

Textbook

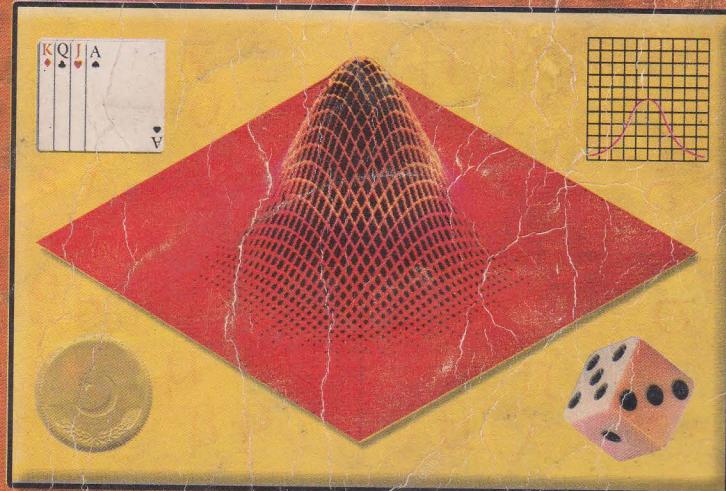
BASIC STATISTICS

For Intermediate Classes

Part-II

by

Ghulam Hussain Kiani
Muhammad Saleem Akhtar



MAJEED BOOK DEPOT

Lahore - Faisalabad - Rawalpindi

STATISTICS

Muhammad Rafiqat.

BASIC STATISTICS

(A Text Book for Intermediate Classes)

PART - II

Pie chart component Bar chart.

By

GHULAM HUSSAIN KIANI
Ex. Associate Professor of Statistics

MUHAMMAD SALEEM AKHTAR
Department of Statistics
Govt. Gordon College Rawalpindi,

MAJEED BOOK DEPOT

DISPLAY CENTRE

Head Office
22-Urdu Bazar, Lahore
Ph:042-37311484, 37355187

UNIQUE PUBLICATIONS Aminpur Bazar, Faisalabad Ph: 041-2643322	Al-Mustafa Plaza, 6 th Road, Rawalpindi Ph: 051-4423948	212-C Sunny Center Satellite Town, Gujranwala Ph: 055-3825612	Al-Hanif Plaza, University Road, Sargodha. Ph: 048-3740043
---	---	--	---

All Rights Reserved

No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying recording, or any information storage and retrieval system, without written permission of the authors or publishers.

REVISED EDITION 2011-2012

Publishers:

Kashif Mukhtar

Majeed Book Depot

Urdu Bazar, Lahore.

Printers:

Z.A. Printers, Lahore.

Copy:

500

Composing:

Muhammad Khurshid Khan &
Shahid Ayub Channa

Price:

190/-
Rs.225/-

Preface

Basic Statistics – Part II has been written to serve as the text for the students of Intermediate level class XII. It has been written strictly according to the new syllabus approved by the Ministry of Education (Curriculum Wing), Government of Pakistan, Islamabad. The book will meet the requirements of all the Education Boards in Pakistan. The students of M.A. Economics, B. Com., M. Sc. Geography, M. Sc. Psychology, B.B.A., Business Administration and the students of many other areas of social sciences can read their courses from this book. They can benefit a lot from this book because the lessons in the book have been discussed in a straightforward, simple and lucid manner.

For the students, who do not have the class-room facility, this book will be a good gift in their hands. The students of Allama Iqbal Open University who are taking up courses of BBA and B.A. will find this book of tremendous value to them. They can prepare their lessons from this book without attending intensive class-room lectures. The book is really 'basic', material-wise, and language-wise and anybody who is interested to learn the basic theory of Statistics will find the book a beneficial guide.

The entire book has been written in a simple manner. Special attention has been given to theory of sampling, hypothesis testing and estimation. The theoretical concepts have been made clear with illustrative examples. Efforts have been made to keep the subject matter close to the situations of practical life so that greater interest is created among the students.

We are extremely grateful to our colleagues who have done the arduous task of the proof reading. Nonetheless some errors might still appear here and there in the book. We shall be grateful if such errors or omissions are brought to our notice for prompt rectification. For the improvement of the book, we always need suggestions and healthy criticism of our readers and the teachers.

A sincere note of appreciation goes to our colleagues in Rawalpindi and Islamabad who have helped us a great deal in completing the task. We also express our gratitude to the members of our family and our friends for their encouragement and cooperation in this endeavour. We wish to thank all the respectable teachers of Statistics who have strengthened us by appreciating our first attempt 'Basic Statistics Part I'.

We would like to close by expressing our thanks to our publishers Messr. Majeed Book Depot, Urdu Bazar, Lahore for their maximum possible efforts in bringing out the book in time. We offer special words of thanks and appreciation for Muhammad Khurshid Khan, the computer operator who worked with us with patience and devotion.

August 1, 2011

Ghulam Hussain Kiani
Muhammad Saleem Akhtar

CONTENTS

Page No.

Chapter 10

Normal Distribution	1-28
10.1. Introduction	1
10.2. Normal Distribution	1
10.3. Properties of the Normal Distribution	2
10.4. Standard Normal Distribution	3
10.5. Use of the Area Table	4
10.6. Normal Frequency Distribution	9
10.7. The Normal Approximation to the Binomial Distribution	10
10.8. Inverse Use of the Area Table	11
☞ Short Definitions	16
☞ Multiple Choice Questions	16
☞ Short Questions	23
Exercises	26-28

Chapter 11

Sampling and Sampling Distributions	29-78
11.1. Introduction	29
11.2. Population	29
11.2.1. Finite Population	29
11.2.2. Infinite Population	30
11.2.3. Target and Sampled Population	30
11.3. Sample	30
11.3.1 Parameter and Statistic	30
11.3.2 Sampling Fraction	30
11.4. Complete Count	31
11.4.1. Population Census	31
11.5. Sample Survey	31
11.5.1. Advantages of Sampling	31
11.5.2. Limitations of Sampling	32
11.5.3. Sample Design	32
11.5.4. Sampling Frame	32
11.5.5. Equal Probability	33

11.5.6. Known Probability	33
11.5.7. Non-Zero Probability	33
11.6. Probability and Non-Probability Sampling	33
11.6.1. Sampling with Replacement	34
11.6.2. Sampling without Replacement.....	35
11.6.3. Combinations.....	35
11.6.4. Permutations	35
11.6.5. Simple Random Sample	35
11.6.6. Difference between Random Sample and Simple Random Sample	36
11.6.7. Selection of Simple Random Sample	36
11.7. Errors	37
11.7.1. Sampling Errors	38
11.7.2. Reducing the Sampling Errors	38
11.7.3. Non-Sampling Errors	39
11.8. Sampling Distributions.....	40
11.8.1. Standard Error	40
11.8.2. Sampling Distribution of \bar{X}	40
11.8.3. Sampling Distribution of s^2 and S^2	47
11.8.4. Sampling Distribution of Difference between two Means	50
11.8.5. Proportion	54
11.8.6. Sampling Distribution of Proportion	55
11.8.7. Sampling Distribution of Difference between \hat{p}_1 and \hat{p}_2	58
☞ Short Definitions	60
☞ Multiple Choice Questions	64
☞ Short Questions	71
Exercises.....	73-78

Chapter 12

Statistical Inference – Estimation.....	79-116
12.1. Introduction	79
12.2. Statistical Inference	79
12.2.1. Approaches of Statistical Inference	79
12.3. Estimation	80

12.3.1. Point estimator and Point Estimate.....	80
12.3.2. Point Estimation	81
12.3.3. Unbiasedness.....	81
12.3.4. Importance of Unbiasedness.....	81
12.4. Interval Estimation	82
12.4.1. Confidence Coefficient.....	82
12.5. Construction of Confidence Interval.....	82
12.5.1. Selection of Proper Confidence Interval.....	83
12.6. Confidence Interval Estimate of Population Mean μ (Large Sample)	84
12.6.1. Meaning of the Confidence Interval	86
12.7. Confidence Interval Estimate for Population Mean μ Population Normal (Small Sample).....	89
12.8. Confidence Interval Estimate for the Difference between two Population Means (Large Samples).....	91
12.9. Confidence Interval Estimate for the Difference between two Population Means – Populations Normal (Small Samples)	93
12.10. Confidence Interval for the Difference between two Population Means – Dependent Samples.....	96
12.11. Proportion	98
12.12. Confidence Interval Estimate for Population Proportion p (Large Sample).....	98
12.13. Confidence Interval Estimate for the Difference between two Population Proportions (Large Samples)	100
☞ Short Definitions	102
☞ Multiple Choice Questions	103
☞ Short Questions	110
Exercises	113-116

Chapter 13

Statistical Inference – Testing of Hypotheses	117-164
13.1. Introduction	117
13.2. Statistical Hypotheses	117
13.2.1. Null Hypothesis.....	118
13.2.2. Alternative Hypothesis	118
13.2.3. Simple Hypothesis.....	119
13.2.4. Composite Hypothesis	119

13.2.5. Acceptance and Rejection of Null Hypothesis	119
13.2.6. Test Statistic.....	119
13.2.7. Acceptance and Rejection Regions.....	120
13.2.8. Two – Tailed Test	120
13.2.9. One – Tailed Test	120
13.3. Errors in Testing of Hypothesis.....	121
13.3.1 Type I – Error	122
13.3.2. Type II – Error.....	122
13.3.3. Relation between α and β	123
13.4. Level of Significance.....	124
13.5. Formulating H_0 and H_1 and Making Critical Region	125
13.6. General Procedure for Testing of Hypothesis	126
13.7. Hypothesis Testing – Population Mean μ , σ Known (Large Sample)	128
13.8. Hypothesis Testing – Population Mean μ , σ not Known (Large Sample)	131
13.9. Hypothesis Testing – Population Mean μ , σ Known Normal Population (Small Sample).....	132
13.10. Hypothesis Testing – Population mean μ , σ Unknown Normal Population (Small Sample).....	132
13.11. Hypothesis Testing – Difference between two Population Means $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Known (Large Samples).....	135
13.12. Hypothesis Testing – Difference between two Population Means $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Unknown (Large Samples)	137
13.13. Test about $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Known, Populations Normal (Small Samples).....	138
13.14. Test about $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 not Known, Populations Normal (Small Samples)	139
13.15. Test about $\mu_1 - \mu_2$, Dependent Samples, Populations Normal	140
13.16. Test of Population Proportion p (Large Sample)	143
13.17. Test of Difference between two Population Proportions, $p_1 - p_2$ (Large Samples).....	146
13.18. Choice of Proper Test – Statistic	149

	Short Definitions	150
	Multiple Choice Questions	152
	Short Questions	157
	Exercises.....	160-164
Chapter 14		
.	Regression and Correlation	165-216
14.1.	Introduction	165
14.2.	Mathematical Model or Equation.....	165
14.3.	Non-Linear Model.....	168
14.4.	Statistical Model	169
14.4.1.	Independent and Dependent Variables.....	171
14.4.2.	Cause and Effect Relation.....	172
14.5.	Regression.....	172
14.5.1.	Simple Linear Regression	173
14.5.2.	Purpose of Regression Analysis.....	173
14.5.3.	Scatter Diagram	173
14.6.	Fitting a Linear Regression Line—the Method of Least Squares.....	175
14.6.1.	Properties of the Regression Line.....	179
14.6.2.	Regression Equation of X on Y	179
14.7.	Introduction.....	184
14.8.	Correlation.....	184
14.8.1.	Measurement of Correlation	184
14.8.2.	Perfect Positive Correlation	185
14.8.3.	Perfect Negative Correlation	185
14.8.4.	No Correlation	185
14.8.5.	Scatter Diagrams	186
14.9.	Correlation Coefficient for Sample Data	187
14.9.1.	Causation in Correlation.....	191
14.9.2.	Spurious Correlation	191
14.9.3.	Change of Origin	191
14.9.4.	Change of Scale	192
14.9.5.	Change of Origin and Scale	193
14.9.6.	'r' in a Linear Regression Relation	193
14.9.7.	'r' for Random Variables	193

14.10.	Relation between b_{yx} , b_{xy} and r	193
14.11.	Properties of Correlation Coefficient r	194
	Short Definitions	197
	Multiple Choice Questions	199
	Short Questions	205
	Exercises	210-216
Chapter 15		
	Association	217-254
15.1.	Variable and Attribute.....	217
15.1.1.	Notation for Attributes	217
15.1.2.	One Attribute	218
15.1.3.	Two Attributes.....	218
15.1.4.	Positive and Negative Classes	219
15.1.5.	Order of Classes.....	220
15.1.6.	Ultimate Class Frequencies.....	220
15.1.7.	Lower Order Frequencies in Terms of Higher Order Frequencies	220
15.1.8.	Higher Order Frequencies into Lower Order Frequencies.....	221
15.2.	Consistency.....	223
15.3.	Independence of Attributes.....	224
15.3.1.	Independence Defined	226
15.3.2.	Another Definition of Independence.....	227
15.4.	Coefficient of Association	229
15.5.	χ^2 -Distribution	230
15.5.1.	Test of Independence	230
15.5.2.	Direct Formula for Calculating χ^2 in 2×2 Contingency Table.....	234
15.6.	Contingency Table of Higher Order	235
15.7.	Limitations of χ^2	236
15.8.	Rank Correlation	240
	Short Definitions	243
	Multiple Choice Questions	244
	Short Questions	247
	Exercises	250-254

Chapter 16

Time Series.....	255-288
16.1. Introduction	255
16.2. Purpose of Time Series.....	255
16.2.1. Graph of the Time Series	255
16.3. Components of a Time Series	256
16.3.1. Secular Trend	256
16.3.2. Seasonal Variation	258
16.3.3. Cyclical Variations	259
16.3.4. Irregular Variations	260
16.4. Analysis of Time Series.....	261
16.5. Measurement of Secular Trend	261
16.5.1. The Method of Free-hand Curve.....	262
16.5.2. The Method of Semi-Averages	263
16.5.3. The Method of Moving Averages	265
16.5.4. Method of Least Squares.....	269
16.5.5. Fitting a Straight Line.....	269
16.5.6. Coding of the Time Periods.....	269
16.5.7. Change of Origin in Coding	270
16.6. Fitting of Second Degree Parabola	273
☞ Short Definitions	275
☞ Link with Time Series Components	277
☞ Multiple Choice Questions	278
☞ Short Questions	282
Exercises.....	285-288

