answers to Review Questions
Practice Problems

INTRODUCTION

Managers have to make decisions with minimum risk in an environment characterized by uncertainty. Acceptance or rejection of a decision depends on acceptance or rejection of a hypothesis. For example, a marketing manager is facing a decision whether to introduce a new product in the market or not. If his company could get a market share of 15 percent or more, then the new product would be introduced in the market. A suitable hypothesis formulation and testing it would help the manager take the right decision. This chapter covers the various tests of hypothesis that are useful in making sound decisions.



Figure 8.1

8.1 STATISTICAL HYPOTHESIS-A CONCEPTUAL FRAMEWORK

What is a Statistical Hypothesis?

A statistical hypothesis is a statement about a population parameter. It may or may not be true. The manager has to ascertain the truth of the hypothesis. Consider the following example.

Statement 1 Not more than 20% of the adults watch children's program in the television.

Statement 2 More than 20% of the adults watch children's program in the television.

First, it should be noted that these two hypotheses cannot be simultaneously true. Only one of them will be true. Likewise, these two hypotheses cannot be simultaneously false. Only one of them will be false. The acceptance or rejection of a particular hypothesis leads to the acceptance or rejection of a particular decision. Statement 1 given above is called the **Null Hypothesis** (H_0). Statement 2 is called the **Alternative Hypothesis** (H_1).

The decision maker normally formulates the **null hypothesis** with a view to getting it rejected. The null hypothesis is in a way, the negation of the truth. The decision maker in his mind wants the **alternative hypothesis** to happen.

Supposing the two hypotheses stated above are in the context of an advertising manager who has to decide whether to invest in a new advertising campaign, or not. His decision depends upon the outcome of the two hypotheses. Let us say that if not more than 20% of the adults watch children's program in the TV, then it is not viable for the company to make investment in the campaign. If statement 1 is correct, the advertising manager will not invest in the new advertising campaign. If statement 2 is correct, he will invest in the new campaign. Therefore, the acceptance or rejection of either of the hypotheses enables the advertising manager to take the right decision. It is pertinent to point out here that formulating and doing exercises on hypothesis are not meant to delight the statisticians but to really help managers in organizations take sound decisions involving minimum risk.

The example that is used here highlights the role of statistical hypothesis in the context of decision-making. To put it succinctly, statistical hypothesis is a statement about the population parameter and in this example, the advertising manager makes a statement about the population proportion of adults watching children's program in the television. He is not simply making a statement and stopping there; but formulates two hypotheses and then basing his final decision on the outcome of these hypotheses. The moral of the story is "**statistical hypothesis is decision oriented**".

The Type I and Type II Errors

If you go through the example discussed above, you will realize that the advertising manager is perhaps testing the hypotheses using sample survey data of the target audience collected from the adult population watching children's program in the television. So, the manager is taking some risk in his decision to invest in the campaign. As you know, the sample data always contain some element of uncertainty and therefore, it is not possible for the manager to be 100 percent certain that he has made the right decision. In a way, the manager makes two types of errors. See the diagram in the next page.

Hypothesis Testing Type I and Type II Errors

	Null Hypothesis	
	True	False
Reject	Type I Error (α)	No Error
Accept	No Error	Type II Error (β)

Figure 8.2

Type I error says that you reject the null hypothesis when it is true. **Type II error** says that you accept the null hypothesis when it is false. Strictly speaking you don't accept the null hypothesis. When you cannot reject the null hypothesis, it only means that the sample evidence does not warrant rejection of the null hypothesis. At any rate in practice, using the terminology "accept the null hypothesis" helps you take the right decision. This little bit of dilution in terminology is permitted in the context of managerial decision-making.

Some further insights into Type I and Type II Errors

Hypothesis Testing Type I and Type II Errors

	Null Hypothesis	
	True	False
Reject	Type I Error (α)	No Error
Accept	No Error	Type II Error (β)

Figure 8.3

You have seen α somewhere isn't? Of course, you have seen it in the previous chapter on estimation in the context of setting up the confidence interval.

The probability of making a Type I error is called the level of significance of the test. It is designated by the Greek letter *alpha* (α). $(1-\alpha)$ is the confidence level that says that you are right in your assessment $(1-\alpha)\%$ of the times. If you set $\alpha = 0.05$ and happen to reject the null hypothesis at this level, there is a 5% probability that you have rejected the null hypothesis when in fact it is true. This also means that you are 95% confident that you have accepted the null hypothesis when it is true. Significance level desired will depend on how much risk you want to take in rejecting the null hypothesis when it is true. Are you now seeing the connection between the confidence interval and level of significance? Good.

The probability of making a Type II error is symbolized by the Greek letter β . $1-\beta$ is called the **power of the test**. The power of the test is the probability of rejecting the null hypothesis when in fact it is false. Suppose you keep $\beta = 10\%$, it means that the power of the test is 90%. That is, the probability of rejecting the null hypothesis when it is false is 90% and only 10% of the time you commit the error of accepting the null hypothesis when it is false. It is desirable to keep both α and β at minimum level. However, a decrease in α will lead to an increase in β , and an increase in α will lead to a decrease in β . It is a general practice to fix α , and let β vary. It is convention to set $\alpha = 0.05$ that corresponds to the confidence level of 95%. Some practitioners at times keep $\alpha = 0.01$

Hypothesis testing - procedure Look at the diagram below giving the steps in Hypothesis Testing.

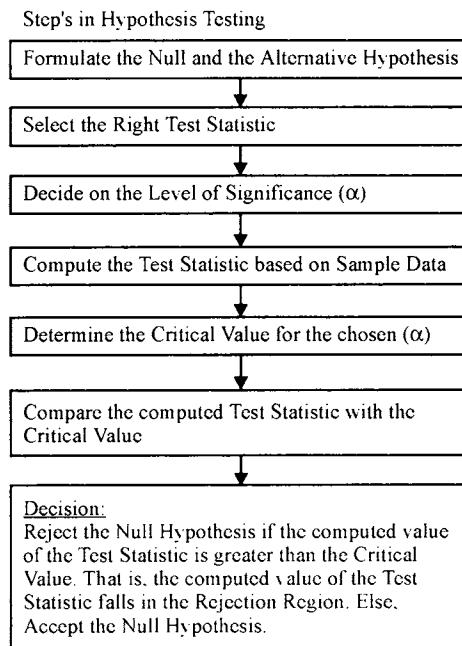


Figure 8.4

The discussion that follows will focus on formulating and testing hypotheses. First, we will focus on the formulation of the null and the alternative hypothesis that are very crucial for decision making. This is discussed next.

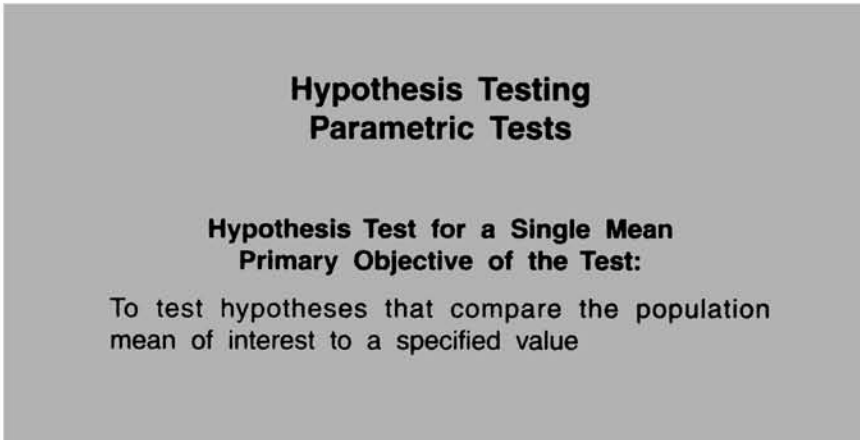
Glimpses into formulating the null and the alternative hypothesis The trickiest part in the entire exercise of hypothesis testing is the formulation of the null and the alternative hypothesis. You must spend a great deal of time in mastering the formulation part.

A hypothesis could be directional or non-directional. A **directional hypothesis** is one in which the population parameter is structured to be greater than or equal to or less than or equal to a specified value. This is known as a **one-tailed test (one-sided test)** in the parlance of statistical hypothesis. A non-directional hypothesis is one in which the population parameter is structured to be equal to a specified value. This is known as a **two-tailed test(two-sided test)**.

To clarify these ideas, first, let us explain the formulation of the null and alternative hypothesis for the univariate case through some examples. Then, we will move on to the procedures for conducting hypothesis testing. This will involve the entire gamut of structuring the hypotheses (null and alternative), step-by-step procedure for accepting or rejecting the hypotheses. As you know, acceptance or rejection of a hypothesis leads to acceptance or rejection of a particular decision.

8.2 HYPOTHESIS TESTING -UNIVARIATE CASE (ONE SAMPLE)

Hypothesis Testing - Population Mean (Single Mean)



**Hypothesis Testing
Parametric Tests**

Hypothesis Test for a Single Mean
Primary Objective of the Test:

To test hypotheses that compare the population mean of interest to a specified value

Figure 8.5

Look at the illustration below. Can you say it is a one-tailed test or a two-tailed test? It is a one-tailed test because the population waiting time of customers at the checkouts of Smart Supermarket is specified to be greater than 15 minutes. Can you formulate the null and the

alternative hypothesis for this illustration? First, try on your own. Of course, it takes practice. The formulation is given below:

Illustration for Test for a Single Mean

Is the average waiting time for the customers of Smart Supermarket at the checkouts greater than 15 minutes?

Figure 8.6

Structure of the null and alternative hypothesis for the illustration

Statement of the Null and Alternative Hypothesis

- $H_0 : \mu \leq 15$
- $H_1 : \mu > 15$

μ represents the Population Mean waiting time of the customers at the checkouts

Figure 8.7

The null hypothesis H_0 is generally, the negation of the truth. The question that is raised about the population mean is, "is the waiting time greater than 15 minutes". So, it is a one sided test and the right structure is that the null hypothesis must specify the population waiting time at the checkouts to be less than or equal to 15 minutes. Consequently, the alternative hypothesis H_1 is that the mean waiting time is more than 15 minutes. This is precisely what is formulated in the diagram above. The formulation of the directional hypothesis (one-tailed) is a frequent problem among the students and practitioners. You will be correct in your formulation, if you **remember two things**. 1) The null hypothesis is formulated with a view to getting it rejected. 2) We believe that the alternative hypothesis will happen.

Investigators intuitively think that the alternative hypothesis is going to materialize. Take this example. It is felt that the waiting time in the checkouts is more than 15 minutes. That is you strongly feel that this is happening. So, this is the alternative hypothesis. In other words, you must formulate the null hypothesis in such a fashion that its rejection automatically leads to the preferred conclusion. Are you getting it right? Good. The other principle to be followed is that read the problem situation given meticulously. You will never go wrong in formulation. I am 99% confident that you can formulate the null and the alternative hypothesis correctly if you faithfully follow the tips given above.

Progressive Test Question In the supermarket example given above, if the question were rephrased as "is the average waiting time of the customers at the checkouts different from 15 minutes", how would you formulate the null and the alternative hypothesis?

Solution $H_0 : \mu = 15$ (the average waiting time of the customers is equal to 15)

$H_1 : \mu \neq 15$ (the average waiting time of the customers is not equal to 15)

This formulation is undoubtedly a **two-tailed test of hypothesis**. Can you say why? The population mean waiting time is specified to be equal to the value 15. Therefore, it's a clear case of two-tailed test. Are you getting more and more confident in formulating the null and the alternative hypothesis? Good.

With the experience you have gained with regard to the formulation of the null and the alternative hypothesis, let us now move on to formulating the null and the alternative hypothesis for problems relating to the population proportion.

Hypothesis testing-population proportion (One sample)

Hypothesis Testing Parametric Tests

Hypothesis Test for a Single Proportion

Primary Objective of the Test:

To test hypotheses that compare the population proportion of interest to a specified value.

Figure 8.8

Let us look at an example for formulation of the hypotheses in the case of the population proportion.

Population proportion-hypothesis formulation an illustration

Illustration for Test for a Single Proportion

Is the proportion of households owning Color TVs in Chennai less than 0.4?

Figure 8.9

Can you say the illustration above is an example of what type of test? You must of course, say that it is a one-tailed test. First, try to formulate the null and the alternative hypothesis yourself. The correct structuring of the null and the alternative hypothesis is given below:

Statement of the Null and Alternative Hypothesis

$H_0 : P \geq 0.4$

$H_1 : P < 0.4$

P represents the Population proportion of households in Chennai owning color television

Figure 8.10**How does hypothesis testing work in practice?****1) Population Mean - (Univariate)**

The best way to understand and apply the concept of hypothesis testing for the population mean is through an example. First, we take the case of large sample.

The marketing manager of a large restaurant has been asked to conduct a survey of its customers belonging to a particular income class. The president of the restaurant is interested in the mean income of its customers. He is further interested in comparing this mean income with that of a recently concluded census study by the government. The government study shows a mean income of Rs. 300000 per year for this class of customers with a standard deviation of Rs. 30000. The president is desirous of finding out whether the population mean of its customers in this category is Rs. 300000 per year, or not. The marketing manager has picked up a random sample of 100 customers of this class from the

customer database. The sample data show a mean income of Rs. 293000 per year. Perform a comprehensive statistical hypothesis testing procedure and state your conclusions.

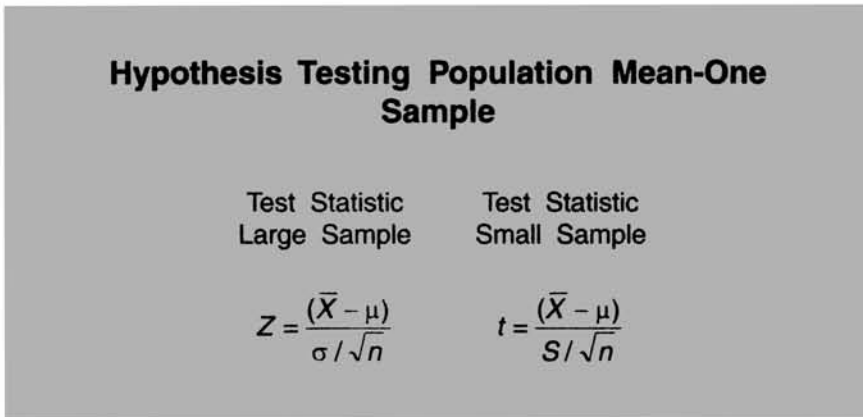


Figure 8.11

Solution By carefully reading the situation, you will be able to recognize this problem as a two-tailed test of hypothesis. Let us solve this problem step-by-step.

Step 1 Formulate the null and the alternative hypothesis.

$H_0 : \mu = 300000$ (the mean income of the population is equal to Rs. 300000)

$H_1 : \mu \neq 300000$ (the mean income of the population is not equal to Rs.300000)

Step 2 Select the right test statistic. The correct test statistic to be used here is the Z test. Why? Because, the sample size is large. The formula for the Z test is given below. Please note that Z follows a standard normal distribution with

mean 0 and standard deviation 1. $Z = \left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right)$

Step 3 Decide on the level of significance α . When the value of the level of significance is not specified in a problem, it is a convention to set the value equal to 0.05. What we're saying is that only 5% of the time we make the mistake of rejecting the null hypothesis when it is true.

Step 4 Compute the test statistic based on sample data. The formula to be used is

$Z = \left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right)$. Under the assumption of the null hypothesis being true, you

can substitute the value Rs. 300000 in the place of μ . $\bar{X} = 293000$. $n = 100$. $\sigma = 30000$.

Upon substitution of these values in the formula, we have,

$$Z = \left(\frac{293000 - 300000}{\frac{30000}{\sqrt{100}}} \right) = -2.33.$$

- Step 5** Determine the Critical Value for the chosen level of significance. Here, $\alpha = 0.05$. The critical value corresponding to the two-tailed test where each tail contains an area of $\alpha/2$ can be easily worked out by using Microsoft Excel. The methodology is already covered in the previous chapters. Here, $\alpha/2 = 0.025$. The critical value of $Z = 1.96$ for positive Z and -1.96 for negative results. Since the normal distribution is symmetrical, we can ignore the sign of Z and just take the positive value of Z . That is the critical value of Z is 1.96. Incidentally, if you choose $\alpha = 0.01$, then the critical value of $Z = 2.58$.
- Step 6** Compare the computed test statistic with the critical value. Here, computed $Z = -2.33$. Since the normal distribution is symmetrical, take the positive value of Z and compare it with the Critical $Z = 1.96$. If you take the negative computed Z , then compare it with -1.96 . Take the positive Z . Simple is best. Why bother?
- Step 7** Decision. If the computed Z is greater than the table Z , reject the null hypothesis H_0 and accept H_1 . Else, accept H_0 . This is same as finding out whether the computed Z falls in the acceptance region or the rejected region. In our case, computed value of Z (take just the positive value) 2.33 is greater than the critical value of $Z = 1.96$. Hence, it falls in the rejection region. Reject H_0 and accept H_1 . For better grasp, see the following diagram giving the acceptance region and rejection region for a level of significance of 0.05 for a two-tailed test.

Identifying the Acceptance and Rejection Region for a level of significance = 0.05

$$\text{Standard Normal Distribution } Z = \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right).$$

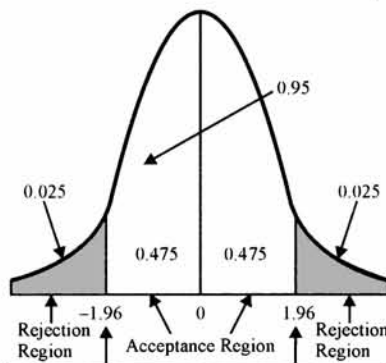


Figure 8.12

Are you able to see now that the computed Z for our example problem = 2.33 falls in the rejection region in the diagram? You must say yes with 95% confidence level if not 99%. The simple rule is that if the computed Z is greater than the critical Z , reject the null hypothesis.

Interpretation of the Results for our example We have rejected H_0 and accepted H_1 . What does this mean? This means that the population mean income of the category of interest to the president of the restaurant is not equal to Rs. 300000 per year at 5% level of significance. Is it more than or less than Rs. 300000? The sample mean suggests that it may be less than Rs. 300000 per year. Can you do this exercise now as a one-tailed hypothesis test? This is a **Progressive Test Question** for you.

P-Value

Some software packages like SPSS and Excel talk about a **P-Value**. This says that if the level of significance is greater than the P-value, reject the null hypothesis. What does it mean? You see, you have chosen $\alpha = 0.05$. This means that you are rejecting the null hypothesis when it is true only 5% of the time. When the P-value is smaller than 5%, is it not obvious that you have to reject H_0 ? Yes. Why? Because, it implies that the probability of rejecting the null hypothesis when it is true, is smaller than 0.05. You are willing to take a risk of 5%. Now, the risk is even smaller. Naturally, you will reject H_0 . Both the methods of deciding whether to accept or reject H_0 are the same. It is a matter of your own convenience.

Progressive Test Question What is the P-value corresponding to the computed $Z = -2.33$ in our example?

Solution Using the Microsoft Excel you just compute the cumulative probability using $Z = -2.33$ in the paste function. As you can see from the output below, the P-Value is = 0.0099. This is smaller than the level of significance = 0.05. Hence, Reject H_0 .

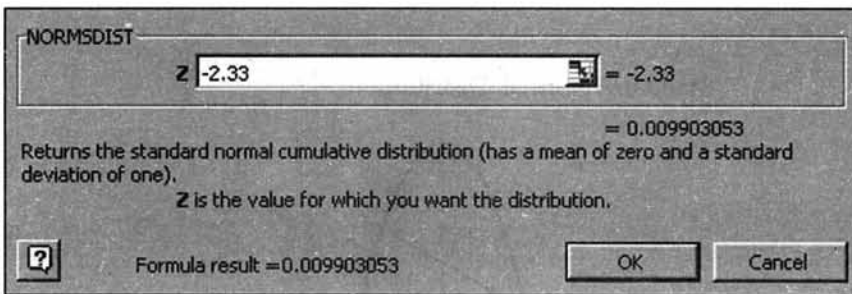


Figure 8.13

Alternatively if you use the normal probability distribution table given in Appendix D, apply the property of symmetry. Take the z value = 2.33 and look at the area between 0 to 2.33 in the table. The value obtained is 0.4901. Since you want probability of z to exceed 2.33, subtract 0.4901 from 0.50. That is $0.5 - 0.4901 = 0.0099$. This is same as what you have got from Microsoft Excel displayed above.

Progressive Test Question: For the same illustration discussed, if the president of the restaurant wants to find out whether the mean income of the customers is less than 300000 per year, perform the hypothesis testing using the same data.

Solution This is clearly a case of one-tailed test. The test statistic to be used is Z statistic. Up to computing Z, every thing remains same. While comparing the computed Z with the critical Z, you need to get the critical Z for the one-tailed test.

The hypothesis formulation is as follows:

$$H_0 : \mu \geq 300000$$

$$H_1 : \mu < 300000$$

$$\text{As before } Z = \left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right) = Z = \left(\frac{293000 - 300000}{\frac{30000}{\sqrt{100}}} \right) = -2.33$$

If we take the one-tailed test, the critical Z from Microsoft Excel corresponding to $\alpha = 0.05 = 1.65$. The computed value of Z taking only the positive value = 2.33 is greater than the critical Z = 1.65. Reject H_0 and accept H_1 . This means that the population mean income of the customers is less than Rs. 300000 per year. See the following diagram depicting the acceptance and rejection region for the one- tailed test.

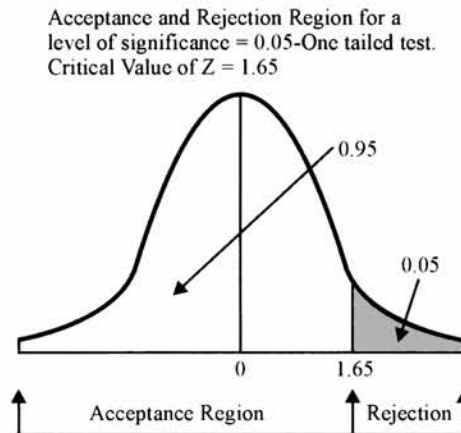


Figure 8.14

2. Population Proportion-Large Sample (Univariate)

A marketing manager of an enterprise is facing a decision whether to introduce a new product into the market or not. Consumer acceptance measured in a blind comparison test is agreed upon as an appropriate basis for evaluation. Marketing of the new product will be pursued only if the acceptance rate exceeds 30%. Otherwise, the new product will not be introduced in

the market. A random sample of 200 consumers reveals that the acceptance rate is 32%. Using a level of significance of 0.05, perform the hypothesis testing and recommend your action. Solution is given below.

Hypothesis Testing Population Proportion

Test Statistic:

$$Z = \frac{(p - P)}{\sqrt{(P(1 - P)/n)}}$$

Figure 8.15

Solution This is a classical hypothesis-testing problem of the population proportion. This is also a one-tailed test. This involves a large sample in which $n = 200$. The Z test for the proportion is the appropriate test.

$H_0 : P \leq 0.30$ (The population proportion of consumer acceptance is less than or equal to 0.30)

$H_1 : P > 0.30$ (The population proportion of consumer acceptance is greater than 0.30)

The Z test for proportion is to be used here. It is given by $Z = \frac{p - P}{\sqrt{\frac{P(1 - P)}{n}}}$. Please note

that Z follows a standard normal with mean 0 and standard deviation 1. Under the null

hypothesis being true, $P = 0.30$. $p = 0.32$. Substituting, we have $Z = \frac{0.32 - 0.30}{\sqrt{\frac{0.30(1 - 0.30)}{200}}} = 0.62$.

The critical value of Z for a one-tailed test is 1.65. Please see the diagram giving acceptance and rejection region we have drawn for testing the population mean. The same diagram holds true in the case of proportion as well. Since, the computed Z is less than critical Z, accept H_0 . What do you conclude?

We have no evidence to reject the null hypothesis based on the sample data at 5% level of significance. This implies that you accept H_0 , and conclude that the population proportion of consumer acceptance is less than or equal to 30%. Hence, the manager should not introduce the new product in the market. You may wonder how come when the sample

proportion is 32%, you say that you should not introduce the new product? Is not 32% better than the 30% stipulated? Yes, but you see statistically speaking, 32% sample proportion has arisen due to chance and not a real one. This is why you say, statistically not significant. As long as statistical significance does not take place, you cannot reject the null hypothesis. This is a real beauty of testing of hypothesis. Unless the manager wants to gamble, he should not venture to introduce the new product.

3. Population Mean-Small Sample (Univariate)

As you may recall from the previous chapter, the role of t distribution in setting up the confidence interval for the population mean, when the sample size is small and the standard deviation is unknown. The same t distribution with $n-1$ degrees of freedom will apply in hypothesis testing as well. Please also see the test statistic for t given in the picture under small sample category.

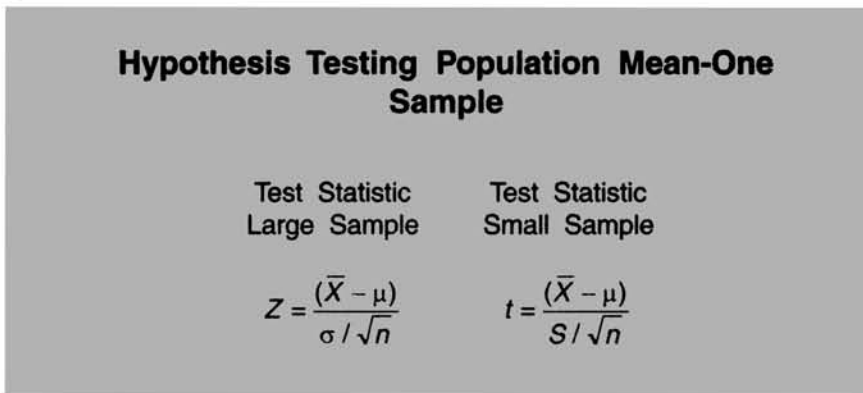


Figure 8.16

The procedure for conducting the hypothesis testing will be exactly same as the one we followed for the large sample case except the test statistic. Let us look at an example.

Example An investigator took a random sample of eight pieces of aluminum die-castings and observed the sample mean strength to be 31.5. Before taking the measurement, the investigator knew that the population mean strength for an older type of aluminum die-casting was 33. The standard deviation of the sample measurements was 1.3. The investigator would like to know whether the population mean strength of the aluminum die-casting is 33. Setup the null and the alternative hypothesis, perform the test, and comment on the results.

Solution This is a small sample case with unknown population standard deviation. The appropriate test is the t test. Please note also from the wording of the problem, you need to perform a two-tailed test. Just like the normal distribution, t is also symmetrical and it is enough if you compare the positive value of the computed t with the critical t for $n-1$ d.f. at 5% level of significance.

$$H_0 : \mu = 33$$

$$H_1 : \mu \neq 33$$

$$t = \frac{(\bar{X} - \mu)}{S/\sqrt{n}} = \frac{(31.5 - 33)}{\frac{1.3}{\sqrt{8}}} = -3.26. \text{ The positive value of the computed } t = 3.26$$

The critical t value for 7d.f ($n - 1 = 8 - 1$) at 5% level of significance from Excel paste function is 2.36(2 places of decimal). Since the calculated t value is greater than the critical t , reject H_0 and accept H_1 . The conclusion is that the mean strength of aluminum die casting of the population is not 33 at 5% level of significance.

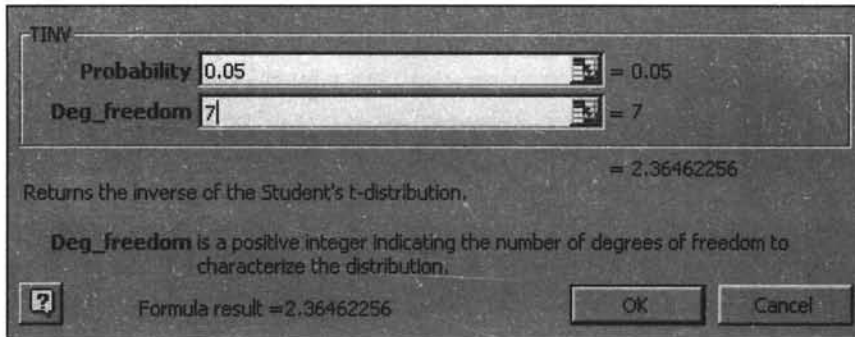


Figure 8.17

Progressive Test Question For the same illustration above if the investigator wants to find out whether the population mean strength of the die-casting is less than 33, how would you perform the t test?

Solution From the wording, it is clearly a one-tailed test. The hypotheses structuring should be as follows: The investigator suspects that the mean strength is less than 33. (This will be H_1)

$$H_0 : \mu \geq 33$$

$$H_1 : \mu < 33$$

$$\text{As before, } t = \frac{(\bar{X} - \mu)}{S/\sqrt{n}} = \frac{(31.5 - 33)}{\frac{1.3}{\sqrt{8}}} = -3.26. \text{ The positive value of the computed } t = 3.26$$

The critical value of t at 5% level of significance for a one tailed test with 7 d.f is to be found out from Excel. First, click paste function, then click Statistical, then click TINV, you get:

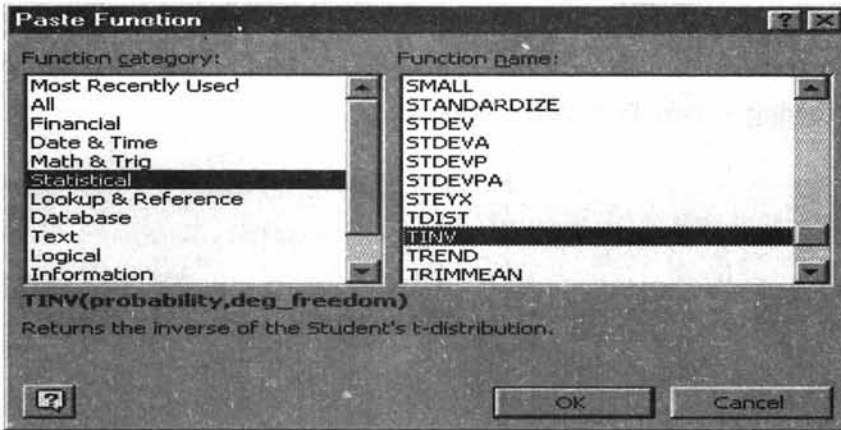


Figure 8.18

Click OK, and then enter the values in the asked cells. When you want the t value at 5% level for a one-tailed test, enter double the value of 5%, that is, 0.10 in the probability cell to get the t value for the one-tailed test. This is the trick you have to play. In other words, for a two-tailed test, enter the probability value as it is (= level of significance) and for the one-tailed test, just double it. You are able to do this because t is symmetrical. Since, the computed positive $t(3.26)$ is greater than the critical $t(1.89)$, reject H_0 and accept H_1 . That is, the population mean strength of die-cast appears to be less than 33 at 5% level of significance.

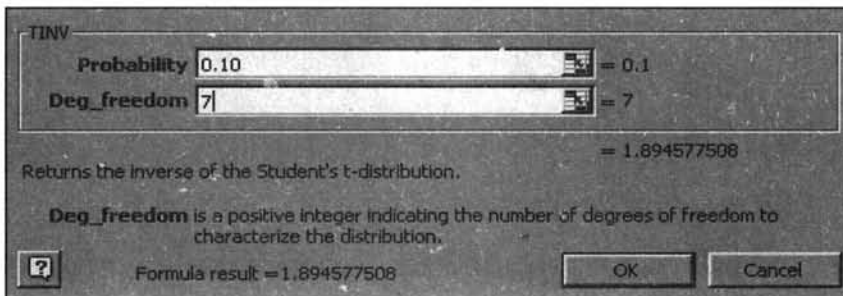


Figure 8.19

8.3 HYPOTHESIS TESTING -BIVARIATE CASE (TWO SAMPLE)

In many decision-making situations, we need to examine whether the means of two populations are same or different. The production manager in a company may wish to test whether the female workers give lower outputs than the male workers. A human resource director may be interested in finding out whether the hourly professional charges by two trainers are the same in two cities. A marketing manager may wish to estimate the difference between men and women with regard to the per capita consumption of a product. Therefore, you need hypothesis testing procedures for comparing two population parameters.

Extending the discussion to more than two parameters is possible and will be discussed in the next chapter.

Hypothesis Testing - Two Population Means

Hypothesis Testing Parametric Tests

Test of Two means

Primary Objective of the Test:

To test hypothesis that compare the Population Mean of Interest for two separate populations (Samples are independent)

Figure 8.20

Hypothesis Testing Difference between Two Population Means

Test Statistic Large Sample	Test Statistic Small Sample
$\frac{Z = (\bar{X}_1 - \bar{X}_2)}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}}$	$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{(S^2(1/n_1 + 1/n_2))}}$

Figure 8.21

A Brief Note on the Test Statistic(s)

- The sampling distribution of the difference between the means of the two populations is of interest to us.
- We take a random sample of n_1 observations from population 1, and n_2 observations from population 2 and compute the difference between the sample means $\bar{X}_1 - \bar{X}_2$. If we generate the distribution of all possible $\bar{X}_1 - \bar{X}_2$, $\bar{X}_1 - \bar{X}_2$ will follow a normal distribution with mean $\mu_1 - \mu_2$ and standard deviation

deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ when the sample size is large (both n_1 and n_2 are greater than or equal to 30). $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ is called the standard error of $\bar{X}_1 - \bar{X}_2$.

- If the two population standard deviations are not known, we can estimate the standard error of $\bar{X}_1 - \bar{X}_2$ by substituting with the sample standard deviations.

That is, $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ will be used in the place of $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

- The Test Statistic $Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$. Z follows normal with mean 0 and standard deviation 1.
- If the original populations are assumed to be normal with standard deviations are equal, the appropriate test statistic is t . This is also to be used when the sample size is small.

- The test statistic $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$. This follows a t distribution with

$n_1 + n_2 - 2$ degrees of freedom. S^2 is called the pooled sum of square variance given by, $S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$. S_1^2 and S_2^2 are the variances of the samples taken from the two populations.

Difference between two population means- large sample

Example A test in computer course was conducted for a group of students, consisting of 70 boys and 60 girls. The marks scored by the students are given below:

Boys	Girls
$n_1 = 70$	$n_2 = 60$
$\bar{X}_1 = 70$	$\bar{X}_2 = 65$
$\sum (X_1 - \bar{X}_1)^2 = 7,500$	$\sum (X_2 - \bar{X}_2)^2 = 7,800$

Is there a significant difference between the performance of the boys and the girls?

Solution From the wording of the question, it is clear that the problem could be structured as a two-tailed test. The sample sizes for both the populations are large, and hence the Z test is appropriate to use. Let us take a level of significance of 5%.

$H_0: \mu_1 = \mu_2$ (The population mean score of boys = The population mean score of girls)

$H_1: \mu_1 \neq \mu_2$ (The population mean score of boys \neq The population mean score of girls)

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The standard deviation of the two populations are not given. We can

use the sample standard deviations in the place of the population standard deviations. That

is, use $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ in the place of $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. In the data for the problem, we are given

$$\sum (X_1 - \bar{X}_1)^2 = 7500 \text{ and } \sum (X_2 - \bar{X}_2)^2 = 7800. \text{ By definition, } S_1^2 = \frac{\sum (X_1 - \bar{X}_1)^2}{n_1 - 1} \text{ and}$$

$$S_2^2 = \frac{\sum (X_2 - \bar{X}_2)^2}{n_2 - 1}. \text{ Substituting and simplifying, we have } S_1^2 = (7500/69) = 108.70$$

$$S_2^2 = (7800/59) = 132.20.$$

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{70 - 65}{\sqrt{\frac{108.70}{70} + \frac{132.20}{60}}} = 2.58.$$

What is your decision? The procedure is exactly the same as the single mean case. If the calculated value of Z is greater than the critical value of Z at 5% level, reject H_0 and accept H_1 . The critical value of Z at 5% level is 1.96 for a two-tailed test.

Identifying the Acceptance and Rejection Region for a level of significance=0.05 (two-tailed test) Standard Normal Distribution

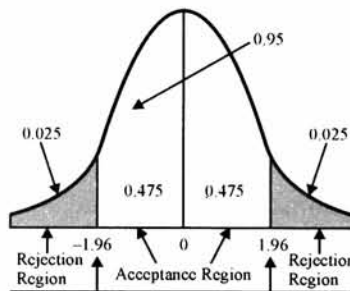


Figure 8.22

In our example, calculated Z is greater than the table Z . Reject H_0 and accept H_1 . That is, the performance of the boys and the girls are not identical. The null hypothesis of equally good performance is rejected. The difference in the mean scores between boys and girls is significant.

Progressive Test Question In the same example discussed, if the objective is to examine whether the boys are better than girls in performance, how will you modify the null and alternative hypothesis? What are your conclusions?

Answer The situation becomes a one-tailed test.

$H_0: \mu_1 \leq \mu_2$ (Mean score of boys is less than or equal to girls)

$H_1: \mu_1 > \mu_2$ (Mean score of boys is greater than girls)

Every thing remains same except the critical value of Z . For a one-tailed test the critical value of Z for a level of significance of 5% =1.65.

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{70 - 65}{\sqrt{\frac{108.70}{70} + \frac{132.20}{60}}} = 2.58$$

The calculated value of Z is greater than the critical value of Z . Reject H_0 at 5% level and accept H_1 . The conclusion is that boys perform better than girls.

Difference between two population means - small sample

Example An Aptitude test was conducted for two groups of executives-Group1 consists of engineers and Group2 consists of accountants. The scores obtained by the candidates are given below:

Engineers	125	115	119	85	97	107	125	125	118
Accountants	112	98	109	96	77	70	114	100	

Do you find any significant difference between the scores of these two groups?

Solution This is situation requires a two-tailed hypothesis testing. The sample sizes are small. $n_1 = 10$ and $n_2 = 9$.

$H_0: \mu_1 = \mu_2$ (The scores of both the groups are equal)

$H_1: \mu_1 \neq \mu_2$ (The scores of both the groups are not equal)

Let us take a level of significance of 0.05. This type of hypothesis problem can be easily solved using Microsoft Excel when the actual sample measurements are given. This is a real beauty.