Chapter 17

Orientation of Computers.....	289-306
17.1. Introduction to Computers.....	289
17.1.1. Computer Capabilities and its Uses	289
17.2. Computer History.....	290
17.3. Types of Computer	291
17.3.1. Analog Computer.....	291

17.3.2. Digital Computer.....	291
17.3.3. Hybrid Computer.....	291
17.4. Classifications of Computers	292
17.4.1. Mainframe Computers	292
17.4.2. Minicomputers.....	292
17.4.3. Microcomputers	292
17.4.4. Super Computers.....	293
17.5. Computer Components.....	293
17.6. Computer Hardware	293
17.6.1. Input Unit.....	293
17.6.2. Central Processing Unit.....	294
17.6.3. Secondary Storage.....	297
17.6.4. Output Unit	298
17.7. Computer Software	299
17.7.1. Programming Languages.....	299
17.7.2. System Software.....	299
17.7.3. Application Software	301
17.8. Basic Idea of Writing and Running a Computer Program	301
17.8.1. Program Design.....	301
17.8.2. Program Writing	301
17.8.3. Testing and Debugging	301
17.8.4. Documentation, Implementation and Maintenance	302
17.9. Number System	302
17.9.1. Decimal Number System	302
17.9.2. Binary Number System	302
17.9.3. Octal Number System	302
17.9.4. Hexadecimal Number System	302
17.10. Binary Number System as a Foundation of Computer	302
Multiple Choice Questions	304
Short Questions	306
Statistical Tables	307-312

Chapter

10

NORMAL DISTRIBUTION

10.1 INTRODUCTION

Historically, the discovery of normal distribution goes back to the seventeenth and eighteenth centuries and is associated with the names of De Moivre (1667 – 1754), Laplace (1749 – 1827) and Gauss (1777 – 1855). At that time, it received the attention of mathematicians and natural and social scientists. Its application to biological data was pioneered at a later date by Sir Francis Galton (1822 – 1911). The normal distribution, also called the normal law of error, is widely used in research in the biological, physical and social sciences. In practical life we quite often come across the distributions close to this distribution and hence the "normal" is used for it. The word normal is not to be used as something opposite to the word abnormal. Normal distribution is also called mother of distributions because various other distributions are generated from this distribution. This distribution makes the base for inferential statistics, a branch of statistics in which we draw conclusions about the populations on the basis of information gained from the sample study.

10.2 NORMAL DISTRIBUTION

Normal distribution was first described in 1733 by De Moivre as being the limiting form of the binomial density as the number of trials become infinite. This discovery did not get much attention and the distribution was "discovered" again by both Laplace and Gauss about a half – century later. Both men dealt with problems of astronomy, and each derived the normal distribution as a distribution that seemingly described the behavior of errors in astronomical measurements. The distribution is often referred to as the "Gaussian" distribution.

One of the most important examples of a continuous probability distribution is the normal distribution also called normal curve or Gaussian distribution. The curve is defined by the equation

$$Y = f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < X < \infty \quad \dots \quad (10.1)$$

where, μ = mean of the distribution — a parameter.

σ = standard deviation of the distribution — a parameter.

π = a constant approximately equal to 3.14159

e = a constant approximately equal to 2.71828

X = abscissa, measurement or score marked on horizontal axis

Y = ordinate, height of curve corresponding to an assigned value of X

The total area bounded by the curve (10.1) and the X-axis is one. The area under the curve between two ordinates $X = a$ and $X = b$, where $a < b$, represents the probability that X lies between a and b and this probability is denoted by $P(a < X < b)$. When the variable X is expressed in terms of standard units or standard normal variate $Z = \frac{X - \mu}{\sigma}$, then equation (10.1) is replaced by the so-called standard form

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}} \quad \dots\dots(10.2)$$

In this case we say that Z is normally distributed. The mean of standard normal variate Z is zero and its variance is one. The value of Z is zero when $X = \mu$. A graph of this standardized normal curve is shown in Fig. 10.1. In this graph we have indicated the areas included between $Z = -1$ and $+1$, $Z = -2$ and $+2$, $Z = -3$ and $+3$ which are 68.27%, 95.45% and 99.73% respectively. The area under this curve bounded by the ordinates at $Z = 0$ and any positive value of Z are given in table. From this table the area between any two ordinates can be found by using the symmetry of the curve about $Z = 0$.

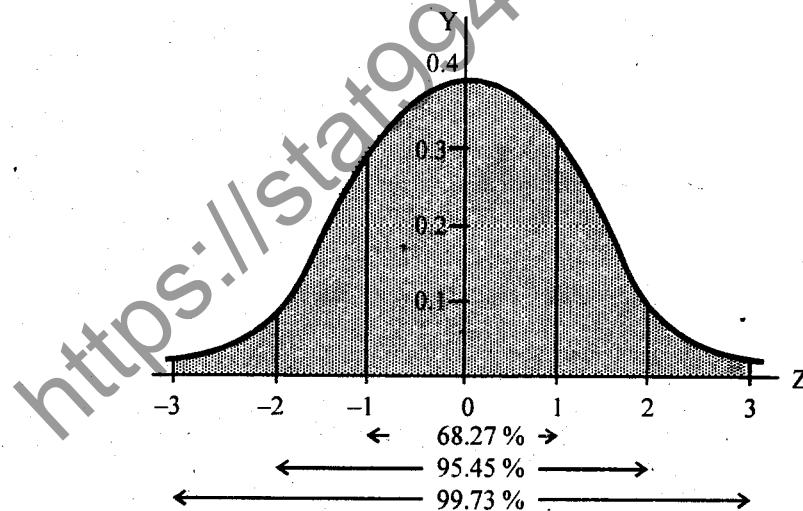


Fig. 10.1

10.3 PROPERTIES OF THE NORMAL DISTRIBUTION

- 2(1) It is symmetrical about ordinate at $X = \mu$. It means that the central ordinate at $X = \mu$ divides the curve into two equal parts.
- 1(2) The arithmetic mean, median and mode coincide.
- (3) The lower and upper quartiles are equidistant from the mean and are at a distance of 0.6745σ .

$$Q_1 = \mu - 0.6745 \sigma$$

$$Q_3 = \mu + 0.6745 \sigma$$

- (4) Mean deviation is 0.7979σ or $4/5\sigma$ (approximately). $\mu \pm 0.6745\sigma$
- (5) Semi-interquartile range or quartile deviation is equal to the probable error which is equal to 0.6745σ or $2/3\sigma$ (approximately).
- (6) The ordinate is highest at the mean μ . $\mu - \sigma, \mu + \sigma$
- (7) It has two points of inflection (the points where the curvature changes its direction) which lie at a distance of one σ above the mean μ and one σ below the mean μ . $(\mu - \sigma, \mu + \sigma)$
- (8) The curve is asymptotic to the base line. It means that it continues to approach but never reaches the base line.
- (9) In normal curve, if n th moment is odd, the value of this odd moment will always be zero. This is because the normal curve is symmetrical and for symmetrical distribution sum of the positive deviations from μ will always be equal to the sum of the negative deviations from μ and thus will cancel out each other. If n th moment is even, we have the relation

$$\mu_n = \frac{n!}{(2)^{\frac{n}{2}} \left(\frac{n}{2}\right)!} \sigma^n \quad (\text{where } n \text{ is even})$$

All odd moments
will always be zero

It follows that $\mu_2 = \sigma^2$, $\mu_4 = 3\sigma^4$, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \gamma_1^2 = 0$, i.e. skewness is zero and

$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3$, $\gamma_2 = 0$, i.e. normal curve has zero kurtosis. The normal distribution is also called mesokurtic.

- (10) The total area under the normal curve is unity.
 (11) Area properties of the normal distribution.

In a normal distribution:

$\mu \pm 0.6745\sigma$ covers 50% area

$\mu \pm 1\sigma$ covers 68.27% area

$\mu \pm 2\sigma$ covers 95.45% area

$\mu \pm 3\sigma$ covers 99.73% area

10.4 STANDARD NORMAL DISTRIBUTION

The properties of the normal curve permit us to define a standardized distribution in terms of the variable Z defined as

$$Z = \frac{X - \mu}{\sigma}$$

This is equivalent to measuring the distance X from the mean μ using the standard deviation σ as the unit of measuring distance. The variable Z is termed as the standard normal variate and plays a very important role in statistics. The probability density function in terms of Z is

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

The mean of random variable Z is zero and its variance is unity. If we know the mean μ and the standard deviation σ , we can calculate Z corresponding to any value of X and corresponding from the central ordinate to the value of Z.

10.5 USE OF THE AREA TABLE

The table "areas under the standard normal curve" gives the areas for various values of Z. For example $Z = -1$ to 0 and 0 to +1 gives the area 0.34134 as shown in the following figure.

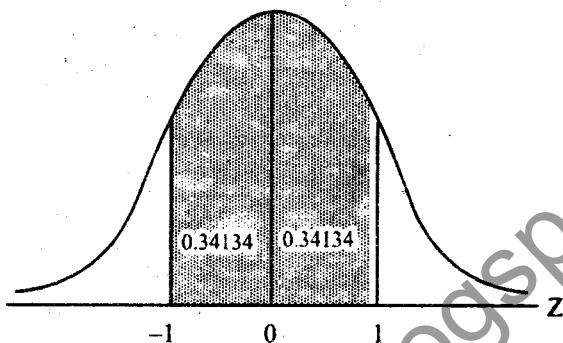


Fig. 10.2

As the curve is symmetrical, the same area table can be used for negative values of Z . The area from $Z = 0$ to $Z = 1$ is 0.34134, similarly the area between $Z = -1$ and $Z = 0$ is also 0.34134.

Example 10.1.

In a normal distribution mean is 100 and standard deviation is 10. Find:

Solution:

Here, $\mu = 100$, $\sigma = 10$, $\sigma^2 = \mu_2 = 100$.

- (i) Mean deviation = 0.7979σ = $0.7979 (10)$ = **7.979**
(ii) Quartile deviation = 0.6745σ = $0.6745 (10)$ = **6.745**
(iii) Third moment about mean = $\mu_3 = 0$, because all odd order moments about mean in a normal distribution are zero, i.e. $\mu_1 = \mu_3 = \mu_5 = \dots = 0$.

$$\text{Fourth moment about mean} = \mu_4 = 3\sigma^4 = 3(10)^4 = 30000$$

$$(iv) \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0)^2}{(100)^3} = 0 \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{30000}{(100)^2} = 3$$

$$(v) \quad Q_1 = \mu - 0.6745 \sigma = 100 - 0.6745(10) = 93.255$$

$$Q_3 = \mu + 0.6745 \sigma = 100 + 0.6745 (10) = 106.745$$

- (vi) Mean = Median = Mode = 100, because in a normal distribution the mean, median and mode coincide.
- (vii) Normal distribution has two points of inflection which lie at a distance of one σ above the mean μ and one σ below the mean μ . i.e.
 $\mu - \sigma = 100 - 10 = 90$ and $\mu + \sigma = 100 + 10 = 110$
- (viii) Maximum ordinate $= \frac{1}{\sigma\sqrt{2\pi}} = \frac{1}{10\sqrt{2(3.1416)}} = 0.0399$

Example 10.2.

In a normal distribution, mean is zero and the standard deviation is 1. Write down its equation and find the value of the maximum ordinate correct to four places of decimal.

Solution:

The equation of the normal curve with mean μ and standard deviation σ is

$$Y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < X < \infty$$

When $\mu = 0$ and $\sigma = 1$, the equation of the normal curve will be

$$Y = f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

We know that the maximum ordinate is at $X = \mu$ and $\mu = 0$, the value of the maximum ordinate is

$$\begin{aligned} Y &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(0)^2} = \frac{1}{\sqrt{2\pi}} e^0 = \frac{1}{\sqrt{2\pi}} \quad (\text{since } e^0 = 1) \\ &= \frac{1}{\sqrt{2(3.1416)}} = \frac{1}{2.5066} = 0.3989 \end{aligned}$$

Example 10.3.

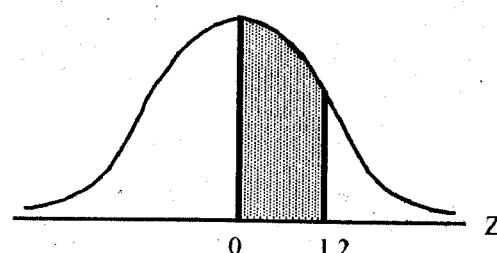
Find the area under the normal curve in each of the cases:

- (i) Between $Z = 0$ and $Z = 1.2$
- (ii) Between $Z = -0.68$ and $Z = 0$
- (iii) Between $Z = -0.46$ and $Z = 2.21$
- (iv) Between $Z = 0.81$ and $Z = 1.94$
- (v) To the left of $Z = -0.6$
- (vi) To the right of $Z = -1.28$
- (vii) To the right of $Z = 2.05$ and to the left of $Z = -1.44$.

Solution:

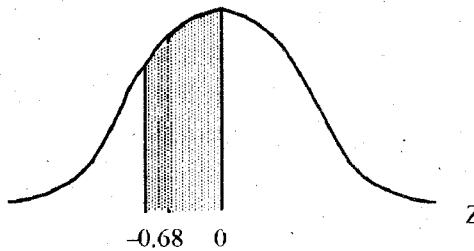
- (i) Between $Z = 0$ and $Z = 1.2$

Required area = Area between
 $Z = 0$ and $Z = 1.2$ is 0.3849



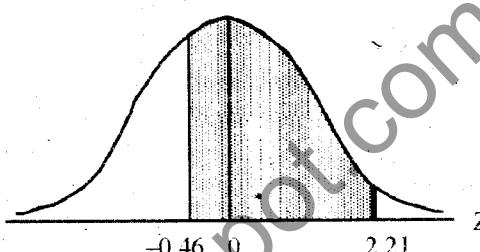
- (ii) Between $Z = -0.68$ and $Z = 0$

Required area = Area between
 $Z = -0.68$ and $Z = 0$ is 0.2518



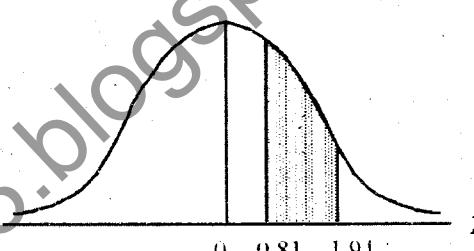
- (iii) Between $Z = -0.46$ and $Z = 2.21$

Required area = (Area between
 $Z = -0.46$ and $Z = 0$)
+ (Area between $Z = 0$ and $Z = 2.21$)
= $0.1772 + 0.4864 = 0.6636$



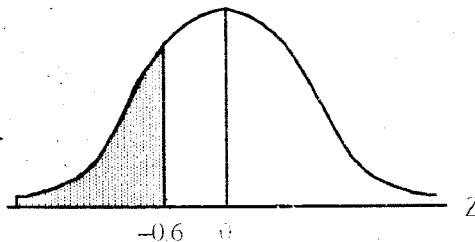
- (iv) Between $Z = 0.81$ and $Z = 1.94$

Required area
= (Area between $Z = 0$ and $Z = 1.94$)
- (Area between $Z = 0$ and $Z = 0.81$)
= $0.4738 - 0.2910 = 0.1828$



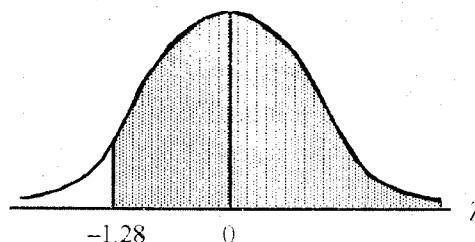
- (v) To the left of $Z = -0.6$

Required area = (Area to left of $Z = 0$)
- (Area between $Z = -0.6$ and $Z = 0$)
= $0.5 - 0.2258 = 0.2742$



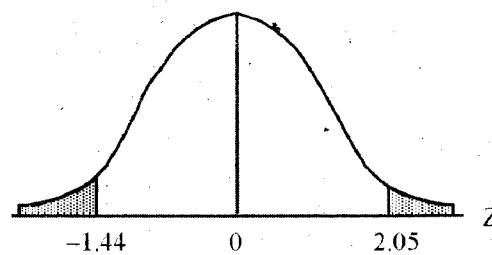
- (vi) To the right of $Z = -1.28$

Required area
= (Area between $Z = -1.28$ and $Z = 0$)
+ (Area to right of $Z = 0$)
= $0.3997 + 0.5 = 0.8997$



- (vii) To the right of $Z = 2.05$ and to the left of $Z = -1.44$

Required area = Total area
- (Area between $Z = -1.44$ and $Z = 0$)
- (Area between $Z = 0$ and $Z = 2.05$)
= $1 - 0.4251 - 0.4798 = 0.0951$



Example 10.4.

Given a normal distribution with $\mu = 40$ and $\sigma = 6$, find

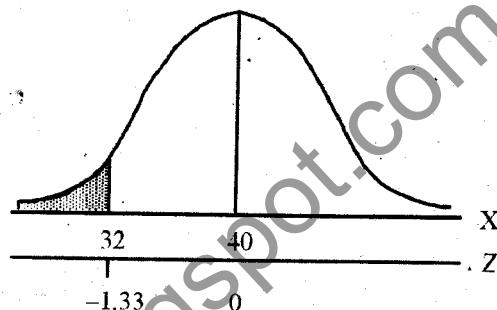
- (a) the area below 32 (b) the area above 27 (c) the area between 42 and 51.

Solution:

$$\text{Here, } \mu = 40, \sigma = 6, Z = \frac{X - \mu}{\sigma} = \frac{X - 40}{6}$$

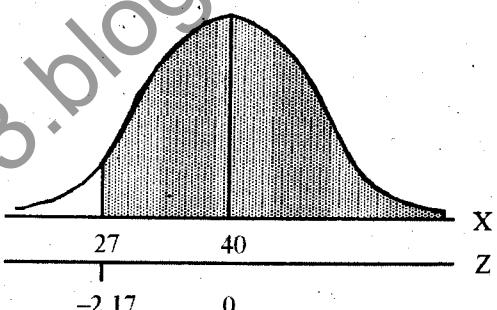
$$(a) Z = \frac{32 - 40}{6} = -\frac{8}{6} = -1.33$$

$$\begin{aligned} P(X < 32) &= P(Z < -1.33) \\ &= P(-\infty \leq Z \leq 0) - P(-1.33 \leq Z \leq 0) \\ &= 0.5 - 0.4082 = 0.0918 \end{aligned}$$



$$(b) Z = \frac{27 - 40}{6} = -\frac{13}{6} = -2.17$$

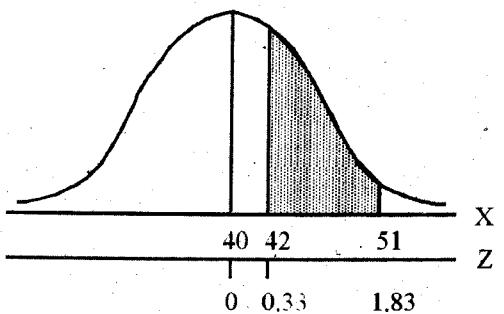
$$\begin{aligned} P(X > 27) &= P(Z > -2.17) \\ &= P(-2.17 \leq Z \leq 0) + P(0 \leq Z \leq \infty) \\ &= 0.4850 + 0.5 = 0.9850 \end{aligned}$$



$$(c) Z_1 = \frac{X_1 - 40}{6} = \frac{42 - 40}{6} = \frac{2}{6} = 0.33$$

$$Z_2 = \frac{X_2 - 40}{6} = \frac{51 - 40}{6} = \frac{11}{6} = 1.83$$

$$\begin{aligned} P(42 \leq X \leq 51) &\Rightarrow P(0.33 \leq Z \leq 1.83) \\ &= P(0 \leq Z \leq 1.83) - P(0 \leq Z \leq 0.33) \\ &= 0.4664 - 0.1293 = 0.3371 \end{aligned}$$

**Example 10.5.**

A newspaper stall sells an average of 400 papers per day. Assume that these sales are normally distributed with a standard deviation of 25. For each of the following probability questions use graphs with both X and Z axes and indicate the corresponding areas under the normal curve. What is the probability that on a given day:

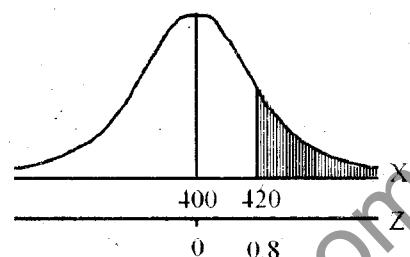
- (a) more than 420 papers will be sold? (b) at most 410 papers will be sold?
 (c) less than 395 papers will be sold? (d) between 390 and 405 papers will be sold?

Solution:

$$\text{Here, } \mu = 400, \sigma = 25, Z = \frac{X - \mu}{\sigma} = \frac{X - 400}{25}$$

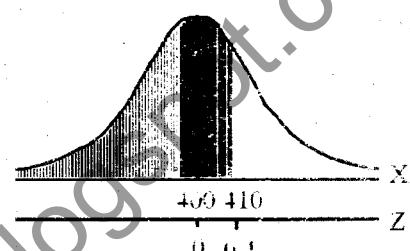
$$(a) Z = \frac{420 - 400}{25} = 0.8$$

$$\begin{aligned} P(X > 420) &= P(Z > 0.8) \\ &= P(0 < Z < \infty) - P(0 < Z < 0.8) \\ &= 0.5 - 0.2881 = 0.2119 \end{aligned}$$



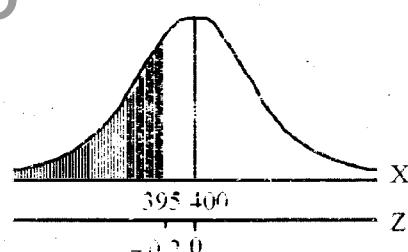
$$(b) Z = \frac{410 - 400}{25} = 0.4$$

$$\begin{aligned} P(X \leq 410) &= P(Z \leq 0.4) \\ &= P(-\infty < Z \leq 0) + P(0 \leq Z \leq 0.4) \\ &= 0.5 + 0.1554 = 0.6554 \end{aligned}$$



$$(c) Z = \frac{395 - 400}{25} = -0.2$$

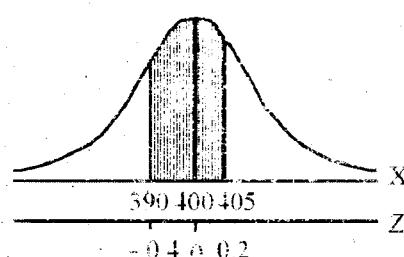
$$\begin{aligned} P(X < 395) &= P(Z < -0.2) \\ &= P(-\infty < Z < 0) - P(-0.2 < Z < 0) \\ &= 0.5 - 0.0793 = 0.4207 \end{aligned}$$



$$(d) Z_1 = \frac{X_1 - 400}{25} = \frac{390 - 400}{25} = -0.4$$

$$Z_2 = \frac{X_2 - 400}{25} = \frac{405 - 400}{25} = 0.2$$

$$\begin{aligned} P(390 < X < 405) &= P(-0.4 < Z < 0.2) \\ &= P(-0.4 < Z < 0) + P(0 < Z < 0.2) \\ &= 0.1554 + 0.0793 = 0.2347 \end{aligned}$$



Example 10.6.

The heights of freshmen students at a military academy are normally distributed with a mean of 5 feet 10 inches and a standard deviation of 2 inches.

- (a) What is the proportion of freshmen at the academy who are taller than 6 feet 3 inches?
- (b) What is the proportion of freshmen who are less than 5 feet 7 inches?
- (c) What is the proportion of freshmen between 5 feet 8 inches and 6 feet 0 inches?

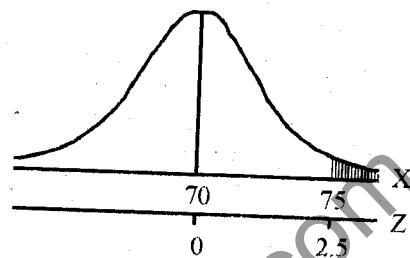
Solution:

$$(a) Z = \frac{75 - 70}{2} = 2.5$$

$$P(X > 75) = P(Z > 2.5)$$

$$= P(0 < Z < \infty) - P(0 < Z < 2.5)$$

$$= 0.5 - 0.4938 = 0.0062 \text{ or } 0.62\%$$

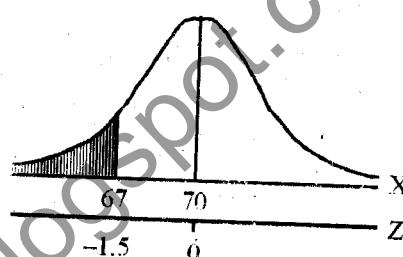


$$(b) Z = \frac{67 - 70}{2} = -1.5$$

$$P(X < 67) = P(Z < -1.5)$$

$$= P(-\infty < Z < 0) - P(-1.5 < Z < 0)$$

$$= 0.5 - 0.4332 = 0.0668 \text{ or } 6.68\%$$



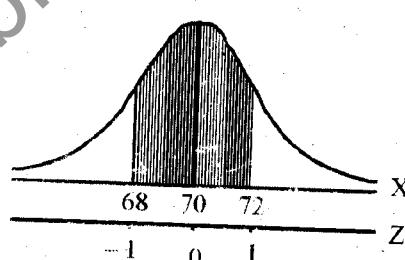
$$(c) Z_1 = \frac{X_1 - 70}{2} = \frac{68 - 70}{2} = -1$$

$$Z_2 = \frac{X_2 - 70}{2} = \frac{72 - 70}{2} = +1$$

$$P(68 < X < 72) = P(-1 < Z < 1)$$

$$= P(-1 < Z < 0) + P(0 < Z < 1)$$

$$= 0.3413 + 0.3413 = 0.6826 \text{ or } 68.26\%$$



10.6 NORMAL FREQUENCY DISTRIBUTION

Sometimes we have to convert the normal probability distribution into normal frequency distribution. When the probability distribution is multiplied with the total number of observations ($\Sigma f = N$), we get the normal frequency distribution. The normal probability distribution is

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

whereas the normal frequency distribution is

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

For example, we know that the probability is 0.6827 that the random variable X will fall between the interval $\mu - \sigma$ to $\mu + \sigma$. The probability 0.6827 can be converted into percentage of observations which lie between $\mu - \sigma$ and $\mu + \sigma$. This percentage is 68.27 %. If the total number of observations are 1000, the interval $\mu - \sigma$ to $\mu + \sigma$ will contain 683 observations, i.e. $0.6827 \times 1000 = 683$.

Example 10.7.

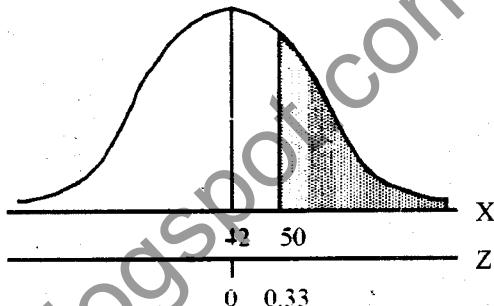
In an intelligence test administered on 1000 children, the average I.Q. was 42 and standard deviation 24.

- Find the number of children exceeding a score of 50.
- Find the number of children lying between the scores 30 and 54.