Step 1 Enter the data in the spreadsheet. The screen will be:

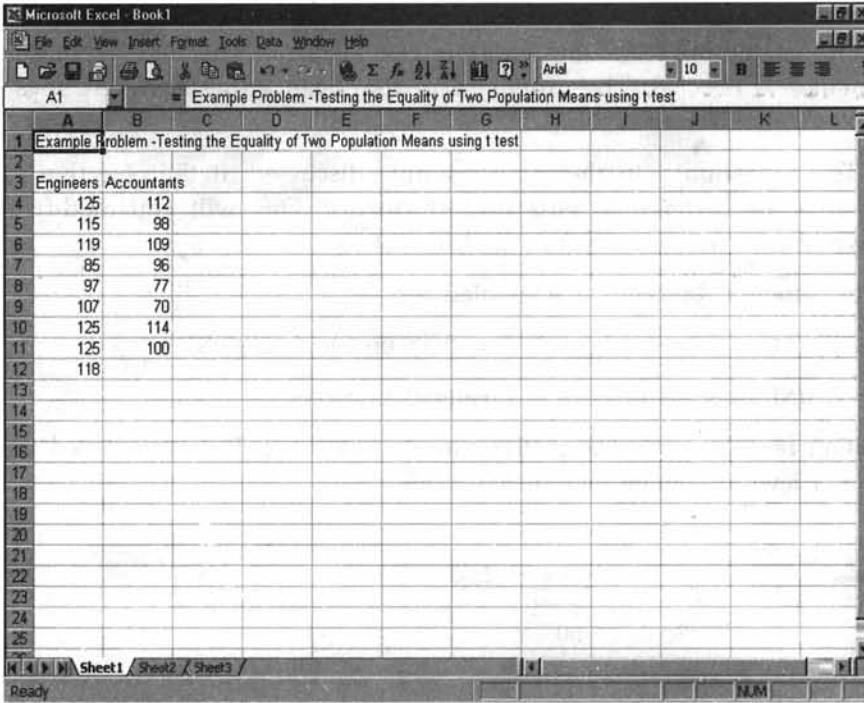


Figure 8.23

Step 2 Click Tools, click Data Analysis and then click t-test two samples assuming equal variances. You now get:

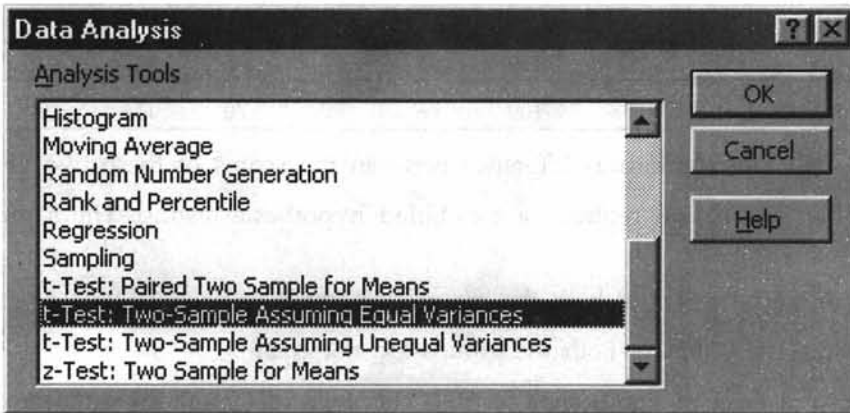


Figure 8.24

Step 3 Click OK. Excel will ask you to enter input range for variable 1 and variable 2. For illustration purpose, how to enter input for variable 1 is given.

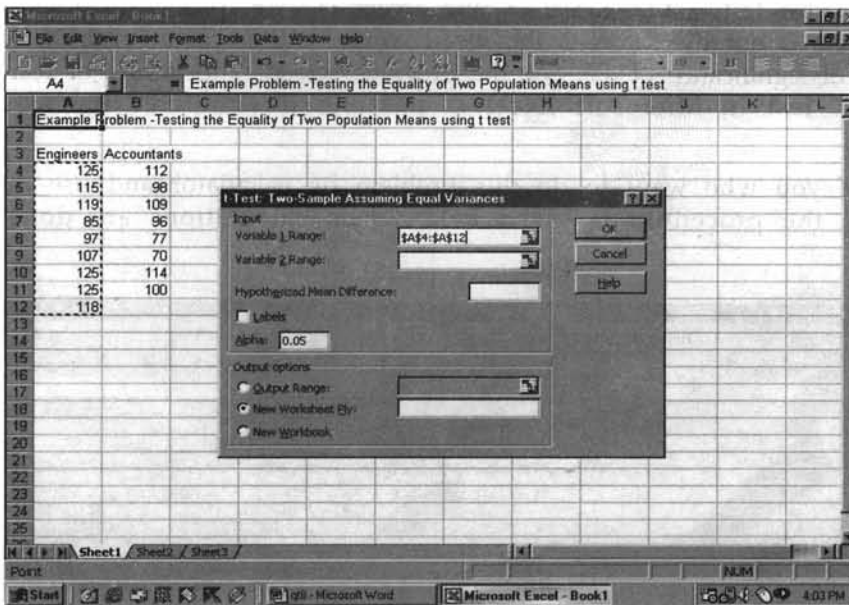


Figure 8.25

After entering the input range for variable 2, click OK. You now get the answer.

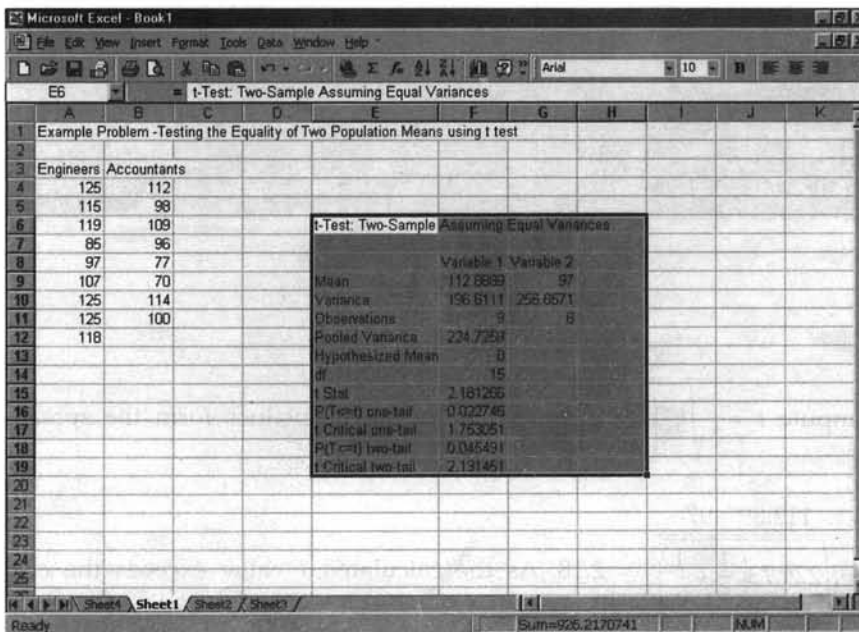


Figure 8.26

You name any thing. It is there. Hypothesized mean = 0 implies ($\mu_1 - \mu_2 = 0$ that is same as $\mu_1 = \mu_2$). As you can see that the computed t is 2.18. It is greater than the critical

t for the two-tailed test = 2.13. Reject H_0 and accept H_1 . We conclude that the mean scores of the engineers is not the same as the mean scores of the accountants. By default, Excel takes a level of significance of .05. See the previous screen. The facility of using the t test as well as Z test in Microsoft Excel is possible if you have the actual sample measurements.

Those of you who want to do this problem by calculator and use the t -table in Appendix E, the procedure is given below. The calculations are done using the spreadsheet.

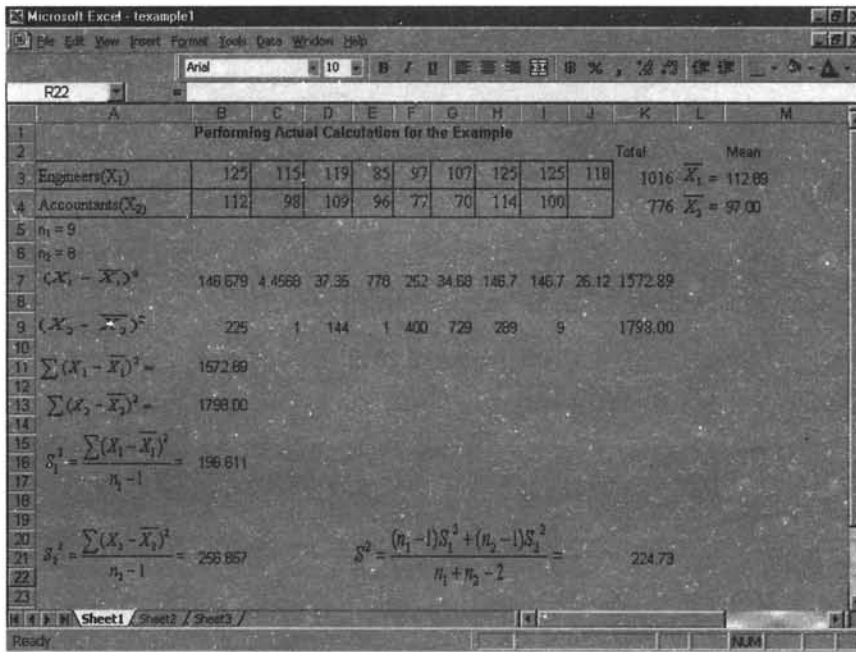


Figure 8.27

Now compute $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$. Substituting the values from the spreadsheet above,

we have $t = \frac{112.89 - 97}{\sqrt{224.73 \left(\frac{1}{9} + \frac{1}{8} \right)}} = 2.18$. As the calculated t value exceeds the critical t for 15

degrees of freedom at 5% level of significance (2.13 from the t table in Appendix E), reject the null hypothesis and accept the alternative hypothesis. This is same as what you have got using the t test under the Data Analysis pack of Microsoft Excel.

Progressive Test Questions

1. For the same example, if you want to find out whether engineers are doing better than the accountants, how will you frame the null and alternative hypothesis? What is your conclusion?
2. What is the P-value corresponding to the one tailed-test?
3. What is the P-value corresponding to the two-tailed test?

For better clarity, the output of Excel is again given below:

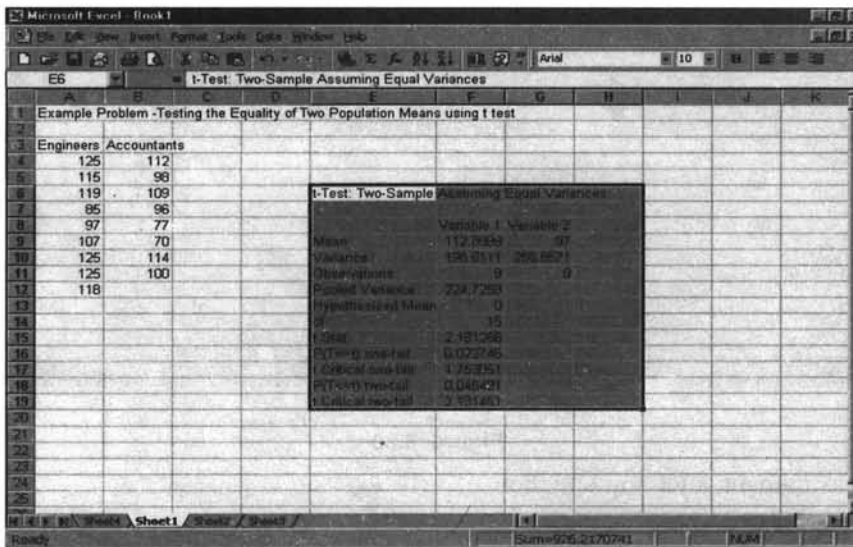


Figure 8.28

Solution

1. $H_0 : \mu_1 \leq \mu_2$
 $H_1 : \mu_1 > \mu_2$

The computed t Statistic = 2.18. The critical t for 15 d.f for the one-tailed test is 1.75. Reject H_0 and accept H_1 . We conclude that the engineers have faired better than the accountants at 5% level of significance.

2. P-value for the one-tailed test is 0.0227 from the output above
3. P-value for the two-tailed test is 0.0455 from the output above

Difference in two population means - small sample (Dependent Sample)

A situation, like the one shown in figure 8.29, can be answered by the t -test when the sample size is small and the measurements are taken before and after situation for the same sample. This t is called the paired t and follows a t distribution with $n-1$ d.f. Please note that the sample size for both the measurements must be equal to n .

Illustration

Based on the sample data collected from a panel of households before and after a special promotion campaign for Surf, has the mean purchase quantity of Surf per household been higher after the campaign than before the campaign?

Figure 8.29

The hypothesis formulation and the test statistic are given below:

Statement of the Null and Alternative Hypothesis

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_1 : \mu_1 - \mu_2 \neq 0$

Test Statistic:

$$t = \frac{\bar{D}}{S / \sqrt{n}} \text{ where } D = X_A - X_B$$

Figure 8.30

X_A = Measurement after the situation

X_B = Measurement before the situation

Paired t test-example A company conducted a promotional campaign in 10 randomly chosen retail outlets. Monthly sales in 1000 units are shown before and after the campaign. Is there any significant difference in sales before and after the campaign?

Outlet No	Before the Campaign	After the Campaign
1	240	270
2	225	245
3	250	260
4	280	290
5	200	190
6	150	160
7	165	160
8	100	130
9	130	135
10	170	175

Solution This is a two-tailed hypothesis problem. Let us take a level of significance of 0.05. The appropriate test statistic is a paired t for the dependent sample.

$H_0: \mu_1 = \mu_2$ (The mean sales are same before and after the campaign)

$H_1: \mu_1 \neq \mu_2$ (The mean sales are not same before and after the campaign)

We can obtain the solution easily from Excel. After the data entry, the screen looks as follows:

Outlet No	Before the Campaign	After the Campaign
1	240	270
2	225	245
3	250	260
4	280	290
5	200	190
6	150	160
7	165	160
8	100	130
9	130	135
10	170	175

Figure 8.31

Click Tools, Click Data Analysis, and then click t-test for paired two sample for means, you get:

t-Test: Paired Two Sample for Means

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

Labels

Alpha:

Output options

Output Range:

New Worksheet Ply:

New Workbook

OK Cancel Help

Figure 8.32

Enter the input for variable 1 range and variable 2 range, and then click OK. You get the answer. Now look at the output. You infer the following.



Figure 8.33

The calculated t value is -2.51197 . As Excel subtracts the sample mean 2 from sample mean 1, the t statistic is negative. As t distribution is symmetrical, it is enough if you take the positive computed t . The positive computed value of $t = 2.51$. Critical t for a two tailed test = 2.26. Reject H_0 and accept H_1 . The mean sales are not the same before and after the campaign. Using a one sided test, it is easy to conclude that the promotion is effective in increasing the sales.

You can also do this same problem by using the formula for a paired t test. This is shown below:

Performing Calculation for the Paired t Test

Outlet No	Before the Campaign (X_A)	After the Campaign (X_B)	$D = X_A - X_B$	$D - \bar{D}$	
1	240	270	30	19.5	
2	225	245	20	9.5	
3	250	260	10	-0.5	
4	280	290	10	-0.5	
5	200	190	-10	-20.5	$S^2 = 174.7222$
6	150	160	10	-0.5	$S = 13.21825$
7	165	160	-5	-15.5	
8	100	130	30	19.5	
9	130	135	5	-5.5	
10	170	175	5	-5.5	
			\bar{D}	10.5	

$$t = \frac{\bar{D}}{(S/\sqrt{n})} = 2.5119745$$

As you can see the value of t calculated is exactly same as that of Data Analysis Pack of Microsoft Excel. The table value of t from Appendix E for 9 degrees of freedom at 5% level of significance is 2.26. Since the calculated t is greater than critical t , reject the null hypothesis and accept the alternative.

Difference between Two Population Proportions:

Hypotheses Testing Parametric Tests

Test of Two Population Proportions

Primary Objective of the Test: To test statistical hypotheses that compare the population proportion of interest for two separate populations.

Figure 8.34

Just like the sampling distribution of the difference in means, the difference between two sample proportions follows a normal distribution when the sample size of both the populations is large (greater than or equal to 30). The test statistic to be used is given by:

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \text{ This follows a normal with mean 0 and standard deviation 1.}$$

p_1 is the sample proportion to estimate the population proportion P_1 .

p_2 is the sample proportion to estimate the population proportion P_2 .

p is the combined mean of p_1 and p_2 with weights n_1 and n_2 . That is, $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$.

The decision rule is the same. If the calculated value of positive Z is greater than the critical Z , reject the null hypothesis and accept the alternative hypothesis. The critical Z for 5% level and 1% level for the two-tailed test are 1.96 and 2.58 respectively. Also for a one-tailed test, the critical value of Z is 1.65 for a level of significance of 5%. Don't memorize. Use Microsoft Excel to get these values.

Example Two random sample surveys, conducted with two months gap between the two, assessed public opinions on the outcome. The question that was posed was, "If the general election was going to take place tomorrow, would you cast your vote for or against the ruling party?"

The results of the two surveys are tabulated below:

	1st survey	2nd Survey
Sample size.....	1000	800
For the ruling party.....	520	380
Against the ruling party.....	480	420

Set up the appropriate hypotheses, test and draw your conclusions.

Solution This requires a structuring of the null hypothesis as no change in pattern of voting between the two months by the public. Symbolically,

$H_0: P_1 = P_2$ (The population proportion favoring the ruling party in the two months gap is the same)

$H_1: P_1 \neq P_2$ (No. It is not the same).

Let us take a level of significance of 5%. It is a two-tailed test and therefore, critical value of $Z = 1.96$.

The test statistic $Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \left(p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \right)$ First, let us calculate p .

$$p = \frac{1000(520/1000) + 800(380/800)}{1000 + 800} = 0.50.$$

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.52 - 0.475}{\sqrt{0.5(1-0.5)\left(\frac{1}{1000} + \frac{1}{800}\right)}} = 1.90$$

The calculated value of Z is less than the table value of Z . Accept H_0 . The inference is that the population proportion of the public favoring the ruling party in the two months gap has not changed at a level of significance of 5%.

Discussion Topic

Analyze, criticize, and explain the following statement.

"What is significant to a manager may not be statistically significant. What is not significant to a manager may be statistically significant".

8.4 CHAPTER SUMMARY

This chapter has introduced you to the nitty-gritty of Hypotheses Testing with conceptual rigor and decision orientation. In particular, this chapter focused on:

- The definition and meaning of statistical hypothesis
- Conceptual foundation of the null and alternative hypothesis
- Type I and Type II error and their relationship with confidence level and power of the test respectively
- How to formulate the null and alternative hypothesis for the one-tailed and two-tailed test
- How the hypothesis works in practice as a decision making tool
- One sample tests for the population mean using Z test for the large sample and t test for the small sample.
- One sample test for the population proportion using Z test for the large sample
- Two sample tests for the difference between two population means, and two population proportions using the Z test.
- Two sample tests for the difference in population means using the t test for small sample when the samples are independent
- Paired t test for the dependent sample (before and after situation type)
- How to apply Microsoft Excel in appropriate tests.

GLOSSARY

Alpha (α) The probability of making a Type I error is called the level of significance of the test. It is designated by the Greek letter *alpha* (α).

Alternative Hypothesis The conclusion we accept when the null hypothesis is rejected.

Beta (β) The probability of making a Type II error is symbolized by the Greek letter (β).

Critical Value It is the standard value of the test statistic for a given level of significance that is used to decide whether to reject the null hypothesis or not.

Hypothesis A hypothesis is a statement about a population parameter. It may or may not be true.

Level of Significance It is the probability of rejecting the null hypothesis when in fact it is true.

Null Hypothesis It is the hypothesis that is always tested. It is so formulated that its rejection leads to the desired conclusion.

One-Tailed Test It is a directional hypothesis in which the population parameter is structured to be greater than or equal to, or less than or equal to a specified value.

Power of the Test It is the probability of rejecting the null hypothesis when in fact it is false.

Two-Tailed Test It is a non-directional hypothesis in which the population parameter is structured to be equal to a specified value.

Type I Error Rejecting a null hypothesis when in fact it is true.

Type II Error Accepting a null hypothesis when in fact it is false.

REVIEW QUESTIONS

1. If level of significance is chosen to be 5% and the P-value is 0.03, then the null hypothesis should be rejected. True or False.
2. For a sample size of 20, the critical t value when $\alpha = 0.05$ for a two-tailed test is -----.
3. A random sample of size 81 with mean 12 and standard deviation 18 is drawn from a normal population. The null hypothesis is that the population mean is 7. What is your conclusion?
 - (a) Null Hypothesis is rejected at 5% level
 - (b) Null Hypothesis is accepted at 5% level
 - (c) Test is inconclusive at 5% level

Questions 4 to 9 will have to be answered based on the following situation:

You are working as a financial analyst for an investment firm. Your boss has asked you to find out whether there is a significant difference in return on investment per annum, between stocks listed on the NSE & BSE? You have collected the following data based on random sample of stocks from these two exchanges. Use a level of significance of 5%.

	NSE	BSE
Sample number of stocks	18	20
Mean return in %	12.5	14.0
Standard deviation in %	2.5	3.0

4. The computed t statistic is
 - (a) 2.66
 - (b) 0.67
 - (c) -1.67
 - (d) 1.83
5. The degrees of freedom associated with the t distribution is -----.
6. The hypotheses to be formulated involve a two-tailed test. True or False.
7. The critical t value for this problem is -----.

8. The conclusion is:
- Reject the null hypothesis
 - Accept the null hypothesis
 - Can't say
9. If the sample sizes of stocks in NSE and BSE were changed to 35 and 45, the conclusion will be:
- Reject the null hypothesis
 - Accept the null hypothesis
 - Can't say

ANSWERS TO REVIEW QUESTIONS

- Answer** The statement is true because this is the basic criterion to decide whether to reject the null hypothesis or not. Whenever the P-value is less than the level of significance, reject the null hypothesis.
- Answer** From Microsoft Excel paste function, t value for 19 d.f. for a two tailed test is 2.09. See diagram below:

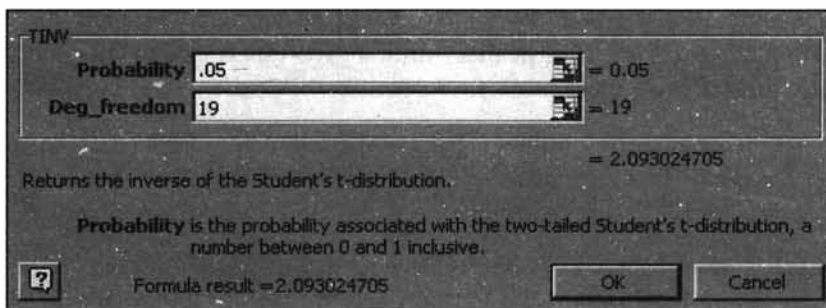


Figure 8.35

- Answer** (a) is the right choice. From the wording, the problem is a two-tailed test. The appropriate test statistic is Z because the sample size is large. At 5% level, the critical Z is 1.96.

$$H_0: \mu = 7$$

$$H_1: \mu \neq 7$$

$$Z = \left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right) = \left(\frac{12 - 7}{\frac{18}{\sqrt{81}}} \right) = 2.5. \text{ The computed } Z \text{ value is greater than the critical } Z.$$

Reject H_0 and accept H_1 . That is, the population mean is not equal to 7.

4. **Answer** The right choice is (c)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(18 - 1)2.5^2 + (20 - 1)3.0^2}{18 + 20 - 2} = 7.70.$$

$$t = \frac{12.5 - 14}{\sqrt{7.70 \left(\frac{1}{18} + \frac{1}{20} \right)}} = -1.67.$$

5. **Answer** For a two sample test, the degrees of freedom = $n_1 + n_2 - 2 = 18 + 20 - 2 = 36$
6. **Answer** True. The wording of the question in the problem is to find out whether the two mean returns differ significantly. So, the two-tailed test is the right choice.
7. **Answer** Using the Microsoft Excel we have the critical value for the t test to be = 2.03 (36 d.f., 5% level of significance, and a two-tailed test). This value can also be taken from the t distribution table in Appendix E.
8. **Answer** (b) is the right choice. The computed t is -1.67 . Just take the positive computed $t = 1.67$. The critical t for 36 d.f = 2.03. Since the computed t is less than critical t at 5% level of significance, accept the null hypothesis of equal means.
9. **Answer** (a) is the right choice. Since the sample sizes are large, Z test is the right statistic. The sample standard deviations can be put in the place of the population standard deviations. The critical Z for 5% level = 1.96.

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{12.5 - 14}{\sqrt{\frac{2.5^2}{35} + \frac{3^2}{45}}} = -2.42.$$

Take the positive $Z = 2.42$. Since the computed Z

is greater than critical Z , reject the null hypothesis of equal means.

PRACTICE PROBLEMS

1. Case Study- New Product Introduction

A marketing manager of an enterprise is facing a decision whether to introduce a new product into the market or not. Consumer acceptance measured in a blind comparison

test is agreed upon as an appropriate basis for evaluation. Marketing of the new product will be pursued only if the acceptance rate exceeds 35%. Otherwise, the new product will not be introduced in the market. A random sample of 300 consumers reveals that the acceptance rate is 38%. Should the new product be introduced? Using a level of significance of 0.05, perform the appropriate hypothesis test and recommend your action.

2. Case Study -Test of Analytical Ability

A leading business school situated close to many corporate houses in a city, conducts on behalf of companies, specially designed tests that would increase the analytical ability of their executives. In this context, it runs a short certificate course that is developed to enhance the ability to analyze data. At the end of this course, a standard written test is conducted. From the past data, it is known that the scores obtained by the executives are normally distributed with mean 350 and standard deviation 70. A new computer-based online test for this same course has just been evolved. 36 executives participated in this online test. Their average score was 375. Is this online test superior to the current written test? Perform an appropriate test of hypothesis and recommend your decision.

3. Case Study - Readymade Garment

A Readymade Garment manufacturer buys various types of yarn. One particular yarn, according to specification should have a tensile strength of 14 kg on an average. The company supervisor who is highly knowledgeable in this area, strongly believes that recent supplies are inferior in quality. A random sample of 100 standard specimen of this yarn was taken. It was found that the average tensile strength in the sample was 12.5 kg with a standard deviation of 1.75 kg. Is the supervisor right in saying that the yarn supplied was of inferior quality? Justify your answer by performing an appropriate test of hypothesis.

4. Case Study - Sales Incentive Scheme

The Vice President (Marketing) of consumer Product Company has just implemented a new sales incentive scheme for his salespeople. He wants an early feedback on the level of success of the new scheme.

It is five months since the incentive scheme has now been implemented. To test the effectiveness of the scheme, the company has taken a random sample of 20 salespeople, measured their sales in the first five months before implementing the scheme, and then measured the sales in the last five months in which the incentive scheme has been in operation. The sales value of the salespeople are shown below:

Sales achieved (000 Rs)

<i>Salesperson</i>	<i>Existing Scheme</i>	<i>New Scheme</i>
1.	60	67
2.	123	140
3.	62	60
4.	85	88
5.	90	90
6.	83	95
7.	45	40
8.	120	125
9.	50	48
10.	90	110
11.	75	95
12.	70	80
13.	80	100
14.	50	60
15.	50	45
16.	60	80
17.	90	95
18.	80	70
19.	85	90
20.	75	80

Has the new incentive scheme been effective in boosting the sales? Justify your answer by performing a suitable test of hypothesis.

Chi-Square Test and Analysis of Variance (ANOVA)

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Explain and use Chi-Square test
- Explain and use ANOVA
- Formulate and test hypothesis using Chi-Square and ANOVA

CHAPTER OUTLINE

- 9.1 Chi-Square (χ^2) Analysis-Basics
- 9.2 Chi-Square Test-Goodness of Fit
- 9.3 Chi-Square Test of Independence
- 9.4 ANOVA-Basics
- 9.5 ANOVA-One-Way Classification
- 9.6 ANOVA-Two-Way Classification
- 9.7 Chapter Summary
- Glossary
- Review Questions
- Answers to Review Questions
- Practice Problems

INTRODUCTION

In the previous chapter, we have made inferences about difference between two population means based on the corresponding sample means. Suppose we are interested in testing the equality of means involving more than two populations, we have an elegant technique known as ANOVA, developed by Ronald Fisher, the father of statistics in the year 1920. The specialty of ANOVA is that it is part of the domain called "Experimental Design", which deals with cause-effect relationship in an effective manner. Cause-effect relationship is also reflected in association of attributes. Association of attributes is effectively answered by the chi-square test. This chapter covers the basic models of chi-square test and ANOVA.

Examining Cause and Effect Relationship



Figure 9.1

9.1 CHI-SQUARE (χ^2) ANALYSIS-BASICS

Chi-square analysis is widely used in research studies for testing hypothesis involving nominal data. Nominal data are also known by two names - categorical data and attribute data. The symbol χ^2 statistics is used to designate the chi-square distribution whose value depends on the number of degrees of freedom (d.f.). A chi-square distribution is a skewed distribution particularly with smaller d.f. As the sample size and therefore the d.f. increases, the χ^2 distribution becomes a symmetrical distribution approaching normality. The general shape of the χ^2 distribution for smaller d.f. is given below:

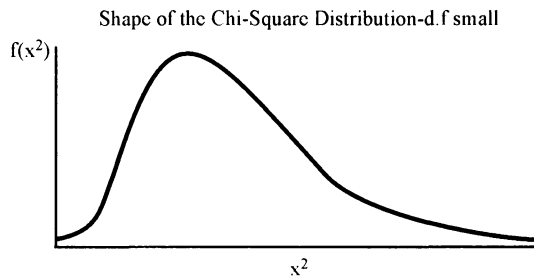


Figure 9.2

The χ^2 test is a **nonparametric** test. Nonparametric means no assumption needs to be made about the form of the original probability distribution from which the samples are drawn. It is a classic nonparametric test involving data measurement in nominal scale. Please note that all parametric tests make the assumption that the samples are drawn from a specified or assumed population. Thus, nonparametric methods are also called "*distribution free*" methods.

Conditions for using the χ^2 Test

- The sample observations drawn from a population must be independent and random.
- The data must be in *frequency* (counting) form. If the original data are in percentages, they must be converted into frequency.
- No frequency in any cell/category must be less than 5. If the frequency is less than 5 for a category, you have to do some regrouping.

When the degrees of freedom is greater than or equal to 30, the χ^2 distribution approaches the normal distribution.

9.2 CHI-SQUARE TEST-GOODNESS OF FIT

Hypothesis testing procedure is exactly the same as what you have seen in chapter 8. For a level of significance, the critical χ^2 will be always the upper χ^2 value from the standard table. Of course, our table and bible is Microsoft Excel. Please note that χ^2 value is always positive. You will understand the use of χ^2 test-goodness of fit by going through the example discussed on the next page.

χ^2 Test Goodness of Fit: Nominal Data

This test is used to examine whether a set of observed frequencies comes from a universe that has a particular distribution (e.g. normal distribution).

This can also be used to know whether some observed pattern of frequencies fit well with an expected pattern of frequencies.

Figure 9.3

Nominal Data: χ^2 Test Univariate Goodness of Fit

Test Statistic:

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$

Where O_i = Observed Frequency for the i th category

E_i = Expected Frequency for the i th category

Figure 9.4

Example In consumer marketing, a common problem the marketing manager encounters is the selection of the appropriate package design. Assume that a marketer wishes to compare five different package designs. He is interested in knowing which is the most preferred one so that the same can be introduced in the market. A random sample of 200 consumers gives the following picture:

<i>Package Design</i>	<i>Preference by consumers</i>
A	36
B	52
C	40
D	35
E	37
Total	200

Do the consumer preferences for the designs show any significant differences?

Solution If you look at the table, you may be tempted to feel that B is the most preferred design. Statistically, you have to find out whether this preference could have arisen due to chance. The appropriate test statistic is the χ^2 test of goodness of fit.

Null Hypothesis All package designs are equally preferred.

Alternative Hypothesis No, they are not equally preferred.

Package Design	Observed(O)	Expected(E)	(O - E) ²	$\left(\frac{(O - E)^2}{E}\right)$
A	36	40	16	0.400
B	52	40	144	3.600
C	40	40	0	0.000
D	35	40	25	0.625
E	37	40	9	0.225
Total	200	200		4.850

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right) = 4.850$$

The critical value of χ^2 at 5% level of significance from Excel is obtained as follows:

Solution continues Click paste function, click Statistical, and then click CHIINV.

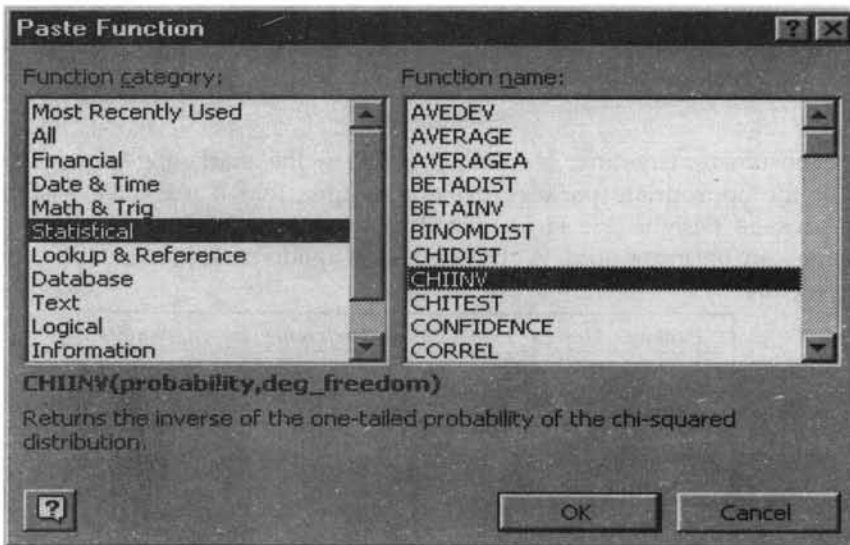


Figure 9.5

Click OK. Then enter the values in the probability cell and d.f, you get:

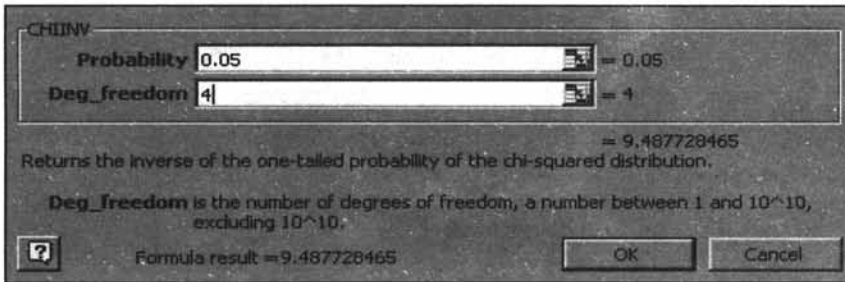


Figure 9.6

The critical χ^2 for 4 d.f at 5% level of significance is 9.49(see above). Since the calculated value of χ^2 is less than critical χ^2 at 5% level, accept the null hypothesis of equal preference. The conclusion is that all packages are equally preferred and difference in preference in the sample survey may have arisen due to chance.

Progressive Test Question Chi-Square Test of goodness of fit is the appropriate test to use when the data measurements are ordinal. True or False

Answer The statement is false. This is because Chi-Square test is possible only when the data measurements are nominal(categorical).

9.3 CHI-SQUARE TEST OF INDEPENDENCE

The goodness-of-fit test explained above is suitable for situations involving one categorical variable (e.g. package design). If there are two categorical variables, and our interest is to find out whether these two variables are associated with each other, the χ^2 test of independence is the appropriate technique to use. This test is very popular for analyzing cross-tabulations in which an investigator is keen to find out whether the two categorical variables are having any relationship with each other.

The cross-tabulation is called a contingency table containing frequency data corresponding to the categorical variable in the row and the column. The marginal totals of the rows and columns are used to calculate the expected frequencies that will be part of the computation of the χ^2 statistic. Have you heard this term "contingency table" earlier? You must say 'yes'! Please go through chapter 4 on probability once more, where the contingency table is discussed and used in the context of computing joint and marginal probability.

A marketing manager may be curious to find out whether consumers in different income strata prefer different brands. Is there any association between the brand preferred and the income strata? The null hypothesis is that there is no association between brand and income; they are independent. χ^2 test of independence in a contingency table is used here. The procedure for the hypothesis testing is same as the one variable case except the calculation of expected frequencies and the degrees of freedom.

Example In a market survey conducted to examine whether the choice of a brand is related

to the income strata of the consumers, a random sample of 600 consumers reveal the following:

Observed frequencies (O_i)

<i>Income Strata (Income Per month)</i>	<i>Brand1</i>	<i>Brand2</i>	<i>Brand3</i>	<i>Total</i>
Less than Rs.10000	132	128	50	310
Rs 10000–15000	62	60	28	150
Rs 15001–20000	30	30	26	86
Above Rs 20000	16	22	16	54
Total	240	240	120	600

The manger who conducted this survey wants to know whether the brand preference is associated with the income strata. If so, he has to evolve a creative advertising strategy. Solution is given next.

Solution The null hypothesis is that there is no association between the brand preference and the income level (These two are independent). The alternative hypothesis is that the brand and income level are associated (dependent).

Let us take a level of significance of 5%.

In order to calculate the χ^2 value, you need to work out the expected frequency in each cell in the contingency table. In our example, there are 4 rows and 3 columns amounting to 12 elements. There will be 12 expected frequencies.

Observed frequencies (O_i)

<i>Income Strata (Income Per month)</i>	<i>Brand1</i>	<i>Brand2</i>	<i>Brand3</i>	<i>Total</i>
Less than Rs.10000	132	128	50	310
Rs 10000–15000	62	60	28	150
Rs 15001–20000	30	30	26	86
Above Rs 20000	16	22	16	54
Total	240	240	120	600

The expected frequency corresponding to row1 and column 1 is computed as follows:

Marginal total of row1 multiplied by marginal total of column1 divided by grand total. This = $(310)(240)/600 = 124$. What is the rationale for this procedure? Simple. The probability that the consumer will have income less than Rs.10000 is $310/600$. The probability that brand1 is preferred by the consumer is $240/600$. Under the null hypothesis of the attributes being independent, the joint probability that the income is less than 10000 and Brand1 is preferred = $(310/600)(240/600)$. Therefore, the expected number of consumers out of 600 who are having income less than 10000 and prefer Brand1 = $(310/600) \cdot (240/600) \cdot 600$. This is = $(310)(240)/600$. Hence, the formula is, "the expected frequency corresponding to a particular row and column = marginal total of that row multiplied by

marginal total of that column divided by the grand total". This is how all the expected frequencies are calculated and filled in the table of Expected Frequency below:

Solution continues

Observed Frequency (O)

<i>Income Strata</i>	<i>Brand1</i>	<i>Brand2</i>	<i>Brand3</i>
Less than 10000	132	128	50
10000 to 15000	62	60	28
15001 to 20000	30	30	26
Above 20000	16	22	16

Expected Frequency (E)

<i>Income Strata</i>	<i>Brand1</i>	<i>Brand2</i>	<i>Brand3</i>
Less than 10000	124	124	62
10000 to 15000	60	60	30
15001 to 20000	34.4	34.4	17.2
Above 20000	21.6	21.6	10.8

Note The fractional expected frequencies are retained for accuracy purpose. Do not round them off.

Compute $\chi^2 = \sum \left(\frac{(O-E)^2}{E} \right)$. There are 12 observed frequencies (O) and 12 expected

frequencies (E). As in the case of the goodness of fit, calculate this χ^2 value. In our case, the computed $\chi^2 = 12.76$.

The critical value of χ^2 depends on the degrees of freedom. The degrees of freedom = (the number of rows-1) multiplied by (the number of cols-1). In our case, there are 4 rows and 3 columns. So the degrees of freedom = (4 - 1). (3 - 1) = 6. You may wonder why this formula is used for computing the degrees of freedom in a contingency table. You have 4 rows. Suppose for the first three rows all entries are known to you. Then the fourth row total can be obtained by subtraction. In our example, in the original table, the first three rows are known cell by cell. The fourth row total = Grand total -total of the first three rows = 600 - 546 = 54. How to get the entries for each cell in row 4? If we know the entries in this row for two columns, the third column entry corresponding to this row4 can be worked out by subtraction. The entries in the first two columns in row 4 are 16 and 22. So, the entry for row4, column3 must be = 54 - 38 = 16. So, if we know the entries for 3 rows and two columns, the remaining cells can be filled in. Extending this logic, if there are m rows and n columns, the degrees of freedom associated with the Chi-square test = (m - 1) × (n - 1).

Using Excel, you can get the upper χ^2 value at 5% level for 6 d.f. This value=12.59. See paste function in the next page.

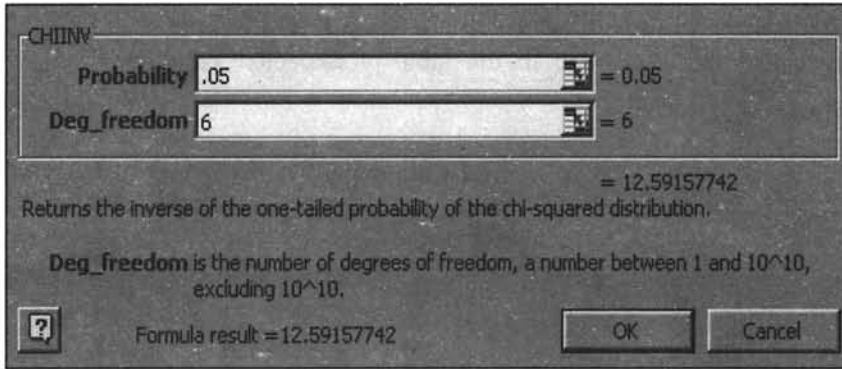


Figure 9.7

Since the calculated value of χ^2 is greater than the critical value of χ^2 , reject the null hypothesis and accept the alternative hypothesis. The conclusion is that the brand preference and income level are associated. A creative advertising strategy will have to be worked out to take care of this relationship.

Progressive Test Question Which of the following are true about the χ^2 test?

- (a) The sample observations drawn must be independent and random
- (b) The samples drawn must be from a normal population
- (c) The data measurements must be categorical
- (d) The data must be expressed in percentage form

Answer (a) and (c) are true because they are part of the conditions for applying the test. (b) is wrong because χ^2 is a nonparametric test and as such does not require the assumption of the samples drawn from a normal distribution. (d) is wrong because the data must be expressed in frequency form.

Progressive Test Question For the distribution with 10 d.f, what is the upper critical value for a level of significance of 5%. Is it?

- (a) 8.31
- (b) 19.30
- (c) 18.31

Solution The right choice is (c). Using the table of Chi-Square distribution in Appendix F, you will find that the upper χ^2 value for 5% level of significance is 18.31.

The best way is you use paste function of Microsoft Excel and get the critical value. It is given below:

Using the Microsoft Excel paste function critical value of χ^2 for 10 d.f at 5% = 18.31.

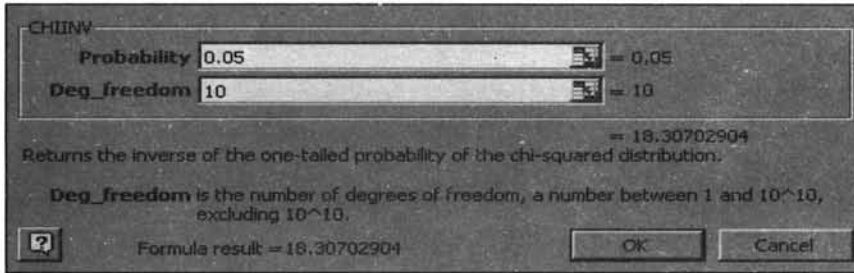


Figure 9.8

9.4 ANOVA-BASICS

- Very often an investigator would like to compare more than two population means in a problem situation. For example, a human resource manager may be interested in the aptitude test scores from a standard test for executives having four types of background. Such comparisons could be accomplished using the t distribution by looking at samples two at a time and then comparing the means. Although this method is feasible, it is an inefficient method of doing the comparison for more than two population means.
- The first reason for its inefficiency is that the standard deviation for the differences between the two sample means is not computed based on observations from all the samples. Instead, it utilizes samples only from the two populations that are under immediate consideration. The second reason is that, intuitively, we will almost always find a significant difference that exists between at least one pair of means when there are enough identical populations. This implies that we cannot trust our level of significance. That is, there is every possibility that the type I error will get inflated in the process.
- We need a technique that will enable us to test the hypothesis of equality of more than two population means that will overcome the deficiencies of the t test. It is in this context that the role of analysis variance, popularly called ANOVA, looms large. Ronald Fisher, considered the father of statistics, was the architect of a fascinating discipline called the "**experimental design**". Fisher's objective was to establish the cause and effect relationship between variables. By performing experiments in his agricultural field, Fisher could test the hypothesis involving more than two population means and the fundamental technique that he used was the analysis of variance. Thus, analysis of variance is part of this great domain of experimental design.
- The beauty of ANOVA is that it performs the test of equality of more than two population means by actually analyzing the variance. In simple terms, ANOVA decomposes the total variation into components of variation. That is, explaining the changes in the response variable caused by these components. To put it succinctly, the total sum of squares is equal to the sum of squares due to causes. We will shortly explain how ANOVA works in practice by first focusing on the one-way classification.

The following visual succinctly captures the essence of ANOVA.

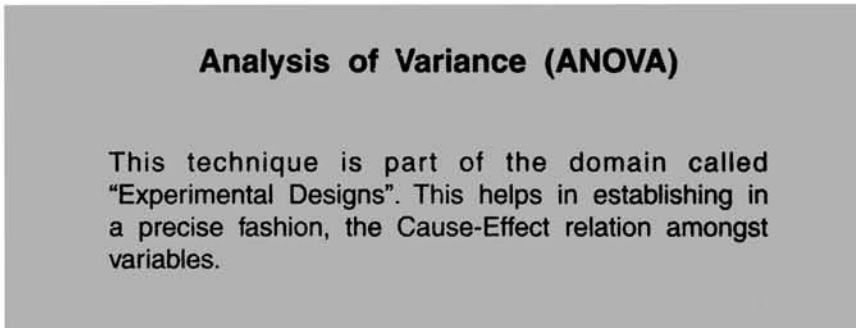


Figure 9.9

9.5 ANOVA-ONE-WAY CLASSIFICATION

The following visual gives one typical application of ANOVA using one-way classification.

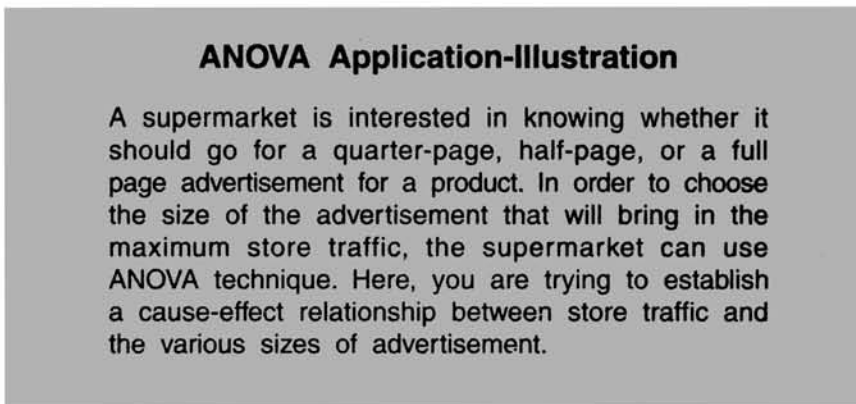


Figure 9.10

In the parlance of experimental design, one-way classification is called *Completely Randomized Design*. The treatments are randomly allocated to the experimental units.

How One-Way Classification Works in Practice?

You are going to first decompose the total sum of squares into sum of squares due to causes. Here, you are assuming that the **Total Sum of Squares = Treatment Sum of Squares + Error Sum of Squares**. The word treatment is generic and as such may denote different methods, machines, different advertisement copy platforms, different strategies, different brands and the like. The variation in sum of squares of the response variable (dependent variable) is caused only by treatment and any thing unexplained by the treatment is attributed to error term.

$$\text{Total Sum of Squares (TSS)} = \sum (X_{ij} - \bar{\bar{X}})^2$$

$$\text{Treatment Sum of Squares (TRSS)} = n \times \sum (\bar{X}_j - \bar{\bar{X}})^2$$

$$\text{Error Sum of Squares (ESS)} = \text{Total Sum of Squares} - \text{Treatment Sum of Squares}$$

Where

X_{ij} is the i th observation under j th treatment

$\bar{\bar{X}}$ is the Overall Mean (Grand Mean)

\bar{X}_j is the j th treatment mean

You may be bewildered by these complex formulas and get worried. Don't worry. A good explanation of the formulas is given below.

Meaning of the formulas

Total Sum of Squares (TSS) = $\sum (X_{ij} - \bar{\bar{X}})^2$. This says, you subtract the grand mean from each observation in the data set, then square each of them, and then add all of them. Grand mean is obtained by adding all observations in the data set and then dividing the sum by the number of observations in the data set.

Treatment Sum of Square (TRSS) = number of observations in each treatment $\times \sum (\bar{X}_j - \bar{\bar{X}})^2$. This says that first work out the mean for each treatment. Subtract the grand mean from each treatment mean and then square the same. Multiply each of these squares by the number of observation for each treatment. Now add all these. You get the treatment sum of squares. Please note that the number of observations under each treatment need not be the same.

Error Sum of Squares (ESS) is obtained by subtracting the treatment sum of squares from total sum of squares. $\text{ESS} = \text{TSS} - \text{TRSS}$

Once you have all the basic sum of squares computed, the next step is to prepare the ANOVA table and test the hypothesis of equality of means.

I hope you are getting some sense out of these formulas. The best way to understand how ANOVA works in practice is through an illustration.

Illustration Problem A consumer marketing group desired to examine whether supermarket chains operating in a city differed in their "out of stock" levels for advertised specials. The group identified the relevant response variable as the percentage of the items advertised not in stock. The following table provides the data collected from three supermarket chains in the city.

Percentage of the items "Out of Stock" on Advertised Specials

<i>Chain1</i>	<i>Chain2</i>	<i>Chain3</i>
15	10	17
14	14	12
20	9	14
15	10	15
16	11	12

The marketing group would like to know whether there are significant differences among the three chains with regard to mean percentage out of stock on advertised specials. How would you analyze this situation?

The basic calculations are given below:

(a) Table showing Grand Mean and Treatment Means.

Percentage of the Items "Out of Stock" on Advertised Specials			
	<i>Chain1</i>	<i>Chain2</i>	<i>Chain3</i>
	15	10	17
	14	14	12
	20	9	14
	15	10	15
	16	11	12
Treatment Means =	16	10.8	14

Grand Mean = 13.6

Please note that the treatment means here denote the mean of each chain because the chains are the treatments. These are given in the bottom row for each column.

(b) Table showing Total Sum of Squares is given below:

Total Sum of Squares Calculation			
	<i>Chain1</i>	<i>Chain2</i>	<i>Chain3</i>
	1.96	12.96	11.56
	0.16	0.16	2.56
	40.96	21.16	0.16
	1.96	12.96	1.96
	5.76	6.76	2.56

TSS = 123.6

In the table above, each cell is nothing but the square of (original value-grand mean). Adding all of these, you get TSS = 123.6.

(c) Table showing Treatment Sum of Squares is given below:

5.76	7.84	0.16	TRSS = 68.8
------	------	------	--------------------

In the table above, each cell is nothing but the square of (treatment mean -grand mean). Adding all of these after multiplying by 5, you get TRSS = 68.8.

The Error Sum of Squares (ESS) = TSS-TRSS = 123.6 – 68.8 = 54.8

The next step is to prepare the ANOVA table. This is given below:

ANOVA Table					
Source of Variation	SS	df	MS	F computed	F critical
Treatment (Between Groups)	68.8	2	34.40	7.53	3.89
Error (Within Groups)	54.8	12	4.57		
Total	123.6	14			

Explanation of the ANOVA table

- The first column is labeled as "source of variation". This tells us the factors responsible for causing variation. You see underneath Treatment (**Between Groups**) and Error (**Within Groups**). The Treatment variation represents the variation caused by the chains with regard to the out of stock situation. The error variation represents the unexplained residual part. As mentioned earlier, the word treatment is generic and as such may denote different methods, machines, different advertisement copy platforms, different strategies, different brands and the like.
- The second column is labeled as SS meaning sum of squares. The sum of squares is worked out corresponding to each source of variation. This is already done and you have simply entered them. Please note that **Total Sum of Squares (TSS)** is the addition of **Treatment Sum of Squares (TRSS)** and **Error Sum of Squares (ESS)**.
- The third column is labeled as df denoting the degrees of freedom for each factor. The easy way to understand is that there are 3 chains and so the chain factor has two degrees of freedom (always the original number-1 that are free to move). The total number of observations = 15 and so the total has 14 d.f. The Error (within group) d.f has been obtained by subtraction $14 - 2 = 12$ d.f (d.f for error = d.f for total - d.f for treatment).
- The fourth column is labeled as MS, meaning the Mean Square. You will notice that there are two mean squares. One is for the treatment (between groups) and another for the error (within Groups). Each mean square is obtained by dividing the sum of squares by the corresponding d.f. Please note that the mean squares are not additive. That is, adding the mean squares of treatment and error will not be equal to total mean square.
- The fifth column is labeled as F, meaning the F statistic that we will use for interpretation of result. This is the computed F. Statistically speaking, the F distribution is a ratio of two independent Chi-Square and so it has a pair of degrees of freedom one for the numerator and another for the denominator. Please note that in our computation of F, we have divided the treatment mean square (between group mean square) by the error mean square (within group mean square) to get the value of F. Both these mean squares follow a chi-square distribution. The computed F has (2, 12) df meaning that the numerator has 2 df and the denominator has 12 df. In short form, it is written as F(2,12).
- The sixth column is labeled as F critical meaning the critical F value for (2, 12) df for a given level of significance (0.05 in our problem). In our problem, F critical = 3.89.

This can be obtained either from Microsoft Excel or from the F table provided in the Appendix G.

Formulation of the null and alternative hypothesis

H_0 : The population means of percentage stock out position for all the three chains are equal

H_1 : The population means of percentage stock out position for all the three chains are not equal

Decision Rule If the computed F is greater than the critical F, reject the null hypothesis H_0 and accept the alternative H_1 .

At 5% level from the ANOVA output of Excel, we have the computed $F = 7.53$ and the critical $F(2,12) = 3.89$. So, reject the null hypothesis and accept the alternative. The inference is that the population means of percentage stock out are not the same for all the three chains. So, what do you do? Now, look at the point estimates from the summary table. Chain 1 has a mean stock out of 16%, chain 2 has a mean stock out of 10.8% and chain 3 has a mean stock out of 14%. Chain 2 has the least stock out percentage followed by chain 3 and then chain 1.

ANOVA with pleasure from Microsoft Excel It is indeed a beauty to get ANOVA worked using *Data Analysis* pack of Excel. It is very tempting to take advantage of Excel. Once you have the taste of it, you will never feel like doing ANOVA with the help of a calculator. The step-by step approach to doing ANOVA in Excel is detailed below:

Step 1 Enter the data in the Microsoft Excel, then click Tools, click Data Analysis, and click Anova: Single Factor. You will get the following:

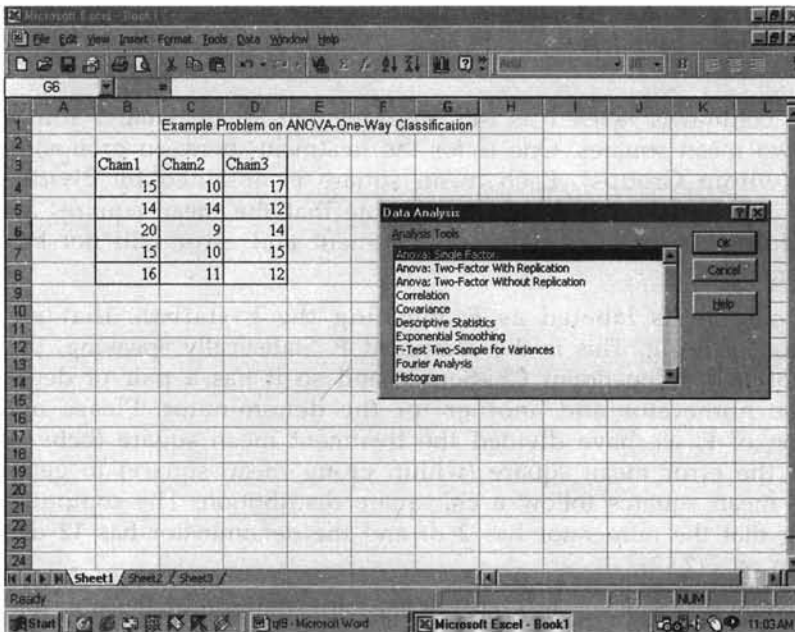


Figure 9.11

Step 2 Click OK and you will get:

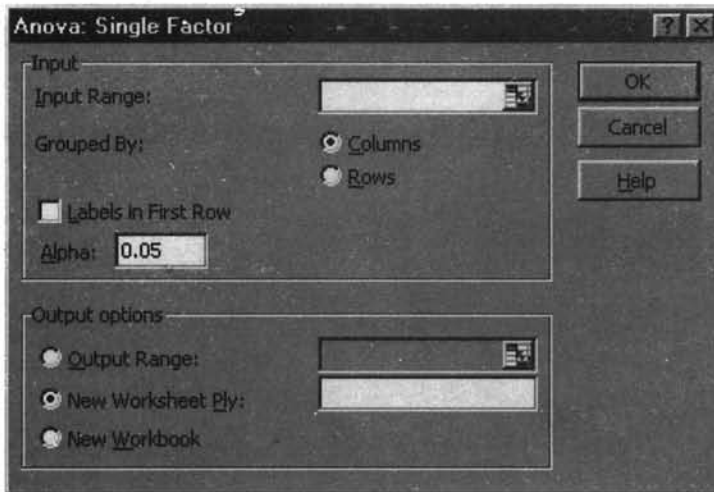


Figure 9.12

Please see the screen above.

By default, the level of significance is taken as 5% and the grouping is done column-wise. You can, of course, do it row wise. In our problem, the chains are the columns.

Step 3 Using the mouse pointer, fill in the Input Range by highlighting the data matrix excluding the labels. Now you get:

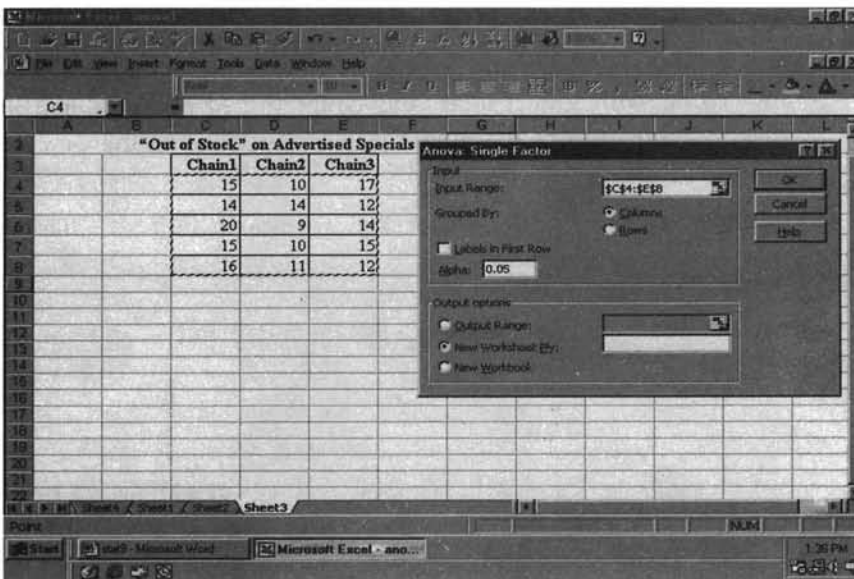


Figure 9.13

Step 4 Click OK and you get the ANOVA table output as follows:

The screenshot shows the ANOVA: Single Factor output table in Microsoft Excel. The table is structured as follows:

ANOVA: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Varianca		
Column 1	5	80	16	5.5		
Column 2	5	54	10.8	3.7		
Column 3	5	79	14	4.5		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	68.8	2	34.4	7.852847	0.007596	3.68229
Within Groups	54.8	12	4.566667			
Total	123.6	14				

Figure 9.14

If you compare the ANOVA output of Excel with what you have got earlier by actual calculation, you notice the following additional features:

- The summary table provides the group means and variances corresponding to each column along with group total, and count (no of observations). In our case, the columns represent the chains and the averages are for the percentage out of stock position.
- The sixth column is labeled as P-value that we know. Refresh Chapter 8.

Assumptions involved in using ANOVA

- The samples drawn from different populations are independent and random. In our case, the samples are independently and randomly drawn from the three supermarket chains.
- The response variables of all the populations are normally distributed. In our example, the response variable namely the percentage stock out is normally distributed.
- The variances of all the populations are equal. In our example, the variances of the three chains are equal.

Discussion Analyze, criticize, and explain the following statement:

"The pattern of differences revealed by ANOVA must be carefully examined in order for a decision maker to decide whether the results are really helpful or only academic in nature".

9.6 ANOVA-TWO -WAY CLASSIFICATION

The two-way classification is a mere extension of the one-way ANOVA. You will have two factors for which the population means will have to be compared. Microsoft Excel has the necessary model to tackle the two-way ANOVA. The testing procedure is exactly the same except for the fact that you will be testing the population means for two factors. Excel does this ANOVA as a row and column comparison. Please note that columns can be taken as **treatments**, and rows as **blocks**. In the parlance of experimental design, two-way ANOVA is called a *Randomized Block Design*. Let us look at an example.

Example A supermarket that has a chain of stores is concerned about its service quality reputation perceived by its customers. The Table below shows the perceived service quality with regard to politeness of the staff. The number in each cell of the table is the percentage of people who have said that the staff is polite. Perform the two-way ANOVA and draw your inferences about the population means of politeness corresponding to the days, as well as, the stores.

Day \ Store	A	B	C	D	E
Monday	79	81	74	77	66
Tuesday	78	86	89	97	86
Wednesday	81	87	84	94	82
Thursday	80	83	81	88	83
Friday	70	74	77	89	68

Step 1 Enter the data in Microsoft Excel, click Tools, click Data Analysis and click Anova: Two factors with out replication, you get:

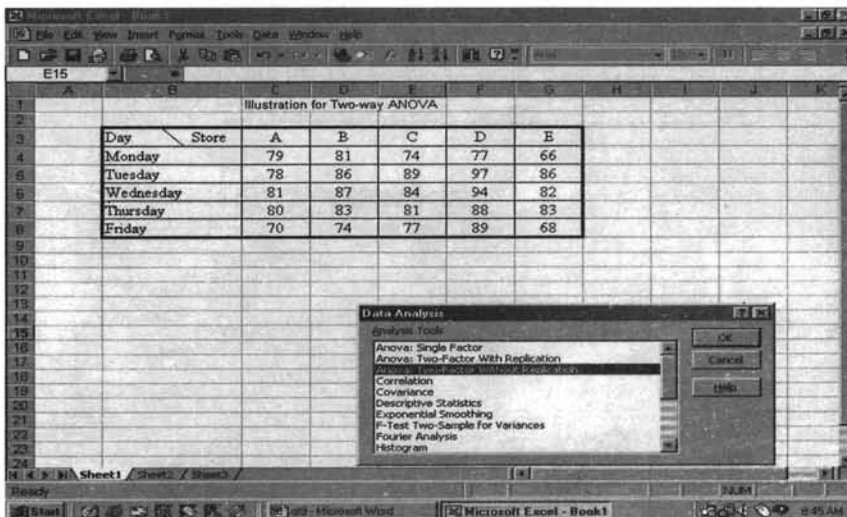


Figure 9.15

Step 2 Click OK and Excel will display, as before, the screen on Input Range.

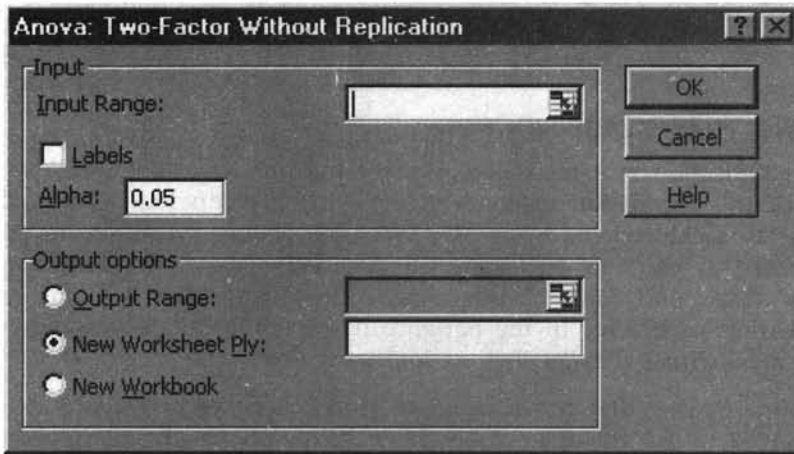


Figure 9.16

Solution continues Please note that Excel can do ANOVA with unequal sample sizes in rows or columns, as well as with and without replication. Here, we have taken without replication meaning that each cell has only one observation.

Step 3 You highlight the entire data matrix using the mouse pointer. You now get the following:

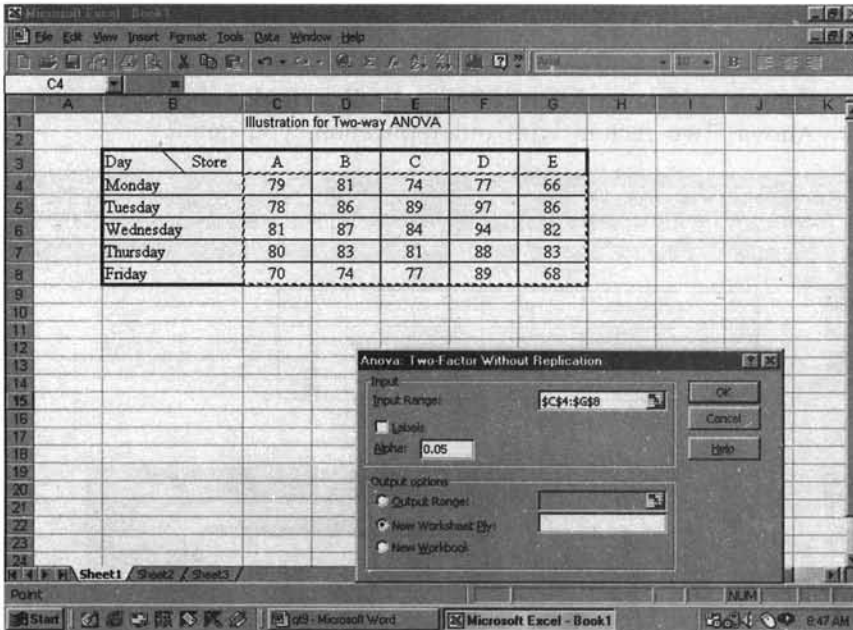


Figure 9.17

Step 4 You click OK and you get the ANOVA output.

Anova: Two-Factor Without Replication						
SUMMARY	Count	Sum	Average	Variance		
Row 1	5	377	75.4	34.3		
Row 2	5	436	87.2	46.7		
Row 3	5	428	85.6	27.3		
Row 4	5	415	83	9.5		
Row 5	5	378	75.6	68.3		
Column 1	5	368	73.6	19.3		
Column 2	5	411	82.2	26.7		
Column 3	5	405	81	34.5		
Column 4	5	446	89	53.5		
Column 5	5	365	73	66		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	Fcrit
Rows	617.36	4	154.34	8.737051	0.000614	3.006917
Columns	451.76	4	112.94	6.534998	0.002575	3.006917
Error	282.64	16	17.665			
Total	1351.76	24				

Figure 9.18

Rows are the days and columns are the stores. The F value computed in both cases is greater than the critical F. So reject the null hypothesis of equality of means in both the cases. The conclusion is that the stores (columns), as well as the days (rows) reveal different patterns in politeness level. The highest politeness level is witnessed on Tuesday and Store D extends the maximum politeness level.

Progressive Test Question In the output above, what are the degrees of freedom for the computed F corresponding to stores and Days?

Answer Degrees of freedom for the F corresponding to stores = F(4,16). Degrees of freedom for the F corresponding to days is also = F(4,16)

Just like the one-way ANOVA if you want to do the two-way ANOVA by formulas, you can take help of the following:

$$\text{Total Sum of Squares (TSS)} = \sum (X_{ij} - \bar{X})^2$$

$$\text{Column Sum of Squares (CSS)} = n \times \sum (\bar{X}_j - \bar{X})^2$$

$$\text{Row Sum of Squares (RSS)} = m \times \sum (\bar{X}_i - \bar{X})^2$$

Error Sum of Squares (ESS) = Total Sum of Squares - Column Sum of Squares - Row Sum of Squares

Where

X_{ij} is the observation corresponding to row j and column i

\bar{X} is the Overall Mean(Grand Mean)

\bar{X}_j is the j th column mean

\bar{X}_i is the i th row mean

The procedure is exactly same as one-way ANOVA except that you also compute the row sum of squares. Go through the one-way ANOVA formula approach once again. For your ready reference, the formula based procedure is given below:

(a) Table showing Row Mean, Column Mean, and Grand Mean

<i>Day \ Store</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>Row Mean</i>
Monday	79	81	74	77	66	75.4
Tuesday	78	86	89	97	86	87.2
Wednesday	81	87	84	94	82	85.6
Thursday	80	83	81	88	83	83
Friday	70	74	77	89	68	75.6
Column Mean =	77.6	82.2	81	89	77	
Grand Mean =	81.36					

(b) Table Showing Total Sum of Squares (TSS)

<i>Day</i>	<i>Store</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
Monday		5.5696	0.1296	54.1696	19.0096	235.9296
Tuesday		11.2896	21.5296	58.3696	244.61	21.5296
Wednesday		0.1296	31.8096	6.9696	159.77	0.4096
Thursday		1.8496	2.6896	0.1296	44.0896	2.6896
Friday		129.05	54.1696	19.0096	58.3696	178.4896
TSS = 1361.76						

Please note that each entry in the above table is the square of (each original value - grand mean). Total Sum of Squares (TSS) = 1361.76

(c) Table showing Column Sum of Squares(CSS)

70.69	3.53	0.65	291.85	95.05	CSS = 461.76
-------	------	------	--------	-------	--------------

Please note that each entry in the above table is the square of (each column mean - grand mean) multiplied by 5. You add all these to get Column Sum of Squares (CSS) = 461.76.

(d) Table showing Row Sum of Squares(RSS)

	177.61
	170.53
	89.888
	13.448
	165.89
RSS =	617.36

Please note that each entry in the above table is the square of (each row mean - grand mean) multiplied by 5. You add all these to get Row Sum of Squares (RSS) = 617.36

(e) Error Sum of Squares(ESS) =TSS-CSS-RSS =1361.76-461.76-617.36 =282.64

You now formulate the ANOVA table by using the knowledge you have acquired in the one-way model. You just add here one more source of variation namely the Rows. Rest of them all are same.

ANOVA Table

Source of Variation	SS	df	MS	F	P-Value	F critical
Rows	617.36	4	154.34	8.737051	0.000614	3.006917
Columns	461.76	4	115.44	6.534956	0.002575	3.006917
Error	282.64	16	17.665			
Total	1361.76	24				

Interpretation of the Results

Rows are the days and columns are the stores. The F value computed in both cases is greater than the critical F. So reject the null hypothesis of equality of means in both the cases. The conclusion is that the stores (columns) as well as the days (rows) reveal different patterns in politeness level. The highest politeness level is witnessed on Tuesday and Store D extends the maximum politeness level.

9.7 CHAPTER SUMMARY

This chapter has introduced you to the basic models of the chi-square tests and ANOVA. The cause-effect relationship has been brought to your attention along with the role of Chi-square and ANOVA. In particular, this chapter focused on:

- > Conceptual framework of the chi-square analysis
- > Basic models of chi-square namely the goodness of fit and test of association in a contingency table through examples
- > Basics of ANOVA

- One-way and Two-way classifications of ANOVA with interpretations in practical problems
- Applying Microsoft Excel to work out the critical value of chi-square
- Applying Microsoft Excel to perform one-way and Two-way ANOVA
- How to get ANOVA using formula approach

GLOSSARY

Analysis of Variance (ANOVA) It is a technique that is used to test the hypothesis of equality of more than two population means.

Chi-Square Distribution It is a probability distribution that is widely used in research studies for testing hypothesis involving nominal data. The value of chi-square distribution depends on the number of degrees of freedom (d.f.).

Contingency Table It is a cross-tabulation table containing frequency data corresponding to the categorical variable in the row and the column.

Expected frequencies These are frequencies that are computed assuming the null hypothesis is true. Thus, we expect these frequencies to take place in a contingency table when the null hypothesis is true.

F Distribution Statistically speaking, the F distribution is a ratio of two independent Chi-Squares and so it has a pair of degrees of freedom one for the numerator and another for the denominator. This is used in ANOVA for testing the equality of more than two population means.

Goodness of Fit It is a test using Chi- Square distribution to determine whether there are significant differences between the observed frequencies and expected frequencies involving categorical data.

Observed Frequencies Frequencies that are actually observed in survey research both for one way table, as well as for contingency table involving nominal (categorical data).

Test of Independence This is a test that is used for analyzing cross-tabulation data in a contingency table in which an investigator is keen to find out whether the two categorical variables are independent of each other.

Treatment Treatment is a generic term that could represent a row or a column in an ANOVA table. Treatments could be different methods, different machines, different markets, and the like.

REVIEW QUESTIONS

Mini Case

A marketing manager of a company selling consumer non-durables was interested in finding out whether the sales of a particular product were influenced by color and size of package.

He performed an experiment in which he generated the likely sales that the company would get by the different colors and sizes of the package. The experiment lasted one month in selected retail outlets. The following data emerged from the study in which the response variable is the likely sales (in 1000units) that the company could get per month.

<i>Color \ Size</i>	<i>Small</i>	<i>Medium</i>	<i>Large</i>
Blue	30	29	41
Red	20	21	30
Yellow	36	38	46
White	26	25	40

- The Total Sum of Squares = -----.
- Sum of Squares due to Color = -----.
- Sum of Squares due to Size = -----.
- Calculated F for testing the equality of mean sales due to color = -----.
- Calculated F for testing the equality of mean sales due to size = -----.
- The null hypothesis of equality of mean sales due to color is rejected. True or False.
- The null hypothesis of equality of mean sales due to size is accepted. True or False.
- Which one of the following cannot be true of the critical value of chi-square?
 - 3.84
 - 23.364
 - 21.89
 - 18.493
- For testing the independence of attributes in a contingency table, the correct test procedure is to:
 - Use Z test
 - Use ANOVA
 - Use t test
 - Use χ^2 test
- In a cross-tabulation (contingency table) of 5 rows and 4 columns, the χ^2 test of independence is used. How many degrees of freedom this χ^2 value will have?
 - 20
 - 8
 - 12

ANSWERS TO REVIEW QUESTIONS

Solution to Mini Case:

The ANOVA output for this problem from Microsoft Excel is as follows. All the questions in this problem can be answered just by looking at the output.

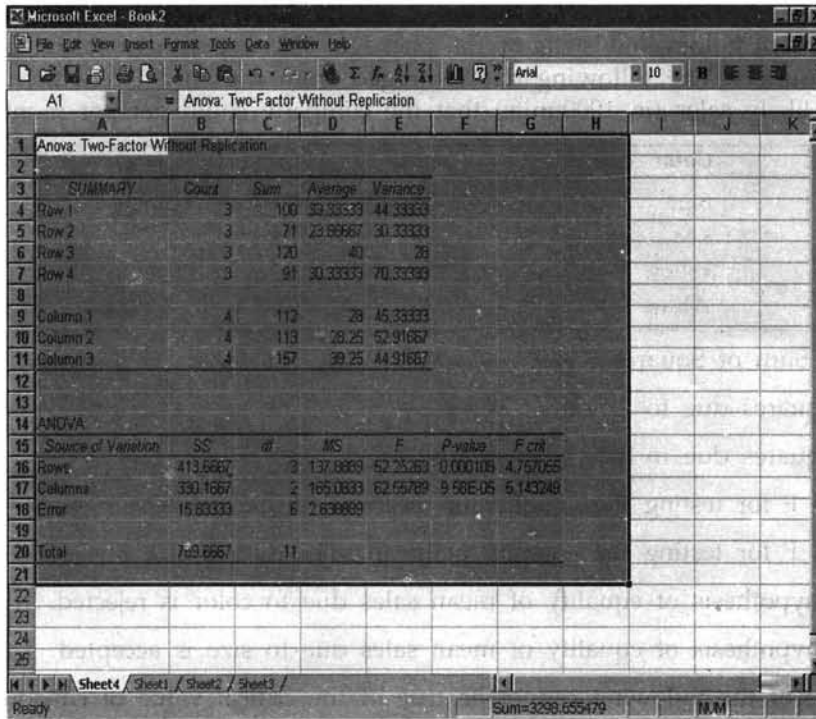


Figure 9.19

1. The Total Sum of Squares = 759.6667
2. Sum of Squares due to Color = 413.6667(Rows are colors)
3. Sum of Squares due to Size = 330.1667(columns are sizes)
4. Calculated F for testing the equality of mean sales due to color = 52.25263
5. Calculated F for testing the equality of mean sales due to size = 62.55789
6. The null hypothesis of equality of mean sales due to color is rejected. True. This is because the calculated F value (52.25263) is greater than critical F value (4.757055).
7. The null hypothesis of equality of mean sales due to size is accepted. False. The calculated F value (62.55789) is greater than critical F value (5.143249). Hence, reject the null hypothesis.
8. **Answer** (c) is the right choice. χ^2 value can never be negative.
9. **Answer** (d) is the right choice. (a) is incorrect because categorical data are involved here and a nonparametric situation exists. (b) is incorrect because ANOVA is used for testing the equality of population means. (c) is incorrect because it is a parametric test used to test the difference in two population means.
10. **Answer** (c) is the right choice. The number of degrees of freedom = (number of rows - 1) multiplied by (number of columns - 1) = (5 - 1) (4 - 1) = 12.

PRACTICE PROBLEMS

1. A refrigerator manufacturer is interested in assessing how many different colors of the refrigerator are to be produced in the coming quarter. Past data indicate that blue, gray, pink, and white are fast moving colors and they are moving in the market in the ratio of 25:35:20:20. The number of refrigerators sold as on date is 4000 in which 1200 are blue, 1300 are gray, 1000 are pink, and 500 are white. The manager would like to find out whether the pattern of sales have changed compared to the ratio of 25:35:20:20. Perform the χ^2 test of goodness of fit and draw your conclusions.
2. **Case Study- Ointment to Treat Fungus Problem on Human Skin**

A large and reputed hospital conducted a test on five leading brands of skin ointment that is used for treating fungus formation on the human skin. Particularly in humid summer, many people suffer from problem of fungus on skin. It causes irritation and gives itching sensation.

The test applied to these five leading brands involves 50 patients who prefer to use traditional herbal-based cream for the fungus problem. The fifty patients were divided into five groups each consisting of 10 patients. One brand of ointment was given to one group and all the five groups had received different brands of ointment. The brands were allotted to these groups in a randomized fashion. The patients were advised to apply the ointment for a period of 2 weeks. At the end of the two weeks, the effectiveness index that measures the ability of the ointment to get rid of the fungus problem was measured. The data matrix containing the effectiveness index is given below:

Effectiveness Index for Five Leading Brands of Ointment

Brand 1	Brand 2	Brand 3	Brand 4	Brand 5
92	118	109	110	111
97	121	95	98	112
90	112	98	104	107
94	116	98	101	110
93	115	102	104	115
95	113	100	106	109
90	111	99	102	111
91	115	102	104	113
98	117	105	106	115
93	111	98	100	110

Perform the one-way ANOVA to test the hypothesis whether the mean effective indices are equal for all the five brands. Interpret the results.