Solution: Here, $\mu = 42$, $\sigma = 24$, $N = 1000$, $Z = \frac{X - \mu}{\sigma} = \frac{X - 42}{24}$.

$$(a) Z = \frac{50 - 42}{24} = 0.33$$

$$\begin{aligned} P(X > 50) &= P(Z > 0.33) \\ &= P(0 \leq Z \leq \infty) - P(0 \leq Z \leq 0.33) \\ &= 0.5 - 0.1293 = 0.3707 \end{aligned}$$



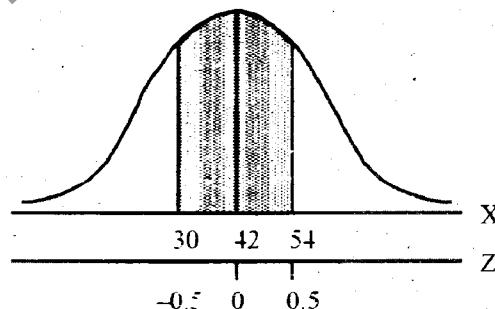
Hence the expected number of children exceeding a score of 50

$$= N.P(X > 50) = 1000(0.3707) = 370.7 \text{ or } 371 \text{ approximately.}$$

$$(b) Z_1 = \frac{X_1 - 42}{24} = \frac{30 - 42}{24} = -0.5$$

$$Z_2 = \frac{X_2 - 42}{24} = \frac{54 - 42}{24} = +0.5$$

$$\begin{aligned} P(30 \leq X \leq 54) &= P(-0.5 \leq Z \leq 0.5) \\ &= P(-0.5 \leq Z \leq 0) + P(0 \leq Z \leq 0.5) \\ &= 0.1915 + 0.1915 = 0.3830 \end{aligned}$$



Hence the expected number of children lying between the scores 30 and 54

$$= N.P(30 \leq X \leq 54) = 1000(0.3830) = 383.$$

10.7 THE NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

The (continuous) normal distribution provides a close approximation to the (discrete) binomial distribution when n , the number of trials is very large and p , the probability of a success on an individual trial is close to 1/2. To provide a theoretical foundation for this argument, let us make the following statement, a proof of which can be found in most of the texts in mathematical statistics.

If X is a random variable having a binomial distribution with the parameters n and p , then $Z = (X - np)/\sqrt{npq}$ approaches the standard normal distribution when n approaches infinity. Strictly speaking, this statement applies when n approaches infinity, but the normal distribution is often used to approximate binomial probabilities even n is fairly small. A good rule of thumb is to use this approximation only when np and nq are both equal to or greater than 5. The procedure to follow in using a normal approximation to the binomial is as follows:

Step 1. Compute $\mu = np$ and $\sigma = \sqrt{npq}$

Step 2. Apply a continuity correction factor to convert a discrete (binomial) random variable into a (normal) continuous random variable, so that the standardized normal Z transformation is

$$Z = \frac{(X - 1/2) - \mu}{\sigma} \quad \text{or} \quad Z = \frac{(X + 1/2) - \mu}{\sigma}$$

Step 3. Use a standard normal table to find the probabilities corresponding to Z in order to obtain the binomial $b(x; n, p)$. For example,

$$P(X = a) \approx P\left[\frac{(a - 1/2) - \mu}{\sigma} \leq Z \leq \frac{(a + 1/2) - \mu}{\sigma}\right]$$

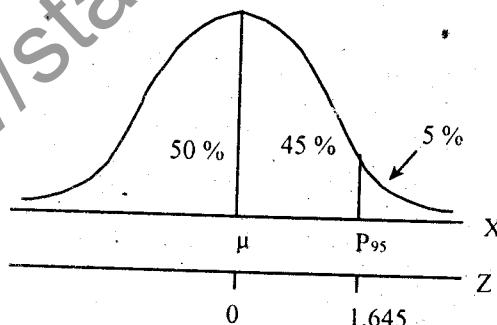
$$P(X \leq b) \approx P\left[Z \leq \frac{(b + 1/2) - \mu}{\sigma}\right]$$

$$P(X \geq c) \approx P\left[Z \geq \frac{(c - 1/2) - \mu}{\sigma}\right] \quad \text{or} \quad 1 - P\left[Z \leq \frac{(c + 1/2) - \mu}{\sigma}\right]$$

where a, b and c are some values of random variable X.

10.8 INVERSE USE OF THE AREA TABLE

The area table of normal distribution is designed to give the areas for various values of the standard normal variate Z. But this area table can also be used to read the values of Z for a certain given area under the normal curve. This is called inverse use of the area table. Suppose there are 95% observations less than a certain point say, $X(P_{95})$. Clearly the area between μ and X is 45 % (0.45). Using the area table in inverse order we can read the value of Z. Corresponding to the area equal to 0.45, the value of Z is 1.645.



Example 10.8.

A random variable X is normally distributed with mean = 40 and standard deviation = 4. (i) Find a point that has 97 % of the distribution below it.

(ii) Find a point that has 62.2 % of the distribution below it.

(iii) Find a point that has 90 % of the distribution above it.

(iv) Find two points containing the middle 98 % area.

(v) Find two points containing the middle 95 % area.

(vi) Find P_{20} , P_{80} , P_{95} and P_{99} .

Solution:

$$\text{Here, } \mu = 40, \sigma = 4, Z = \frac{X - \mu}{\sigma} = \frac{X - 40}{4}$$

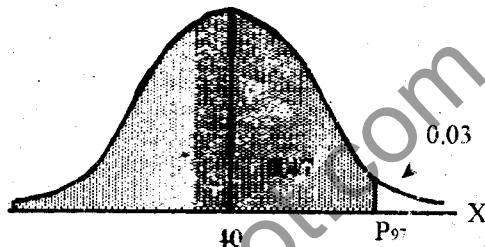
- (i) P_{97} is a point having 97 percent of the area below it. Area table shows this point to be

$$1.88 = \frac{X - 40}{4}$$

$$X - 40 = 4(1.88)$$

$$X = 40 + 7.52 = 47.52$$

$$\text{Thus, } P_{97} = 47.52$$



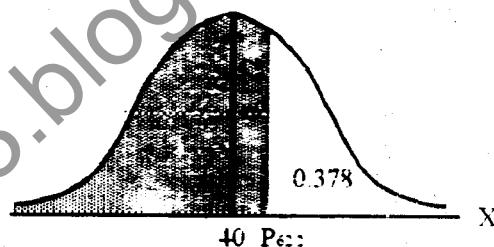
- (ii) $P_{62.2}$ is a point having 62.2 percent of the area below it. Area table shows this point to be

$$0.3108 = \frac{X - 40}{4}$$

$$X - 40 = 4(0.3108)$$

$$X = 40 + 1.2432 = 41.2432$$

$$\text{Thus, } P_{62.2} = 41.2432$$

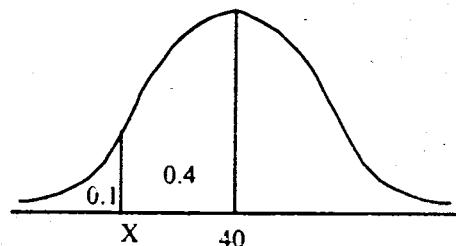


- (iii) 90 % of the area above it, Area table shows this point to be

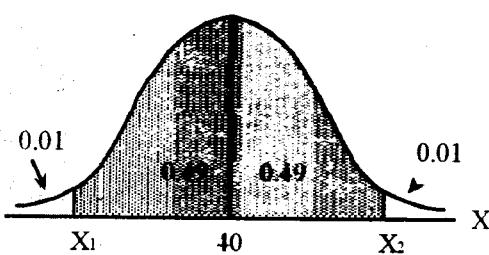
$$-1.28 = \frac{X - 40}{4}$$

$$X - 40 = 4(-1.28)$$

$$X = 40 - 5.12 = 34.88$$



- (iv) 98 % area under the normal curve means that 0.49 area each to the left and right of mean. From the area table, the value of Z for which the area between 0 and Z is 0.49 is 2.33. Since X_1 will be on the left of the mean, therefore Z_1 will be negative and X_2 is on the right of the mean, therefore Z_2 will be positive



$$Z_1 = \frac{X_1 - 40}{4}$$

$$-2.33 = \frac{X_1 - 40}{4}$$

$$X_1 - 40 = 4(-2.33)$$

$$X_1 = 40 - 9.32 = 30.68$$

$$Z_2 = \frac{X_2 - 40}{4}$$

$$+ 2.33 = \frac{X_2 - 40}{4}$$

$$X_2 - 40 = 4(2.33)$$

$$X_2 = 40 + 9.32 = 49.32$$

Hence two points are 30.68 and 49.32.

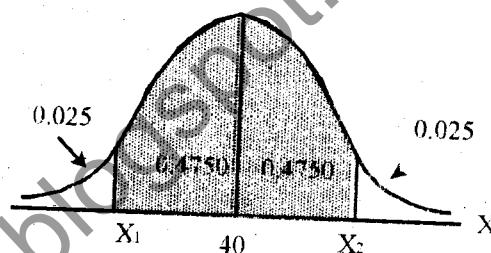
- (v) 95 % area under the normal curve means that 0.4750 area each to the left and right of mean. From the area table, the value of Z for which the area between 0 and Z is 0.4750 is 1.96. Since X_1 will be on the left of the mean, therefore Z_1 will be negative and X_2 is on the right of the mean, therefore Z_2 will be positive

$$Z_1 = \frac{X_1 - 40}{4}$$

$$-1.96 = \frac{X_1 - 40}{4}$$

$$X_1 - 40 = 4(-1.96)$$

$$X_1 = 40 - 7.84 = 32.16$$



$$Z_2 = \frac{X_2 - 40}{4}$$

$$+ 1.96 = \frac{X_2 - 40}{4}$$

$$X_2 - 40 = 4(1.96)$$

$$X_2 = 40 + 7.84 = 47.84$$

Hence two points are 32.16 and 47.84.

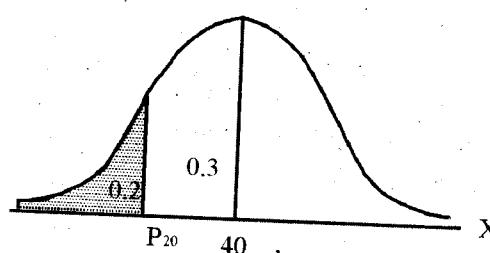
- (vi) P_{20} is a point having 20 percent of the area below it. Area table shows this point to be

$$-0.8415 = \frac{X - 40}{4}$$

$$X - 40 = 4(-0.8415)$$

$$X = 40 - 3.366 = 36.634$$

Thus, $P_{20} = 36.634$.



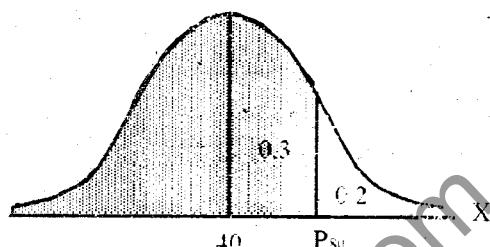
P_{80} is a point having 80 percent of the area below it. Area table shows this point to be

$$0.8415 = \frac{X - 40}{4}$$

$$X - 40 = 4(0.8415)$$

$$X = 40 + 3.366 = 43.366$$

Thus, $P_{80} = 43.366$.



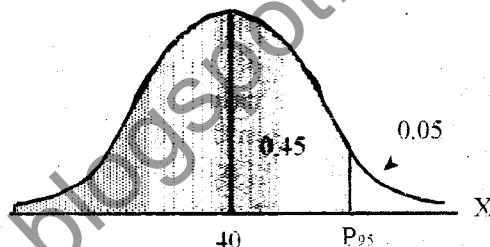
P_{95} is a point having 95 percent of the area below it. Area table shows this point to be

$$1.645 = \frac{X - 40}{4}$$

$$X - 40 = 4(1.645)$$

$$X = 40 + 6.58 = 46.58$$

Thus, $P_{95} = 46.58$.



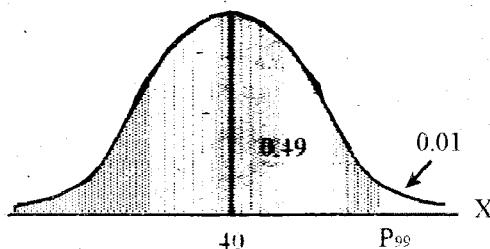
P_{99} is a point having 99 percent of the area below it. Area table shows this point to be

$$2.33 = \frac{X - 40}{4}$$

$$X - 40 = 4(2.33)$$

$$X = 40 + 9.32 = 49.32$$

Thus, $P_{99} = 49.32$.



Example 10.9.

In a normal distribution 31% of the items are under 54 and 8% are over 76. Find the mean and the standard deviation of the distribution.

Solution:

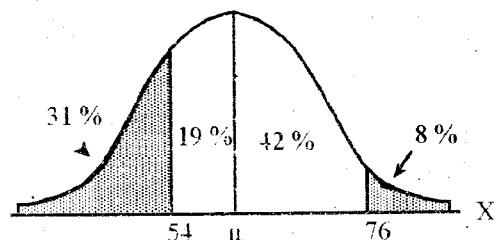
Let μ = Mean, σ = Standard deviation of the normal distribution.

$$Z = \frac{X - \mu}{\sigma} \quad Z_1 = \frac{54 - \mu}{\sigma} = \frac{54 - \mu}{\sigma} \quad Z_2 = \frac{76 - \mu}{\sigma} = \frac{76 - \mu}{\sigma}$$

Since 31% of the items are under 54, the area to the left of the ordinate at $X = 54$, is 0.31. Therefore, the area between $X = 54$ and the mean μ is $0.5 - 0.31 = 0.19$.

Then the corresponding value of Z_1 is 0.4958

$$\therefore Z_1 = -0.4958 = \frac{54 - \mu}{\sigma}$$



(∴ We have taken Z_1 to be negative because it falls on the left side of the ordinate at mean)

$$54 - \mu = -0.4958 \sigma \text{ or } \mu - 0.4958 \sigma = 54 \quad \dots \dots \quad (1)$$

Again

It is given that 8 % of the items are over 76. Therefore, the area under the normal curve between μ and 76 is 0.42 (or 42 %).

The corresponding value of Z_2 is 1.4053 i.e. $Z_2 = 1.4053 = \frac{76 - \mu}{\sigma}$

(We have taken Z_2 to be positive because it falls on the right of the mean ordinate)

$$76 - \mu = 1.4053 \sigma \text{ or } \mu + 1.4053 \sigma = 76 \quad \dots \dots \quad (2)$$

Solving equations (1) and (2), we get $1.9011 \sigma = 22$ or $\sigma = 11.57$

Substituting $\sigma = 11.57$ in equation (1), we get

$$\mu - 0.4958(11.57) = 54 \text{ or } \mu = 54 + 5.7364 = 59.7364 \text{ or } 59.74$$

Hence, Mean = 59.74 and S.D. = 11.57

Example 10.10.