3. **Case Study-Do Color and Size of Package Design Boost the Sales?**

The marketing manager of a consumer product company wanted to know whether it is worth investing money and efforts in designing different sizes of package design with different colors. He was wondering if the factors color and size of package could

enhance the sales significantly. He performed the following experiment. The data matrix containing the response variable in 1000 Rs is given below:

<i>Color</i>	<i>Size of Package</i>		
	<i>Large</i>	<i>Medium</i>	<i>Small</i>
Blue	73	96	116
Red	77	98	118
Pink	98	111	149

Perform the two-way ANOVA and test whether the mean sales are influenced by package size and color. What are your findings?

4. Case Study-Comparison of Life of Different Brands of Tire

A marketing manager of company producing tires was interested in knowing the comparative picture of the average life of various brands of tire. An experiment was carried out in which the life of 5 brands of tire was estimated by actually running the cars for 10000 kms. The experiment was done in four cities to take care of the road conditions. The following data were collected.

Tire life in 1000 km

<i>City / Brand</i>	<i>Brand 1</i>	<i>Brand 2</i>	<i>Brand 3</i>	<i>Brand 4</i>
City 1	40	39	51	45
City 2	30	31	40	48
City 3	46	48	56	50
City 4	36	35	50	55

Perform two-way ANOVA and test whether significant differences in tire life exist among cities and among brands. What are your findings?

Correlation and Regression

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Define Correlation Coefficient with its properties
- Calculate Correlation Coefficient and Interpret
- Appreciate the role of Regression
- Formulate the Regression Equation and use it for estimation and prediction

CHAPTER OUTLINE

- 10.1 What is Correlation?
 - 10.2 Insights into Correlation
 - 10.3 Basics of Regression
 - 10.4 Regression Model
 - 10.5 Chapter Summary
- Glossary
Review Questions
Answers to Review Questions
Practice Problems

INTRODUCTION

Managers very often have to assess the nature and degree of relationship between variables. For example, a marketing manager would like to know the degree of relationship between advertising expenditure and the sales volume. Normally, you expect a positive relationship between sales and advertising expenditure. The manager would like to know whether money spent on advertising is justified in terms of sales generated; flat 10 percent increase in advertisement expenditure will result in how much extra sales volume? This type of question could be answered by Correlation and Regression. This Chapter covers the nitty-gritty of correlation and regression.

Correlation between Advertising Outlay and \$ Sales

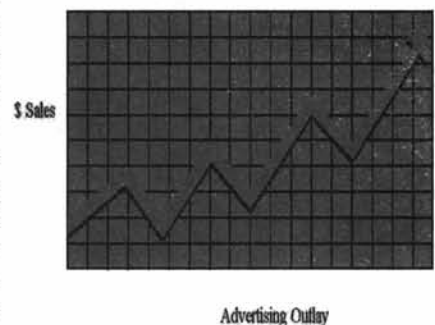


Figure 10.1

10.1 WHAT IS CORRELATION?

The manager of the business environment of today is very often interested in finding out whether there is any association between two or more variables and if it is true, he would like to know the strength of relationship between the variables. The strength of relationship is also known as the degree of relationship. In the previous Chapter, we have provided a conceptual framework of the Chi Square distribution that does try to provide some answer to the question of finding out whether the two attributes in a contingency table or associated or not. The degree of relationship between two variables can be elegantly worked out by correlation coefficient when the variables are intervally scaled.

What is the correlation between demand and price of a product? For all normal commodities we know that when price increases, the demand decreases and when price decreases, the demand increases. Economists call this inverse relationship between demand and price, as the price elasticity of demand. So, logically speaking the correlation coefficient between demand and price must be negative.

Let's see another example. As a manager of marketing if you are interested in finding out the degree of relationship between the advertising budget and the sales generated, the correlation coefficient can help you very much. You expect the correlation coefficient between the sales and the advertising budget will be positive.

In a number of situations where managers are interested in establishing the strength of relationship between variables, correlation coefficient can help them a great deal. Examples include the correlation between motivation of workers and productivity, customer satisfaction and profitability, percentage defectives and cost, and preventive maintenance vs. machine breakdowns.

10.2 INSIGHTS INTO CORRELATION

Scatter diagram In order to understand the nature of correlation, scatter diagram is the appropriate tool to use. Scatter diagram enables the investigator to know whether there is positive correlation, negative correlation or no correlation in a problem situation. In simple terms, scatter diagram depicts in graphical manner any two variables plotted in the X-Y plane. The following visual outlines the purpose of scatter diagram.

Scatter Diagram

Purpose:

To understand the relationship between two variables and the degree of relationship indicating the presence of positive correlation, negative correlation, or no correlation between the variables under study.

Figure 10.2

Positive Correlation As the value of one variable increases, the value of the other variable also increases. For example, you normally expect a positive correlation between **advertising** and **sales**. As you increase the amount spent on advertising, the sales volume will also increase.

Another example is the correlation between customer satisfaction and company profitability. Other things remaining same, you expect this correlation to be positive.

The following visual captures the essence of positive correlation.

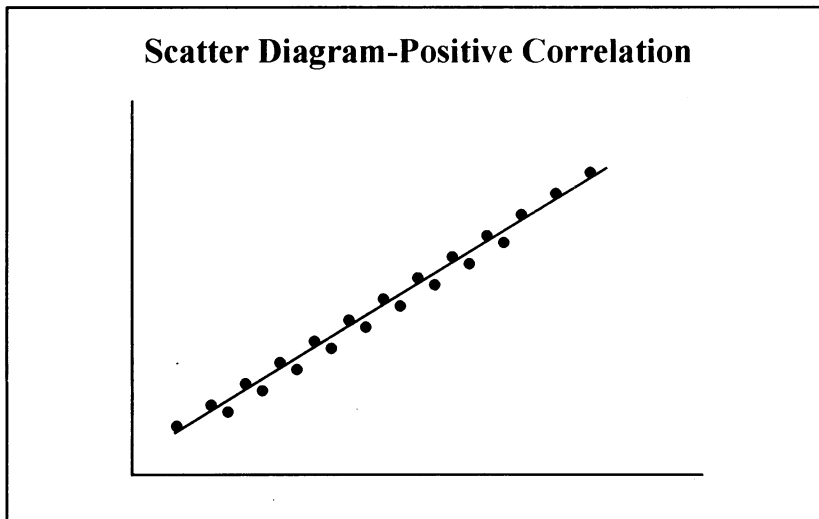


Figure 10.3

Please note that all the points are not falling exactly on the line in the above diagram. Yet, it shows a high degree of positive correlation. When all the points fall exactly on the line, it is called a **perfect positive correlation**. Very rarely, you will come across situations in which the *perfect positive correlation* is witnessed. Incidentally, the straight line that is superimposed in the picture is called the *line of best fit*. You will see the importance of this line in regression topic.

Negative Correlation As the value of one variable increases, the value of the other variable decreases. For example, the correlation between demand and price is negative for all normal commodities. The economists say the price elasticity is negative, meaning the relationship between demand and price is negative.

Another example is in the context of measuring productivity in a factory. If you want to correlate productivity with absenteeism of workers, you will find this relationship to be negative. More is the absenteeism, less is the productivity and vice-versa. Thus, productivity versus absenteeism will show a negative correlation. If all the sample points fall exactly on the line of best fit, it is a case of *perfect negative correlation*.

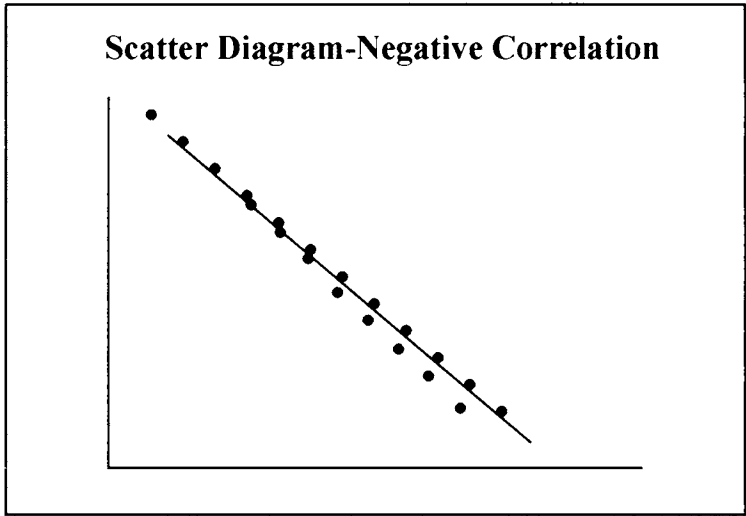


Figure 10.4

No Correlation At times, we may not be able to find any correlation pattern. It may be a case of absence of correlation. We say that no linear correlation is observed. This is because the correlation coefficient that we apply in practice is based on a linear relationship. The linear correlation coefficient was developed by Karl Pearson. See visual below. As you can see, no useful relationship can be established and it points to the fact that no correlation in the linear sense exists.

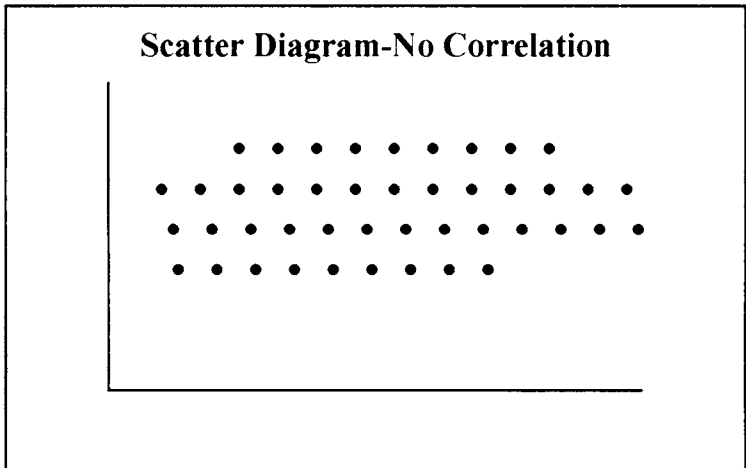


Figure 10.5

Pearson's correlation coefficient For a sample of n observations selected on two variables X and Y , the sample correlation coefficient of Karl Pearson is defined as follows:

Pearson's Correlation Coefficient:

The Pearson Correlation coefficient measures the degree to which there is a linear association between two intervally scaled variables. Correlation could be positive, negative or zero. The correlation coefficient is always between -1 and $+1$.

Figure 10.6

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}. \text{ This is also known as Product Moment Correlation.}$$

Here, r represents the sample correlation coefficient.

Properties of Correlation Coefficient

- The correlation coefficient is a pure number independent of unit of measurement and scale. The value of r will not change if X and Y are converted into U and V by transformation of scale.
- The correlation coefficient always lies between -1 and $+1$
- The three extreme positional values of r are shown below:

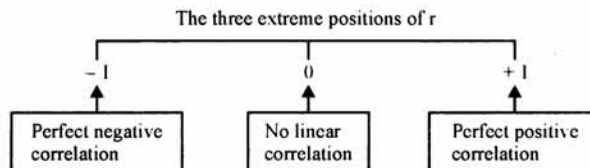


Figure 10.7

Example The following data refer to two variables-promotional expenses (Rs. Lakhs) and sales (1000 units) collected in the context of a promotional study. Calculate the correlation coefficient and comment.

Promotional Expenses	Sales
7	12
10	14
9	13
4	5
11	15
5	7
3	4

The calculations are shown below using the spreadsheet.

Promotional Expenses (X)	Sales (Y)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
7	12	0	2	0	0	4
10	14	3	4	12	9	16
9	13	2	3	6	4	9
4	5	-3	-5	15	9	25
11	15	4	5	20	16	25
5	7	-2	-3	6	4	9
3	4	-4	-6	24	16	36
7	10			83	58	124

In the spreadsheet calculations shown above, in the first two columns, the numbers 7 and 10 in the bottom row are the mean of X and Y. That is $\bar{X} = 7$ and $\bar{Y} = 10$. Likewise, $\sum(X - \bar{X})(Y - \bar{Y}) = 83$, $\sum(X - \bar{X})^2 = 58$ and $\sum(Y - \bar{Y})^2 = 124$.

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}} = \frac{83}{\sqrt{(58)(124)}} = 0.9787.$$

Comments The promotional expense is strongly associated with sales and the correlation is very close to 1.

Now, we will demonstrate how easy it is to do this exercise on Microsoft Excel. No hassles that involve complex formulas and memorizing expressions. Just follow the step-by-step procedure given.

Step 1 Enter the data in the spreadsheet. Then click Paste Function, click Statistical, and click CORREL. The following will appear:

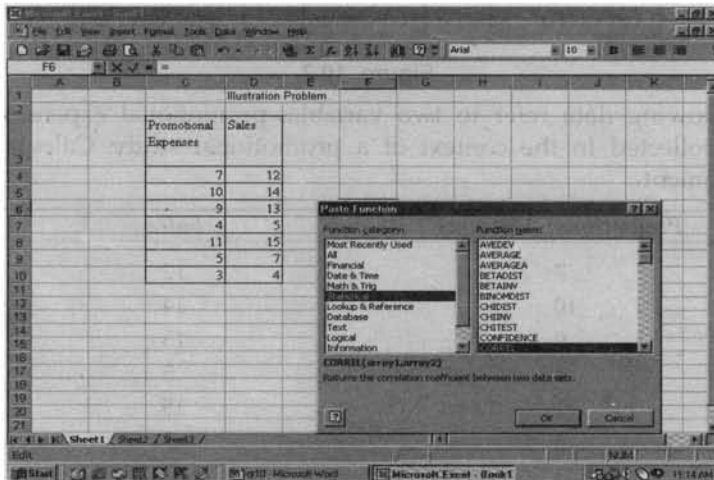


Figure 10.8

Step 2 Click OK. You will get:

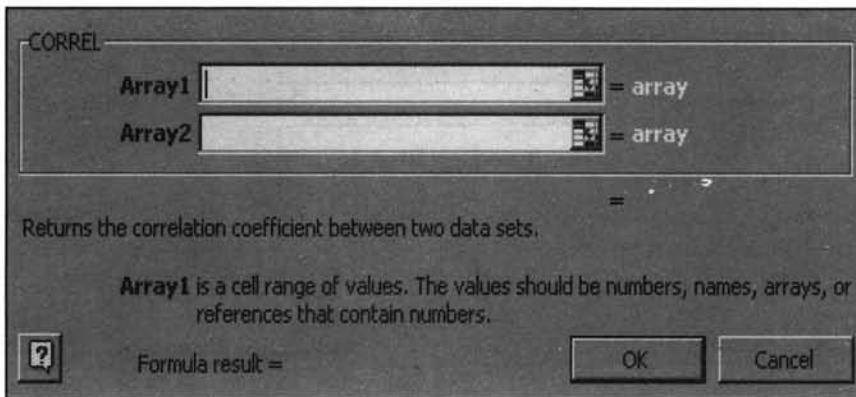


Figure 10.9

Highlight the data for array 1 and array 2 using the mouse. Then click OK. You get the answer. This screen is shown below:

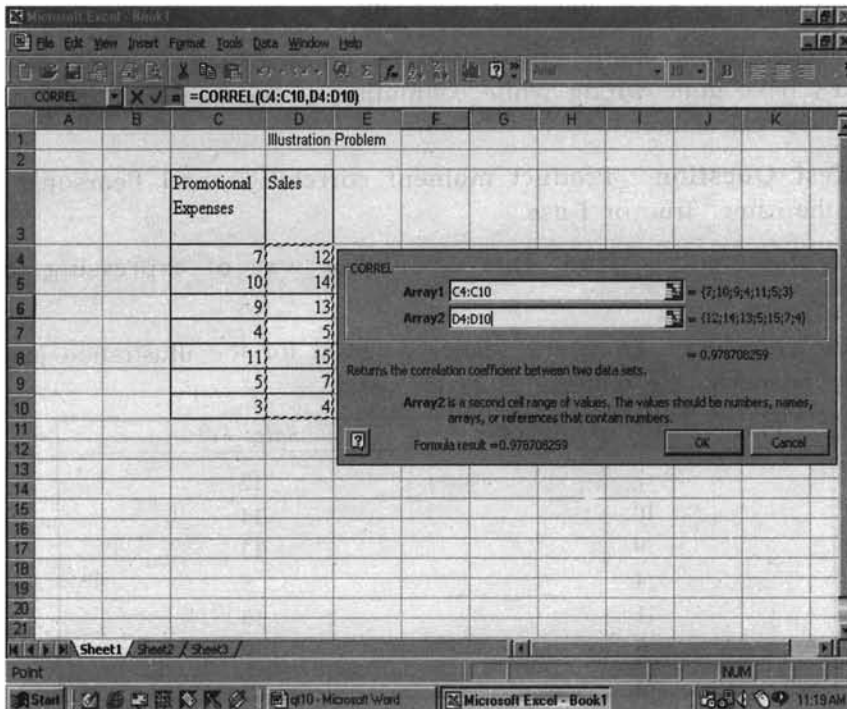


Figure 10.10

In order to see the computed correlation sharply, the output screen is exclusively shown below:

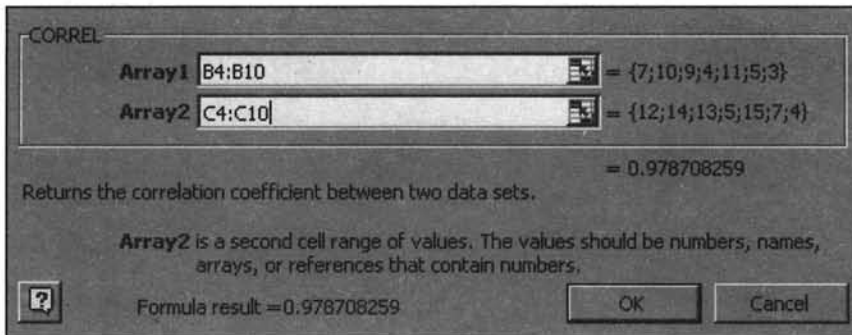


Figure 10.11

Comments The correlation coefficient is 0.9787(see Excel output above). The promotional expense is strongly associated with sales and the correlation is very close to 1.

Progressive Test Question In a market research study, the correlation coefficient between income and consumption is 2.8. Interpret the result.

Answer The correlation coefficient cannot be greater than 1. How can it be 2.8? Something must have gone wrong while computing r . The analyst must find out what went wrong in the calculation.

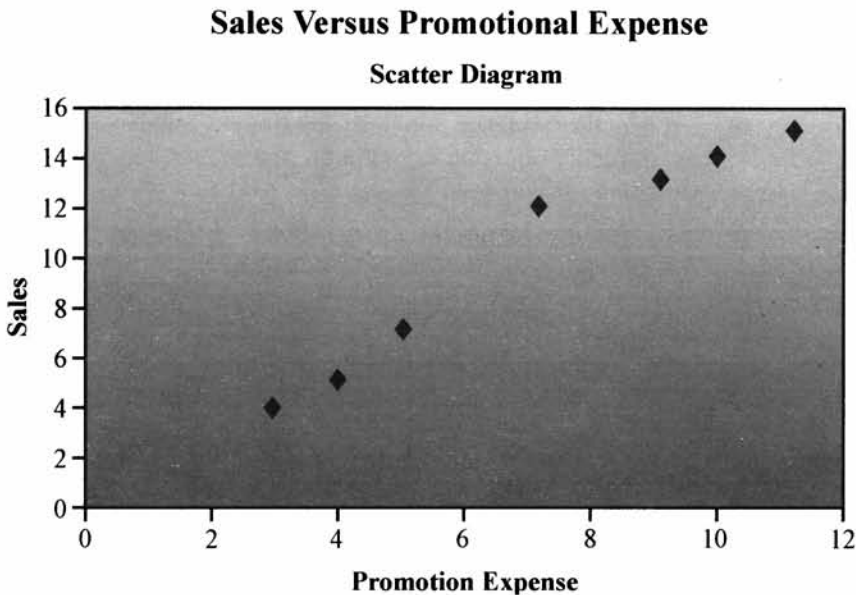
Progressive Test Question Product moment correlation and Pearson's correlation coefficient are the same. True or False.

Answer The statement is True. This is another way of expressing the Pearson Correlation.

Progressive Test Question Draw the scatter diagram for the illustration just discussed. For your ready reference, the data set is given again.

Promotional Expenses (X)	Sales (Y)
7	12
10	14
9	13
4	5
11	15
5	7
3	4

Solution We can use Microsoft Excel to draw the scatter diagram. Please brush up again the Chart wizard.

Scatter diagram: Output from Excel**Figure 10.12**

The scatter diagram clearly brings into sharp focus the high degree of positive correlation between promotional expense and sales. The pattern shows a linear correlation of a very high magnitude. Just to refresh your memory, for drawing the scatter diagram, click **Chart Wizard**, click **XY scatter** under chart type, click Next, enter the data in the Data Range cell reference by highlighting with the mouse the data in the spreadsheet, then click Next. Enter under the **Chart Options**, Title for the X-axis and Y-axis, and then click to **Finish**. You have the scatter diagram ready. Fine-tune and embellish the same according to your taste.

10.3 BASICS OF REGRESSION**Need for Regression**

The Pearson's correlation coefficient gives you just the degree of relationship or association. It cannot help you estimate or predict the response variable for a given independent variable. The response variable is called the dependent variable. In the example we have taken for the correlation coefficient, 'promotional expense' is the independent variable and 'sales' is the dependent variable. Sales depend on promotional expense. Using regression analysis, it is possible to predict sales for a given promotion expense. For business planning and forecasting, regression is much more useful than correlation.

The following visual gives succinctly the objectives of regression.

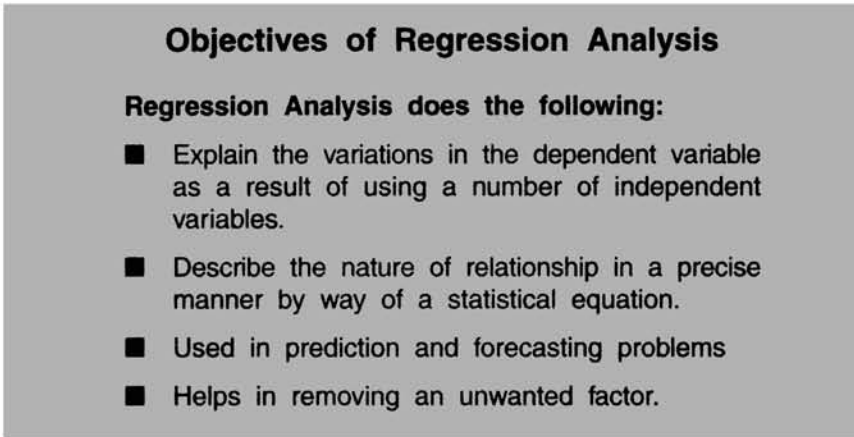


Figure 10.13

10.4 REGRESSION MODEL

Simple Linear Regression Model In this model, dependent variable is a linear function of one independent variable. For example, demand may be structured as a linear function of price. Based on sample data collected for the dependent and independent variable, a model is postulated connecting the dependent variable with the independent variable in a linear equation form. Symbolically, we write the sample regression line as follows:

$$Y = a + bX$$

where

Y is the dependent variable

X is the independent variable

a and b are constants.

a and b are determined by statistical least square method. b is called the regression coefficient(slope), and a is the constant term (intercept).

Historical Perspective

Just for knowledge sake, it is worth pointing out here that the estimates for a and b obtained by least square method are called 'Best Linear Unbiased Estimates' (BLUE), first pioneered by Gauss and Markoff in the context of General Linear Models that take care of Multiple Linear Regression as well.

Values of a and b in the case of simple linear regression model The values of a and b are obtained by solving the normal equations that are given below:

$$\sum Y = na + b \sum X$$

$$\sum YX = a \sum X + b \sum X^2$$

Here, Y is the dependent variable, X is the independent variable, and n is the sample size.

Solving these two normal equations,

You will find:

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

Multiple Linear Regression Model Whenever we are interested in the combined influence of several independent variables upon one dependent variable, our model is that of multiple regression. Demand, for example, may be a function of price, income of the consumer, advertising expense, industrial growth, and competitor's price. When all these independent variables change, what happens to the demand is a study of multiple linear regression.

We will first understand how to set up the simple linear regression model and interpret the results before moving on to multiple regression model. This is because simple linear regression lays the foundation for multiple regression analysis.

How does Simple Linear Regression work in practice?

To understand the nitty-gritty of simple regression, let us take the same example for which we have worked out the correlation coefficient.

Example The following data refer to two variables-promotional expenses(Rs. Lakhs) and sales(1000 units) collected in the context of a promotional study. Set up the simple linear regression model and predict sales when promotional expense is Rs.13 lakhs.

Promotional Expenses	Sales
7	12
10	14
9	13
4	5
11	15
5	7
3	4

Solution The basic calculations are shown in the spreadsheet below:

Promotional Expenses (X)	Sales (Y)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
7	12	0	2	0	0	4
10	14	3	4	12	9	16
9	13	2	3	6	4	9
4	5	-3	-5	15	9	25
11	15	4	5	20	16	25
5	7	-2	-3	6	4	9
3	4	-4	-6	24	16	36
7	10			83	58	124

In the spreadsheet calculations shown above, in the first two columns, the numbers 7 and 10 in the bottom row are the mean of X and Y. That is $\bar{X} = 7$ and $\bar{Y} = 10$. In the 5th, 6th, and 7th columns, you find $\sum (X - \bar{X})(Y - \bar{Y}) = 83$, $\sum (X - \bar{X})^2 = 58$ and $\sum (Y - \bar{Y})^2 = 124$.

You postulate the model in the standard form as follows:

$$Y = a + bX$$

where

Y is the dependent variable

X is the independent variable

a and b are constants.

As already worked out by solving the two normal equations,

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = (83/58) = 1.4310$$

$$a = \bar{Y} - b\bar{X} = 10 - 1.4310(7) = -.017$$

So the fitted equation is:

$$Y = -0.017 + 1.4310X. \text{ This is the line of best fit.}$$

To predict the sales when promotional expense = 13, put $X = 13$ in the fitted equation, you will get the answer. $Y = -0.017 + 1.4310(13) = 18.59$. The estimated sales when promotional expense is Rs. 13 lakhs is = 18.59(1000) units = 18590.

The concept of Coefficient of Determination for Statistical Validity

R^2 is called the **coefficient of determination**. This gives the contribution made by regression in explaining the variations in the dependent variable. This is worked out as a ratio between the regression sum of square and the total sum of square. In other words, R^2 measures the percent variation in the dependent variable as explained by the independent variable. Closer the value of R^2 to 1, greater is the veracity of the model. To calculate, you need the following terms.

$$\text{Regression Sum of Squares} = \sum (Y_e - \bar{Y})^2$$

$$\text{Error Sum of Squares} = \sum (Y - Y_e)^2$$

$$\text{Total Sum of Squares} = \sum (Y - \bar{Y})^2$$

Where Y_e is the estimated value of Y for a given X . This is obtained from the fitted line of regression.

Please note:

$$\text{Total Sum of Squares} = \text{Regression Sum of Squares} + \text{Error Sum of Squares}$$

For our problem, the basic calculations in the context of finding R^2 is given below using the spreadsheet.

Promotional Expenses (X)	Sales (Y)	Y_e	$(Y_e - \bar{Y})^2$	$(Y - Y_e)^2$	$(Y - \bar{Y})^2$
7	12	10.00	0.00	4.00	4.00
10	14	14.29	18.43	0.09	16.00
9	13	12.86	8.19	0.02	9.00
4	5	5.71	18.43	0.50	25.00
11	15	15.72	32.77	0.52	25.00
5	7	7.14	8.19	0.02	9.00
3	4	4.28	32.77	0.08	36.00
	10		118.78	5.22	124.00

From the spreadsheet above,

$$\text{Total Sum of Squares} = \sum (Y - \bar{Y})^2 = 124.00$$

$$\text{Regression Sum of Squares} = \sum (Y_e - \bar{Y})^2 = 118.78$$

$$\text{Error Sum of Squares} = \sum (Y - Y_e)^2 = 5.22$$

$$R^2 = (\text{Regression Sum of squares} / \text{Total Sum of Squares}) = (118.72/124) = 0.9579$$

The interpretation is 95.79% of the variations in sales is explained by promotional expense and only about 4.21% is explained by the error or residual term. So, the model fitted is fairly accurate.

Regression with Pleasure from Microsoft Excel

We are going to use Excel for solving this problem. We will avoid all derivations and mathematical expressions that must have taxed you very much. When you have Excel software, why bother? Follow the steps meticulously.

Step 1 Enter the given Data in the spreadsheet. Click Tools and then click Data Analysis.

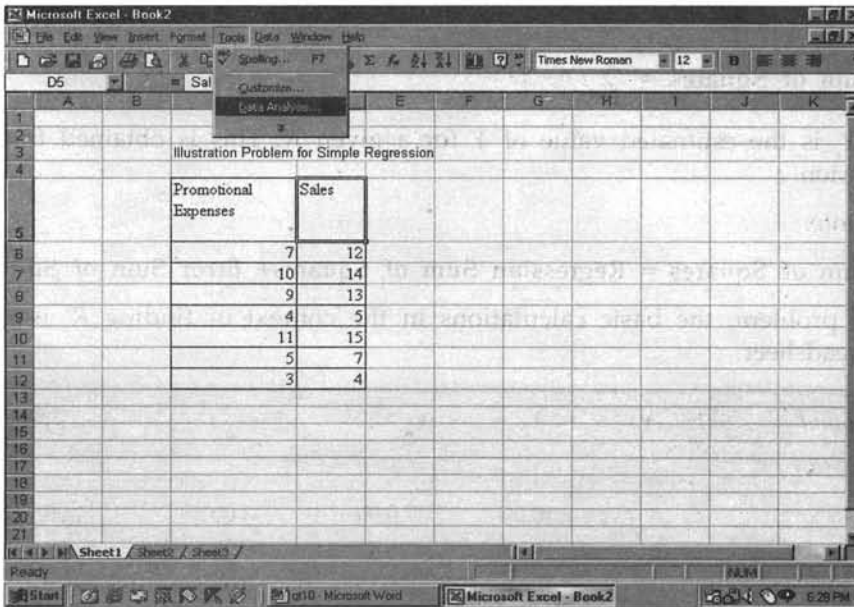


Figure 10.14

Step 2 In Data Analysis, highlight Regression with the mouse.

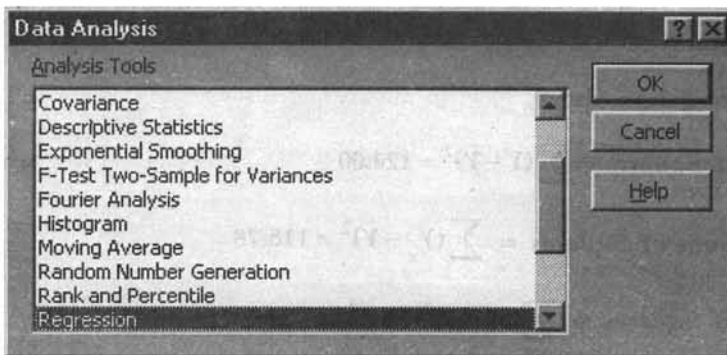


Figure 10.15

- Step 3** Click OK, and then enter the input range for Y and X in the following screen that will appear. Use the mouse to highlight the data you have for X and Y in the spreadsheet.

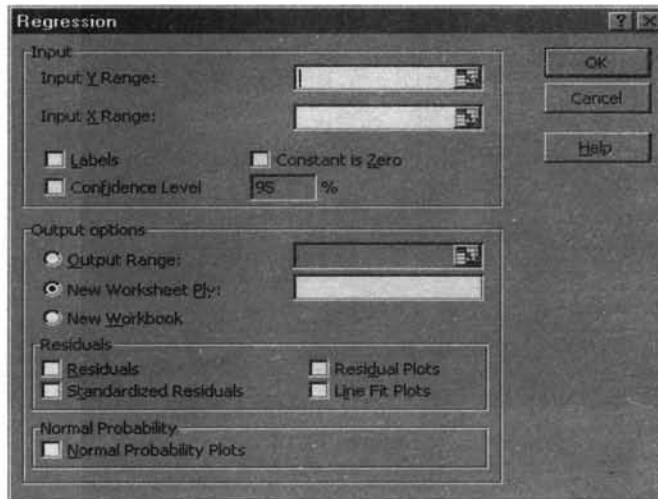


Figure 10.16

After highlighting the data for Input Y Range and Input X range in the above screen, you get:

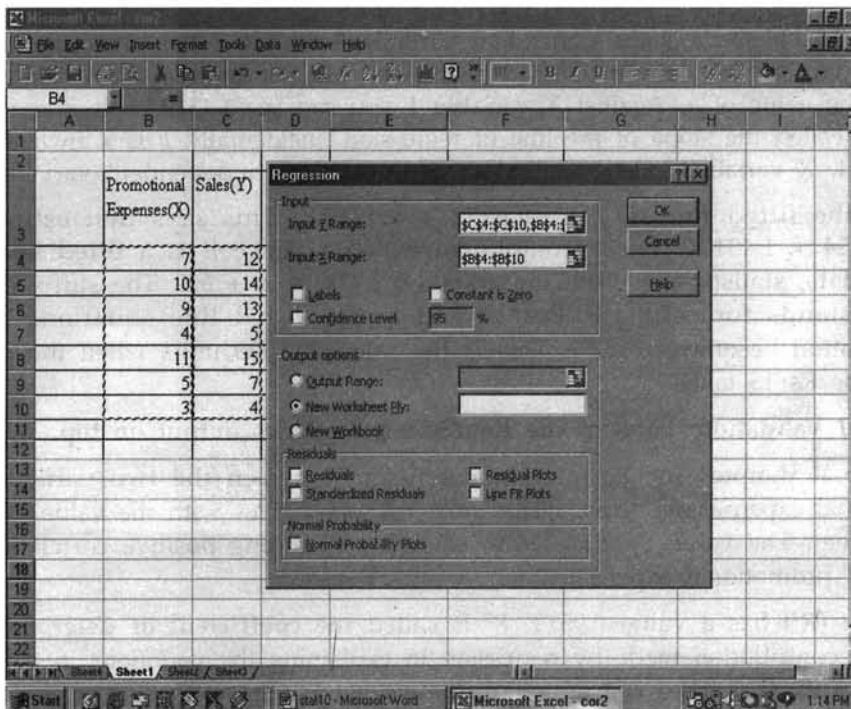


Figure 10.17

Step 4 Click OK. You now get the regression results for the given problem.

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.9787								
R Square	0.9579								
Adjusted R Square	0.9494								
Standard Error	1.0222								
Observations	7.0000								
ANOVA									
		df	SS	MS	F	Significance F			
Regression		1	118.7759	118.7759	113.6799	0.0001			
Residual		5	5.2241	1.0448					
Total		6	124.0000						
		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept		-0.0172	1.0159	-0.0170	0.9871	-2.6286	2.5941	-2.6286	2.5941
X Variable 1		1.4310	0.1342	10.6621	0.0001	1.0960	1.7761	1.0960	1.7761

Figure 10.18

Explanation on the Output

The output of Excel for the regression analysis is very detailed giving perhaps more than what we actually need. Let us focus on the important ones.

1. Regression Equation:

$Y = a + bX$. Here, Y is the dependent variable (sales). a = the intercept value seen in the last part. Under column labeled coefficient, against **Intercept** you can see **-0.0172**. This is the value of a . Against **X variable 1** you can see **1.4310**. This is the value of b . b is also called the slope of the line of regression. Incidentally, b is known as regression coefficient. X variable 1 denotes that there is only one independent variable.

Hence, the fitted line is $Y = -0.0172 + 1.4310X$. This says that estimated sales = $-0.01724 + 1.43104$ times promotional expense. Since it is a fitted line based on sample data, statisticians write the equation as $Y_e = a + bX$. The suffix e associated with Y stands for estimate. Put $x = 13$ in the fitted line equation. Y value on simplification becomes = 18.59. This is the sales in 1000 units when the promotional expense is Rs 13 lakhs

2. Statistical Validation: Look at the Regression Statistics output on top.

Multiple R denotes the correlation coefficient between the two variables, namely promotional expense and sales. Please note that this tallies with the value of r we have done earlier. The value of $R = 0.9787$ (shows very strong positive correlation between sales and promotional expense).

R Square (R^2) has a value 0.9579. R^2 is called the **coefficient of determination**. This gives the contribution made by regression in explaining the variations in the dependent variable. This is worked out as a ratio between the regression sum of square and the total sum of square. Closer the value of R^2 to 1, greater is the veracity of the model.

In our case $R^2 = 0.9579$. The interpretation is 95.79% of the variations in sales is explained by promotional expense and only about 4.21% is explained by the error or residual term. So, the model fitted is fairly accurate.

Adjusted R^2 : When more independent variables are added in the regression model, R^2 value will increase. It needs to be corrected to reflect the reality. This is achieved by

Adjusted R^2 . It is computed by the formula $(\text{Adjusted } R^2) = 1 - (1 - R^2) \frac{(n-1)}{(n-k)}$. n is the

number of observations and k is the number of constants in the regression equation. Here $n = 7$ and $k = 2$. If you substitute and simplify, you get the adjusted R^2 value = 0.9494 that is given in the output. Excel will take care of this. Don't worry.

Standard Error The standard error of the sample dependent variable is given by the square root of the mean square corresponding to the Residual term in the ANOVA table that just follows the Regression Statistics. The value of the standard error (1.0222) is given in Regression Statistics.

Item of interest on the ANOVA output

- **Regression Sum of Squares** = $\sum (Y_e - \bar{Y})^2 = 118.7759$
- **Residual Sum of Squares** = $\sum (Y - Y_e)^2 = 5.2241$ (same as Error Sum of Squares)
- **Total Sum of Squares** = $\sum (Y - \bar{Y})^2 = 124.0000$
- Mean Squares due to regression and error are worked out by dividing the sum of squares by the corresponding degrees of freedom. They are respectively = 118.7759, and 1.0448. Please see them in Excel ANOVA output.
- **F statistics** computed is nothing but the ratio between the mean squares of regression and residual. That is calculated $F = (118.7759/1.0448) = 113.6799$. This is also given in the Excel ANOVA output.

Null Hypothesis There is no linear relationship between Y and X in the population regression line

Alternative Hypothesis There is linear relationship between Y and X in the Population Regression Line

Look at the calculated F value. It is 113.61 and is greater than the critical $F(6.61)$. Reject the null hypothesis. Please note that you can get the critical F using FINV function of Excel. Here significance F value (**P-value**) is given to be 0.0001. It is less than the level of significance 0.05. Reject the null hypothesis. The conclusion is that sales is a linearly related to promotion expense.

Incidentally, you can construct ANOVA by using the calculator. All formulas in this regard are given to you. For getting the critical value of F at 5% level of significance, you please use the F distribution table in Appendix G. The hypothesis testing procedure is same regardless of which approach you prefer. But let me tell you once you use Excel, you may not like to use the calculator.

Progressive Test Question Calculate R^2 using the ANOVA table.

Solution $R^2 = (\text{Regression Sum of Squares})/\text{Total Sum of Square} = (118.7759)/124 = 0.95787$

Progressive Test Question For testing the hypothesis of the linear relationship in the above problem, what are the degrees of freedom of the F statistic?

Solution Look at the ANOVA output. The degrees of freedom for the regression is 1, and the error is 5. So the F ratio will have 1 degree of freedom for the numerator and 5 degrees of freedom for the denominator. The answer is $F(1,5)$.

Progressive Test Question What is the standard error of the distribution of sample sales (Y)?

Solution From the ANOVA, Residual Mean square(Error Mean Square) = 1.044828. Take the square root of this. That is the standard error required. The value is $\sqrt{(1.044828)} = 1.0222$. Please note that this value tallies with the value of standard error under Regression Statistics of Excel in the summary output.

Things to do in a Simple Linear Regression Model

- > Postulate the model $Y = a+bX$.
- > Enter the sample data for X and Y in Microsoft Excel.
- > Perform the Regression Analysis and get the summary output from Excel.
- > Write the Regression Equation using the intercept and coefficient of X from Excel summary output. Predict Y for a given X .
- > Validate the model statistically by looking at R^2 as well as F statistic in the ANOVA that tests the null hypothesis of no linear relationship.
- > After statistical validation, use the model for estimation and prediction.

Please note that you can do all the steps mentioned above using the calculator. Every formula and step in this regard are comprehensively explained.

Assumptions involved in a Regression Model

- > The sample observations drawn are independent and random.
- > For a given value of X , the conditional distribution of Y is normally distributed.
- > The population regression of Y on X is assumed to be linear.

Progressive Test Question What are the residual terms for the regression example discussed by us? Plot the residuals.

Promotional Expenses	Sales
7	12
10	14
9	13

Promotional Expenses	Sales
4	5
11	15
5	7
3	4

Solution In the options, click Residuals, and Residual Plots after entering the input range for Y and X. You get the following:

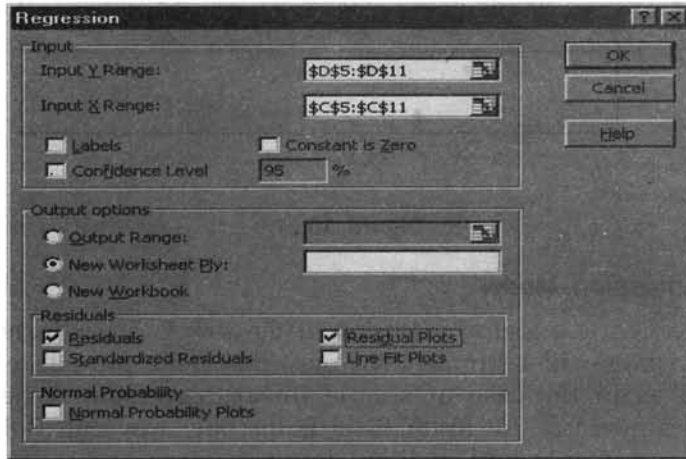


Figure 10.19

Click OK, and you get the residuals and the residuals plotted. Please note that residual = $Y - Y_e$ (Actual Value-Estimated value).

Observation	Predicted Y	Residuals
1	10	0
2	14.2931034	-0.2931034
3	12.3033333	2.6966667
4	6.7788955	0.2211045
5	15.7241579	-0.7241579
6	7.13793102	0.13793102
7	4.27666667	0.72333333

Figure 10.20

Solution continues

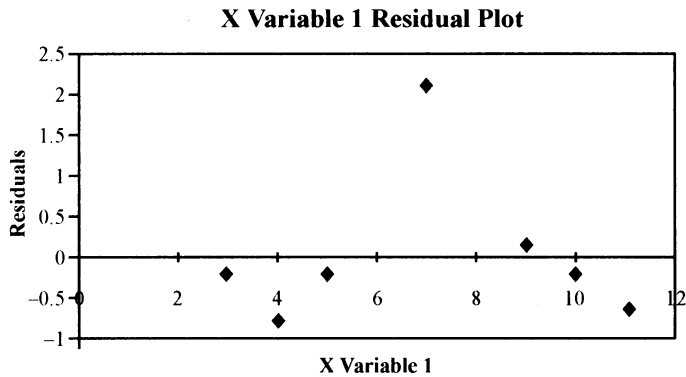


Figure 10.21

Multiple Linear Regression Model

Multiple linear regression is a logical extension of the simple linear regression. The number of independent variables will be more than one. The same procedure of setting up the model is followed as in the case of simple linear regression. When the number of independent variables increases, Microsoft Excel is the only way out. Doing the calculations using a calculator is not only very tedious, but also error prone. If you want to do a multiple regression model involving 10 independent variables using a calculator, you must be crazy! The best way to understand how multiple regression works in practice is through an example.

Example Eight patients underwent an operation in a hospital. Measurements of weight(kg), duration of operation(minutes), and blood loss(ml) were taken. The hospital authorities would like to know whether the blood loss was related to weight and duration of operation. The data are as follows:

<i>Weight(X1)</i>	<i>Duration of Operation (X2)</i>	<i>Blood Loss (Y)</i>
44	108	505
42	85	492
70	88	472
45	114	506
50	110	484
51	101	492
36	97	515
53	121	466

Questions:

1. Write the regression equation of blood loss in terms of weight and duration of operation
2. What is the R^2 value? Interpret.
3. Validate the model using ANOVA.

Solution Proceed exactly like the simple regression model in Microsoft Excel. Enter the Data, click Tools, click Data Analysis, click Regression. You get:

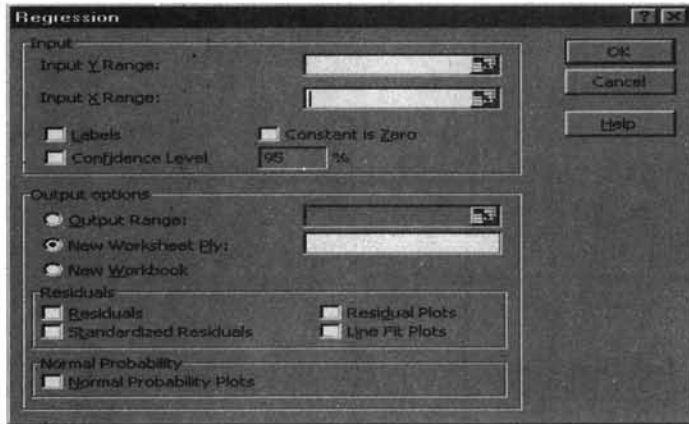


Figure 10.22

Solution continues Enter the input range for Y exactly as before. For input range for X, you need to highlight the two columns containing X1 and X2 without the label. That is all. You get now the following. See the two columns of X1 and X2 are highlighted to convey to Excel that it is multiple regression model.

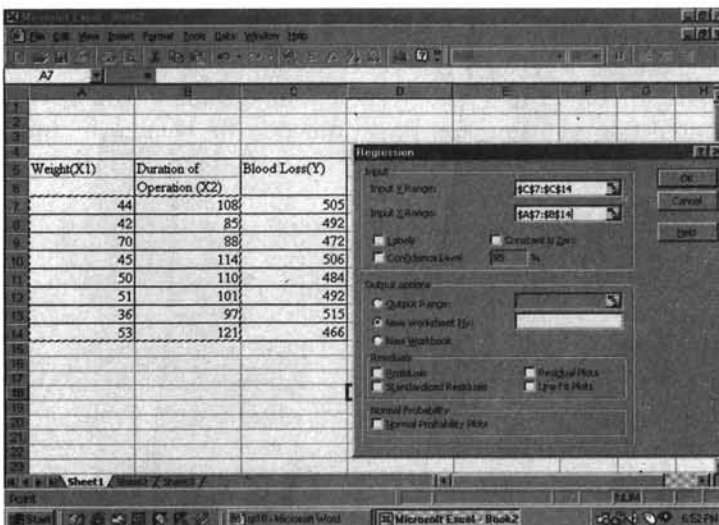


Figure 10.23

Solution continues Now click OK. You get the answer. The output appears as follows:

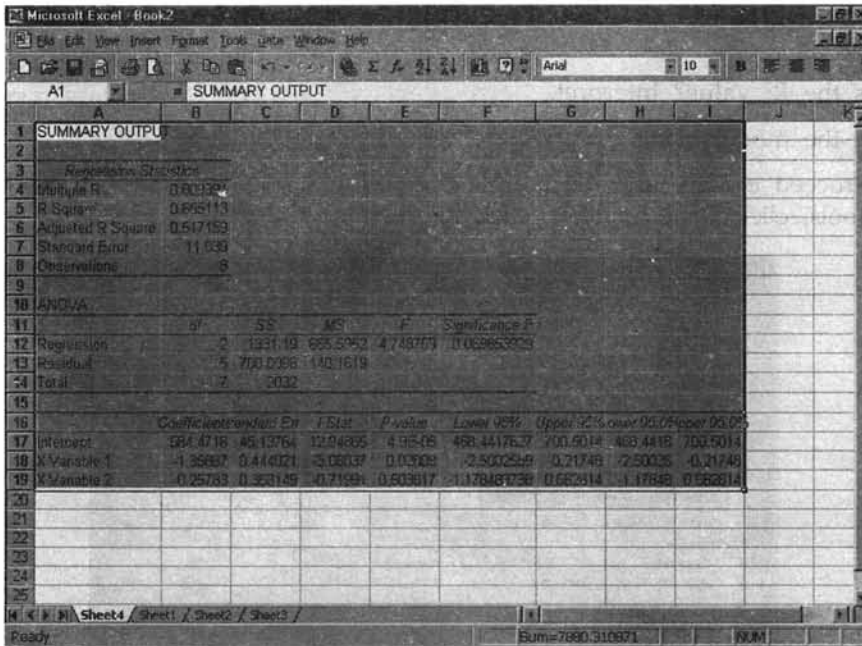


Figure 10.24

1. Regression equation is $Y = 584.4716 - 1.35887X_1 - 0.25783X_2$
2. From regression Statistics on top $R^2 = 0.6551$. This means that 65.51% of variations in blood loss is explained by weight and duration. About 35.49% are accounted by error. The R^2 value suggests that the model is not robust and more factors will have to be added. Let us see what ANOVA concludes.
3. In ANOVA, calculated F value is 4.75 and F significance is 0.0699 (P-value). Since the P value is more than the level of significance 0.05, accept the null hypothesis of no linear relationship between Blood loss and weight and duration. You get the same conclusion by working out F critical using the paste function or F table in Appendix G. Critical F for F(2,5) for 5% is = 5.79. Calculated F is less than critical F. So, accept the null hypothesis.

Limitations of Multiple Regression Model

- The most crucial assumption made is that the independent variables are not correlated with each other. If they are correlated, then the regression coefficients cannot be estimated. This problem is called multicollinearity. The procedure followed for resolving multicollinearity is to drop the independent variable that has the highest standard deviation and then rework the model again. You may also like to use two-stage least square method that is part of econometrics. The other way is to transform a set of correlated independent variables into an uncorrelated set of variables by the technique called principal component analysis. This is an advanced technique requiring the help of advanced statistical software, like SPSS.

- When there are wild fluctuations in one or more of the independent variables, multiple regression model crumbles and will be highly unreliable.
- In order to use the multiple regression model for prediction, you have to first predict the values of the independent variables using some other prediction method.
- In forecasting problems, multiple regression at best can work for short and medium term only. It cannot be successfully used for long term forecasting.

Discussion Topic

Analyze, criticize, and explain the following statement:

"Regression analysis is much more useful than correlation analysis in providing solution to business problems".

10.5 CHAPTER SUMMARY

This Chapter has introduced you to the essentials of correlation and regression with their important features. Specifically, this Chapter focused on:

- Meaning and role of correlation
- Understanding correlation through scatter diagram
- Definition and properties of Pearson's correlation coefficient
- How to compute correlation coefficient
- Basics of regression analysis
- Simple linear regression model and statistical validation
- Multiple linear regression model and statistical validation
- Extensive use of Microsoft Excel to get solution for correlation and regression

GLOSSARY

Coefficient of Determination This gives the contribution made by regression in explaining the variations in the dependent variable. Closer this value to 1, greater is the veracity of the linear model postulated.

Correlation It is a study that focuses on the strength of association or relationship between variables.

Correlation Coefficient It measures the degree to which two intervally-scaled variables are linearly associated. It is a pure number that lies in the interval -1 to $+1$. There could be zero correlation, positive correlation, or negative correlation.

Dependent Variable It is the response variable that we are estimating or predicting in a regression model. It is a function of one or more independent variables.

Independent Variables These are variables that are known and are found in the right side of the regression equation. The response variable is postulated as a function of

these independent variables. In other words, the behavior of the independent variables affect the value of the response (dependent) variable.

Least Square Method This is used to determine the best linear unbiased estimates of the regression coefficients in a regression model.

Multiple Linear Regression It is a study focusing on the combined influence of several independent variables upon one dependent variable using a linear model.

Regression Analysis It is a process of predicting the value of the response variable (dependent variable) that depends on one or more number of independent variables. Prediction is achieved through a statistical equation based on the least-square method.

Residual It is the difference between the actual value of the dependent variable and the predicted value of the dependent variable obtained from regression equation. This measures the backtracking ability of the model to predict.

Scatter diagram It is diagram that depicts in a graphical manner any two variables plotted in the X-Y plane. Scatter diagram enables the user understand the nature of correlation between two variables.

Simple Linear Regression In this model, dependent variable is a linear function of one independent variable.

Slope It is the value of the regression coefficient in the regression equation. It gives for every unit increase in its value, the resultant change in the dependent variable.

Y-Intercept It is the constant term in the regression equation. It is obtained by letting the values of the independent variable equal to zero.

REVIEW QUESTIONS

1. If increasing values of one variable are associated with decreasing values of another variable and vice versa, then the two variables are negatively correlated. True or False.
2. The residuals of a regression line are:
 - (a) Positive always
 - (b) The difference between actual Y values and estimated Y values
 - (c) Are all zero for the perfect fit

Questions 3 to 8 refer to the following data:

x	6	8	11	12	13
y	3	5	6	7	8

3. The slope of the regression line is -----.
4. The intercept of the regression line of y on x is -----.
5. The value of the correlation coefficient is -----.
6. The value of R^2 is -----.

7. Total sum of squares = -----.
8. Regression sum of squares = -----.

Questions 9 and 10 will have to be answered based on the following data.

In a multiple regression study involving three independent variables, the adjusted R^2 is 0.8650. The error sum of squares = 350. The number of observations = 10.

9. R^2 value is -----.
10. Regression sum of squares is -----.

ANSWERS TO REVIEW QUESTIONS

1. **Answer** The statement is true. This is the very property of negative correlation. See scatter diagram in the Chapter for better clarity.
2. **Answer** (b) and (c) are correct. (a) is incorrect because residuals can be negative also. (b) is correct because the residual is defined as the difference between actual Y and estimated Y ($Y - Y_e$). (c) is correct because when there is a perfect fit (perfect correlation), all sample points will exactly fall on the regression line. So, the residuals are zero.

Answers to questions 3 to 8 can be answered based on the following regression output of Excel. Alternatively, if you work out on your own using a calculator you will get the same output.

The screenshot shows an Excel window titled 'Microsoft Excel - Book2' with a 'SUMMARY OUTPUT' table. The table contains the following data:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.960737							
R Square	0.922914							
Adjusted R Square	0.949125							
Standard Error	0.433061							
Observations	5							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	14.23579	14.23579	75.525	0.00320719			
Residual	3	0.564206	0.188235					
Total	4	14.8						
Coefficients, Standard Error, t Stat, P-value, Lower 95%, Upper 95%, Lower 95.0%, Upper 95.0%								
Intercept	-0.67093	0.766948	-0.87229	0.447342	-3.11774701	1.776548	-3.11772	1.776548
X Variable 1	0.647068	0.074407	8.696284	0.0032	0.410253305	0.883804	0.410254	0.883804

Figure 10.25

Answers to questions 3 to 8.

3. The slope of the regression line is 0.647059
4. The intercept of the regression line of y on x is -0.67059
5. The value of the correlation coefficient is 0.980737
6. The value of R^2 is 0.961844
7. Total sum of squares = 14.8
8. Regression sum of squares = 14.23529
9. **Answer** Adjusted $R^2 = 1 - (1 - R^2) \left(\frac{n-1}{n-k} \right) = 0.8650$ (given)

$$1 - (1 - R^2) \left(\frac{10-1}{10-3} \right) = 0.8650$$

$$1 - 0.8650 = (1 - R^2) \left(\frac{9}{7} \right)$$

$$1 - R^2 = \left(\frac{7}{9} \right) (0.1350)$$

$$R^2 = 0.8950.$$

10. **Answer** If R^2 is 0.8950, then error sum of square as % to total sum of squares must be = 0.1050. Error sum of squares = 350. Therefore total sum of squares = $350 / (0.1050) = 3333.33$. Regression sum of squares + Error sum of squares = Total sum of squares. Hence, regression sum of squares = $3333.33 - 350 = 2983.33$

PRACTICE PROBLEMS

1. Fit a regression model using simple linear regression of Y on X for the following data. What is the R^2 value? What is the product moment correlation between X and Y ? Plot the actuals of Y along with the fitted regression line.

(X)	(Y)
14	24
20	30
18	26
8	10
22	30
10	14
6	8

2. **Case Study-Monthly Sales Forecast**

A supermarket that has a chain of 15 retail outlets in a city wants to predict the monthly sales of its entire operation based on the performance of all its retail outlets. The number of customers who visited during one typical month in all the outlets were recorded along with the money value purchase made by them. Based on the data that are given in the next page, answer the following:

- (a) Identify the dependent and independent variable in this forecast model.
- (b) Fit a regression model to the data.
- (c) Validate it statistically.
- (d) Comment on the reliability of the model to forecast.

Supermarket Data on its Retail Outlets		
<i>Retail Outlet</i>	<i>Consumers</i>	<i>Sales (LakhRs)</i>
1	1814	22.4
2	1852	22.1
3	1012	13.68
4	1482	18.42
5	1578	18.84
6	1778	20.16
7	1748	18.9
8	1020	13.46
9	1058	14.48
10	840	12.24
11	1358	15.26
12	1744	18.86
13	1848	18.92
14	1214	15.28
15	904	13.84

3. Case Study- Fuel Consumption for Car

A market research firm on behalf of a client who is planning to develop a new fuel-efficient car is studying the fuel consumption of a popular make of car when it runs with unleaded petrol. The initial diagnostic data were collected from 10 trips of same distance covered under similar road conditions using the same car. The hypothesis was that the fuel consumption in kms per liter was dependent on the average speed in kms per hour, and the total weight (including passengers and luggage loaded in the car measured in 1000 kg).

Initial diagnostic Data			
<i>Trip</i>	<i>Km per Liter</i>	<i>Average Speed (km)</i>	<i>Total Weight ('000 kg)</i>
1	14	50	2
2	12	40	3
3	15	45	2
4	16	55	1.5
5	12	35	3
6	16	60	1
7	13	55	5
8	14	55	2
9	13	40	2
10	11	30	3

QUESTIONS

- (a) Fit a multiple regression model by taking fuel consumption as the dependent variable and average speed and total weight as the independent variables.
- (b) Validate the model using R^2 and ANOVA table.
- (c) What are your final conclusions about the model fitted?
- (d) What other variables you think that should be included in the model for better prediction?

4. Case Study- Are Sales influenced by Sales Promotion and Advertising?

To study the impact of advertising and sales promotion expenditure on sales, the following data were collected from ten companies.

(Figures in Rs. Lakhs)

<i>Company</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>
Advertising expense	50	15	40	80	65	35	75	50	60	70
Sales Promotion expense	25	30	40	60	50	45	25	50	30	50
Sales	210	180	200	400	300	230	350	300	280	380

- (a) Set up and validate the multiple linear regression model for this problem. (Sales is the dependent variable, advertising and promotion expense as independent variables)
- (b) Find the expected sales if advertising expense is Rs. 40 lakhs and the promotion expense is Rs. 80 lakhs?
- (c) What are your reservations regarding the model fitted?

Decision Analysis

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Structure any decision problem
- Construct and use decision trees
- Calculate and use Expected Monetary Value (EMV) for decision making
- Assess and appreciate value of information
- Revise the decision using Posterior Probability Analysis

CHAPTER OUTLINE

- 11.1 Steps in Systematic Problem Solving
 - 11.2 How to Structure a Decision Problem
 - 11.3 Expected Monetary Value (EMV)
 - 11.4 Decision Tree
 - 11.5 Value of Sample Information
 - 11.6 Chapter Summary
- Glossary
Review Questions
Answers to Review Questions
Practice Problems

INTRODUCTION

Managers must be capable of making decisions under conditions of uncertainty. They must also be capable of using information, which may be inadequate for decision-making. The question that must be asked before making the final decision is, "is it worth gathering additional information?". This Chapter provides a framework for decision making with minimum risk under uncertain environment.



Figure 11.1

11.1 STEPS IN SYSTEMATIC PROBLEM SOLVING

Many decisions will have to be made very often with a little bit of incomplete data in an environment of uncertainty. It is in this context that the role of systematic problem solving looms large. The schematic diagram succinctly describes the steps involved in problem solving.

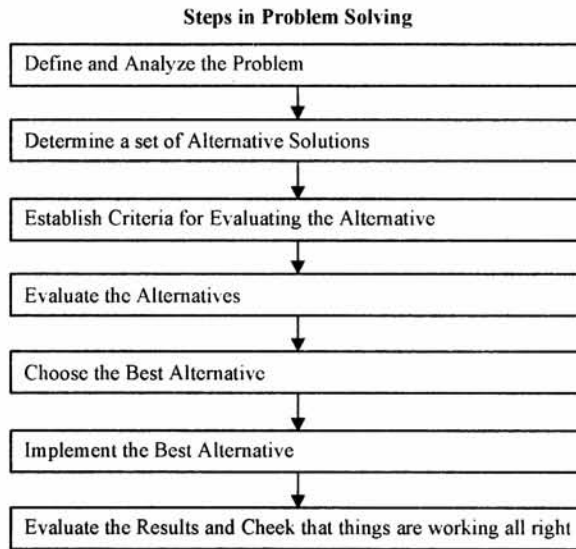


Figure 11.2

Let us understand these steps of systematic problem solving using a case study given below:

Case Study-Product Mix Decision

A manufacturer produces two decorative models (standard and deluxe) of chairs. In the light of forthcoming peak selling season, he must decide how many of each model to produce. Variable cost of the standard model is Rs. 500, and selling price is Rs. 1000; variable cost of the deluxe model is Rs. 1000, and selling price is Rs. 2000. Demand is uncertain and has the following discrete distribution:

<i>Standard Model</i>		<i>Deluxe Model</i>	
<i>Demand</i>	<i>Probability</i>	<i>Demand</i>	<i>Probability</i>
3000	0.40	1000	0.30
4000	0.60	2000	0.70

Further, it is known from market research that the probabilities of demand for these two models are independent. Production capacity is limited to a total of 5000 chairs. Chairs not sold during the peak season are disposed of at salvage prices of Rs. 250 for the standard model and Rs. 500 for the deluxe model.

Steps in Systematic Problem Solving Explained for the Case

For this case study, we will explain logically all the steps one by one. Please note that we are not going to provide here the actual solution that needs the help of decision analysis. This will be covered shortly in this chapter. Here, we will provide the basis for systematically solving a problem that is the foundation of management science.

Define and Analyze the Problem

The problem is to determine how many standard and deluxe model decorative chairs to be produced for the peak-selling season. It is a product mix decision problem involving the number of standard and deluxe model to be produced so as to maximize the overall financial contribution to the manufacturer. Primarily, the factors that affect this product mix decision are the demand pattern for the models that are uncertain, capacity of the plant that is restricted to a total of 5000 chairs (both models put together), and chairs not sold during the season. The probability distribution of demand for both the models are given. The present selling price and variable cost for the models can be assumed to remain constant during the short-term time horizon. The demand distributions for the models are independent.

Determine a Set of Alternative Solutions

The possible product mix combinations in numbers are given in the table below:

<i>Standard</i>	<i>Deluxe</i>	<i>Total</i>
3000	1000	4000
3000	2000	5000
4000	1000	5000
4000	2000	6000

Obviously, the last combination (4000, 2000) is ruled out because capacity of the plant is limited to 5000 numbers. So the alternatives are 3000 standard and 1000 deluxe, or 3000 standard and 2000 deluxe, or 4000 standard and 1000 deluxe.

Establish Criteria for Evaluating the Alternatives

The criterion here is naturally the financial contribution that the manufacturer will get for each alternative. Contribution per unit for each model could be worked out using the formula, contribution = selling price - variable cost. The details are given, and so you can work them out.

Evaluate the Alternatives

Evaluate each of the product mix using the criterion of contribution. You have to be careful while working out the financial contribution. You should subtract from the contribution obtained for any alternative, the adverse financial cost caused by the excess production if any that could not be sold during the peak season. For example, take the combination of producing 4000 standard and 1000 deluxe. Suppose the actual demand is 3000 standard and 1000 deluxe. There is no problem as far as deluxe model is concerned. But you have produced 4000 standard and the demand is only 3000. So, you have an excess production of

1000 numbers of standard. The adverse impact is that these 1000 numbers can be sold at a throwaway salvage price of Rs. 250 in the off- season. The actual contribution for selling it in the off-season is =selling price in off season-variable cost incurred = $250-500 = -250$ or a loss of 250 rupees. The financial impact caused by the excess stock of 1000 deluxe = $(1000)(250) = \text{Rs } 2.5 \text{ lakhs}$. This loss will have to be taken out of the gross contribution. Whenever demand is more than production, the question of excess stock does not arise.

Choose the Best Alternative

Find out which of the alternatives gives the highest expected net contribution. The word net is important because from the financial contribution obtained for each alternative, the excess production impact worked out financially will have to be taken out. The word 'expected' is also important because demand is a random variable, and the expected contribution is calculated using probabilities of demand as weights.

Implement the Best Alternative

This is a very crucial part. If the best solution is not properly implemented, then the solution is an exercise in academics. The successful implementation of the best alternative requires the active coordination of many departments. For example, there should be proper coordination between marketing and production; likewise coordination between finance and marketing when it comes to price and cost details.

Evaluate the Results and Check that things are working all right

This step is the feedback and control to ensure that the objective of the decision problem is realized. When some bottlenecks are taking place like breakdown of a machine, delay in receipt of raw materials etc, they have to be set right so that the targeted objective of the highest contribution is achieved. Likewise, when there is a sudden shift in demand, the manufacturer must be able to revise his product mix.

Progressive Test Questions With reference to the case study discussed for explaining the steps involved in systematic problem solving, answer the following questions:

1. What is the contribution for a standard chair during peak season? -----.
2. What is the contribution for the deluxe chair during off-season? -----.
3. The probability of the demand being 3000 standard and 2000 deluxe is-----.

Solution

1. Contribution for standard model during peak season = selling price - variable cost = $1000 - 500 = \text{Rs. } 500$ per chair.
2. Contribution for deluxe model during off season is = salvage price in off season - variable cost = $500-1000 = -500$ or a loss of Rs. 500.
3. Since the demand distribution of the models are independent, probability of the demand being 3000 standard and 2000 deluxe = $P(3000 \text{ standard}) \cdot P(2000 \text{ deluxe}) = (0.40)(0.70) = 0.28$ or 28%.

11.2 HOW TO STRUCTURE A DECISION PROBLEM

Using the steps in systematic problem solving that is the foundation stone, you can easily structure a decision problem with a little bit of fine-tuning. A decision problem involves a set of alternatives. You have to establish a criterion or a set of criteria for evaluating the alternatives. You have to then evaluate alternatives using criterion/criteria defined. The best decision corresponds to the best alternative chosen. In this Chapter on decision analysis, the alternatives are evaluated using monetary value as the criterion. Monetary value for each alternative depends on the "states of nature"(uncertain events). The consequence that result from each alternative and for each state of nature in monetary value is called a payoff matrix. The following visual succinctly portrays how to structure a decision problem.

Please note that a decision problem involves a scientific and systematic problem solving approach in which a quantitative analysis using monetary value becomes a basis for decision-making.

How to Structure A Decision Problem

- A decision problem is characterized by a set of alternatives states of nature, and the consequences expressed in monetary value.
- The alternatives pave the way for possible strategies the decision maker can implement.
- The states of nature refer to events that are not under the control of the decision maker. States of nature should be so defined that they become mutually exclusive and collectively exhaustive.
- For each alternative and state of nature, there is a resulting monetary value. These are displayed in matrix form called a **payoff table**.

Figure 11.3

Next, we will discuss the structure of decision problem with an example. The example picked up for discussion is a very simple one. Problems in real life situations could be very complex. However, for the purpose of clarity in thinking and understanding, this example will be very useful and it will lay the foundation for solving more challenging problems.

Example A small fruit merchant has got a problem on hand. He has to decide how many dozens of a particular type of fruit to stock on a given day. Total demand per day is uncertain. He has analyzed the past data and found the following pattern of demand distribution based on 360 days.

Total demand per day (In dozens)	Number of days each demand level was recorded	Probability of demand (based on relative frequency)
25	72	0.20
30	90	0.25
35	108	0.30
40	90	0.25

Fruits not sold on any day perish and have to be thrown out. Selling price of the fruit per dozen is \$30. Cost of procurement and other incidentals add to \$20 per dozen. How many dozens per day should the merchant stock?

Solution In this decision problem, states of nature refer to uncertain demand patterns that could take values 25, 30, 35, or 40 with associated probabilities 0.20, 0.25, 0.30, or 0.25 respectively. Please note that the states of nature are mutually exclusive and collectively exhaustive. The total probability adds to one.

The alternatives are whether to stock 25, 30, 35, or 40. For each alternative, demand could be 25, or 30, or 35, or 40. Can you work out the pay off matrix? Try your best. Otherwise follow the steps.

Pay off Matrix for the Example:

Pay off Matrix (Figures in \$)

Demand (States of Nature)	If you decide to stock (Decision alternatives)			
	25	30	35	40
25(0.20)	250	150	50	-50
30(0.25)	250	300	200	100
35(0.30)	250	300	350	250
40(0.25)	250	300	350	400

Let us explain how the pay off matrix is worked out. Profit per dozen = Selling price - Cost = $30 - 20 = \$10$. Fruits not sold on the same day will have loss of \$20 per dozen. If you stock 25 and the demand is also 25, the profit will be $(25)(10) = \$250$. If the demand is more than 25, even then the profit will be \$250 because you have stocked only 25. Of course, you have lost the opportunity of making a profit of more than \$250. You are wiser after the event. So you see under column 25 dozens, the pay offs are 250 all through. Let us come to the next column that has a decision of stocking 30 dozens. If you stock 30 and the demand is 25, you can sell only 25. The profit in this transaction is \$250. There is excess stock of 5 dozens for which you will incur a loss of $(5)(20) = \$100$. This will have to be subtracted from 250 resulting in net profit of \$150. For the remaining elements in this column, you will have a profit of \$300 all through because either demand equals 30 or exceeds 30. Extending this logic, you can fill up the other elements in the pay off matrix.

From the pay off matrix, can you say what is the optimal decision? That is, how many dozens should you stock? This can be answered by using the criterion called Expected Monetary Value (EMV). This is explained next.

11.3 EXPECTED MONETARY VALUE (EMV)

The following visual explains the importance and method of computing EMV.

Expected Monetary Value Criterion

- If probabilities regarding the states of nature are available, one can use the expected monetary value (EMV) criterion to select the best alternative.
- EMV for each decision is computed by summing the products of the pay off of each state of nature and the probability of the respective state of nature taking place
- The alternative giving the highest expected monetary value (EMV) is selected as the best decision.

Figure 11.4

Progressive Test Question Work out EMV for each decision alternative.

Solution EMV for the illustration

	Demand (States of Nature)	Probability of Demand	If you decide to stock (Decision alternatives)			
	25	0.20	25	30	35	40
	30	0.25	250	300	200	100
	35	0.30	250	300	350	250
	40	0.25	250	300	350	400
	EMV		250	270	252.5	190

Figure 11.5

Let me show you the calculation of EMV for one decision alternative. Take the case of stocking 30 dozens. The spreadsheet displays the pay offs under the option of 30 dozens. Multiply these pay offs by probabilities of states of nature (probabilities of demand occurring) and then add them to get EMV. EMV corresponding to a stock of 30 dozens =

$(150)(0.20) + (300)(0.25) + (300)(0.30) + 300(0.25) = 30 + 75 + 90 + 75 = \270 . Likewise all EMVs are computed in the spreadsheet.

As you can see the highest EMV occurs when you stock 30 dozens. Hence the fruit merchant should stock 30 dozens and $EMV = \$270$. Call this EMV as EMV optimum. Please note that EMV is a weighed average of monetary consequences using probability of each state of nature as weight.

Expected Value of Perfect Information (EVPI)

Very often in a decision problem, any manager of an enterprise has to answer the question "whether to take action now, or gather additional information and act later. Is the cost of additional information justified in terms of the additional profit you may get? This issue can be resolved using EVPI. EVPI is calculated as the difference between two factors- expected profit with perfect information (EPPI) and EMV optimum.

Symbolically, $EVPI = EPPI - EMV \text{ optimum}$. We know already what is EMV optimum. To repeat again, EMV optimum is the highest EMV that you get among many decision alternatives. To appreciate and calculate EPPI, let us go back to our example of fruit merchant problem. The pay off matrix is again displayed for clarity.

Pay off Matrix (Figures in \$)

<i>Demand (States of Nature)</i>	<i>If you decide to stock (Decision alternatives)</i>			
	25	30	35	40
25(0.20)	250	150	50	-50
30(0.25)	250	300	200	100
35(0.30)	250	300	350	250
40(0.25)	250	300	350	400

If you know demand is 25, then the best decision is to stock 25 because this gives the highest pay off of \$ 250(maximum of 250, 150, 50, -50). Likewise, if demand is 30, the best decision is to stock 30. Proceeding in this manner, you can complete the pay off matrix for the best decision when demand is known in advance. The following table emerges.

Pay off Matrix with perfect information (Figures in \$)

<i>Demand (States of Nature)</i>	<i>(Best Decision alternatives) Stock</i>			
	25	30	35	40
25(0.20)	250			
30(0.25)		300		
35(0.30)			350	
40(0.25)				400

Now use probabilities of state of nature as weights and compute the expected value as before. You get EPPI.

$$EPPI = (250)(0.20) + (300)(0.25) + (350)(0.30) + (400)(0.25) = 330.$$

Hence $EVPI = EPPI - EMV \text{ optimum} = 330 - 270 = \60 . EVPI acts as an upper bound measuring worth of gathering additional information.

Progressive Test Question For each decision alternative, if you multiply the probabilities of states of nature with the corresponding pay offs, you get:

1. Expected profit with perfect information (EPPI)
2. Expected value of perfected information (EVPI)
3. Expected monetary value (EMV)

Solution Correct choice is 3). This is the very definition of EMV. EMV is a weighted average of monetary value in which probabilities of states of nature are used as weights. They are multiplied with pay off of corresponding states of nature and summed to get EMV.

Progressive Test Question $EVPI = \text{-----}$.

Solution $EVPI = EPPI - EMV \text{ optimum}$.

Progressive Test Question EMV optimum gives the lowest monetary value. True or False

Answer The statement is false because EMV optimum gives the highest monetary value.

Discussion Topic

Analyze, criticize, and explain the following statement:

'Decision-making using EMV criterion is possible only on paper because probabilities of states of nature in real business problems cannot be obtained'.

Opportunity Loss Table

If the pay off matrix is expressed in terms of opportunity loss for every consequence, it is called a loss table. Opportunity loss is based on the concept of opportunity cost that measures the cost of sacrificing one alternative against another.

For the fruit merchant problem discussed, let us work out the pay off matrix in terms of opportunity cost of revenue foregone and verify the best decision obtained using EMV optimum will be the same as one obtained from minimum expected opportunity loss (EOL).

For illustration sake, let us take the decision alternative of 30 dozens to stock. If the demand is 25 dozens, there is excess stock of 5 dozens for which you incur a loss of $(5)(20) = \$100$. If the demand is 30 dozens, you can sell all 30 dozens and the opportunity loss is \$0. If the demand is 35 dozens, then you have lost an opportunity of making an additional profit of $(5)(10) = \$50$ had you stocked 35 dozens. If the demand is 40 dozens, you have lost an opportunity of making an additional profit of $(10)(10) = \$100$ had you stocked 40 dozens. Using this logic the following opportunity loss table is constructed. The optimal decision is one that corresponds to the minimum expected opportunity loss (EOL). This will of course tally with the optimum decision based on the one that corresponds to maximum EMV.

Opportunity Loss Table (\$)

Demand (States of Nature)	If you decide to stock (Decision alternatives)			
	25	30	35	40
25(0.20)	0	100	200	300
30(0.25)	50	0	100	200
35(0.30)	100	50	0	100
40(0.25)	150	100	50	0
EOL	\$80	\$60	\$77.5	\$140

Computing EOL is identical to computing EMV. Opportunity loss in each column in the above table is multiplied by the corresponding probabilities of demand and then added to get EOL. Please note that the optimal decision is to stock 30 dozens because EOL for this decision is \$60, which is the least. Thus minimum EOL decision tallies with maximum EMV decision. Please note that **Minimum EOL = EVPI!**

11.4 DECISION TREE

Decision tree is a very useful graphical tool for structuring and solving decision problems. The following visual succinctly portrays a decision tree.

Decision Tree

- A decision tree is a graphical representation of a decision problem.
- Each decision tree has two types of nodes; circles represent the states of nature and squares represent the decision alternatives.
- The branches emanating from each circle denote different states of nature. The branches emanating from squares denote the different decision alternatives.
- At the end of each extremity of a tree, the payoffs from all the branches are displayed.

Figure 11.6

Example For drawing the decision tree, let us take the same example of fruit merchant. For ready reference, the problem is stated below:

A small fruit merchant has got a problem on hand. He has to decide how many dozens of a particular type of fruit to stock on a given day. Total demand per day is uncertain. He has analyzed the past data and found the following pattern of demand distribution based on 360 days.

Total demand per day (In dozens)	Number of days each demand level was recorded	Probability of demand (based on relative frequency)
25	72	0.20
30	90	0.25
35	108	0.30
40	90	0.25

Fruits not sold on any day perish and have to be thrown out. Selling price of the fruit per dozen is \$30. Cost of procurement and other incidentals add to \$20 per dozen. Draw the decision tree and answer how many dozens per day should the merchant stock?