In a normal distribution the lower quartile is 10 and the upper quartile is 22. Find mean and standard deviation of the distribution.

Solution: Here, $Q_1 = 10$ and $Q_3 = 22$

The two quartiles are given by

$$Q_1 = \mu - 0.6745 \sigma \text{ and } Q_3 = \mu + 0.6745 \sigma$$

Substituting the values of Q_1 and Q_3 , we get

$$\mu - 0.6745 \sigma = 10 \quad \dots \dots \quad (1) \quad \mu + 0.6745 \sigma = 22 \quad \dots \dots \quad (2)$$

Solving equations (1) and (2), we get $2\mu = 32$ or $\mu = 16$

Substituting $\mu = 16$ in equation (1), we get

$$16 - 0.6745 \sigma = 10 \text{ or } \sigma = 8.9$$

Thus, the mean and standard deviation of the normal distribution are 16 and 8.9 respectively.

SHORT DEFINITIONS

Normal Distribution

A normal distribution is a particular idealized, smooth, bell-shaped histogram with all of the randomness removed. It represents an ideal data set that has lots of numbers concentrated in the middle of the range and trails off symmetrically on both sides. A data set follows a normal distribution if it resembles the smooth, symmetric, bell-shaped normal curve, except for some randomness. The normal distribution plays an important role in statistical theory and practice.

Standard Normal Distribution

The distribution of a normal random variable with mean zero and standard deviation one is called a standard normal distribution.

or

A normal distribution that has a mean of zero and standard deviation of one is called the standard normal distribution. If Z is the standard normal random variable, then

Z has the probability distribution $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ for $-\infty < z < +\infty$.

MULTIPLE - CHOICE QUESTIONS

1. A normal distribution has the mean $\mu = 200$. If 70 percent of the area under the curve lies to the left of 220, the area to the right of 220 is:

(a) 0.3	(b) 0.5
(c) 0.2	(d) 0.7
2. Given a normal distribution with $\mu = 100$ and $\sigma^2 = 100$, the area to the left of 100 is:

(a) one	(b) equal to 0.5
(c) less than 0.5	(d) greater than 0.5
3. A random variable has a normal distribution with the mean $\mu = 400$. If 80 percent of the area under the curve lies to the left of 500, the area between 400 and 500 is:

(a) 0.5	(b) 0.2
(c) 0.3	(d) zero
4. In a normal distribution mean is 100 and standard deviation is 10. The values of points of inflection are:

(a) 100 and 110	(b) 80 and 120
(c) 90 and 110	(d) none of the above
5. If X is a normal variate with mean 20 and variance 16. The respective values of β_1 and β_2 are:

(a) 0 and 3	(b) 3 and 1
(c) 0.5 and 1	(d) 3 and 3

6. A random variable X is normally distributed with $\mu = 70$ and $\sigma^2 = 25$. The third moment about arithmetic mean is:
- zero
 - less than zero
 - greater than zero
 - none of the above
7. If X is $N(100, 25)$, the fourth central moment is: $\text{Ans} = 35^2$
- 65
 - 75
 - 85
 - 100
- For the standard normal distribution, $P(Z > \text{mean})$ is:
- more than 0.5
 - less than 0.5
 - equal to 0.5
 - difficult to tell
9. Given a standardized normal distribution (with a mean of zero and a standard deviation of one), $P(Z < \text{variance})$ is equal to:
- 0.8413
 - 0.3413
 - 0.1587
 - 0.5000
10. If $Y = 5X + 10$ and X is $N(10, 25)$, then mean of Y is: $E_Y = a + b\mu$
- 50
 - 60
 - 70
 - 135
11. If X is a normal random variable with mean $\mu = 50$ and standard deviation $\sigma = 7$, if $Y = X - 7$ then standard deviation of Y is: $E(Y) = E(X) + 10 = 60$
- 7
 - 14
 - 0
 - 49
12. The area to the left of $(\mu + \sigma)$ for a normal distribution is approximately equal to: $Z = \frac{(Y+5)-X}{7}$
- 0.16
 - 0.34
 - 0.50
 - 0.84
13. For a normal distribution with $\mu = 10$, $\sigma = 2$, the probability of a value greater than 10 is: $Z = \frac{(Y+5)-X}{7}$
- 0.1915
 - 0.3085
 - 0.6915
 - 0.5000
14. For a normal distribution with mean μ and standard deviation σ : $Z = \frac{(Y+5)-X}{7}$
- Approximately 5 % of values are outside the range $(\mu - 2\sigma)$ to $(\mu + 2\sigma)$
 - Approximately 5 % of values are greater than $(\mu + 2\sigma)$
 - Approximately 5 % of values are outside the range $(\mu - \sigma)$ to $(\mu + \sigma)$
 - Approximately 5 % of values are less than $(\mu - 3\sigma)$
15. The normal distribution is a proper probability distribution of a continuous random variable, the total area under the curve $f(x)$ is: $Z = \frac{(Y+5)-X}{7}$
- equal to one
 - less than one
 - more than one
 - between -1 and +1
16. In a normal probability distribution of a continuous random variable, the value of standard deviation is: $Z = \frac{(Y+5)-X}{7}$
- zero
 - less than zero
 - greater than zero
 - none of the above

17. The value of e is approximately equal to:
- (a) 2.7183
 - (b) 2.1783
 - (c) 2.8173
 - (d) 3.1416
18. The value of π is approximately equal to:
- (a) 3.4116
 - (b) 3.1416
 - (c) 3.1614
 - (d) 3.6416
19. The normal probability distribution with mean np and variance npq may be used to approximate the binomial distribution if $n \geq 50$ and both np and nq are:
- (a) greater than 5
 - (b) less than 5
 - (c) equal to 5
 - (d) difficult to tell
20. The parameters of the normal distribution are:
- (a) μ and σ^2
 - (b) μ and σ
 - (c) np and nq
 - (d) n and p .
21. The median of a normal distribution corresponds to a value of Z is:
- (a) 0
 - (b) 1
 - (c) 0.5
 - (d) -0.5
22. The mean and standard deviation of the standard normal distribution are respectively:
- (a) 0 and 1
 - (b) 1 and 0
 - (c) μ and σ^2
 - (d) π and e
23. If a normal distribution with $\mu = 200$ has $P(X > 225) = 0.1587$, then $P(X < 175)$ equal to:
- (a) 0.3413
 - (b) 0.8413
 - (c) 0.1587
 - (d) 0.5000
24. Given a random variable X which is normally distributed with a mean and variance both equal to 100. The value of mean deviation is approximately equal to:
- (a) 7
 - (b) 8
 - (c) 8.5
 - (d) 9
25. If X is a normal variate with mean 50 and standard deviation 3. The value of quartile deviation is approximately equal to:
- (a) 1
 - (b) 1.5
 - (c) 2
 - (d) 2.5
26. In normal probability distribution for a continuous random variable, the value of mean deviation is approximately equal to :
- (a) $2/3$
 - (b) $2/3 \sigma$
 - (c) $4/5$
 - (d) $4/5 \sigma$

27. In a normal distribution whose mean is μ and standard deviation σ , the value of quartile deviation is approximately :
- $4/5$
 - $4/5 \sigma$
 - $2/3 \sigma$
 - $2/3$
28. In a normal distribution, the lower and upper quartiles are equidistant from the mean and are at a distance of :
- 0.7979
 - 0.7979σ
 - 0.6745
 - 0.6745σ
29. In a normal curve, the ordinate is highest at:
- mean
 - variance
 - standard deviation
 - Q_1
30. The total area of the normal probability density function is equal to:
- 0
 - 0.5
 - 1
 - 0.25
31. The normal curve is symmetrical and for symmetrical distribution, the values of all odd order moments about mean will always be:
- 1
 - 0.5
 - 0.25
 - 0
32. In a normal curve $\mu \pm 0.6745 \sigma$ covers:
- 50 % area
 - 68.27 % area
 - 95.45 % area
 - 99.73 % area
33. The skewness and kurtosis of the normal distribution are respectively:
- zero and zero
 - zero and one
 - one and zero
 - one and one
34. In a normal curve, the highest point on the curve occurs at the mean, μ , which is also the:
- median and mode
 - geometric mean and harmonic mean
 - lower and upper quartiles
 - variance and standard deviation
35. The normal probability density function/curve is symmetrical about the mean, μ , i.e. the area to the right of the mean is the same as the area to the left of the mean. This means that $P(X < \mu) = P(X > \mu)$ is equal to:
- 0
 - 1
 - 0.5
 - 0.25

36. The shape of the normal curve depends upon the value of:
 (a) standard deviation (b) Q_1
 (c) mean deviation (d) quartile deviation
37. In a standard normal distribution, the value of mode is:
 (a) equal to zero (b) less than zero
 (c) greater than zero (d) exactly one
38. In a standard normal distribution, the area to the left of $Z = 1$ is:
 (a) 0.6413 (b) 0.7413
 (c) 0.8413 (d) 0.3413
39. The semi-inter quartile range for a standard normal random variable Z is:
 (a) 0.6745 (b) 0.6745σ
 (c) 0.7979 (d) 0.7979σ
40. The lower and upper quartiles for a standardized normal variate are respectively:
 (a) -0.6745σ and 0.6745σ (b) -0.6745 and 0.6745
 (c) -0.7979σ and 0.7979σ (d) -0.7979 and 0.7979
41. The value of the standard deviation σ of a normal distribution is always:
 (a) equal to zero (b) greater than zero
 (c) less than zero (d) equal to 0.5
42. The maximum ordinate of a normal curve is at:
 (a) $X = \mu$ (b) $X = \mu + \sigma$
 (c) $X = \mu - 2\sigma$ (d) $X = \sigma^2$
43. If $X \sim N(100, 64)$, then standard deviation σ is:
 (a) 100 (b) 64
 (c) 8 (d) $100 - 64 = 36$
44. The value of second moment about the mean in a normal distribution is 5. The fourth moment about the mean in the distribution is:
 (a) 5 (b) 15
 (c) 25 (d) 75
45. Most of the area under the normal curve with parameters μ and σ lies between:
 (a) $\mu - 0.5\sigma$ and $\mu + 0.5\sigma$ (b) $\mu - \sigma$ and $\mu + \sigma$
 (c) $\mu - 2\sigma$ and $\mu + 2\sigma$ (d) $\mu - 3\sigma$ and $\mu + 3\sigma$
46. If X is a normal random variable having mean μ , then $E|X - \mu|$ is equal to:
 (a) variance (b) standard deviation
 (c) quartile deviation (d) mean deviation

47. If X is a normal random variable having mean μ , then $E(X - \mu)^2$ is equal to:

- | | |
|-----------------|---------------|
| (a) σ^2 | (b) σ |
| (c) $3\sigma^4$ | (d) β_1 |

48. Which of the following is possible in normal distribution:

- | | |
|------------------|------------------|
| (a) $\sigma < 0$ | (b) $\sigma = 0$ |
| (c) $\sigma > 0$ | (d) $\sigma > n$ |

49. The range of normal distribution is:

- | | |
|--------------|----------------------------|
| (a) 0 to n | (b) 0 to ∞ |
| (c) -1 to +1 | (d) $-\infty$ to $+\infty$ |

50. The range of standard normal distribution is:

- | | |
|--------------|----------------------------|
| (a) 0 to n | (b) 0 to ∞ |
| (c) 0 to k | (d) $-\infty$ to $+\infty$ |

51. In the normal distribution, the value of the maximum ordinate is equal to:

- | | |
|-----------------------------------|-----------------------------------|
| (a) $\frac{1}{\sqrt{2\pi}}$ | (b) $\frac{1}{\sqrt{2\pi e}}$ |
| (c) $\frac{1}{\sqrt{2\pi\sigma}}$ | (d) $\frac{1}{\sigma\sqrt{2\pi}}$ |

52. The value of the ordinate at points of inflection of the normal curve is equal to:

- | | |
|-----------------------------------|-----------------------------------|
| (a) $\frac{1}{\sqrt{2\pi}}$ | (b) $\frac{1}{\sqrt{2\pi e}}$ |
| (c) $\frac{1}{\sqrt{2\pi\sigma}}$ | (d) $\frac{1}{\sigma\sqrt{2\pi}}$ |

53. $P(\mu - \sigma < X < \mu + \sigma)$ is equal to:

- | | |
|------------|------------|
| (a) 0.5000 | (b) 0.6827 |
| (c) 0.9545 | (d) 0.9973 |

54. In a normal curve $\mu \pm 2\sigma$ covers:

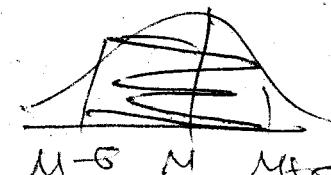
- | | |
|-----------------|-----------------|
| (a) 50% area | (b) 68.27% area |
| (c) 95.45% area | (d) 99.73% area |

55. In X is $N(\mu, \sigma^2)$, the percentage of the area contained within the limits $\mu \pm 3\sigma$ is:

- | | |
|------------|------------|
| (a) 50% | (b) 68.27% |
| (c) 95.45% | (d) 99.73% |

56. The probability density function of the standard normal distribution is:

- | | |
|--|--|
| (a) $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$ | (b) $\frac{1}{\sigma\sqrt{2\pi e}} e^{-\frac{Z^2}{4}}$ |
| (c) $\frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$ | (d) $\frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{4}}$ |



57. The equation of the normal frequency distribution is:

(a) $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

(b) $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

(c) $\frac{1}{\sigma\sqrt{2\pi}e} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

(d) $\frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

58. If X is $N(\mu, \sigma^2)$ and if $Y = a + bX$, then mean and variance of Y are respectively:

(a) μ and σ^2

(b) $a + \mu$ and $b\sigma^2$

(c) $a + b\mu$ and σ^2

(d) $a + b\mu$ and $b^2\sigma^2$

59. If Z is a standard normal variate, then $P(|Z| < 1.96)$ is equal to:

(a) 0.0250

(b) 0.4750

(c) 0.95

(d) 0.9750

60. If Z is a standard normal variate, then $P(-2.575 \leq Z \leq +2.575)$ is equal to:

(a) 0.9951

(b) 0.99

(c) 0.4951

(d) 0.4949

61. If Z is a standard normal variate, then $P(-1.645 \leq Z \leq +1.645)$ is equal to:

(a) 0.90

(b) 0.95

(c) 0.98

(d) 0.99

62. If Z is a standard normal variate, then $P(-2.33 \leq Z \leq +2.33)$ is equal to:

(a) 0.4901

(b) 0.6827

(c) 0.9545

(d) 0.9802

63. In normal distribution:

(a) mean = median = mode

(b) mean < median < mode

(c) mean > median > mode

(d) mean ≠ median ≠ mode

64. In a normal distribution $Q_1 = 20$ and $Q_3 = 40$, then mean is equal to:

(a) 20

(b) 30

(c) 40

(d) 60

65. The value of maximum ordinate in standard normal distribution is equal to:

(a) $\frac{1}{\sqrt{2\sigma}}$

(b) $\frac{1}{\sqrt{2\pi}\sigma}$

(c) $\frac{1}{\sqrt{2\pi}}$

(d) $\frac{1}{\sigma\sqrt{2\pi}}$

ANSWERS

1. (a)	2. (b)	3. (c)	4. (c)	5. (a)	6. (a)	7. (b)	8. (c)
9. (a)	10. (b)	11. (a)	12. (d)	13. (d)	14. (a)	15. (a)	16. (c)
17. (a)	18. (b)	19. (a)	20. (b)	21. (a)	22. (a)	23. (c)	24. (b)
25. (c)	26. (d)	27. (c)	28. (d)	29. (a)	30. (c)	31. (d)	32. (a)
33. (a)	34. (a)	35. (c)	36. (a)	37. (a)	38. (c)	39. (a)	40. (b)
41. (b)	42. (a)	43. (c)	44. (d)	45. (d)	46. (d)	47. (a)	48. (c)
49. (d)	50. (d)	51. (d)	52. (b)	53. (b)	54. (c)	55. (d)	56. (c)
57. (d)	58. (d)	59. (c)	60. (b)	61. (a)	62. (d)	63. (a)	64. (b)
65. (c)							

SHORT QUESTIONS

1. Guests at a large hotel stay for an average of 9 days with a standard deviation of 2.4 days. Among 1000 guests how many can be expected to stay less than 7 days. Assume that length of stay is normally distributed.

Ans. 203

2. A machine which automatically packs potatoes into bags is known to operate with a mean of 20 kg. and standard deviation of 0.5 kg. Assuming normality, find the percentage of bags weight more than 20 kg.

Ans. 50%

3. The heights of a large sample of men were found to be approximately normally distributed with mean 67.56 inches and standard deviation 2.57 inches. Find the height exceeded by 5 % of the men.

Ans. 71.79

4. Records from a dental practice show that the probability of waiting to go into the surgery for more than 20 minutes is 0.0299. If the waiting time is normally distributed with standard deviation 3.78 minutes, find the mean waiting time.

Ans. 12.52

5. If X is normally distributed with a mean of 4 and a standard deviation of 4, find the probability that X is less than 6.

Ans. 0.6915

6. Find the probability that the value of a standard normal variable is less than 2.

Ans. 0.9772

7. Find the probability that the value of a standard normal variable exceeds 1.5.

Ans. 0.0668

8. The scores made by candidates in a certain test are normally distributed with mean equal to 500 and standard deviation equal to 100. Find the probability that a score will greater than 700.

Ans. 0.0228

9. A manufacturer of pipe knows that the pipe lengths it produces vary in diameter and that the diameters are normally distributed. The mean diameter is 1 inch, and the probability that a length of pipe will have a diameter exceeding 1.1 inches is 0.1587. Find the variance of the diameters.

Ans. 0.01 (inches)²

10. The mean wages of a certain group of workers working in a factory is Rs.285 with standard deviation of Rs.50. Find the percentage of workers who get above Rs.200.

Ans. 95.54%

11. For a normal distribution the first moment about 10 is 40 and the fourth moment about 50 is 48. Find the standard deviation of the distribution.

Ans. 2

12. In a normal distribution $\mu = 163$ and $Q_3 = 171.094$. Find the standard deviation.

Ans. 12

13. In a normal distribution the lower quartile is 10 and the standard deviation is 10. Find the mean of the distribution.

Ans. 16.745

14. A normal distribution has the mean $\mu = 85$, standard deviation $\sigma = 4.5$. Find the value of Q_3 .

Ans. 88.04

15. If X is a normal random variable with mean $\mu = 113.49$ and standard deviation $\sigma = 20$. Find the value of Q_1 .

Ans. 100

16. Define normal distribution.

17. Write down any five properties of normal distribution.

18. Write down the equation of the normal curve

- (i) with mean μ and standard deviation σ
- (ii) with mean 50 and standard deviation 10.

19. Define the normal probability density function.

20. Define the normal frequency distribution.

21. Define the standard normal distribution.

22. What is the relationship between the binomial distribution and the normal distribution?

23. Describe the important properties of the normal distribution.
24. What is a standardized normal variate.
25. Write down the ordinates of the standard normal curve at
(i) $Z = 1$ (ii) $Z = -1$.
26. Explain why odd order moments about mean equals zero for the normal distribution.
27. Explain why β_1 equals zero for the normal distribution.
28. The normal curve is defined by the equation

$$Y = f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < X < +\infty$$

where $\mu = \underline{\hspace{2cm}}$ $\sigma = \underline{\hspace{2cm}}$ $\pi = \underline{\hspace{2cm}}$ $e = \underline{\hspace{2cm}}$ $X = \underline{\hspace{2cm}}$ $Y = \underline{\hspace{2cm}}$

29. Define the points of inflection in a normal distribution.
30. Sketch the normal curve and then place the values for the means on the respective X and Z scales. Verify that the area under the normal curve between the mean and 2 standard deviations above and below it is 0.9544.
31. Sketch and verify the area under the normal curve between the mean and 3 standard deviations above and below it is 0.9973.
32. When is it appropriate to use a normal approximation to the binomial distribution?
33. Write down the basic properties of the standard normal curve.
34. Complete the following table for the normal curve with parameters μ and σ . Draw four graphs illustrating your results.

Given X-values	Corresponding Z-scores	Area between X-values
$\mu - 0.6745\sigma$ and $\mu + 0.6745\sigma$	-0.6745 and +0.6745	0.50
$\mu - 1\sigma$ and $\mu + 1\sigma$		
$\mu - 2\sigma$ and $\mu + 2\sigma$		
$\mu - 3\sigma$ and $\mu + 3\sigma$		

EXERCISES

- 1.** If the random variable Z has the standard normal distribution, find:
- (i) $P(Z < 1.46)$
 - (ii) $P(Z > 1.46)$
 - (iii) $P(Z < -1.48)$
 - (iv) $P(Z > -1.96)$
 - (v) $P(0.65 < Z < 1.99)$
 - (vi) $P(0 < Z < 1.15)$
 - (vii) $P(-1.32 < Z < 1.65)$
 - (viii) $P(-1.25 < Z < 0)$.

Ans. (i) 0.9279 (ii) 0.0721 (iii) 0.0694 (iv) 0.9750 (v) 0.2345
 (vi) 0.3749 (vii) 0.8571 (viii) 0.3944.

- 2.** In a normal distribution, M.D. = 3.9895, then find standard deviation, quartile deviation, second and fourth moments about mean of the normal distribution.

Ans. S.D. = 5, Q.D. = 3.3725, $\mu_2 = 25$, $\mu_4 = 1875$.

- 3.** In a normal distribution with $\mu = 20$ and $\sigma = 5$. Find the area:
- (i) between 20 and 30
 - (ii) less than 25
 - (iii) more than 30
 - (iv) between 12 and 18
 - (v) between 30 and 42

Ans. (i) 0.4772 (ii) 0.8413 (iii) 0.0228 (iv) 0.2898 (v) 0.0228

- 4.** The mean sales of all the different branches of a big cloth shop is Rs.10000 with a standard deviation of Rs.3000. You are required to determine the percentage/proportion of shops the sales of which are between Rs.11000 and Rs.12000.

Ans. 11.93 %.

- 5.** X is a normal variate with mean 1 and standard deviation 3, find the probability that: (i) $3.43 \leq X \leq 6.19$ (ii) $-1.43 \leq X \leq 6.19$.

Ans. (i) 0.1672 (ii) 0.7492

- 6.** In a normal distribution the mean is five and the variance is one. Write down its equation. Also find the value of maximum ordinate correct to two places of decimals.

Ans. $Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X-5)^2}$, 0.40

- 7.** The scores made by candidates in a certain test are normally distributed with mean 500 and standard deviation 100. What percent of the candidates received scores: (i) greater than 700 (ii) less than 400 (iii) between 400 and 600 (iv) which differ from mean by more than 150.

Ans. (i) 2.28 % (ii) 15.87 % (iii) 68.26 % (iv) 13.36 %

- 8.** If the average height of miniature poodles is 30 centimeters, with a standard deviation of 4.4 centimeters, what percentage of miniature poodles exceeds 35 centimeters in height, assuming that the heights follow a normal distribution and can be measured to any desired degree of accuracy?

Ans. 11.12 %

9. The diameters of bolts manufactured by a company are normally distributed with mean 0.25 inches and standard deviation 0.02 inches. A bolt is considered defective if its diameter is ≤ 0.20 or ≥ 0.28 inches. Find the percentage of defective bolts manufactured by the company.

Ans. 7.3%

10. If the weights of ball bearings are normally distributed with mean 0.6140 newtons and standard deviation 0.0025 newtons, determine the percentage of ball bearings with weights

- between 0.610 and 0.618 newtons inclusive
- greater than 0.617 newtons
- less than 0.608 newtons

Ans. (i) 89.04% (ii) 11.51% (iii) 0.82%

11. Let X be a normal random variable with mean = 16 and standard deviation = 5.

- Determine:
- $P(\text{between } 11 \text{ and } 21)$
 - $P(\text{at least } 26)$
 - $P(\text{less than or equal to } 6)$
 - $P(\text{at most } 21)$

Ans. (i) 0.6826 (ii) 0.0228 (iii) 0.0228 (iv) 0.8413

12. In a certain examination 3000 students appeared. The average marks obtained were 50% and standard deviation was 5%. How many students do you expect who obtain:
- More than 60% marks
 - Less than 40% marks
 - Between 40% and 60% marks?

Ans. (i) 68 (ii) 68 (iii) 2863

13. Assume the mean height of soldiers to be 69 inches with a variance of 9 inches. How many soldiers in a regiment of 1000 would you expect to be over six feet tall?

Ans. 159

14. The mean life of stockings used by an army was 40 days, with a standard deviation of 8 days. Assume the life of the stockings follows a normal distribution. If 100000 pairs are issued, how many would need replacement

- before 35 days?
- after 46 days?

Ans. (i) 26600 (ii) 22660

15. Given that $\mu = 300$ and $\sigma^2 = 100$. Find

- the area above 314
- the two values that contain the middle 75% area
- Q_1 and Q_3 of the normal distribution.

Ans. (i) 0.0808 (ii) 288.5, 311.5 (iii) 293.255, 306.745

16. A random variable X is normally distributed with mean = 70 and S.D. = 5.

- Find a point that has 87.9% of the distribution below it.
- Find a point that has 81.7% of the distribution above it.
- Find two such points between which the central 70% of the distribution lies.

- Find two such points between which the central 90% of the distribution lies.

Ans. (i) 75.85 (ii) 65.48 (iii) 64.815, 75.185 (iv) 61.775, 78.225

28

~~17.~~ In a normal distribution $\mu = 40$ and $\sigma = 3.8$. Find:

- two points such that the curve has a 98 % chance of falling between them
- the chance that a single observation will be less than 38.6.

Ans. (i) 31.146, 48.854 (ii) 0.3557

~~18.~~ If X is $N(24, 16)$, then find the: (i) 33rd percentile (ii) 9th decile.

Ans. (i) 22.24 (ii) 29.12.

~~19.~~ In a normal distribution 30 % of the values are under 50 and 10 % are over 70. Find the mean and standard deviation of the distribution.

Ans. (55.81, 11.09)

~~20.~~ Given a normal distribution with $\mu = 40$ and $\sigma = 6$, find the value X that has:

- 38 % of the area below it.
- 5 % of the area above it.

Ans. (i) 38.167 (ii) 49.87

~~21.~~ Assume that we have a large number of students whose average weight is 150 lbs. and that the weights are normally distributed. If we know that 36.4 % of the students have weight between 137 and 163 lbs. What is the standard deviation of the weights.

Ans. 27.66

~~22.~~ In a normal distribution 25 % of the items are under 60 and 15 % are over 90. Find the mean and standard deviation of the distribution.

Ans. (71.82, 17.53)

~~23.~~ In a "normal distribution" the lower and upper quartiles are 18 and 26 respectively. Find its mean and standard deviation. Also find mean deviation.

Ans. 22, 5.93, 4.73

~~24.~~ In a normal distribution 7 % of the items are under 35 and 89 % are under 63. What is the mean and standard deviation of the distribution?

Ans. 50.29, 10.33

~~25.~~ The heights of a large sample of men were found to be approximately normally distributed with mean 67.56 inches and standard deviation 2.57 inches. What height is exceeded by 5 % of the men?

Ans. 71.79.

Chapter

11

SAMPLING AND SAMPLING DISTRIBUTIONS

11.1 INTRODUCTION

In our daily life it is quite often that we have to examine some given material. We examine fruit before we purchase it, we make a small study of the material whenever we have to purchase something. Even the children check the sweets, pencils, bats, rubbers and other items when they have to purchase them. This approach is applied in different fields of life. The products of the factories are inspected to ensure the desired quality of the products. The medicines are manufactured on commercial scale when their effects have been tested on the patients. The different fertilizers are tested on agricultural plots and different foods are tested on animals. Small dams are constructed in the laboratories to study the life and other characteristics of the big dams before they are actually constructed. Some colour may be applied on a wall, on a door or cloth etc., and the result of the colour is observed before it is applied on large scale. Cement, steel and bricks are examined before using them in different places. This process of inspection is very wide and is commonly used on various occasions. But this job is never done on very large scale. This process is carried out on a small scale. On the basis of this small study, we make an opinion about the entire material under study.