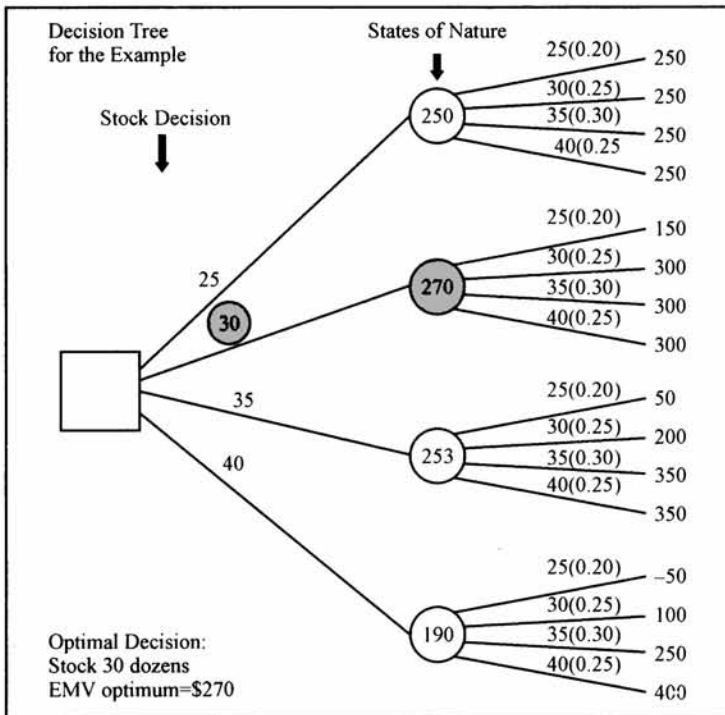


Figure 11.7

Explanation on the Decision tree for the Example

1. Branches emanating from the square represent decision alternatives of stocking 25, 30, 35, and 40 dozens.
2. For each decision alternative, there are four states of nature (demand pattern) with associated probabilities.

3. The branches that emanate from the circle represent the states of nature of the uncertain demand that could be 25, 30, 35, or 40 dozens. The figures with in brackets for each demand represent the probability of demand occurring.
4. The pay offs for each decision alternative is given at the extremity of the tree for every branch of states of nature.
5. For every decision option, working backwards, expected monetary value (EMV) is calculated as the sum product of two terms namely pay off of state of nature (demand) and probability of state of nature (demand). The optimal decision is obtained by folding back the tree. All EMV values obtained by folding back the tree are displayed in each circle. EMV optimum is one that gives the highest monetary value. This is highlighted in the tree diagram in the beginning decision node. The other decision alternatives are therefore discarded. The principle of folding back is elegant and convenient especially in complex problems including sequential decisions.
6. EMV optimum = \$270. So, stock 30 dozens is the optimal solution to the problem.

Solution using decision tree approach tallies with what you have got earlier. The advantage of decision tree approach is that it is very versatile when you have complicated problems including sequential decisions. You will also appreciate the utility of a decision tree in assessing value of sample information. This is given next.

Decision Tree - A comprehensive case problem

In order to reinforce the conceptual framework of decision tree and its utility in solving a business problem, the following case problem is discussed below.

Case Study-Product Mix to Maximize Expected Contribution

A manufacturer produces two decorative models (standard and deluxe) of chairs. In the light of forthcoming peak selling season, he must decide how many of each model to produce. Variable cost of the standard model is Rs. 500, and selling price is Rs. 1000; variable cost of the deluxe model is Rs. 1000, and selling price is Rs. 2000. Demand is uncertain and has the following discrete distribution:

<i>Standard Model</i>		<i>Deluxe Model</i>	
<i>Demand</i>	<i>Probability</i>	<i>Demand</i>	<i>Probability</i>
3000	0.40	1000	0.30
4000	0.60	2000	0.70

Further, it is known from market research that the probabilities of demand for these two models are independent. Production capacity is limited to a total of 5000 chairs. Chairs not sold during the peak season are disposed of at salvage prices of Rs. 250 for the standard model and Rs. 500 for the deluxe model.

QUESTION

Draw the decision tree and obtain the optimal product mix using expected monetary value (EMV) criterion.

Solution Let us first structure the problem in terms of a decision tree.

Facts given :

During Season

Contribution margin per standard chair(S) = Selling price -variable cost = 1000 – 500 = Rs. 500

Contribution margin per deluxe chair(D) = Selling price -variable cost = 2000 – 1000 = Rs. 1000

Off Season:

Contribution margin per standard chair(S) = Selling price -variable cost = 250 – 500 = Rs. -250

Contribution margin per deluxe chair(D) = Selling price -variable cost = 500 – 1000 = Rs. -500

Production capacity is limited to a total of 5000 chairs.

Possible Decision Alternatives involving production of Standard and Deluxe Chairs

<i>Standard (S)</i>	<i>Deluxe(D)</i>	<i>Total</i>
3000	1000	4000
3000	2000	5000
4000	1000	5000

Please note that the product mix combination of 4000 standard and 2000 deluxe is not considered because the production capacity is limited to 5000 chairs. Before drawing the tree, let us construct the pay off matrix for the states of nature.

Pay off matrix for the problem- (Figures in Rs Lakhs)

<i>States of Nature (Demand)</i> <i>(Standard, Deluxe)</i>	<i>Probability</i>	<i>Product Mix Decision Alternatives</i>		
		<i>S , D</i> <i>(3000, 1000)</i>	<i>S , D</i> <i>(3000, 2000)</i>	<i>S , D</i> <i>(4000, 1000)</i>
(3000, 1000)	0.12	25	20	22.5
(3000, 2000)	0.28	25	35	22.5
(4000, 1000)	0.18	25	20	30
(4000, 2000)	0.42	25	35	30

How the pay offs are computed? Take the first decision alternative of 3000 standard and 1000 deluxe. Since demand for each variety is either equal to or exceeding the product mix, the contribution you will get = 3000(500) + 1000(1000) = Rs 25 lakhs. So, the entire column is shown = 25. Look at the second alternative of 3000 standard and 2000 deluxe. If the demand is 3000 standard and 1000 deluxe, there is an excess production of 1000 deluxe that will fetch a contribution of Rs -500 per chair, and for the 1000 chairs this would

amount to a loss of Rs 5 lakhs. This is taken out from the total contribution of Rs. 25 lakhs obtained from selling 3000 standard and 1000 deluxe. Suppose the demand is 3000 standard and 2000 deluxe, your contribution will $3000(5000) + 2000(1000) = \text{Rs } 35$ lakhs. Applying this logic by taking care of excess production, if any, all the cells are filled up. The decision tree is shown below. The optimum product mix is to produce 3000 standard and 2000 deluxe that gives the highest EMV of Rs. 30.5 lakhs.

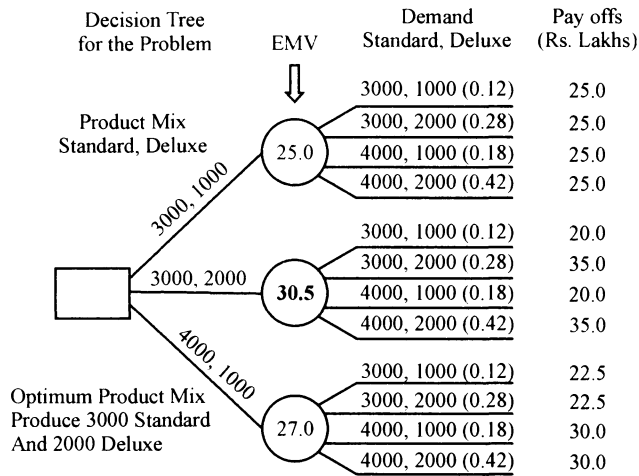


Figure 11.8

11.5 VALUE OF SAMPLE INFORMATION

Expected value of perfect information (EVPI) discussed earlier in this Chapter just gives an upper bound on the value of gathering additional information in a decision problem. Suppose we propose to conduct a sample survey to obtain additional information. You must ask the question, "is it worth doing this?". The answer to this depends on the value of information we gather. This is calculated using expected value of sample information (EVSI). EVSI, intuitively and logically, must be = Expected profit with sample information (EPSI) - EMV optimum. In order to get EPSI, we have to revise the probabilities of the states of nature, based on sample information we hope to get. The probabilities that are associated with the states of nature when you begin the exercise are called prior probabilities. For example, probabilities of the demand distribution in the fruit merchant problem are **prior probabilities**. Once sample information is used to revise these probabilities, they become **posterior probabilities**. Posterior probabilities are used to rework expected monetary values. The highest monetary value among the revised EMVs will give EPSI. From this if you subtract EMV optimum, you get EVSI. An example will clarify concepts.

Example The Marketing manager of a company producing consumer durable wants to decide whether to launch a new product or not. The sales quantity of the new product has two states of nature namely high sales or low sales. The pay off table is as follows.

Pay Off Table (Figures in \$ million)

<i>States of Nature</i>	<i>Probability</i>	<i>Launch</i>	<i>Do not launch</i>
High sales	0.4	5	0
Low sales	0.6	-4	0

The marketing manager does not want to decide just on the basis of EMV and would like to gather additional information before taking the final decision. A market research agency is willing to do a study that will cost \$0.3million. There are two possibilities. The agency may predict high sales for the new product or it may predict low sales for the new product. Past record of the agency in terms of predicting ability is tabulated below:

<i>States of Nature</i>	<i>Prior Probability</i>	<i>Agency predicts High Sales given States of Nature</i>	<i>Agency predicts Low Sales given States of Nature</i>
High Sales	0.4	0.6	0.4
Low Sales	0.6	0.3	0.7

Questions

- (a) Should the product be launched based on EMV criterion?
 (b) Should the marketing manager take the help of the agency or not?

Solution to (a)

The product should not be launched. EMV for launching = $(5)(0.4) + (-4)(0.6) = \$-0.4\text{ml}$. Since this is less than \$ 0(EMV optimum) that you get if product is not launched, the product should not be launched.

Solution to (b) In order to solve this problem, we need to work out the revised probabilities of the states of nature from prior probabilities of the states nature. This means we want to obtain posterior probabilities of high sales and low sales given agency prediction. This, statisticians call as inverse probability approach based on Bayes' theorem. You can easily get this by constructing joint probability table given below:

<i>States of Nature</i>	<i>Agency Predicts High Sales</i>	<i>Agency Predicts Low Sales</i>	<i>Marginal Probability States of Nature</i>
High Sales	0.24	0.16	0.40
Low Sales	0.18	0.42	0.60
Marginal Probability of Agency Prediction	0.42	0.58	1.00

Please note that the joint probability table is obtained by multiplying the probabilities of agency prediction given states of nature and prior probabilities. For example, take the first cell in the matrix above. It is equal to probability agency predicts high sales given state of nature is high sales multiplied by probability of high sales. This is = $(0.6)(0.4) = 0.24$. Likewise, all entries are computed.

Now the calculation of revised probabilities is very simple. Probability of high sales given agency predicts high sales = $0.24/0.42 = 0.57$. Probability of low sales given agency

predicts high sales = $0.18/0.42 = 0.43$. The following table represents posterior probabilities given agency prediction.

Revised Posterior probability table

Posterior Probability of	Given Agency Predicts	
	High Sales	Low Sales
High Sales	0.57	0.28
Low Sales	0.43	0.72

A completed tree for solving this problem is given below.

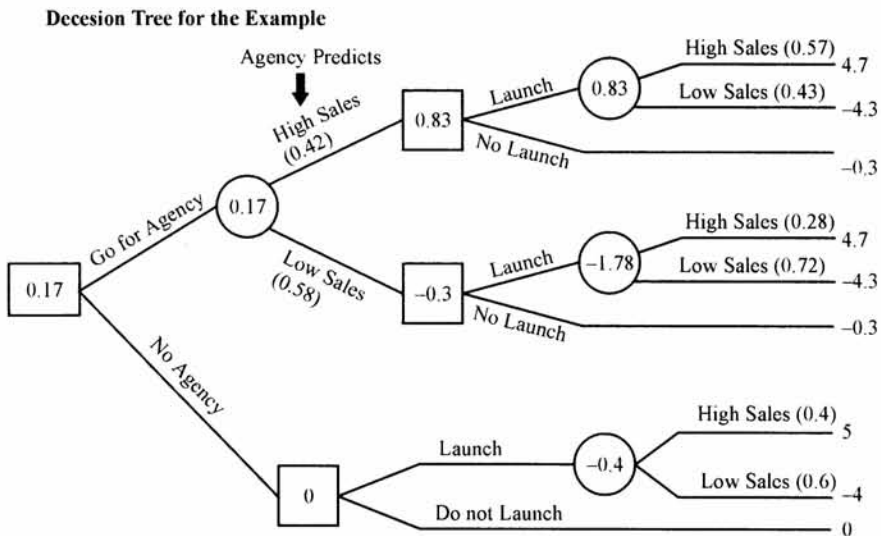


Figure 11.9

Explanation on the tree above

- The manager has two options in the first stage of decision namely "go for agency" or "no agency". This appears in the first square node.
- If you don't go for agency, then you have two decision options-launch the new product or do not launch the new product. These two options are shown in the next square node. If you launch, states of nature (demand) may be high sales or low sales with prior probabilities 0.4 and 0.6 respectively. Monetary values corresponding to these two states of nature are given as 5 and -4 at the extreme end of the branch. These are in \$million. If you don't launch, you get \$0 million. EMV is calculated in the usual way for these two options. They are -0.4 and 0 respectively. So, EMV optimum =0. This appears in the square node corresponding to no agency.
- If you decide to go for agency, there are two states of nature. Agency predicts high

sales with a probability of 0.42 and low sales with a probability of 0.58. These probabilities are from the revised posterior probability table.

- If agency predicts high sales, you have again got two-decision options - launch or no launch. Under each of these options, the states of nature could be high sales or low sales. However, the probabilities of these states of nature are posterior probabilities. These are given in brackets. Please refer again revised posterior probability.
- EMV values are worked out as before by multiplying probabilities of states of nature with corresponding pay offs and then summing them. EMVs are displayed in circle and square nodes in the diagram.
- Please also note that while working out EMV values, cost of doing the study by agency (\$0.3million) is taken out to represent the correct net monetary value. Under the scenario of taking agency help, if you choose not to launch, you still incur cost of \$0.3million(-0.3 EMV).
- When you finally fold back the tree, you find going for agency produces a positive EMV of \$0.17 million. This is better than getting \$0 million of no agency. Hence, go for agency.

Obviously $EVSI = EPSI - EMV \text{ optimum} = 0.17 - 0 = \0.17 million . This is the gain you have got by using sample information.

Looking at the decision tree, you can conclude the following:

Sequential Decisions

- Take the decision of going to the agency.
- If agency predicts high sales, then launch the new product (EMV \$0.83 million).
- If agency predicts low sales, then do not launch the new product (EMV \$-0.3 million).
- Expected value of sample information (EVSI) is \$0.17 million.

Please appreciate the profundity of posterior analysis. If you have gone entirely by EMV criterion without gathering additional information, you would not have introduced the new product and you would have missed an opportunity. The moral of the story is that it is better to gather additional information through sample survey before making the final decision. This is particularly important in crucial decision problems.

11.6 CHAPTER SUMMARY

In this Chapter, you have been introduced to a comprehensive framework on decision-making under conditions of uncertainty using decision analysis tools. Specifically, this Chapter focused on:

- How to structure a decision problem in a systematic fashion
- Pay off matrix and its use
- Expected monetary value (EMV) as a major criterion in selecting the best decision

- Expected value of perfect information (EVPI) in assessing value of information
- Construction and use of decision tree in solving simple as well as complex problems
- Revising probabilities through posterior analysis to improve the power of decision making
- Expected value of sample information (EVSI), an important technique that comes handy before taking a final decision, particularly, in the context of introducing a new product into the market.

GLOSSARY

Decision Alternatives It is a list of all possible choices available to the decision maker.

Decision Tree It is a graphical representation of a decision problem. It has two types of nodes - square and circle. Branches emanating from the square nodes are the decision alternatives. Branches emanating from the circle nodes represent the possible states of nature. The payoffs of all the branches are displayed at the end of each extremity of the tree.

Expected Monetary Value (EMV) It is a weighted average of monetary value in which probabilities of states of nature are used as weights. They are multiplied with pay off of corresponding states of nature and summed to get EMV.

Expected Opportunity Loss (EOL) It is a weighted average of opportunity loss in which probabilities of states of nature are used as weights. They are multiplied with opportunity losses of corresponding states of nature and summed to get EOL.

Expected Value of Perfect Information (EVPI) It is the difference between two factors - Expected profit with perfect information (EPPI) and EMV optimum (EMV for the best decision alternative). EVPI acts as an upper bound measuring worth of gathering additional information.

Expected Value of Sample Information It is the difference between Expected profit with sample information (EPSI) and EMV optimum. It measures the gain from sampling.

Opportunity Loss It is the opportunity cost of sacrificing one alternative against another.

Payoff Matrix This is also called a payoff table that gives the consequence that result from each alternative for each state of nature in monetary value.

Posterior Probabilities These are conditional probabilities that are obtained using Bayes theorem to revise prior probabilities in the light of additional information.

Prior probabilities These are the probabilities of the states of nature that are known to us in the beginning.

States of Nature These are events that are not under the control of the decision maker.

REVIEW QUESTIONS

1. A payoff matrix takes into consideration the states of nature that will influence the selection of a particular decision alternative. True or False.
 2. Expected value of perfect information (EVPI) acts as upper bound on the amount that can be spent in gathering additional information. True or False.
 3. In a decision problem, an analyst finds highest expected monetary value to be \$ 5000. Expected value of perfect information (EVPI) is \$12000. So, expected profit with perfect information (EPPI) is:
 - (a) \$ 7000
 - (b) \$ 5000
 - (c) \$ 18000
 - (d) \$ 17000
 4. EVSI can be more than EVPI. True or False.
 5. In a decision tree circles represent decision alternatives and squares represent states of nature. True or False.
 6. Posterior probabilities can be calculated without prior probabilities. True or False.
- Questions 7 and 8 should be answered based on the following decision tree:

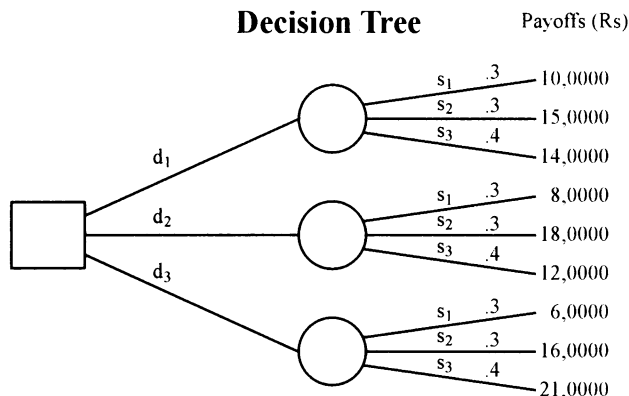


Figure 11.10

7. The best decision alternative based on EMV criterion is:
 - (a) d_1
 - (b) d_2
 - (c) d_3
8. EVPI is -----.

ANSWERS TO REVIEW QUESTIONS

1. **Answer** The statement is true because the financial consequences in a pay off matrix are definitely influenced by uncertain events called states of nature.
2. **Answer:** The statement is true. EVPI is the maximum amount a decision-maker would like to spend to gather additional information. It is a measure of worth of information.

3. **Answer:** (d) is the right choice. $EVPI = EPPI - EMV$ optimum.
That is \$ 12000 = $EPPI - \$5000$. So, $EPPI = \$12000 + \$ 5000 = \$ 17000$
 4. **Answer:** The statement is false because $EVPI$ is the maximum amount that you are willing to spend in gathering additional information. At best, $EVSI$ can be equal to $EVPI$. It can never be more than $EVPI$.
 5. **Answer:** The statement is false. In a decision tree, circles represent states of nature and squares represent decision alternatives.
 6. **Answer:** The statement is false. Posterior probabilities are revised probabilities based on prior probabilities. It uses sample information to revise the prior probabilities.
- Answers** To questions 7 and 8. Look at the diagram again.

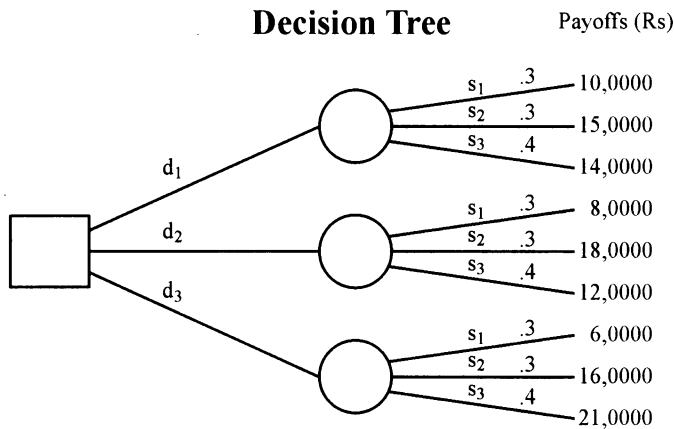


Figure 11.11

7. **Answer** (c) is the right choice. Let us calculate EMV for d_1 , d_2 , and d_3 .
For d_1 $EMV = (10000)(0.3) + (15000)(0.3) + (14000)(0.4) = \text{Rs. } 131000$
For d_2 $EMV = (8000)(0.3) + (18000)(0.3) + (12000)(0.4) = \text{Rs. } 126000$
For d_3 $EMV = (6000)(0.3) + (16000)(0.3) + (21000)(0.4) = \text{Rs. } 150000$
As you will see the highest EMV occurs for decision option d_3 .
8. **Answer** is Rs. 18000. From the decision tree above, we note the following: When state of nature is S_1 , the best pay off is Rs. 100000. When state of nature is S_2 , the best pay off is Rs. 180000. When state of nature is S_3 , the best pay off is Rs. 210000. Using probabilities as weights, we have $EPPI = (100000)(0.3) + (180000)(0.3) + (210000)(0.4) = \text{Rs. } 168000$.
 $EVPI = EPPI - EMV$ optimum = $168000 - 150000 = \text{Rs. } 18000$

PRACTICE PROBLEMS

1. Case Study - How Many Machines to Install?

A large international chemical manufacturing company is building a plant for exporting into the international market a new chemical called Galaxy that is used in agricultural sector. The firm is uncertain about the demand pattern for Galaxy during the initial year of sales. The following is the observed probability distribution based on past frequency distribution of demand for Galaxy.

<i>Galaxy Demand for the First Year (Kgs)</i>	<i>Probability</i>
1200	0.1
1500	0.3
1800	0.3
2100	0.2
2400	0.1

Galaxy is manufactured on a specially designed machine. Each machine gives an output of 300 kgs of Galaxy per year. The total cost of buying and running one machine is \$1200000 per year. This total cost includes supervision, maintenance, other fixed costs, interest and other overheads. The profit from one-year production of one machine without including the total cost of the machine is \$2.0 million. You assume that there is sufficient demand to run the machine for the entire year.

The problem facing the company is to decide how many machines should it install in the plant for the first year. The second and subsequent years give rise to no problems, as the firm feels that it will be able to buy additional machines if required to meet the demand if it zooms. Unused capacity, if any, can be absorbed in future years if the demand grows.

QUESTIONS

Draw the decision tree for this problem and then answer how many machines the firm should install in order to maximize the expected net profit in the first year.

If an agency claims that it could accurately predict the demand in the first year, what is the maximum consulting fee can you offer?

2. Case Study - Should Lambda Maintain Existing Price or Follow Competition?

Lambda Company, which has its headquarters and manufacturing operations in India, has established a marketing office in the U.S to cater to attractive export opportunities for its products. Market demand analysis suggests that a particular item called Product Z has good potential in America. ABC Corporation in the U.S is currently discussing with Lambda about Product Z, and it is likely to buy 10000 units subject to certain price levels stipulated by ABC to be agreed upon by Lambda. Another competitor is also very keen to bag this order and is negotiating with ABC. Product Z is manufactured in Division C and every one unit of it requires one unit of Product Y that is being manufactured in Division B. Every one unit of Product Y requires one unit of Product X that is being manufactured in Division A.

ABC is a tough negotiator, and it is extremely important for Lambda to clinch the

order. The competitor is considering a proposal of significantly lowering the price to edge out Lambda in this race. Exhibit 1 gives the cost details, which will form an important basis for price negotiation with ABC.

Exhibit 1

<i>Standard Cost per Unit</i>	<i>Product X</i>	<i>Product Y</i>	<i>Product Z</i>
Materials purchased outside	\$ 4.00	\$ 6.00	\$ 2.00
Direct labor*	\$ 2.00	\$ 2.00	\$ 4.00
Variable Overhead	\$ 2.00	\$ 2.00	\$ 4.00
Fixed overhead per unit	\$ 6.00	\$ 4.00	\$ 6.00
Standard Volume (units)	10000	10000	10000

*Direct labor be treated as variable cost

The present selling price for Product Z is \$56.00. Listed below are a series of possible price reductions envisaged by competition and the impact of these reductions on the volume of sales if Lambda does not reduce its price to the level of competition.

- Possible competitive price: \$54; \$52; \$50; \$46; \$44. The corresponding probabilities for these prices are 0.10, 0.20, 0.25, 0.30, 0.15 respectively
- Sales volume to Lambda if price of Product Z is stuck to \$56.00: 9000; 7000; 5000; 2000; 0.
- Sales volume to Lambda if price of Product Z is reduced to competitive levels: 10000; 10000; 10000; 10000; 10000.

QUESTIONS

- (a) Draw the decision tree for this problem and find out the best alternative based on EMV criterion?
 - (b) Is Lambda Company better advised to maintain its price at \$ 56.00 or to follow competition?
3. **Case Study- Launch or not to Launch**

The Marketing manager of a multinational company producing consumer durable wants to decide whether to launch a new product or not. The sales quantity of the new product has three states of nature, namely, high sales, medium sales, or low sales. The pay off table is as follows.

Pay Off Table (Figures in \$ million)

<i>States of Nature</i>	<i>Probability</i>	<i>Launch</i>	<i>Do not launch</i>
High sales	0.4	6	0
Medium Sales	0.3	3	0
Low sales	0.3	-4	0

The marketing manager would like to gather additional information before taking the final decision. A market research agency is willing to do a study that will cost \$1

million. There are three possibilities. The agency may predict high sales for the new product, may predict medium sales for the new product, or it may predict low sales for the new product. Past record of the agency in terms of predicting ability is tabulated below:

<i>States of Nature</i>	<i>Prior Probability</i>	<i>Agency predicts High Sales given States of Nature</i>	<i>Agency predicts Medium Sales given States of Nature</i>	<i>Agency predicts Low Sales given States of Nature</i>
High Sales	0.4	0.7	0.2	0.1
Medium sales	0.3	0.2	0.6	0.2
Low Sales	0.3	0.1	0.2	0.7

QUESTIONS

- (a) Should the product be launched based on EMV criterion?
- (b) Should the marketing manager take the help of the agency or not?
- (c) What are the values of EVPI and EVSI?

Forecasting

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Appreciate the role of forecasting
- List and describe the qualitative and quantitative methods of forecasting
- Forecast using time series models
- Measure the forecast error to assess the accuracy of the models

CHAPTER OUTLINE

- 12.1 Forecasting-Basics
 - 12.2 Qualitative Methods of Forecasting
 - 12.3 Quantitative Methods of Forecasting
 - 12.4 Chapter Summary
- Glossary
Review Questions
Answers to Review Questions
Practice Problems

INTRODUCTION

Many crucial decisions made by management depend upon the assessment of the future demand for products and services, sales growth, and cost trends. Management must forecast the future in order to make sound decisions today. So, managers need efficient and reliable forecasting methods for business planning. This chapter presents a few widely used forecasting techniques in practice.

What is the demand for our latest model personal computer?



Figure 12.1

12.1 FORECASTING-BASICS

Why Forecasting?

- Demand, or sales forecasting, is the foundation stone upon which the entire business planning is built. An organization cannot predict its profitability without predicting sales revenue. Sales revenue cannot be predicted without forecasting sales in physical quantities. The entire production program and materials resource planning cannot be achieved without a realistic sales forecast of the various products the organization would like to market. Corporate plans, turnaround plans and competitive business strategies need the help of forecasting. In other words, not to forecast is to assume status quo and do nothing. This will never be acceptable to any manager in any organization.
- We must, of course, recognize the fact that future is uncertain, and therefore, no forecasting can be hundred percent accurate. This is a paradox in forecasting: on one hand you need sales forecasts, and on the other hand no forecast can be accurate.
- Managers in any business enterprise have no choice between forecasting and not forecasting because without a sound forecasting system, the risk of making a wrong decision increases. Managers however have choice amongst the methods of forecasting. In this chapter, we will dwell at length on the popular methods of forecasting in practice.

Forecasting Methods in Practice

The following visual portrays the widely used forecasting techniques in practice.

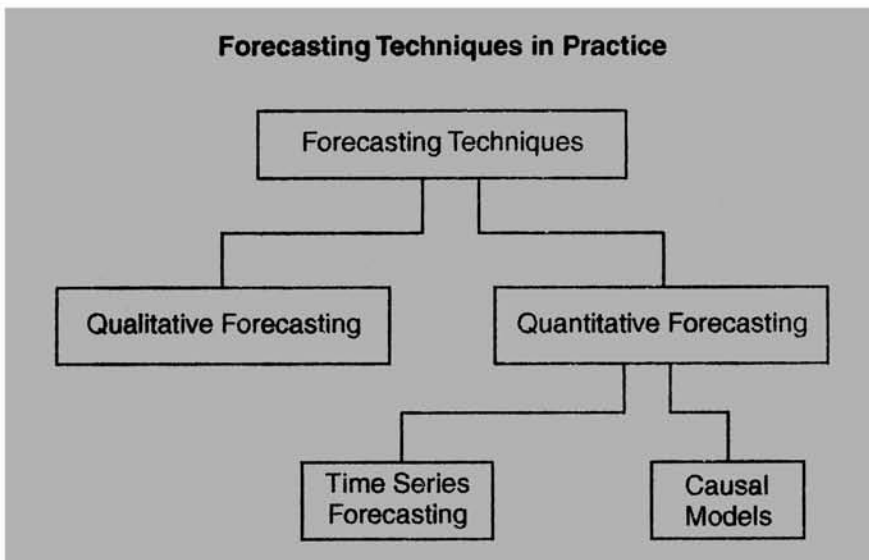


Figure 12.2

Selecting the Right Forecasting Technique-Guidelines

- Availability of data: If no appropriate historical data are available, quantitative techniques of forecasting are not possible. Only qualitative forecasting techniques are possible.
- Accuracy envisaged: Greater the accuracy needed, greater is the need for sophisticated techniques of forecasting.
- Urgency with which the forecast is sought: If forecasts are required urgently, only less sophisticated techniques are possible to use.
- Cost: This includes cost of forecasting exercise, and what it costs the firm if a wrong forecast is made.

12.2 QUALITATIVE METHODS OF FORECASTING

Qualitative Forecasting

Imagine your company is about to introduce a new product that is unknown in the market. In this context, there will be no historical data that you could use to forecast your sales. It is a situation where you will find complete absence of any useful data. Under these circumstances, qualitative forecasting is the only method by which you could forecast your sales for your new product.

Figure 12.3

Methods Used in Qualitative Forecasting

The following visual gives the popular methods of qualitative forecasting.

Four widely used Methods of Qualitative Forecasting

- Expert Opinion
- Market Survey
- Delphi Method
- Historical Analogy

Figure 12.4

Progressive test question Qualitative forecasting is superior to quantitative forecasting because it emphasizes on quality. True or False.

Answer The statement is false. These are the two broad types of forecasting methods. Qualitative is judgmental in nature whereas quantitative relies on hard data. By themselves, these two have nothing to do with the word quality. Quality of the forecast is an entirely different subject and depends on many factors.

Expert Opinion

- In this method, a group of experts from diverse background such as marketing, sales, finance, operations, and purchasing are asked to make forecast for the product under consideration. A consensus is then reached on a forecast figure. Each expert brings with him/her a set of biases, and perspectives that might influence the forecast. Of course, their judgment would be substantiated by a wealth of information that include past data, industry growth rates, competitive strategies and reactions from customers and distributors.
- The advantages of this method: 1) It is fast and efficient. 2) It is timely and based on good information content. 3) It uses the collective knowledge of experts.
- The disadvantages of this method: 1) Experts can make mistakes. 2) Subjectivity and bias of experts can vitiate the forecast. 3) The group dynamics of the experts could be greatly influenced by the degree of dominance of a particular person. He who could shout loudest, might get his way.

Market Survey

In this method, you conduct a market survey of customers' intentions to buy a product. A carefully designed questionnaire is administered to the selected target audience of customers. Customers are selected independently using a representative random sample. This method is very popular and if carefully implemented will give you good results.

- This is the apt technique to use, particularly if you want to forecast sales for a new product or new brand.
- This method of forecasting requires the active cooperation of the target audience.
- The sample size must be reasonably large. Larger the sample size, smaller will be the standard error and sampling error.
- Larger the sample size, the more time consuming and costly the survey will be. So, you have to strike a balance between sample size and cost.

Delphi Method

- In the expert opinion method of forecasting, a consensus forecast is arrived at after eliciting the opinion and views of experts with diverse background. Certainly this method is subject to group dynamics (effects). At times, judgments may be highly influenced by persuasions of some group members who have strong likes and dislikes. Delphi method attempts to retain the wisdom and accumulated knowledge of a group while simultaneously attempting to reduce the group effects.

- In Delphi method, group members are asked to make individual assessment about a forecast. These assessments are compiled and then fed back to the members, so that they get the opportunity to compare their judgment with others. They are then given an option to revise their forecasts. After three or four replications, group members reach their final conclusion.

Historical Analogy

This method is applied when a new product is about to be introduced by a company. Forecasting sales for new products are difficult in view of lack of proper historical data. Historical analogy method attempts to forecast sales for a new product based on the performance of related or similar products in the market place. The database of sales of these products forms the basis for forecasting.

The drawbacks of this method include:

- You cannot precisely say, how your new product is similar or related to a particular product.
- Suppose you have a number of products that you feel are similar to yours. Which of these will you consider as most similar to yours?
- Products that are similar to yours could have failed in the past for a variety of reasons. Let us say a similar product failed in the past because whenever there was an advertisement about this product, it was not available on the shelf. So, the consumers developed a negative perception about this product and became skeptical about its availability. You may not know all these and simply conclude your product will also fail!

Progressive test question Which one of the following will not affect the selection of a forecasting method?

- (a) Accuracy envisaged
- (b) Professional style of the organization
- (c) Data availability
- (d) Cost

Answer: (b) is the right choice. Selection of a forecasting method certainly will depend on (a) Accuracy, (c) Data availability (d) Cost. Selection of a forecasting method will not depend on the professional style of the organization.

Progressive test question What are the four methods of qualitative forecasting?

Answer:

1. Expert Opinion
2. Market Survey
3. Delphi method
4. Historical Analogy

Progressive test question

What is the paradox in forecasting?

Answer Because future is uncertain, no forecasting can be hundred percent accurate. This is a paradox in forecasting; on one hand you need sales forecasts, and on the other hand no forecast can be accurate.

Discussion Topic

Analyze, criticize, and explain the following statement:

" If at all forecasting techniques would be useful, they would be only in predicting sales for products. They have insignificant role in sales of services".

12.3 QUANTITATIVE METHODS OF FORECASTING

Quantitative forecasting uses statistical analysis of data to forecast sales. Time series analysis and causal model fall under the purview of quantitative forecasting. In this chapter, we will discuss time series analysis (projective methods) of forecasting and causal model that uses regression analysis for forecasting. Please note that we have already covered regression analysis extensively in chapter 10. You are expected to go through chapter 10 so that you will be able to appreciate regression method of forecasting when we discuss it.

Time Series Analysis

What is meant by time series data? The following visual captures the essence of time series.

What is Time Series?

Time series are series of observations that are taken at regular intervals of time. Data on weekly sales, monthly sales, and annual sales are examples of time series.

Like many other data sets, if you have a time series data set, the first step in analyzing it is to draw a graph, particularly a simple scatter diagram or a line graph that will reveal sharply any underlying patterns.

Figure 12.5

Time series is made up of four components

- > **Trend (T)** represents the long-term behavior of a time series. This would tell whether the time series data reveal a steady upward or downward movement.
- > **Seasonal Variation (S)** represents variation caused by season. Typically, this shows variation in demand during peak and lean season. For example, demand for snow tires will be at its peak during winter in USA.

- > **Cyclical Variation (C)** represents the typical business cycles that occur sporadically in several years. For example, in stock market, you will witness cycle of buoyancy or boom and cycle of recession that occur once in a while between many years.
- > **Random Variation (R)** represents irregular variations that occur by chance having no assignable cause. Random variation cannot be predicted.

Moving Average

The following visual portrays the essence of moving average:

Moving Average

The pattern revealed in observations vary over a time horizon. Instead of taking the average of all historical data, only the latest n periods of the data are used to get a forecast for the next period. This is the very essence of moving average forecast.

Moving Average (MA) Forecast for the next period = Average of n most recent time series data.

Figure 12.6

An illustration will help you understand how moving average (MA) works in practice.

Example A company is interested in forecasting demand for one of its products. Past data on demand for the last 12 months are available and given below: Using a period of 3 months, make a moving average forecast for period 13(13th month).

<i>Month</i>	<i>Sales (100 units)</i>
1	15
2	9
3	16
4	17
5	11
6	20
7	10
8	17
9	12
10	9
11	18
12	20

Spreadsheet Showing the Moving Average Calculation For the Example Problem

Month	Sales (100 units)	3 Monthly Moving Average
1	15	
2	9	
3	16	13.33
4	17	14.00
5	11	14.67
6	20	16.00
7	10	13.67
8	17	15.67
9	12	13.00
10	9	12.67
11	18	13.00
12	20	15.67

The first moving average is $(15 + 9 + 14)/3 = 13.33$. The second moving average is $(9 + 16 + 17)/3 = 14.00$. Likewise, all entries are filled in the above spreadsheet. As you will notice, the number of moving averages in column 3 are only 10 compared to the original number of observations of 12; so by using moving average method, you have lost 2 observations. This is inevitable in moving average method of forecast.

Moving Average using Microsoft Excel

We will work out the moving averages for the same example problem just discussed using Microsoft Excel.

Step 1 Click tools, click Data Analysis, and then click Moving Average. You get:

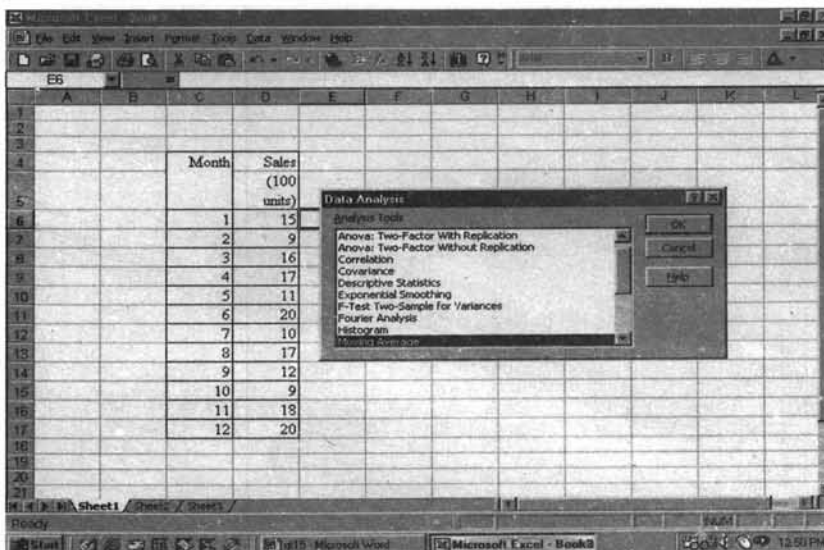


Figure 12.7

Step 2 Click OK, and then enter *Input Range* using the mouse by highlighting cells in column D from D6 to D17. Enter 3 in cell called *Interval* to indicate that 3 period (month) moving average is solicited. Specify the *Output Range* as E6. Click the *Chart Output* option to get the graph also. The screen will look like this.

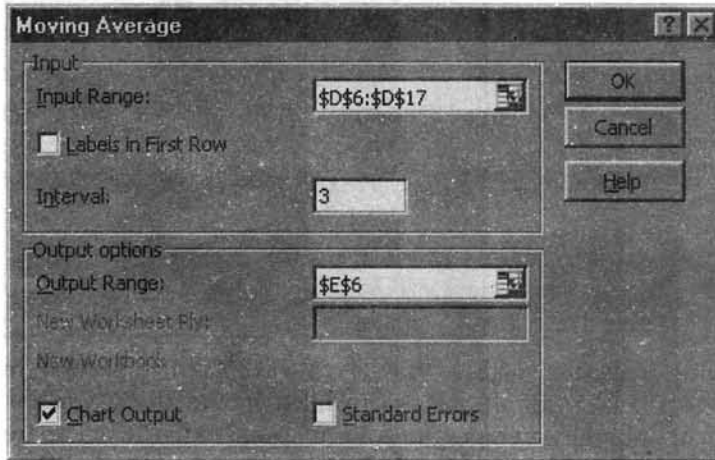


Figure 12.8

Please note that if you don't specify *Output Range*, Excel will give the output in a New Worksheet. So, you face no hassles. The flexibility and versatility provided by Excel is truly outstanding. You name any thing. It is there.

Step 3 Click OK and you get:

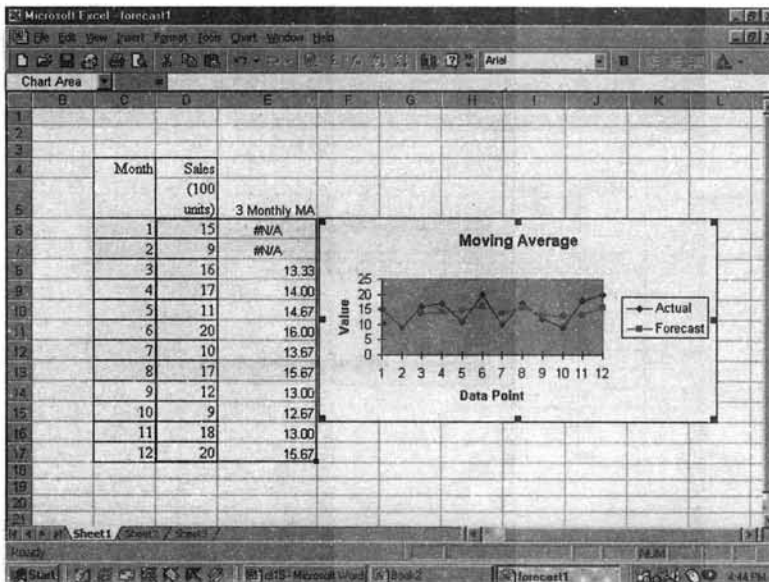


Figure 12.9

In column E, 3 month moving averages are worked out. Moving averages for months 1 and 2 are obviously not possible. The first moving average is $(15 + 9 + 14)/3 = 13.33$. The second moving average is $(9 + 16 + 17)/3 = 14.00$. Likewise all entries are filled in by Excel, automatically. As you will notice that the number of moving averages are only 10 in column E compared to the original number of observations of 12. So by using moving average method, you have lost 2 observations. This is inevitable in moving average method of forecast.

The forecast demand for month 13 = 15.67(moving average corresponding to month 12). Please note that the forecast for the next period is always the most recent moving average.

Progressive test question What is the moving average forecast demand for month 13 if you use a 4 monthly interval (period)?

Solution Using Microsoft Excel, we have the following solution:

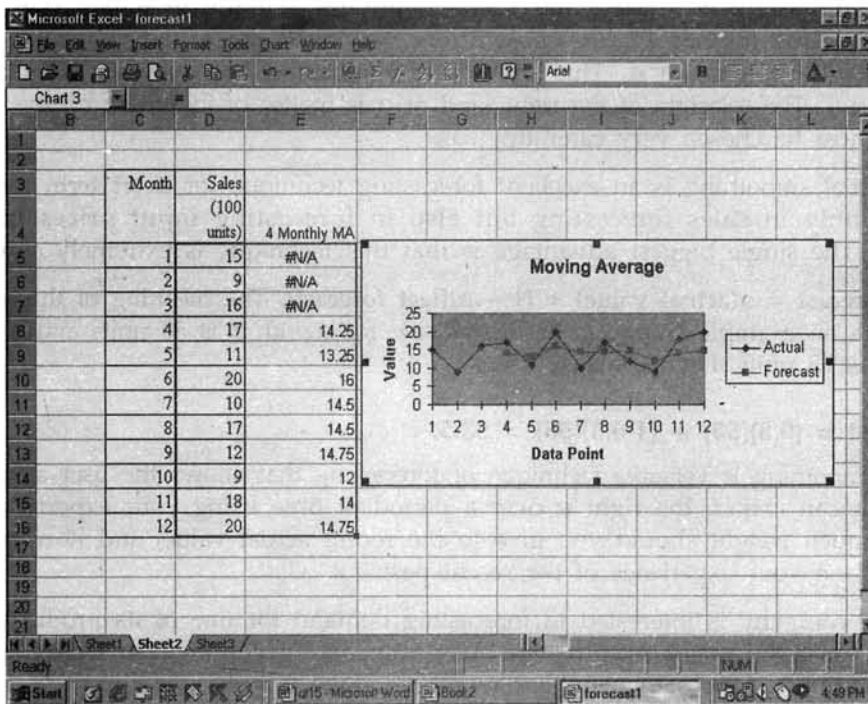


Figure 12.10

Forecast for month 13 is = 14.75 (Moving average corresponding to period 12 is 14.75).

Drawbacks of moving average forecast Moving average forecast is quick, easy and fairly inexpensive to implement. It provides a reasonably good forecast for the immediate future (very short term). However, practicing managers must remember the drawbacks given in the following visual.

Drawbacks of Moving Average Forecast

- Moving averages do not react well to seasonal variations
- All observations considered in a time horizon are given the same weight
- A large amount of historical data should be gathered and maintained to update forecast values
- The choice of the period(n) is generally arbitrary.

Figure 12.11

Exponential Smoothing

Exponential smoothing is a particular case of moving average in which there are three components - 1) the forecast for the most recent period, 2) the actual value for the period, and 3) a smoothing constant α . This smoothing constant α is a weighting factor that lies between 0 and 1. The selection of the right kind of α is matter of judgment by the experienced user. But, it must be chosen very carefully.

Exponential smoothing is an excellent forecasting technique for short term forecasting. It is used not only in sales forecasting but also in forecasting input prices in materials procurement. The single biggest advantage is that this technique is extremely simple to use.

New Forecast = α (actual value) + (1 - α)(last forecast). The meaning of this statement is explained with an example. Suppose, the actual sale for month 2 is 50 units and your forecast for month 2 is 55 units. Let us take $\alpha = 0.3$.

New Forecast = (0.3)(50) + (1-0.3)(55) = 53.5.

Exponential smoothing is versatile technique of forecasting that allows the user a great deal of flexibility. You can choose the right α over a period of time using your experience. You can decide how much weight should you give to the recent actual value, and how much to the forecast based on your experience of the recent past.

Example A company is interested in forecasting demand for one of its products. Past data on demand for the last 12 months are available and given below: Using exponential smoothing technique, forecast demand for month 13. Take $\alpha = 0.2$.

<i>Month</i>	<i>Sales (100 units)</i>
1	15
2	14
3	16
4	17
5	15

<i>Month</i>	<i>Sales (100 units)</i>
6	18
7	20
8	22
9	23
10	21
11	24
12	26

Spreadsheet Showing Basic Calculations

<i>Month</i>	<i>Sales (100 units)</i>	<i>Smoothing Values</i>	$\alpha = 0.2$
1	15		
2	14	15.00	
3	16	14.80	
4	17	15.04	
5	15	15.43	
6	18	15.35	
7	20	15.88	
8	22	16.70	
9	23	17.76	
10	21	18.81	
11	24	19.25	
12	26	20.20	

The smoothing values are appearing in the 3rd column. These are calculated using the formula **New Forecast = α (actual value) + (1 - α)(last forecast)**. Just for clarity, let me show a couple of smoothed values. There will be no smoothing value possible for the first month. For the second month, the smoothed value is taken as the previous month's actual. In this case, it is 15. Now, applying the formula,

$$\text{New Forecast} = \alpha(\text{actual value}) + (1 - \alpha)(\text{last forecast}),$$

you get on substitution,

New Forecast = $0.2(14) + (1 - 0.2)(15) = 14.8$. This is the forecast that appears in column three against month 3. For the 4th month, again apply the formula,

$$\text{New Forecast} = \alpha(\text{actual value}) + (1 - \alpha)(\text{last forecast}) = 0.2(16) + (1 - 0.2)(14.8) = 15.04.$$

This is the forecast for the 4th month. Proceeding in this manner, all the exponential smoothing forecasts for the remaining months have been worked out.

$$\text{Forecast for the 13th Month} = 0.2(\text{actual value}) + (1 - \alpha)(\text{last forecast})$$

$$= 0.2(26) + (1 - 0.2)(20.2) = 21.36. \text{ This is the demand forecast for the month 13.}$$

Smooth Exponential Smoothing Using Microsoft

Solution to the same problem

Step 1 In Microsoft Excel spreadsheet, click *Tools*, click *Data Analysis*, and click *Exponential Smoothing*. The following screen will appear.

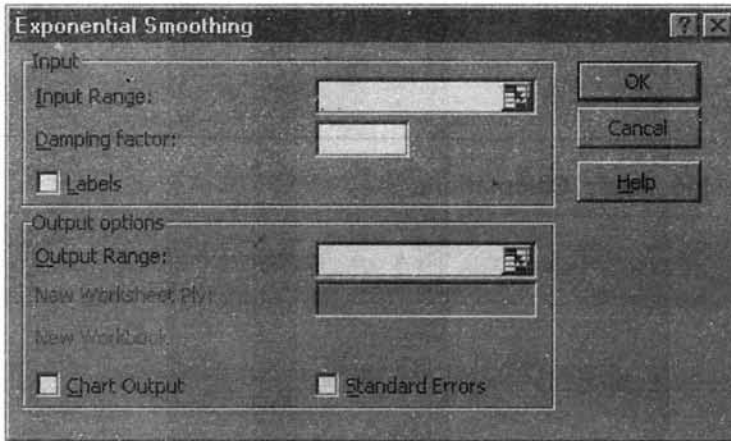


Figure 12.12

Step 2 In the prompt, highlight as usual Input Range. In the cell for *Damping factor*, enter 0.8 (This is Excel's way of asking for the value of $1-\alpha$). Please note that you have been given $\alpha = 0.2$. Highlight range for output, and then click OK. Before clicking OK, you can also click for *Chart Output* so that you get the graph of forecast values with actuals.

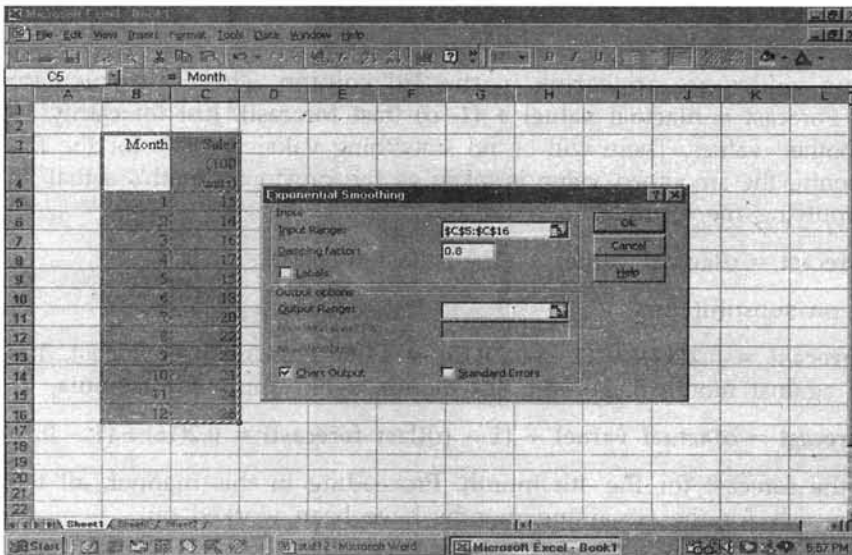


Figure 12.13

Step 3 Click OK, and you get everything you want.

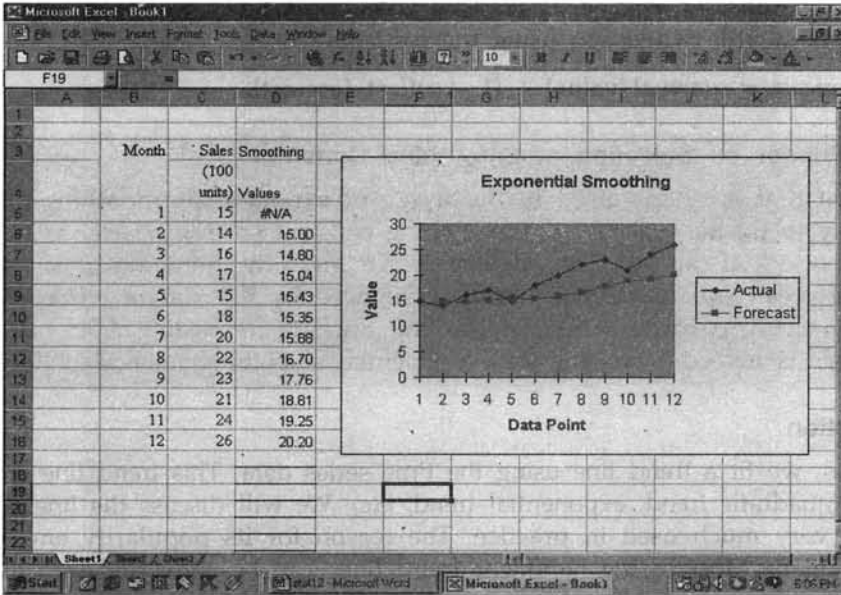


Figure 12.14

This is exactly identical to the results you have got by actual calculation method. Please see exponential smoothing forecast values appearing in column D.

To forecast for the month of 13, bring the mouse to cell D16 and click. Then click *Edit*, click *Copy*. Bring the mouse to cell D17 and then click *Paste* icon. You get the answer. The answer will appear in cell D17. This answer screen is shown below:

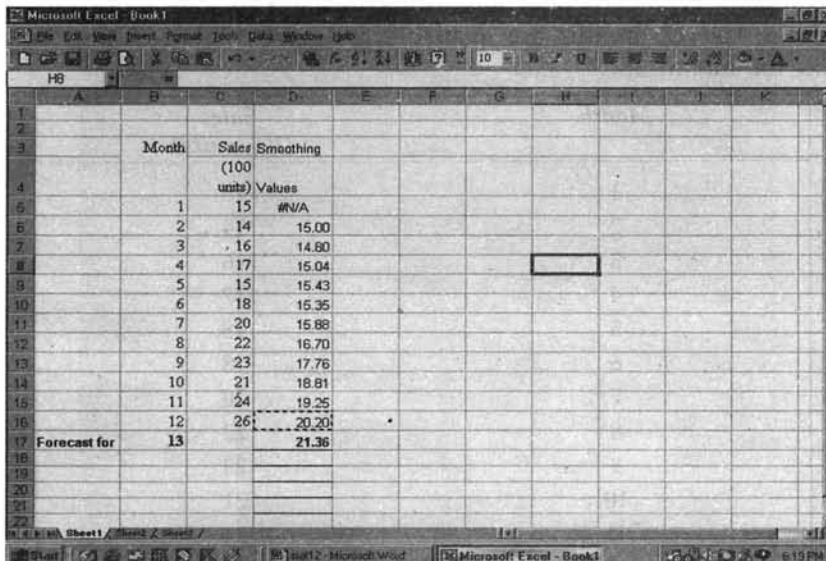


Figure 12.15

You can see in the above Excel output, Forecast for the month of 13, appears in cell D17 in bold. The answer seen is **21.36**. This is same as what you have got by the actual calculation done a little while ago using the formula,

$$\text{New Forecast} = \alpha(\text{actual value}) + (1 - \alpha)(\text{last forecast}).$$

Points to Ponder on Selection of Smoothing Constant

The question that is often raised in the usage of exponential smoothing is, is there a scientific way to fix the value of the smoothing constant α ? The answer is both, Yes and No. **Yes**, because you can get that value of α that gives the minimum mean square error. **No**, because mean square error is highly influenced by the square terms of individual errors. Judgment of α based on experience and tracking forecasting efficiency is the only way out. This is indeed a weakness of exponential smoothing method of forecasting.

Trend Projection

In this method, we fit a trend line using the time series data. This trend line could be linear or nonlinear (quadratic trend, exponential trend, etc). We will discuss the linear trend that is popular, and very much used in practice. The reason for its popularity emerges from the following rationale. Most of the nonlinear trend lines can be converted into linear lines by mathematical transformation. The linear trend is a reasonable good approximation of trend pattern that is revealed by time series data.

In simple terms, we fit an equation of the form $Y = a + bt$ using the method of least square. Please note that least square method is used in regression analysis that was discussed in chapter 10. In this equation, Y is the dependent variable and t is the independent variable. In other words, you assume that forecast values of Y will be shaped by the past pattern only. The historical trend continues.

Example A company is interested in forecasting demand for one of its products. Past data on demand for the last 12 months are available and given below: Using linear trend line, forecast demand for month 13.

<i>Month</i>	<i>Sales (100 units)</i>
1	15
2	14
3	16
4	17
5	15
6	18
7	20
8	22
9	23
10	21
11	24
12	26

You have two options. Option 1 is to use the formula approach involving a simple linear regression model covered in Chapter 10. The second option is to use Data Analysis of Excel.

If you use formula approach, the formula will have t in the place of X as covered in Chapter 10.

$$b = \frac{\sum (t - \bar{t})(Y - \bar{Y})}{\sum (t - \bar{t})^2}$$

$$a = \bar{Y} - b\bar{t}$$

The basic calculations are shown in the following spreadsheet.

Month	Sales (100 units)	$(t - \bar{t})$	$(Y - \bar{Y})$	$(t - \bar{t})(Y - \bar{Y})$	$(t - \bar{t})^2$
1	15	-5.5	-4.25	23.3750	30.2500
2	14	-4.5	-5.25	23.6250	20.2500
3	16	-3.5	-3.25	11.3750	12.2500
4	17	-2.5	-2.25	5.6250	6.2500
5	15	-1.5	-4.25	6.3750	2.2500
6	18	-0.5	-1.25	0.6250	0.2500
7	20	0.5	0.75	0.3750	0.2500
8	22	1.5	2.75	4.1250	2.2500
9	23	2.5	3.75	9.3750	6.2500
10	21	3.5	1.75	6.1250	12.2500
11	24	4.5	4.75	21.3750	20.2500
12	26	5.5	6.75	37.1250	30.2500
6.5	19.25			149.5000	143.0000

Please note that the values in the first two columns of the bottom row are the respective values of \bar{t} and \bar{Y} . The bottom row of the 5th and 6th columns, respectively, represent $(t - \bar{t})(Y - \bar{Y})$, and $(t - \bar{t})^2$.

$$b = \frac{\sum (t - \bar{t})(Y - \bar{Y})}{\sum (t - \bar{t})^2} = (149.50/143.00) = 1.045$$

$$a = \bar{Y} - b\bar{t} = 19.25 - 1.0455 \cdot 6.5 = 12.45$$

So, the fitted line is given by,

$$Y = 12.45 + 1.045t$$

Forecast demand for month 13 = $12.45 + (1.045)(13) = 26.04$ (units of 100)

Solution Using Excel: Click Tools, click *Data Analysis*, and click *Regression*.

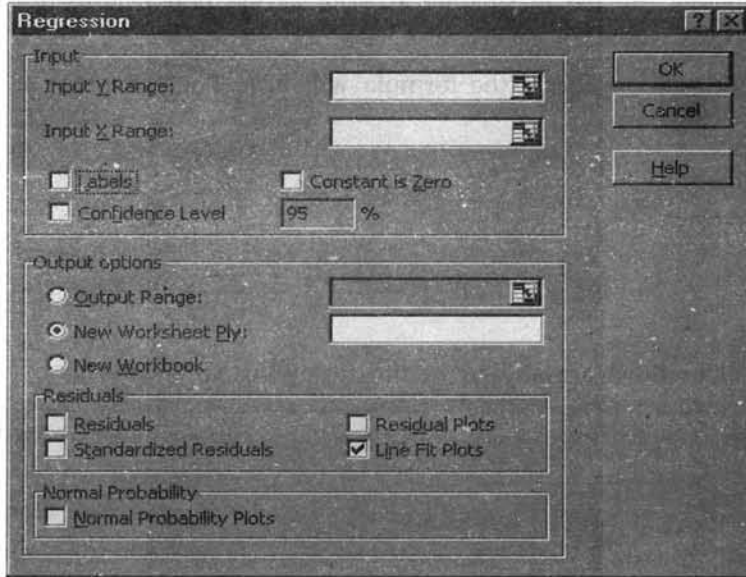


Figure 12.16

Enter data for Y range and X range just like what you have done in chapter 10. Click with the mouse *Line Fit Plots* and click OK, you get:

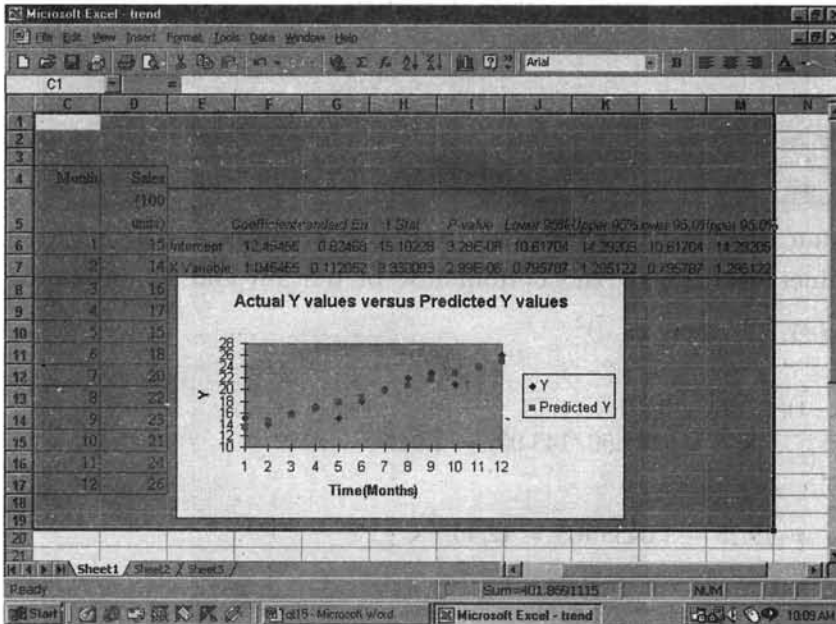


Figure 12.17

As you can see, intercept = 12.45 and slope = 1.045. So, the fitted line is given by $Y = 12.45 + 1.045t$ (Please note that t is the independent variable, while Excel has set X as independent variable by default).

$$\text{Forecast demand for month 13} = 12.45 + (1.045)(13) = 26.04$$

What happens if the trend equation is not linear?

Two Nonlinear trend equations:

1. Suppose the trend equation is of the form $Y = a + bt + ct^2$.

How will you project future trend? You let $t^2 = u$. This will make the equation,

$Y = a + bt + cu$. This is a typical multiple regression model that can be solved using Microsoft Excel.

2. Suppose $Y = ae^{bt}$.

Take Log on both sides, this becomes $\text{Log } Y = \text{Log } a + bt$. This is of the form, $Z = A + bt$ which is a simple linear regression model that can be solved either using formula approach, or Microsoft Excel.

Forecasting Using Multiple Regression Model

Multiple Linear Regression Model Whenever you are interested in the combined influence of several independent variables upon one dependent variable that you want to forecast, your model is that of multiple regression. Demand for example, may be a function of price, income of the consumer, advertising expense, industrial growth, and competitor's price. When all these independent variables change, what happens to the demand projection is a study of multiple linear regression.

Things to do in a Multiple Linear Regression Model

- > Postulate the model $Y = a + bX_1 + cX_2 + dX_3 + \dots$
- > Enter the sample data for X_s and Y in Microsoft Excel.
- > Perform the Regression Analysis and get the summary output from Excel.
- > Write the Regression Equation using the intercept and coefficient of X_s from Excel summary output. Predict Y for given X_s .
- > Validate the model statistically by looking at R^2 as well as F statistic in the ANOVA that tests the null hypothesis of no linear relationship.
- > After statistical validation, use the model for estimation and prediction.

Case Example-Sales Forecasting

To measure the effect of advertising and sales promotional efforts, the following data were collected from a consumer marketing company for the last 10 months. Figures in the following table are in \$1000.

Month	Sales(Y)	Advertisement Expense (X1)	Sales Promotional Expense (X2)
1	200	45	15
2	250	50	20
3	300	55	24
4	650	85	45
5	400	65	30
6	300	55	25
7	320	57	27
8	450	68	32
9	350	60	28
10	550	70	37

Answer the questions below, using Microsoft Excel:

1. Set up a regression model by taking Sales (Y) as the dependent variable and advertisement expense (X1) and sales promotion expense (X2) as independent variables and validate the model using R² value.
2. Forecast X1 and X2 for month 11 by using exponential smoothing technique.
3. Now forecast sales for month 11 using results obtained in 2.

Solution Invoke Data Analysis Pack of Microsoft Excel and enter the data as required under Regression and execute the model. You get the following output:

(Please note that the details of each step are covered in Chapter 10 and are not repeated here. Only the final output is given).

Month	Sales(Y)	Advertisement Expense (X1)	Sales Promotional Expense (X2)																												
1	200	45	15	<table border="0"> <tr><td colspan="3">SUMMARY OUTPUT</td></tr> <tr><td colspan="3"><i>Regression Statistics</i></td></tr> <tr><td>Multiple R</td><td>0.987153</td><td><i>Coefficients</i></td></tr> <tr><td>R Square</td><td>0.974471</td><td>Intercept</td><td>-195.761</td></tr> <tr><td>Adjusted R Square</td><td>0.967177</td><td>X Variable 1</td><td>5.033966</td></tr> <tr><td>Standard Error</td><td>25.16275</td><td>X Variable 2</td><td>9.388292</td></tr> <tr><td>Observations</td><td>10</td><td></td><td></td></tr> </table>			SUMMARY OUTPUT			<i>Regression Statistics</i>			Multiple R	0.987153	<i>Coefficients</i>	R Square	0.974471	Intercept	-195.761	Adjusted R Square	0.967177	X Variable 1	5.033966	Standard Error	25.16275	X Variable 2	9.388292	Observations	10		
SUMMARY OUTPUT																															
<i>Regression Statistics</i>																															
Multiple R	0.987153	<i>Coefficients</i>																													
R Square	0.974471	Intercept	-195.761																												
Adjusted R Square	0.967177	X Variable 1	5.033966																												
Standard Error	25.16275	X Variable 2	9.388292																												
Observations	10																														
2	250	50	20																												
3	300	55	24																												
4	650	85	45																												
5	400	65	30																												
6	300	55	25																												
7	320	57	27																												
8	450	68	32																												
9	350	60	28																												
10	550	70	37																												

You can see, from the output, the following:

$Y = -195.76$ (2 places of decimal taken)

$X1 = 5.03, X2 = 9.39$. So the fitted model is $Y = -195.76 + 5.03X1 + 9.39X2$.

The model has a good accuracy level as evident from the R² value that is quite high,

= 0.97 (two places of decimal). Even the **adjusted $R^2 = 0.97$** , indicating the robustness of the model to predict. In other words, R^2 value is close to 1, and hence, the model is reliable in forecasting. This answers (part 1) of the question.

2) Invoke Exponential Smoothing under Data Analysis Pack. Enter the input data for X1 and X2 separately. Use a dampening factor of 0.7 (same as Alpha = 0.3). The following output emerges. (Please note that the step-by step approach is covered earlier in this Chapter. Only the final output is given)

Exponential Smoothing Forecast Values	
X1	X2
45.00	15.00
46.50	16.50
49.05	18.75
59.84	26.63
61.38	27.64
59.47	26.85
58.73	26.89
61.51	28.42
61.06	28.30

The smoothed values will start from period 2 only in Excel's Exponential smoothing and will not be available for period 1.

Forecast of X1 for period 11 = Alpha (actual value in period 10) + (1-Alpha)(forecast of period 10) = 0.3(70) + 0.7(61.06) = 63.74.

Forecast of X2 for period 11 = Alpha (actual value in period 10) + (1-Alpha) (forecast of period 10) = 0.3(37) + 0.7(28.3) = 30.91.

This answers (part 2) of the question.

3. Forecast of sales for period 11 is obtained by substituting the values of X1 and X2 (obtained in the previous part given above) in the regression equation.

$$Y = -195.76 + 5.03X_1 + 9.39X_2.$$

Forecast of sales for period 11 = $-195.76 + 5.03(63.74) + 9.39(30.91) = 415.1$. Since sales are in \$ 1000, the forecast for period 11 is a sale of **\$415100**.

Limitations of Using Multiple Regression Model for Forecasting

- The most crucial assumption made is that the independent variables are not correlated with each other. If they are correlated, then the regression coefficients cannot be estimated. This problem is called multicollinearity. The procedure followed for resolving multicollinearity is to drop the independent variable that has the highest standard deviation and then rework the model again. You may also like to use two-stage least square method that is part of econometrics. The other way is to transform a set of correlated independent variables into an uncorrelated set of variables by the

technique called principal component analysis. This is an advanced technique requiring the help of advanced statistical software, like SPSS.

- When there are wild fluctuations in one or more of the independent variables, multiple regression model crumbles, and will be highly unreliable.
- In order to use the multiple regression model for prediction, you have to first predict the values of the independent variables using some other prediction method.
- In forecasting problems, multiple regression, at best, can work for short and medium term only. It cannot be successfully used for long term forecasting.

A Brief Note on Accuracy of Forecast

- We have discussed a number of forecasting techniques in this chapter. Needless to say, accuracy of forecast is paramount. Accuracy measures must reflect the closeness of predicted values with the actual values. Closer the predicted value to actual, greater is the accuracy. Backtracking ability of forecast is measured by the behavior of forecast values towards the actual. In all time series forecast methods that we have discussed so far, we have provided graphical display of predicted versus actual values to understand the backtracking ability of forecast model under consideration.
- Another point to be noted is that suitable adjustments should be made in the forecast figures arrived at. This would include adjustments for seasonality, and cycles. For example, you have the trend projections based on least square line. You have made a projection for the coming period. This projection figure will have to be suitably modified if there is a strong seasonality. You can easily establish seasonal index for each calendar month. This is obtained by dividing the actual value by the corresponding trend value. If you continuously maintain a large database, seasonal indices could be updated in a dynamic manner. All you do is, to first project the trend value for the forecast period by using least square method, or moving average, and then multiply this trend value by the corresponding seasonal index; you get a forecast adjusted for seasonality.
- All these measures will improve your accuracy. There are two methods in practice that are used for understanding **forecast error**. 1) **Average Absolute Error**- This is obtained by computing the absolute difference between forecast value and actual value for every element in the time series data set, and then taking the average of all these values. 2) **Average Percentage Relative Error**- In this method, you first compute the absolute difference between forecast value and actual value for every element in the data set, and then divide each one of them by the corresponding actual value. Take the average of such values and multiply by 100. You get average percentage error. Which of these methods you choose, is a matter of judgment.
- Intuitively and logically, the graph of forecast values should be reasonably close to the actual values. If it is not, look for reasons and gather more data. Revise your model completely, if needed.

- Accuracy can be greatly improved if you have a large amount of historical data. This will permit you to use advanced forecasting techniques, like Box-Jenkin method, Adaptive Filtering, and Econometric models of forecasting. The discussion of these is beyond the scope of this chapter. Those interested can refer to books on Econometrics for treatment on these advanced topics.

12.4 CHAPTER SUMMARY

This chapter has provided a conceptual framework on various forecasting techniques with their strengths and limitations. Specifically, this chapter focused on:

- The need for forecasting.
- Schematic diagram giving classification of forecasting techniques in practice.
- Guidelines for selecting a forecasting method.
- Qualitative or judgmental forecasting split into expert opinion, market survey, Delphi method, and historical analogy.
- Quantitative forecasting split into time series analysis and causal method. Detailed coverage on time series analysis as well causal model involving regression.
- Time series models split into moving average, exponential smoothing, and trend projection using least square line.
- Forecasting accuracy and associated measures on forecast error.
- Use of Microsoft Excel to compute moving averages, exponential smoothing, trend line based on least square and multiple regression forecasting.

GLOSSARY

Average Absolute Error This is obtained by computing the absolute difference between forecast value and actual value for every element in the time series data set, and then taking the average of all these values.

Average Percentage Error In this method, we first compute the absolute difference between forecast value and actual value for every element in the data set, and then divide each one of them by the corresponding actual value. We then take the average of such values and multiply by 100 to get average percentage error.

Causal Forecasting Model It is a statistical model that establishes functional relationship between the dependent variable to be forecast, and one or more independent variables. This is same as regression model.

Cyclical Variation It is one of the components of a time series, which represents the typical business cycles that occur sporadically in several years.

Delphi Method It is one of the qualitative forecasting techniques that solicits the opinion and accumulated knowledge of a group, while simultaneously attempting to reduce the group dynamics.

Expert Opinion Method This is one of the qualitative methods of forecasting in which a group of experts from diverse background such as marketing, sales, finance, operations, and purchasing are asked to make forecast for the product under consideration. Then a consensus is reached on a forecast figure. Each expert brings with him/her a set of biases, and perspective that might influence the forecast. This technique gets affected by group dynamics.

Exponential Smoothing It is a time series forecasting technique which can be considered as a particular case of moving average in which there are three components - 1) the forecast for the most recent period, 2) the actual value for the period, and 3) a smoothing constant. This smoothing constant is a weighting factor that lies between 0 and 1.

Forecast Error It is the difference between the actual values and the predicted values in a forecasting problem.

Historical Analogy Method This method is applied when a new product is about to be introduced by a company. Forecasting sales for new products are difficult in view of lack of proper historical data. Historical analogy method attempts to forecast sales for a new product based on the performance of related or similar products in the market place.

Market Survey Method In this method, we conduct a market survey of customers' intentions to buy a product. A carefully designed questionnaire is administered to the selected target audience of customers. Customers are selected independently using a representative random sample. This method is very popular, and if carefully implemented, will give you good results.

Mean Square Error This is computed by the sum squares of deviations between the actual values and the predicted values divided by the number of observations.

Moving Average It is a time series forecasting method in which, the forecast for the next period is the average of the n most recent observations.

Qualitative Forecast It is a subjective forecasting technique based on judgment, expert opinion, and experience.

Quantitative Forecast This forecast is based on statistical analysis of data. Time series analysis and causal model fall under the purview of quantitative forecasting.

Random Variation It represents irregular variation in a time series that occur by chance, having no assignable cause. Random variation cannot be predicted.

Seasonal Variation It represents variation caused by season. Typically, this shows variation in demand during peak and lean season.

Time Series These are series of observations taken at regular intervals of time.

Trend It represents the long-term behavior of a time series. This would tell whether the time series data reveal a steady upward or downward movement.

Trend Projection This is a time series forecasting method in which we fit a function

using the method of least squares. In this function the response variable to be forecast is the dependent variable and time is the independent variable.

REVIEW QUESTIONS

1. Moving average method of forecasting requires a large amount of historical data. True or False.
2. What are the four components of a time series?
3. What are the guidelines for selecting a forecasting technique?
4. Selection of the period in moving average method is generally very scientific. True or False.
5. Exponential smoothing method is very apt for short term forecasting. True or False.
6. What techniques could you use to forecast sales if a new product is to be introduced in the market?
 - (a) Historical analogy
 - (b) Delphi method
 - (c) Market survey
 - (d) Expert opinion
 - (e) All of the above
7. Multiple regression model is suitable for long term forecasting. True or False.

ANSWERS TO REVIEW QUESTIONS

1. The statement is true. The accuracy of the forecast to be obtained using moving average requires a large amount of historical data.
2. 1) Trend 2) Seasonal variation 3) Cyclical variation and 4) Random variation.
3. Availability of data, Accuracy envisaged, Urgency with which the forecast is sought, and Cost.
4. The statement is false because in moving average method, selection of the period is generally arbitrary.
5. The statement is true. Exponential smoothing works very well for short-term projections.
6. (e) is the right choice. All qualitative methods of forecasting (*a, b, c, d*) can be used depending on the situation.
7. The statement is false. Multiple regression at best, can work for short and medium term only. It cannot be successfully used for long term forecasting.

PRACTICE PROBLEMS

1. Case Study-Demand Forecasting

To measure the effect of advertising and price of the product on the demand pattern, the following data were collected from a consumer marketing company for the last 10 months.

<i>Month</i>	<i>Sales(Y) In units</i>	<i>Advertisement Expense (X1) Rs Lacs</i>	<i>Price (X2) In Rs</i>
1	300	45	2400
2	350	50	2200
3	400	55	2100
4	500	85	2000
5	400	65	2150
6	450	60	2000
7	420	57	2150
8	550	68	1950
9	450	60	2050
10	500	70	2000

QUESTIONS

1. Set up a regression model by taking Sales (Y) as the dependent variable and advertisement expense (X1) and price (X2) as independent variables. Validate the model using R^2 value.
2. Forecast X1 and X2 for month 11 by using exponential smoothing technique.
3. Now forecast sales for month 11 using results obtained in 2).

2. Case Study - Sales Forecast Using Time Series

The following data refer to the past 12 months sales of a consumer product.

<i>Month</i>	<i>Demand</i>
1	16
2	14
3	12
4	15
5	18
6	21
7	23
8	24
9	25
10	26
11	37
12	38

QUESTIONS

1. Forecast demand for the 13th month using 3 monthly, 6 monthly, and 9 monthly moving averages.
2. Fit a least square linear trend to the data and project for the month of 13.
3. Use exponential smoothing technique and forecast the demand for month 13. Take $\alpha = 0.3$.

3. Case Study-Commercial Vehicle Sales

The following data refer to the sales of commercial vehicles at the All India Level of a leading automobile company in the country during three financial years (April to March).

Sales Figures in Numbers

	<i>Year 1</i>	<i>Year 2</i>	<i>Year 3</i>
APR	724	1414	1243
MAY	1440	2117	1818
JUN	1606	2199	2880
JUL	1656	2583	1693
AUG	1549	2358	2136
SEP	2285	3677	3707
OCT	1523	1823	1931
NOV	1856	2372	1637
DEC	2135	2301	1746
JAN	2119	2761	2638
FEB	2075	2110	2655
MAR	3850	3996	3576

Questions

1. Draw the time series graph depicting the comparative sales for the three years.
2. Compute a 12 monthly moving averages and plot the graph of the moving averages.
3. What is the forecast for April in the fourth year?
4. **Case Study- Sales Projection Using Past Data**

The following data are the sales of a company in the past sixteen years. The company is interested in analyzing the data in the context of business planning for the next three years. In particular, the company would like to study the pattern of sales emerging from the data in order to project the sales for the next three years in a reasonable manner.