11.2 POPULATION

The word *population* or *statistical population* is used for all the individuals or objects on which we have to make some study. We may be interested to know the quality of bulbs produced in a factory. The entire product of the factory in a certain period is called a *population*. We may be interested in the level of education in primary schools. All the children in the primary schools will make a *population*. The *population* may contain living or non-living things. The entire lot of anything under study is called *population*. All the fruit trees in a garden, all the patients in a hospital and all the cattle in a cattle farm are examples of *populations* in different studies.

11.2.1 FINITE POPULATION

A *population* is called *finite* if it is possible to count its individuals. It may also be called a *countable population*. The number of vehicles crossing a bridge every day, the number of deaths per year and the number of words in a book are *finite populations*. The number of units in a *finite population* is denoted by N . Thus N is the size of the *population*.

11.2.2 INFINITE POPULATION

Sometimes it is not possible to count the units contained in the *population*. Such a *population* is called *infinite* or *uncountable*. Let us suppose that we want to examine whether a coin is true or not. We shall toss it a very large number of times to observe the number of heads. All the tosses will make an *infinite* or *countably infinite population*. The number of germs in the body of a patient of malaria is perhaps something which is *uncountable*.

11.2.3 TARGET AND SAMPLED POPULATION

Suppose we have to make a study about the problems of the families living in rented houses in a certain big city. All the families living in rented houses is our *target population*. The entire *target population* may not be considered for the purpose of selecting a sample from the population. Some families may not be interested to be included in the sample. We may ignore some part of the *target population* to reduce the cost of study. The *population* out of which the sample is selected is called *sampled population* or *studied population*.

11.3 SAMPLE

Any part of the population is called a *sample*. A study of the *sample* enables us to make some decisions about the properties of the population. The number of units included in the *sample* is called the size of the *sample* and is denoted by n . A good *sample* is that one which speaks about the qualities of the population. A *sample* study leads us to make some inferences about the population measures. This process is called *sampling*.

11.3.1 PARAMETER AND STATISTIC

Any measure of the population is called *parameter* and the word *statistic* is used for any value calculated from the sample. The population mean μ is a *parameter* and the sample mean \bar{X} is a *statistic*. The sample mean \bar{X} is used to estimate the population mean μ . Similarly the population variance σ^2 is a *parameter* and the sample variance S^2 is a *statistic*. In general the symbol θ is used for a *parameter* and the symbol $\hat{\theta}$ is used for a *statistic*. The value of the *parameter* is mostly unknown and the sample statistic is used to make some inferences about the unknown *parameter*.

11.3.2 SAMPLING FRACTION

If size of the population is N and size of the sample is n , the ratio $\frac{n}{N}$ is called

the *sampling fraction*. If $N = 100$, $n = 10$, the ratio $\frac{n}{N} = \frac{10}{100} = \frac{1}{10}$. It means that on the average 10 units of the population will be represented by one unit in the sample.

- If the *sampling fraction* $\frac{n}{N}$ is multiplied with 100, we get the *sampling fraction* in percentage form. Thus $\frac{n}{N} \times 100 = \frac{10}{100} \times 100 = 10\%$. It means 10% of the population is included in the sample.

11.4 COMPLETE COUNT

If we collect information about all the individuals in the population, the study is called *complete count* or *complete enumeration*. The word *census* is also used for the entire population study. In statistical studies the *complete count* is usually avoided. If size of the population is large, the *complete count* requires a lot of time and a lot of funds. The *complete count* is mostly difficult for various reasons. Suppose we want to make a study about the cattle in the cattle farms in our country. We are interested in the average cost of their food for a certain period. We want to link their cost of food with their sale price. This is of course, an important study. It is very difficult to collect and maintain the information about each and every cattle in the farms. If at all we are able to do it, the study may not be of much use. The desired information can be obtained from a reasonable sample size of the cattles.

11.4.1 POPULATION CENSUS

A complete count of the human population is called *population census*. In Pakistan, the first *population census* was conducted in 1951 and the second was conducted in 1961. The third *census of population* could not be conducted in 1971 because of agitations in the then East Pakistan. It was conducted in 1972. The 4th *census* was conducted in 1981. The fifth population census was conducted in 1998. A lot of information is collected about the human population through the *population census* conducted regularly after every 10 years. The *census* reports give information about various characteristics of the population e.g., the urban and rural population, the skilled and un-skilled labour force, the agricultural labour force and the industrial workers, level of education and illiteracy in the country, geographical distribution of the population, age and sex distribution of the population etc.

11.5 SAMPLE SURVEY

If it is not essential to conduct the complete enumeration, then a sample of some suitable size is selected from the population and the study is carried out on the sample. This study is called *sample survey*. Most of the research work is done through *sample surveys*. The opinion of the voters in favour of certain proposed election candidates is obtained through *sample surveys*.

11.5.1 ADVANTAGES OF SAMPLING

Sampling has some advantages over the complete count. These are:

(i) Need for Sampling

Sometimes there is a need for sampling. Suppose we want to inspect the eggs, the bullets, the missiles and the tires of some firm. The study may be such that the objects are destroyed during the process of inspection. Obviously, we cannot afford to destroy all the eggs and the bullets etc. We have to take care that the wastage should be minimum. This is possible only in sample study. Thus sampling is essential when the units under study are destroyed.

(ii) Saves Time and Cost

As the size of the sample is small as compared to the population, the time and cost involved on sample study are much less than the complete counts. For complete count huge funds are required. There is always the problem of finances. A small

sample can be studied in a limited time and total cost of sample study is very small. For complete count, we need a big team of supervisors and enumerators who are to be trained and they are to be paid properly for the work they do. Thus the sample study requires less time and less of cost.

(iii) Reliability

If we collect the information about all the units of population, the collected information may be true. But we are never sure about it. We do not know whether the information is true or is completely false. Thus we cannot say anything with confidence about the quality of information. We say that the *reliability* is not possible. This is a very important advantage of sampling. The inference about the population parameters is possible only when the sample data is collected from the selected sample.

(iv) Sometimes the experiments are done on sample basis. The fertilizers, the seeds and the medicines are initially tested on samples and if found useful, then they are applied on large scale. Most of the research work is done on the samples.

(v) Sample data is also used to check the accuracy of the census data.

11.5.2 LIMITATIONS OF SAMPLING

Sometimes the information about each and every unit of the population is required. This is possible only through the complete enumeration because the sample will not serve the purpose. Some examples in which the sampling is not allowed are:

- (i) To conduct the elections, we need a complete list of the voters. The candidates participating in the election will not accept the results prepared from a sample. With increase in literacy, the people may become statistical minded and they may become willing to accept the results prepared from the sample. In advanced countries the opinion polls are frequently conducted and unofficially the people accept the results of sample surveys.
- (ii) Tax is collected from all the tax payers. A complete list of all the tax payers is required. The telephone, gas and electricity bills are sent to all the consumers. A complete list of the owners of land and property is always prepared to maintain the records. The position of stocks in factories requires complete entries of all the items in the stock.

11.5.3 SAMPLE DESIGN

In sample studies, we have to make a plan regarding the size of the sample, selection of the sample, collection of the sample data and preparation of the final results based on the sample study. The whole procedure involved is called the *sample design*. The term sample survey is used for a detailed study of the sample. In general, the term sample survey is used for any study conducted on the sample taken from some real world data.

11.5.4 SAMPLING FRAME

A complete list of all the units of the population is called the *sampling frame*. A *unit* of population is a relative term. If all the workers in a factory make a

population, a single worker is a unit of the population. If all the factories in a country are being studied for some purpose, a single factory is a unit of the population of factories. The *sampling frame* contains all the units of the population. It is to be defined clearly as to which units are to be included in the frame. The frame provides a base for the selection of the sample.

11.5.5 EQUAL PROBABILITY

The term equal probability is frequently used in the theory of sampling. This term is quite often not understood correctly. It is thought to be close to 'equal' in meaning. It is not true always. Suppose there is a population of 50 ($N = 50$) students in a class. We select any one student. Every student has probability $1/50$ of being selected. Then a second student is selected. Now, there are 49 students in the population and every student has $1/49$ probability of being selected. When the first student is selected, all the students have equal ($1/50$) chance of selection and when the second student is selected, again all the students have equal ($1/49$) chance of selection. But $1/50$ is not equal to $1/49$. Thus equal probability of selection means the probability when the individual is selected from the remaining available units in the population. At the time of selecting a unit, the probability of selection is equal. It is called *equal probability* of selection.

11.5.6 KNOWN PROBABILITY

In sampling theory the term *known probability* is used in random (probability) sampling. Let us explain it by taking an example. Suppose there are 300 workers in a certain factory out of which 200 are skilled and 100 are non-skilled. We have to select one sample (sub-sample) out of skilled workers and one sample out of unskilled workers. When the first worker out of skilled workers is selected, each worker has a probability of selection equal to $1/200$. Similarly when the first worker out of un-skilled workers is selected, each worker has a probability of selection equal to $1/100$. Both these probabilities are *known*, though they are not equal.

11.5.7 NON-ZERO PROBABILITY

Suppose we have a population of 500 students out of which 50 are non-intelligent. We have decided to select an intelligent student from the population. The probability of selecting an intelligent student is $1/450$ which is *non-zero*. In this example, we have decided to exclude the non-intelligent students from the population for the purpose of selecting a sample. Thus probability of selecting a non-intelligent student is zero.

11.6 PROBABILITY AND NON-PROBABILITY SAMPLING

The term *probability sampling* is used when the selection of the sample is purely based on chance. The human mind has no control on the selection or non-selection of the units for the sample. Every unit of the population has known non-zero probability of being selected for the sample. The probability of selection may be equal or unequal but it should be non-zero and should be *known*. The *probability sampling* is also called the random sampling (not simple random sampling). Some examples of random sampling are:-

- (i) Simple random sampling.
- (ii) Stratified random sampling.
- (iii) Systematic random sampling.

In *non-probability sampling*, the sample is not based on chance. It is rather determined by some person. We cannot assign to an element of population the probability of its being selected in the sample. Somebody may use his personal judgement in the selection of the sample. In this case the sampling is called *judgement sampling*. A drawback in *non-probability sampling* is that such a sample cannot be used to determine the error. Any statistical method cannot be used to draw inference from this sample. But it should be remembered that judgement sampling becomes essential in some situations. Suppose we have to take a small sample from a big heap of coal. We cannot make a list of all the pieces of coal. The upper part of the heap will have perhaps big pieces of coal. We have to use our judgement in selecting a sample to have an idea about the quality of coal. The *non-probability sampling* is also called non-random sampling.

11.6.1 SAMPLING WITH REPLACEMENT

Sampling is called *with replacement* when a unit selected at random from the population is returned to the population and then a second element is selected at random. Whenever a unit is selected, the population contains all the same units. A unit may be selected more than once. There is no change at all in the size of the population at any stage. We can assume that a sample of any size can be selected from the given population of any size. This is only a theoretical concept and in practical situations the sample is not selected by using this scheme of selection. Suppose the population size $N = 5$ and sample size $n = 2$, and sampling is done *with replacement*. Out of 5 elements, the first element can be selected in 5 ways. The selected unit is returned to the main lot and now the second unit can also be selected in 5 ways. Thus in total there are $5 \times 5 = 25$ samples or pairs which are possible. Suppose a container contains 3 good bulbs denoted by G_1 , G_2 and G_3 and 2 defective bulbs denoted by D_1 and D_2 . If any two bulbs are selected *with replacement*, there are 25 possible samples listed between in Table 11.1.

Table 11.1

	G_1	G_2	G_3	D_1	D_2
G_1	G_1G_1	G_1G_2	G_1G_3	G_1D_1	G_1D_2
G_2	G_2G_1	G_2G_2	G_2G_3	G_2D_1	G_2D_2
G_3	G_3G_1	G_3G_2	G_3G_3	G_3D_1	G_3D_2
D_1	D_1G_1	D_1G_2	D_1G_3	D_1D_1	D_1D_2
D_2	D_2G_1	D_2G_2	D_2G_3	D_2D_1	D_2D_2

The number of samples is given by $N^n = 5^2 = 25$. The selected sample will be any one of the 25 possible samples. Each sample has equal probability $1/25$ of selection. A sample selected in this manner is called simple random sample.

11.6.2 SAMPLING WITHOUT REPLACEMENT

Sampling is called *without replacement* when a unit is selected at random from the population and it is not returned to the main lot. First unit is selected out of a population of size N and the second unit is selected out of the remaining population of $N - 1$ units and so on. Thus the size of the population goes on decreasing as the sample size n increases. The sample size n cannot exceed the population size N . The unit once selected for a sample cannot be repeated in the same sample. Thus all the units of the sample are distinct from one another. A sample *without replacement* can be selected either by using the idea of permutations or combinations. Depending upon the situation, we write all possible permutations or combinations. If the different arrangements of the units are to be considered, then the permutations (arrangements) are written to get all possible samples. If the arrangement of units is of no interest, we write the combinations to get all possible samples.

11.6.3 COMBINATIONS

Let us again consider a lot (population) of 5 bulbs with 3 good (G_1, G_2 and G_3) and 2 defective (D_1 and D_2) bulbs. Suppose we have to select two bulbs in any order there are ${}^5C_2 = \frac{5!}{2!(3!)} = 10$ possible *combinations or samples*. These *combinations (samples)* are listed as: $G_1G_2, G_1G_3, G_2G_3, G_1D_1, G_1D_2, G_2D_1, G_2D_2, G_3D_1, G_3D_2, D_1D_2$.

There are 10 possible samples and each of them has probability of selection equal to $1/10$. The selected sample will be any one of these 10 samples. The sample selected in this manner is also called simple random sample. In general, the number of samples by *combinations* is equal to ${}^N C_n = \frac{N!}{n!(N-n)!}$.