<i>Year</i>	<i>Sales In Rs. Crores</i>
1	26.20
2	35.40
3	39.20
4	45.80
5	49.00
6	50.40

<i>Year</i>	<i>Sales In Rs. Crores</i>
7	54.80
8	60.00
9	71.80
10	77.60
11	81.40
12	84.60
13	96.80
14	100.60
15	107.60
16	112.00

QUESTIONS

1. Forecast the sales for the next three years using the least square linear trend line.
2. Comment on the validity of the model by performing appropriate analysis.

References

RELATED INTERNET SITES

<http://www.davidmlane.com/hyperstat/>

<http://www.prenhall.com/berenson/>

<http://anu.edu.au/nceph/surfstat/surfstat-home/course.html>

<http://www.bmj.com/statsbk/>

<http://www.statsoft.com/textbook/stathome.html>

<http://vassun.vassar.edu/~lowry/webtext.html>

<http://www.stat.berkeley.edu/users/stark/SticiGui/Text/index.htm>

<http://www2.chass.ncsu.edu/garson/pa765/statnote.htm>

<http://www.vishstat.com/>

BOOKS

1. Anderson, Sweeney and Williams, *Statistics for Business and Economics*, South- Western College Publishing
2. Levin and Rubin, *Statistics for Management*, Prentice Hall
3. Levine and Berenson, *Basic Business Statistics*, Prentice Hall
4. Snedecor and Cochran *Statistical Methods*, Iowa University press
5. Levine and Berenson, *Statistics for Managers using Microsoft Excel*, Prentice Hall

Test Your Knowledge on Business Statistics

CONCEPT QUESTIONS

CHAPTER 1

1. What do you mean by the term Statistics?
2. Distinguish between Descriptive Statistics and Inferential Statistics.
3. Explain the terms Population and Sample.
4. Explain with examples the Scales of Measurement.
5. Distinguish between Primary Data and Secondary Data.
6. What are the steps involved in a Statistical Investigation?
7. Justify why should you study Statistics?

CHAPTER 2

1. Distinguish between Data and Information.
2. Why should you Classify Data?
3. What is a Frequency Distribution?
4. What are the guidelines involved in constructing a Frequency Distribution?
5. What is a Histogram? What are its uses?
6. What is meant by the term Relative Frequency?
7. Explain the term Cumulative Frequency Distribution.
8. What is a Frequency Polygon?
9. What tools do you use to identify Pattern in Data Set?

CHAPTER 3

1. What do you mean by Central Tendency?
2. Explain with examples the three popular Measures of Central Tendency.
3. Compare and contrast Mean, Median, and Mode.
4. What is Dispersion?
5. What are the strengths and limitations of Range, Mean Absolute Deviation (MAD), and Standard Deviation?
6. How do you measure Consistency in performance?
7. Explain with examples the term Coefficient of Variation (CV).

CHAPTER 4

1. What do you understand by the term Probability?
2. What are the properties of probability?
3. What is a Sample Space?
4. What are the Types of Probability?
5. What is the meaning of Equally Likely?
6. What is a Random Variable?
7. What do you mean by Mutually Exclusive Events?
8. What do you mean by Independent Events?
9. Explain with the help of an example the Addition Rule of probability.
10. Explain with an example the Multiplication Rule of probability.
11. What is the meaning Conditional Probability?
12. Explain with an example the concept of a Probability Tree.

CHAPTER 5

1. What is a Probability Distribution?
2. What do you mean by Discrete Probability Distribution?
3. What do you mean by Continuous Probability Distribution?
4. What is an Observed Distribution?
5. Describe the conditions for using the Binomial Distribution.
6. Describe the uses of Binomial Distribution.
7. What are the uses of Poisson Distribution?
8. When can Binomial Distribution be approximated to Poisson Distribution?
9. Describe the properties of Normal Distribution.

10. What are the uses of Normal Distribution?
11. What is a Standard Normal Distribution?
12. When can Binomial Distribution be approximated to Normal Distribution?
13. What is the Parameter of Binomial Distribution?
14. What is the Parameter of Poisson Distribution?
15. What are the Parameters of Normal Distribution?

CHAPTER 6

1. What is the connection between Sample and Population (Universe)?
2. What is Census?
3. What is Sampling?
4. Why do you need Sampling?
5. What are the Methods of Sampling that you are aware of ? Describe each one of them briefly.
6. Distinguish between Probability Sampling and Non-Probability Sampling.
7. What are the problems in using Non-Probability Sampling in practice?
8. What do you mean by Sampling Distribution?
9. Explain Central Limit Theorem and its role.
10. What is Standard Error?
11. What is Sampling Error?

CHAPTER 7

1. Why do you need Estimation?
2. What is the difference between an Estimator and an Estimate?
3. What is a Point Estimate? Give an example.
4. What is an Unbiased Estimator?
5. Explain the concept of Interval Estimation with an example in the case of Population Mean and Population Proportion.
6. What is the role of Normal Distribution in setting up a Confidence Interval?
7. What are the similarities and dissimilarities between t Distribution and Normal Distribution?
8. What do you mean by the term "Degrees of Freedom"?
9. What is the role of t Distribution in setting up a Confidence Interval?
10. Explain with examples the method of determining the Sample Size using Confidence Interval in the case of estimating Population Mean and Population Proportion for any statistical survey.

CHAPTER 8

1. What is a Hypothesis? Give an example.
2. Explain Null Hypothesis and Alternative Hypothesis in the context of structuring Hypotheses.
3. What are Type I and Type II Errors?
4. What is Power of a Test?
5. What is Level of Significance?
6. How do you decide whether to Reject or Accept a null hypothesis?
7. Explain what do you mean by One-Tail Test and Two-Tail Test.
8. Explain with examples the Univariate Hypothesis Tests for Large Sample and Small Sample.
9. Explain with examples the Bivariate Hypothesis Tests for Large Sample and Small Sample.
10. Explain the term P-Value.
11. Pictorially explain the Acceptance Region and Rejection Region.

CHAPTER 9

1. Why Chi-Square test is considered Non-Parametric?
2. What are the conditions for applying Chi-Square Test?
3. Explain with an example the role of Chi-Square Test of Goodness in the univariate case.
4. Explain with an example the role of Chi-Square Test in the context of Test of Independence in a Contingency Table.
5. What do you understand by Analysis of Variance(ANOVA)?
6. What is the specialty of ANOVA?
7. Is ANOVA a must?
8. Explain with an example One-way Classification ANOVA.
9. Explain with an example Two-Way Classification ANOVA.
10. What are the assumptions involved in ANOVA?

CHAPTER 10

1. What is Correlation? Give an example.
2. What are the properties of Correlation Coefficient of Karl Pearson?
3. What is Regression? Give an example.
4. What is a Scatter Diagram? What are its uses?
5. Explain Simple Linear Regression with an example.

6. Explain Multiple Linear Regression with an example.
7. What are the uses of Regression?
8. What is Coefficient of Determination? Describe its role in regression.
9. What are the assumptions involved in linear regression?
10. What are the problems in using multiple regression?

CHAPTER 11

1. What are the Steps involved in Systematic Problem Solving?
2. What is Pay-Off Matrix?
3. Explain the role of Expected Monetary Value (EMV) in decision analysis.
4. What is Expected Value of Perfect Information (EVPI)? What is its role?
5. What is Expected Value of Sample Information (EVSI)? What is its role?
6. What is a Decision Tree? What are its uses?
7. Explain the connection between EVPI and EOL.
8. How do you structure a Decision Problem?
9. What do you mean by Folding Back the Tree?

CHAPTER 12

1. What is Forecasting?
2. Why do you need Forecasting?
3. What are the Forecasting Methods available in practice?
4. What are the Guidelines for selecting the Right Forecasting Technique?
5. Distinguish between Qualitative Methods of Forecasting and Quantitative Methods of Forecasting.
6. Describe the Role of each of the qualitative methods of forecasting.
7. What is a Time Series?
8. What are the Components of Time Series?
9. Explain with an example Moving Average Forecasting.
10. Explain with an example Exponential Smoothing Method of Forecasting.
11. What are the Drawbacks of Moving Average Forecasting?
12. Describe Trend Projection.
13. Explain with an example Multiple Regression Forecasting Model.
14. What are the limitations of Multiple Regression Forecasting Model?
15. Write a brief note on Accuracy of Forecast.

		p									
n	x	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
6	0	0.9415	0.8858	0.8330	0.7828	0.7351	0.6899	0.6470	0.6064	0.5679	0.5314
	1	0.0571	0.1085	0.1546	0.1957	0.2321	0.2642	0.2922	0.3164	0.3370	0.3543
	2	0.0014	0.0055	0.0120	0.0204	0.0305	0.0422	0.0550	0.0688	0.0833	0.0984
	3	0.0000	0.0002	0.0005	0.0011	0.0021	0.0036	0.0055	0.0080	0.0110	0.0146
	4	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	0.0008	0.0012
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0	0.9321	0.8681	0.8080	0.7514	0.6983	0.6485	0.6017	0.5578	0.5168	0.4783
	1	0.0659	0.1240	0.1749	0.2192	0.2573	0.2897	0.3170	0.3396	0.3578	0.3720
	2	0.0020	0.0076	0.0162	0.0274	0.0406	0.0555	0.0716	0.0886	0.1061	0.1240
	3	0.0000	0.0003	0.0008	0.0019	0.0036	0.0059	0.0090	0.0128	0.0175	0.0230
	4	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0011	0.0017	0.0026
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0	0.9227	0.8508	0.7837	0.7214	0.6634	0.6096	0.5596	0.5132	0.4703	0.4305
	1	0.0746	0.1389	0.1939	0.2405	0.2793	0.3113	0.3370	0.3570	0.3721	0.3826
	2	0.0026	0.0099	0.0210	0.0351	0.0515	0.0695	0.0888	0.1087	0.1288	0.1488
	3	0.0001	0.0004	0.0013	0.0029	0.0054	0.0089	0.0134	0.0189	0.0255	0.0331
	4	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0013	0.0021	0.0031	0.0046
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0004
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0	0.9135	0.8337	0.7602	0.6925	0.6302	0.5730	0.5204	0.4722	0.4279	0.3874
	1	0.0830	0.1531	0.2116	0.2597	0.2985	0.3292	0.3525	0.3695	0.3809	0.3874
	2	0.0034	0.0125	0.0262	0.0433	0.0629	0.0840	0.1061	0.1285	0.1507	0.1722
	3	0.0001	0.0006	0.0019	0.0042	0.0077	0.0125	0.0186	0.0261	0.0348	0.0446
	4	0.0000	0.0000	0.0001	0.0003	0.0006	0.0012	0.0021	0.0034	0.0052	0.0074
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	0.0008
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0	0.9044	0.8171	0.7374	0.6648	0.5987	0.5386	0.4840	0.4344	0.3894	0.3487
	1	0.0914	0.1667	0.2281	0.2770	0.3151	0.3438	0.3643	0.3777	0.3851	0.3874
	2	0.0042	0.0153	0.0317	0.0519	0.0746	0.0988	0.1234	0.1478	0.1714	0.1937
	3	0.0001	0.0008	0.0026	0.0058	0.0105	0.0168	0.0248	0.0343	0.0452	0.0574

		p									
n	x	0.11	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55
1	0	0.8900	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000	0.4500
	1	0.1100	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5000	0.5500
2	0	0.7921	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500	0.2025
	1	0.1958	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000	0.4950
	2	0.0121	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500	0.3025
3	0	0.7050	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250	0.0911
	1	0.2614	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750	0.3341
	2	0.0323	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750	0.4084
	3	0.0013	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250	0.1664
4	0	0.6274	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625	0.0410
	1	0.3102	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500	0.2005
	2	0.0575	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750	0.3675
	3	0.0047	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500	0.2995
	4	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625	0.0915
5	0	0.5584	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313	0.0185
	1	0.3451	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1563	0.1128
	2	0.0853	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125	0.2757
	3	0.0105	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125	0.3369
	4	0.0007	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1563	0.2059
	5	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0313	0.0503
6	0	0.4970	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156	0.0083
	1	0.3685	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938	0.0609
	2	0.1139	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344	0.1861
	3	0.0188	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125	0.3032
	4	0.0017	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344	0.2780
	5	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938	0.1359
	6	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156	0.0277
7	0	0.4423	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078	0.0037
	1	0.3827	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547	0.0320
	2	0.1419	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1172
	3	0.0292	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734	0.2388
	4	0.0036	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734	0.2918
	5	0.0003	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641	0.2140
	6	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547	0.0872
	7	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078	0.0152

		p									
n	x	0.11	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55
8	0	0.3937	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039	0.0017
	1	0.3892	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0313	0.0164
	2	0.1684	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094	0.0703
	3	0.0416	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188	0.1719
	4	0.0064	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734	0.2627
	5	0.0006	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188	0.2568
	6	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094	0.1569
	7	0.0000	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313	0.0548
9	0	0.3504	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020	0.0008
	1	0.3897	0.3679	0.3020	0.2253	0.1556	0.1004	0.0605	0.0339	0.0176	0.0083
	2	0.1927	0.2597	0.3020	0.3003	0.2668	0.2162	0.1612	0.1110	0.0703	0.0407
	3	0.0556	0.1069	0.1762	0.2336	0.2668	0.2716	0.2508	0.2119	0.1641	0.1160
	4	0.0103	0.0283	0.0661	0.1168	0.1715	0.2194	0.2508	0.2600	0.2461	0.2128
	5	0.0013	0.0050	0.0165	0.0389	0.0735	0.1181	0.1672	0.2128	0.2461	0.2600
	6	0.0001	0.0006	0.0028	0.0087	0.0210	0.0424	0.0743	0.1160	0.1641	0.2119
	7	0.0000	0.0000	0.0003	0.0012	0.0039	0.0098	0.0212	0.0407	0.0703	0.1110
	8	0.0000	0.0000	0.0000	0.0001	0.0004	0.0013	0.0035	0.0083	0.0176	0.0339
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0020	0.0046	
10	0	0.3118	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010	0.0003
	1	0.3854	0.3474	0.2684	0.1877	0.1211	0.0725	0.0403	0.0207	0.0098	0.0042
	2	0.2143	0.2759	0.3020	0.2816	0.2335	0.1757	0.1209	0.0763	0.0439	0.0229
	3	0.0706	0.1298	0.2013	0.2503	0.2668	0.2522	0.2150	0.1665	0.1172	0.0746
	4	0.0153	0.0401	0.0881	0.1460	0.2001	0.2377	0.2508	0.2384	0.2051	0.1596
	5	0.0023	0.0085	0.0264	0.0584	0.1029	0.1536	0.2007	0.2340	0.2461	0.2340
	6	0.0002	0.0012	0.0055	0.0162	0.0368	0.0689	0.1115	0.1596	0.2051	0.2384
	7	0.0000	0.0001	0.0008	0.0031	0.0090	0.0212	0.0425	0.0746	0.1172	0.1665
	8	0.0000	0.0000	0.0001	0.0004	0.0014	0.0043	0.0106	0.0229	0.0439	0.0763
	9	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0016	0.0042	0.0098	0.0207
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010	0.0025	
20	0	0.0972	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000	0.0000	0.0000	0.0000
	1	0.2403	0.1368	0.0576	0.0211	0.0068	0.0020	0.0005	0.0001	0.0000	0.0000
	2	0.2822	0.2293	0.1369	0.0669	0.0278	0.0100	0.0031	0.0008	0.0002	0.0000
	3	0.2093	0.2428	0.2054	0.1339	0.0716	0.0323	0.0123	0.0040	0.0011	0.0002
	4	0.1099	0.1821	0.2182	0.1897	0.1304	0.0738	0.0350	0.0139	0.0046	0.0013
	5	0.0435	0.1028	0.1746	0.2023	0.1789	0.1272	0.0746	0.0365	0.0148	0.0049
	6	0.0134	0.0454	0.1091	0.1686	0.1916	0.1712	0.1244	0.0746	0.0370	0.0150
	7	0.0033	0.0160	0.0545	0.1124	0.1643	0.1844	0.1659	0.1221	0.0739	0.0366
8	0.0007	0.0046	0.0222	0.0609	0.1144	0.1614	0.1797	0.1623	0.1201	0.0727	

APPENDIX

C

Poisson Probability Table

λ										
x	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
0	0.9900	0.9802	0.9704	0.9608	0.9512	0.9418	0.9324	0.9231	0.9139	0.9048
1	0.0099	0.0196	0.0291	0.0384	0.0476	0.0565	0.0653	0.0738	0.0823	0.0905
2	0.0000	0.0002	0.0004	0.0008	0.0012	0.0017	0.0023	0.0030	0.0037	0.0045
3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002
4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

λ										
x	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10
0	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	0.3329
1	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659	0.3679	0.3662
2	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647	0.1839	0.2014
3	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494	0.0613	0.0738
4	0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111	0.0153	0.0203
5	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020	0.0031	0.0045

λ										
x	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	0.0008
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

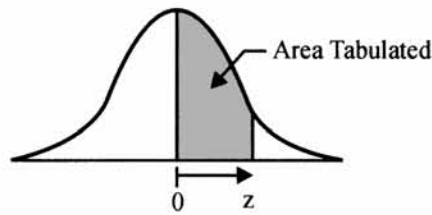
λ										
x	1.20	1.30	1.40	1.50	1.60	1.70	1.80	1.90	2.00	2.10
0	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353	0.1225
1	0.3614	0.3543	0.3452	0.3347	0.3230	0.3106	0.2975	0.2842	0.2707	0.2572
2	0.2169	0.2303	0.2417	0.2510	0.2584	0.2640	0.2678	0.2700	0.2707	0.2700
3	0.0867	0.0998	0.1128	0.1255	0.1378	0.1496	0.1607	0.1710	0.1804	0.1890
4	0.0260	0.0324	0.0395	0.0471	0.0551	0.0636	0.0723	0.0812	0.0902	0.0992
5	0.0062	0.0084	0.0111	0.0141	0.0176	0.0216	0.0260	0.0309	0.0361	0.0417
6	0.0012	0.0018	0.0026	0.0035	0.0047	0.0061	0.0078	0.0098	0.0120	0.0146
7	0.0002	0.0003	0.0005	0.0008	0.0011	0.0015	0.0020	0.0027	0.0034	0.0044
8	0.0000	0.0001	0.0001	0.0001	0.0002	0.0003	0.0005	0.0006	0.0009	0.0011
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0003
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

λ										
x	2.20	2.30	2.40	2.50	2.60	2.70	2.80	2.90	3.00	3.10
0	0.1108	0.1003	0.0907	0.0821	0.0743	0.0672	0.0608	0.0550	0.0498	0.0450
1	0.2438	0.2306	0.2177	0.2052	0.1931	0.1815	0.1703	0.1596	0.1494	0.1397
2	0.2681	0.2652	0.2613	0.2565	0.2510	0.2450	0.2384	0.2314	0.2240	0.2165
3	0.1966	0.2033	0.2090	0.2138	0.2176	0.2205	0.2225	0.2237	0.2240	0.2237
4	0.1082	0.1169	0.1254	0.1336	0.1414	0.1488	0.1557	0.1622	0.1680	0.1733
5	0.0476	0.0538	0.0602	0.0668	0.0735	0.0804	0.0872	0.0940	0.1008	0.1075
6	0.0174	0.0206	0.0241	0.0278	0.0319	0.0362	0.0407	0.0455	0.0504	0.0555
7	0.0055	0.0068	0.0083	0.0099	0.0118	0.0139	0.0163	0.0188	0.0216	0.0246
8	0.0015	0.0019	0.0025	0.0031	0.0038	0.0047	0.0057	0.0068	0.0081	0.0095
9	0.0004	0.0005	0.0007	0.0009	0.0011	0.0014	0.0018	0.0022	0.0027	0.0033
10	0.0001	0.0001	0.0002	0.0002	0.0003	0.0004	0.0005	0.0006	0.0008	0.0010
11	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003

λ										
x	5.20	5.30	5.40	5.50	5.60	5.70	5.80	5.90	6.00	6.10
0	0.0055	0.0050	0.0045	0.0041	0.0037	0.0033	0.0030	0.0027	0.0025	0.0022
1	0.0287	0.0265	0.0244	0.0225	0.0207	0.0191	0.0176	0.0162	0.0149	0.0137
2	0.0746	0.0701	0.0659	0.0618	0.0580	0.0544	0.0509	0.0477	0.0446	0.0417
3	0.1293	0.1239	0.1185	0.1133	0.1082	0.1033	0.0985	0.0938	0.0892	0.0848
4	0.1681	0.1641	0.1600	0.1558	0.1515	0.1472	0.1428	0.1383	0.1339	0.1294
5	0.1748	0.1740	0.1728	0.1714	0.1697	0.1678	0.1656	0.1632	0.1606	0.1579
6	0.1515	0.1537	0.1555	0.1571	0.1584	0.1594	0.1601	0.1605	0.1606	0.1605
7	0.1125	0.1163	0.1200	0.1234	0.1267	0.1298	0.1326	0.1353	0.1377	0.1399
8	0.0731	0.0771	0.0810	0.0849	0.0887	0.0925	0.0962	0.0998	0.1033	0.1066
9	0.0423	0.0454	0.0486	0.0519	0.0552	0.0586	0.0620	0.0654	0.0688	0.0723
10	0.0220	0.0241	0.0262	0.0285	0.0309	0.0334	0.0359	0.0386	0.0413	0.0441
11	0.0104	0.0116	0.0129	0.0143	0.0157	0.0173	0.0190	0.0207	0.0225	0.0244
12	0.0045	0.0051	0.0058	0.0065	0.0073	0.0082	0.0092	0.0102	0.0113	0.0124
13	0.0018	0.0021	0.0024	0.0028	0.0032	0.0036	0.0041	0.0046	0.0052	0.0058
14	0.0007	0.0008	0.0009	0.0011	0.0013	0.0015	0.0017	0.0019	0.0022	0.0025
15	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006	0.0007	0.0008	0.0009	0.0010
16	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004
17	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
λ										
x	6.20	6.30	6.40	6.50	6.60	6.70	6.80	6.90	7.00	7.10
0	0.0020	0.0018	0.0017	0.0015	0.0014	0.0012	0.0011	0.0010	0.0009	0.0008
1	0.0126	0.0116	0.0106	0.0098	0.0090	0.0082	0.0076	0.0070	0.0064	0.0059
2	0.0390	0.0364	0.0340	0.0318	0.0296	0.0276	0.0258	0.0240	0.0223	0.0208
3	0.0806	0.0765	0.0726	0.0688	0.0652	0.0617	0.0584	0.0552	0.0521	0.0492
4	0.1249	0.1205	0.1162	0.1118	0.1076	0.1034	0.0992	0.0952	0.0912	0.0874
5	0.1549	0.1519	0.1487	0.1454	0.1420	0.1385	0.1349	0.1314	0.1277	0.1241
6	0.1601	0.1595	0.1586	0.1575	0.1562	0.1546	0.1529	0.1511	0.1490	0.1468
7	0.1418	0.1435	0.1450	0.1462	0.1472	0.1480	0.1486	0.1489	0.1490	0.1489
8	0.1099	0.1130	0.1160	0.1188	0.1215	0.1240	0.1263	0.1284	0.1304	0.1321
9	0.0757	0.0791	0.0825	0.0858	0.0891	0.0923	0.0954	0.0985	0.1014	0.1042
10	0.0469	0.0498	0.0528	0.0558	0.0588	0.0618	0.0649	0.0679	0.0710	0.0740
11	0.0265	0.0285	0.0307	0.0330	0.0353	0.0377	0.0401	0.0426	0.0452	0.0478
12	0.0137	0.0150	0.0164	0.0179	0.0194	0.0210	0.0227	0.0245	0.0263	0.0283
13	0.0065	0.0073	0.0081	0.0089	0.0099	0.0108	0.0119	0.0130	0.0142	0.0154
14	0.0029	0.0033	0.0037	0.0041	0.0046	0.0052	0.0058	0.0064	0.0071	0.0078
15	0.0012	0.0014	0.0016	0.0018	0.0020	0.0023	0.0026	0.0029	0.0033	0.0037
16	0.0005	0.0005	0.0006	0.0007	0.0008	0.0010	0.0011	0.0013	0.0014	0.0016
17	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0007
18	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001

		λ								
x	7.20	7.30	7.40	7.50	7.60	7.70	7.80	7.90	8.00	8.10
0	0.0007	0.0007	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004	0.0003	0.0003
1	0.0054	0.0049	0.0045	0.0041	0.0038	0.0035	0.0032	0.0029	0.0027	0.0025
2	0.0194	0.0180	0.0167	0.0156	0.0145	0.0134	0.0125	0.0116	0.0107	0.0100
3	0.0464	0.0438	0.0413	0.0389	0.0366	0.0345	0.0324	0.0305	0.0286	0.0269
4	0.0836	0.0799	0.0764	0.0729	0.0696	0.0663	0.0632	0.0602	0.0573	0.0544
5	0.1204	0.1167	0.1130	0.1094	0.1057	0.1021	0.0986	0.0951	0.0916	0.0882
6	0.1445	0.1420	0.1394	0.1367	0.1339	0.1311	0.1282	0.1252	0.1221	0.1191
7	0.1486	0.1481	0.1474	0.1465	0.1454	0.1442	0.1428	0.1413	0.1396	0.1378
8	0.1337	0.1351	0.1363	0.1373	0.1381	0.1388	0.1392	0.1395	0.1396	0.1395
9	0.1070	0.1096	0.1121	0.1144	0.1167	0.1187	0.1207	0.1224	0.1241	0.1256
10	0.0770	0.0800	0.0829	0.0858	0.0887	0.0914	0.0941	0.0967	0.0993	0.1017
11	0.0504	0.0531	0.0558	0.0585	0.0613	0.0640	0.0667	0.0695	0.0722	0.0749
12	0.0303	0.0323	0.0344	0.0366	0.0388	0.0411	0.0434	0.0457	0.0481	0.0505
13	0.0168	0.0181	0.0196	0.0211	0.0227	0.0243	0.0260	0.0278	0.0296	0.0315
14	0.0086	0.0095	0.0104	0.0113	0.0123	0.0134	0.0145	0.0157	0.0169	0.0182
15	0.0041	0.0046	0.0051	0.0057	0.0062	0.0069	0.0075	0.0083	0.0090	0.0098
16	0.0019	0.0021	0.0024	0.0026	0.0030	0.0033	0.0037	0.0041	0.0045	0.0050
17	0.0008	0.0009	0.0010	0.0012	0.0013	0.0015	0.0017	0.0019	0.0021	0.0024
18	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009	0.0011
19	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003	0.0003	0.0004	0.0005
20	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

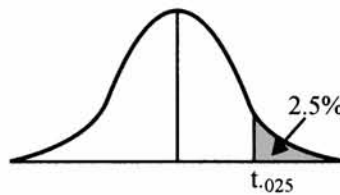
Normal Distribution Table



<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.10	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.20	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.30	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.40	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.50	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.60	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.70	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.80	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.90	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.00	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.10	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.20	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.30	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.40	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.50	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.60	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545

<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.70	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.80	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.90	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.00	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.10	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.20	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.30	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.40	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.50	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.60	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.70	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.80	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.90	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.00	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

***t* Distribution Table**

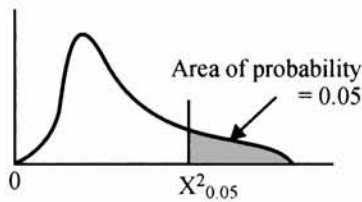


Example If you are using a two-tail test at 5% level of significance with 25 degrees of freedom, look for the *t* value under column 0.025 and 25 degrees of freedom row. The *t* value for this = 2.0595. If your using a one tail test at 5% level of significance with 25 degrees of freedom, look for the *t* value under column 0.05 and 25 degrees of freedom row. The *t* value for this = 1.7081

Upper <i>t</i> Value										
<i>Degrees of freedom</i>	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.3249	1.0000	3.0777	6.3137	12.7062	31.8210	63.6559	127.3211	318.2888	636.5776
2	0.2887	0.8165	1.8856	2.9200	4.3027	6.9645	9.9250	14.0892	22.3285	31.5998
3	0.2767	0.7649	1.6377	2.3534	3.1824	4.5407	5.8408	7.4532	10.2143	12.9244
4	0.2707	0.7407	1.5332	2.1318	2.7765	3.7469	4.6041	5.5975	7.1729	8.6101
5	0.2672	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321	4.7733	5.8935	6.8685
6	0.2648	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074	4.3168	5.2075	5.9587
7	0.2632	0.7111	1.4149	1.8946	2.3646	2.9979	3.4995	4.0294	4.7853	5.4081
8	0.2619	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008	5.0414
9	0.2610	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498	3.6896	4.2969	4.7809

<i>Degrees of freedom</i>	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
10	0.2602	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437	4.5868
11	0.2596	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058	3.4966	4.0248	4.4369
12	0.2590	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545	3.4284	3.9296	4.3178
13	0.2586	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520	4.2209
14	0.2582	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874	4.1403
15	0.2579	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467	3.2860	3.7329	4.0728
16	0.2576	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208	3.2520	3.6861	4.0149
17	0.2573	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.6458	3.9651
18	0.2571	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105	3.9217
19	0.2569	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.5793	3.8833
20	0.2567	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453	3.1534	3.5518	3.8496
21	0.2566	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314	3.1352	3.5271	3.8193
22	0.2564	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188	3.1188	3.5050	3.7922
23	0.2563	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073	3.1040	3.4850	3.7676
24	0.2562	0.6848	1.3178	1.7109	2.0639	2.4922	2.7970	3.0905	3.4668	3.7454
25	0.2561	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874	3.0782	3.4502	3.7251
26	0.2560	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787	3.0669	3.4350	3.7067
27	0.2559	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707	3.0565	3.4210	3.6895
28	0.2558	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633	3.0470	3.4082	3.6739
29	0.2557	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564	3.0380	3.3963	3.6595
30	0.2556	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500	3.0298	3.3852	3.6460
40	0.2550	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045	2.9712	3.3069	3.5510
60	0.2545	0.6786	1.2958	1.6706	2.0003	2.3901	2.6603	2.9146	3.2317	3.4602
120	0.2539	0.6765	1.2886	1.6576	1.9799	2.3578	2.6174	2.8599	3.1595	3.3734
1000	0.2534	0.6747	1.2824	1.6464	1.9623	2.3301	2.5807	2.8133	3.0984	3.3002

Chi-Square Distribution Table



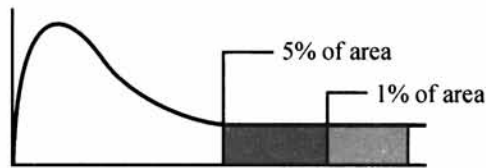
Example The upper χ^2 value at 5% level of significance with 9 degrees of freedom is obtained at the point of intersection of 9 degrees of freedom row and 0.05 column. This χ^2 value is = 16.918960

Upper Chi-Square Value					
<i>Degrees of freedom</i>	<i>0.1</i>	<i>0.05</i>	<i>0.025</i>	<i>0.01</i>	<i>0.005</i>
1	2.705541	3.841455	5.023903	6.634891	7.879400
2	4.605176	5.991476	7.377779	9.210351	10.596530
3	6.251394	7.814725	9.348404	11.344882	12.838073
4	7.779434	9.487728	11.143262	13.276699	14.860166
5	9.236349	11.070483	12.832492	15.086317	16.749648
6	10.644637	12.591577	14.449355	16.811872	18.547513
7	12.017031	14.067127	16.012774	18.475324	20.277738
8	13.361562	15.507312	17.534545	20.090159	21.954861
9	14.683663	16.918960	19.022778	21.666048	23.589275

<i>Degrees of freedom</i>	<i>0.1</i>	<i>0.05</i>	<i>0.025</i>	<i>0.01</i>	<i>0.005</i>
10	15.987175	18.307029	20.483201	23.209287	25.188055
11	17.275007	19.675153	21.920023	24.725022	26.756864
12	18.549340	21.026055	23.336660	26.216964	28.299660
13	19.811933	22.362027	24.735581	27.688184	29.819318
14	21.064141	23.684782	26.118935	29.141163	31.319425
15	22.307121	24.995797	27.488365	30.577951	32.801491
16	23.541821	26.296221	28.845325	31.999861	34.267053
17	24.769028	27.587100	30.190983	33.408717	35.718378
18	25.989418	28.869321	31.526410	34.805237	37.156386
19	27.203565	30.143505	32.852337	36.190775	38.582122
20	28.411970	31.410420	34.169581	37.566272	39.996856
21	29.615086	32.670558	35.478856	38.932232	41.400943
22	30.813285	33.924460	36.780678	40.289448	42.795664
23	32.006890	35.172460	38.075609	41.638334	44.181385
24	33.196235	36.415026	39.364060	42.979781	45.558363
25	34.381583	37.652489	40.646498	44.314014	46.927966
26	35.563164	38.885130	41.923138	45.641636	48.289777
27	36.741228	40.113266	43.194521	46.962837	49.645035
28	37.915907	41.337152	44.460790	48.278166	50.993559
29	39.087475	42.556948	45.722279	49.587829	52.335495
30	40.256017	43.772954	46.979218	50.892181	53.671868
40	51.805044	55.758487	59.341679	63.690771	66.766047
50	63.167113	67.504805	71.420194	76.153802	79.489839
60	74.396999	79.081954	83.297706	88.379430	91.951806
70	85.527036	90.531262	95.023149	100.425051	104.214769
80	96.578196	101.879472	106.628542	112.328791	116.320928
90	107.565010	113.145234	118.135908	124.116195	128.298676
100	118.498002	124.342101	129.561252	135.806891	140.169714

Upper Chi-Square Value					
<i>Degrees of freedom</i>	0.995	0.99	0.975	0.95	0.9
1	0.000039	0.000157	0.000982	0.003932	0.015791
2	0.010025	0.020100	0.050636	0.102586	0.210721
3	0.071723	0.114832	0.215795	0.351846	0.584375
4	0.206984	0.297107	0.484419	0.710724	1.063624
5	0.411751	0.554297	0.831209	1.145477	1.610309
6	0.675733	0.872083	1.237342	1.635380	2.204130
7	0.989251	1.239032	1.689864	2.167349	2.833105
8	1.344403	1.646506	2.179725	2.732633	3.489537
9	1.734911	2.087889	2.700389	3.325115	4.168156
10	2.155845	2.558199	3.246963	3.940295	4.865178
11	2.603202	3.053496	3.815742	4.574809	5.577788
12	3.073785	3.570551	4.403778	5.226028	6.303796
13	3.565042	4.106900	5.008738	5.891861	7.041500
14	4.074659	4.660415	5.628724	6.570632	7.789538
15	4.600874	5.229356	6.262123	7.260935	8.546753
16	5.142164	5.812197	6.907664	7.961639	9.312235
17	5.697274	6.407742	7.564179	8.671754	10.085183
18	6.264766	7.014903	8.230737	9.390448	10.864937
19	6.843923	7.632698	8.906514	10.117006	11.650912
20	7.433811	8.260368	9.590772	10.850799	12.442601
21	8.033602	8.897172	10.282907	11.591316	13.239596
22	8.642681	9.542494	10.982330	12.338009	14.041490
23	9.260383	10.195689	11.688534	13.090505	14.847954
24	9.886199	10.856349	12.401146	13.848422	15.658679
25	10.519647	11.523951	13.119707	14.611396	16.473405
26	11.160218	12.198177	13.843881	15.379163	17.291880
27	11.807655	12.878468	14.573373	16.151395	18.113889
28	12.461281	13.564666	15.307854	16.927876	18.939235
29	13.121067	14.256406	16.047051	17.708381	19.767740
30	13.786682	14.953464	16.790756	18.492667	20.599245
40	20.706577	22.164201	24.433058	26.509296	29.050516
50	27.990825	29.706725	32.357385	34.764236	37.688637
60	35.534397	37.484796	40.481707	43.187966	46.458885
70	43.275305	45.441700	48.757536	51.739263	55.328945
80	51.171933	53.539983	57.153152	60.391459	64.277842
90	59.196327	61.754019	65.646592	69.126018	73.291079
100	67.327533	70.064995	74.221882	77.929442	82.358127

F Distribution Table



Example Upper F value at 5% level of significance with degrees of freedom (6, 8) = 3.58. Please note that n_1 is the degrees of freedom for the numerator and n_2 is the degrees of freedom for the denominator.

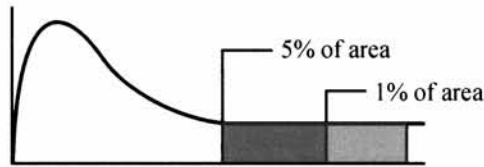
Upper F values for 5% significance

		n_1									
		1	2	3	4	5	6	7	8	9	10
n_2											
1		161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2		18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3		10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4		7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5		6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6		5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7		5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8		5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35

		n_1									
		1	2	3	4	5	6	7	8	9	10
n_2											
9		5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10		4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11		4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12		4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13		4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14		4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15		4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16		4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17		4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18		4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19		4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20		4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21		4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22		4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23		4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24		4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25		4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
30		4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40		4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60		4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120		3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
1000		3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84

Upper F values for 5% significance

		n_1									
		10	12	15	20	24	30	40	60	120	1000
n_2											
1		241.88	243.90	245.95	248.02	249.05	250.10	251.14	252.20	253.25	254.19
2		19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.49
3		8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4		5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5		4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6		4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7		3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8		3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9		3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10		2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11		2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.41
12		2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13		2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14		2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.14
15		2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16		2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.02
17		2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.97
18		2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19		2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20		2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.85
21		2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.82
22		2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.79
23		2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24		2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.74
25		2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.72
30		2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.63
40		2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.52
60		1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.40
120		1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.27
1000		1.84	1.76	1.68	1.58	1.53	1.47	1.41	1.33	1.24	1.11



Example Upper F value at 1% level of significance with degrees of freedom (6, 8) = 6.37. Please note that n_1 is the degrees of freedom for the numerator and n_2 is the degrees of freedom for the denominator.

Upper F values for 1% significance

		n_1									
		1	2	3	4	5	6	7	8	9	10
n_2											
1		4052.18	4999.34	5403.53	5624.26	5763.96	5858.95	5928.33	5980.95	6022.40	6055.93
2		98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40
3		34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23
4		21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5		16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6		13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7		12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8		11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9		10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10		10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11		9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12		9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13		9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14		8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15		8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16		8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17		8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18		8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19		8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20		8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21		8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22		7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23		7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24		7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25		7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
30		7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40		7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80

		n_1									
		1	2	3	4	5	6	7	8	9	10
n_2											
60		7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
120		6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
1000		6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34

Upper F values for 1% significance

		n_1									
		10	12	15	20	24	30	40	60	120	1000
n_2											
1		6055.93	6106.68	6156.97	6208.66	6234.27	6260.35	6286.43	6312.97	6339.51	6362.80
2		99.40	99.42	99.43	99.45	99.46	99.47	99.48	99.48	99.49	99.50
3		27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.14
4		14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.47
5		10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.03
6		7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.89
7		6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.66
8		5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.87
9		5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.32
10		4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.92
11		4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.61
12		4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.37
13		4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.18
14		3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.02
15		3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.88
16		3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.76
17		3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.66
18		3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.58
19		3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.50
20		3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.43
21		3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.37
22		3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.32
23		3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.27
24		3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.22
25		3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.18
30		2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.02
40		2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.82
60		2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.62
120		2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.40
1000		2.34	2.20	2.06	1.90	1.81	1.72	1.61	1.50	1.35	1.16