11.6.4 PERMUTATIONS

Each combination generates a number of arrangements (*permutations*). Thus in general the number of *permutations* is greater than the number of combinations. In the previous example of bulbs, if the order of the selected bulbs is to be considered then the number of samples by *permutations* is given by ${}^5P_2 = \frac{5!}{(5-2)!} = 20$. These samples are:

$G_1G_2 \quad G_2G_1 \quad G_1G_3 \quad G_3G_1 \quad G_2G_3 \quad G_3G_2 \quad G_1D_1 \quad D_1G_1 \quad G_1D_2 \quad D_2G_1$
 $D_2D_1 \quad D_1G_2 \quad G_2D_1 \quad D_2G_2 \quad G_3D_1 \quad D_1G_3 \quad G_3D_2 \quad D_2G_3 \quad D_1D_2 \quad D_3D_1$

Each sample has probability of selection equal to $1/20$. The selected sample keeping in view the order of the bulbs will be any one of these 20 samples. A sample selected in this manner is also called simple random sample because each sample has equal probability of being selected.

11.6.5 SIMPLE RANDOM SAMPLE

Simple random sample (SRS) is a special case of a random sample. A sample is called *simple random sample* if each unit of the population has an equal chance of being selected for the sample. Whenever a unit is selected for the sample, the units

of the population are equally likely to be selected. It must be noted that the probability of selecting the first element is not to be compared with the probability of selecting the second unit. When the first unit is selected, all the units of the population have the equal chance of selection which is $1/N$. When the second unit is selected, all the remaining $(N - 1)$ units of the population have $1/(N - 1)$ chance of selection.

Another way of defining a *simple random sample* is that if we consider all possible samples of size n , then each possible sample has equal probability of being selected.

If sampling is done with replacement, there are N^n possible samples and each sample has probability of selection equal to $1/N^n$. If sampling is done without replacement with the help of combinations then there are ${}^N C_n$ possible samples and each sample has probability of selection equal to $1/{}^N C_n$. If samples are made with permutations, each sample has probability of selection equal to $1/{}^N P_n$. Strictly speaking, the sample selected by without replacement is called *simple random sample*.

11.6.6 DIFFERENCE BETWEEN RANDOM SAMPLE AND SIMPLE RANDOM SAMPLE

If each unit of the population has known (equal or un-equal) probability of selection in the sample, the sample is called a random sample. If each unit of the population has *equal* probability of being selected for the sample, the sample obtained is called simple random sample.

11.6.7 SELECTION OF SIMPLE RANDOM SAMPLE

A *simple random sample* is usually selected by without replacement. The following methods are used for the selection of a *simple random sample*:

(i) Lottery Method

This is an old classical method but it is a powerful technique and modern methods of selection are very close to this method. All the units of the population are numbered from 1 to N . This is called sampling frame. These numbers are written on the small slips of paper or the small round metallic balls. The paper slips or the metallic balls should be of the same size otherwise the selected sample will not be truly random. The slips or the balls are thoroughly mixed and a slip or ball is picked up. Again the population of slips is mixed and the next unit is selected. In this manner, the number of slips equal to the sample size n are selected. The units of the population which appear on the selected slips make the *simple random sample*. This method of selection is commonly used when size of the population is small. For a large population there is a big heap of paper slips and it is difficult to mix the slips properly.

(ii) Using a Random Number Table

All the units of the population are numbered from 1 to N or from 0 to $N - 1$. We consult the random number table to take a *simple random sample*. Suppose the size of the population is 80 and we have to select a random sample of 8 units. The units

of the population are numbered from 01 to 80. We read two-digit numbers from the table of random numbers. We can take a start from any columns or rows of the table. Let us consult *random number table* given in this book. Two-digit numbers are taken from the table. Any number above 80 will be ignored and if any number is repeated, we shall not record it if sampling is done without replacement. Let us read the first two columns of the table. The random number from the table are 10, 37, 08, 12, 66, 31, 63 and 73. The two numbers 99 and 85 have not been recorded because the population does not contain these numbers. The units of the population whose numbers have been selected constitute the *simple random sample*. Let us suppose that the size of the population is 100. If the units are numbered from 001 to 100, we shall have to read 3-digit random numbers. From the first 3 columns of the random number table, the random numbers are 100, 375, 084, 990, 128 and so on. We find that most of the numbers are above 100 and we are wasting our time while reading the table. We can avoid it by numbering the units of the population from 00 to 99. In this way, we shall read 2-digit numbers from the table. Thus if N is 100, 1000 or 10000, the numbering is done from 00 to 99, 000 to 999 or 0000 to 9999.

(III) Using the Computer

The facility of selecting a *simple random sample* is available on the computers. The computer is used for selecting a sample of prize-bond winners, a sample of Haj applicants, a sample of applicants for residential plots and for various other purposes.

11.7 ERRORS

Suppose we are interested in the value of a population parameter, the true value of which is 0 but is unknown. The knowledge about 0 can be obtained either from a sample data or from the population data. In both cases, there is a possibility of not reaching the true value of the parameter. The difference between the calculated value (from sample data or from population data) and the true value of the parameter is called *error*. Thus error is something which cannot be determined accurately if the population is large and the units of the population are to be measured. Suppose we are interested to find the total production of wheat in Pakistan in a certain year. Sufficient funds and time are at our disposal and we want to get the 'true' figure about production of wheat. The maximum we can do is that we contact all the farmers and suppose all the farmers give maximum cooperation and supply the information as honestly as possible. But the information supplied by the farmers will have *errors* in most of the cases. Thus we may not be able to identify the 'true' figure. Inspite of all efforts, we shall be in darkness. The calculated or the observed figure may be good for all practical purposes but we can never claim that a true value of the parameter has been obtained. If the study of the units is based on 'counting' may be we can get the true figure of the population parameter. There are two kinds of *errors* (i) sampling errors or random errors (ii) non-sampling errors.

11.7.1 SAMPLING ERRORS

These are the errors which occur due to the nature of sampling. The sample selected from the population is one of all possible samples. Any value calculated from the sample is based on the sample data and is called sample statistic. The sample statistic may or may not be close to the population parameter. If the statistic is $\hat{\theta}$ and the true value of the population parameter is θ , then the difference $\hat{\theta} - \theta$ is called *sampling error*. It is important to note that a statistic is a random variable and it may take any value. A particular example of *sampling error* is the difference between the sample mean \bar{X} and the population mean μ . Thus sampling error is also a random term. The population parameter is usually not known, therefore the *sampling error* is estimated from the sample data. The *sampling error* is due to the reason that a certain part of the population goes to the sample. Obviously, a part of the population cannot give the true picture of the properties of the population. But one should not get the impression that a sample always gives the result which is full of errors. We can design a sample and collect the sample data in a manner so that the *sampling errors* are reduced. The *sampling errors* can be reduced by the following methods:

- (i) by increasing the size of the sample
- (ii) by stratification.

11.7.2 REDUCING THE SAMPLING ERRORS

(i) By Increasing the size of the sample

The sampling error can be reduced by increasing the sample size. If the sample size n is equal to the population size N , then the sampling error is zero.

(ii) By Stratification

When the population contains homogeneous units, a simple random sample is likely to be representative of the population. But if the population contains dissimilar units, a simple random sample may fail to be representative of all kinds of units in the population. To improve the result of the sample, the sample design is modified. The population is divided into different groups containing similar units. These groups are called *strata*. From each group (stratum), a sub-sample is selected in a random manner. Thus all the groups are represented in the sample and sampling error is reduced. It is called stratified-random sampling. The size of the sub-sample from each stratum is frequently in proportion to the size of the stratum. Suppose a population consists of 1000 students out of which 600 are intelligent and 400 are non-intelligent. We are assuming here that we do have this much information about the population. A stratified sample of size $n = 100$ is to be selected. The size of the stratum is denoted by N_1 and N_2 respectively and the size of the samples from each stratum may be denoted by n_1 and n_2 . It is written as under:

Stratum No.	Size of stratum	Size of sample from each stratum
1	$N_1 = 600$	$n_1 = \frac{n \times N_1}{N} = \frac{100 \times 600}{1000} = 60$
2	$N_2 = 400$	$n_2 = \frac{n \times N_2}{N} = \frac{100 \times 400}{1000} = 40$
	$N_1 + N_2 = N = 1000$	$n_1 + n_2 = n = 100$

The size of the sample from each stratum has been calculated according to the size of the stratum. This is called *proportional allocation*. In the above sample design, the sampling fraction in the population is $\frac{n}{N} = \frac{100}{1000} = \frac{1}{10}$ and the sampling fraction in both the strata is also $1/10$. Thus this design is also called *fixed sampling fraction*. This modified sample design is frequently used in sample surveys. But this design requires some prior information about the units of the population. On the basis of this information, the population is divided into different strata. If the prior information is not available then the stratification is not applicable.

11.7.3 NON-SAMPLING ERRORS

There are certain sources of errors which occurs both in sample survey as well as in the complete enumeration. These errors are of common nature. Suppose we study each and every unit of the population. The population parameter under study is the population mean and the 'true' value of the parameter is μ which is unknown. We hope to get the value of μ by a complete count of all the units of the population. We get a value called 'calculated' or 'observed' value of the population mean. This observed value may be denoted by μ_{cal} . The difference between μ_{cal} and μ (true) is called *non-sampling error*. Even if we study the population units under ideal conditions, there may still be the difference between the observed value of the population mean and the true value of the population mean. *Non-sampling errors* may occur due to many reasons. Some of them are:

- (i) The units of the population may not be defined properly. Suppose we have to carry out a study about skilled labour force in our country. Who is a skilled person. Some people do more than one job. Some do the secretariat jobs as well as the technical jobs. Some are skilled but they are doing the job of un-skilled worker. Thus it is important to clearly define the units of the population otherwise there will be *non-sampling errors* both in the population count and the sample study.
- (ii) There may be poor response on the part of respondents. The people do not supply correct information about their income, their children, their age and property etc. These errors are likely to be of high magnitude in population study than the sample study. To reduce these errors the respondents are to be persuaded.

- (iii) The things in human hand are likely to be mis-handled. The enumerators may be careless or they may not be able to maintain uniformity from place to place. The data may not be collected properly from the population or from the sample. These errors are likely to be more serious in the population data than the sample data.
- (iv) Another serious error is due to 'bias'. Bias means an error on the part of the enumerator or the respondent when the data is being collected. Bias may be intentional or un-intentional. An enumerator may not be capable of reporting the correct data. If he has to report about the condition of crops in different areas after heavy rainfalls, his assessments may be biased due to lack of training or he may be inclined to give wrong reports. Bias is a serious error and cannot be reduced by increasing the sample size. Bias may be present in the sample study as well as the population study.

11.8 SAMPLING DISTRIBUTIONS

Suppose we have a finite population and we draw all possible simple random samples of size n by without replacement or with replacement. For each sample we

calculate some statistic (sample mean \bar{X} or proportion \hat{p} etc.). All possible values of the statistic make a probability distribution which is called the *sampling distribution*. The number of all possible samples is usually very large and obviously the number of statistics (any function of the sample) will be equal to the number of sample if one and only one statistic is calculated from each sample. In fact, in practical situations, the *sampling distribution* has very large number of values. The shape of the *sampling distribution* depends upon the size of the sample and the nature of the population and the statistic which is calculated from all possible simple random samples. Some of the famous *sampling distributions* are:

- (i) Binomial distribution. (ii) Normal distribution. (iii) t-distribution.
- (iv) Chi-square distribution. (v) F-distribution.

These distributions are called the derived distributions because they are derived from all possible samples.

11.8.1 STANDARD ERROR

The standard deviation of some statistic is called the *standard error* of that statistic. If the statistic is \bar{X} , the standard deviation of all possible values of \bar{X} is called *standard error of \bar{X}* which may be written as S.E. (\bar{X}) or $\sigma_{\bar{X}}$. Similarly, if the sample statistic is proportion \hat{p} , the standard deviation of all possible values of \hat{p} is called *standard error of \hat{p}* and is denoted by $\sigma_{\hat{p}}$ or S.E. (\hat{p}).

11.8.2 SAMPLING DISTRIBUTION OF \bar{X}

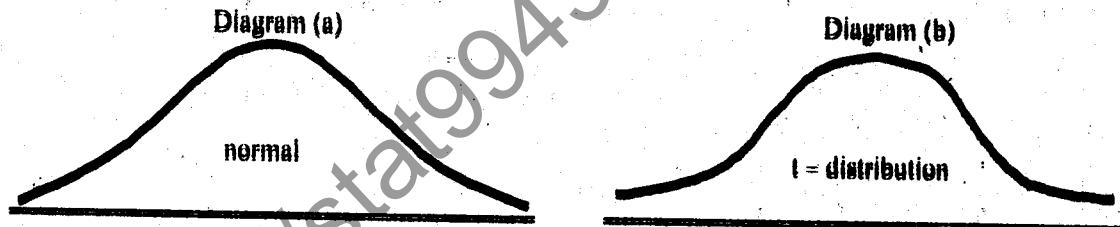
The probability distribution of all possible values of \bar{X} calculated from all possible simple random samples is called the *sampling distribution* of \bar{X} . In brief, we shall call it *distribution of \bar{X}* . The mean of this distribution is called *expected value*

of \bar{X} and is written as $E(\bar{X})$ or $\mu_{\bar{X}}$. The standard deviation (standard error) of this distribution is denoted by $S.E.(\bar{X})$ or $\sigma_{\bar{X}}$ and the variance of \bar{X} is denoted by $Var(\bar{X})$ or $\sigma_{\bar{X}}^2$. The distribution of \bar{X} has some important properties as under:

- (i) An important property of the distribution of \bar{X} is that it is a normal distribution when the size of the sample is large. When the sample size n is more than 30, we call it a large sample size. The shape of the population distribution does not matter. The population may be normal or non-normal, the distribution of \bar{X} is normal for $n > 30$. But this is true when the number of samples is very large.

As the distribution of random variable \bar{X} is normal, \bar{X} can be transformed into standard normal variable Z where $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$.

The distribution of \bar{X} has the t-distribution when the population is normal and $n \leq 30$. Diagram (a) shows the normal distribution and diagram (b) shows the t-distribution.



- (ii) The mean of the distribution of \bar{X} is equal to the mean of the population. Thus $E(\bar{X}) = \mu_{\bar{X}} = \mu$ (Population mean). This relation is true for small as well as large sample size in sampling without replacement and with replacement.
- (iii) The standard error (standard deviation) of \bar{X} is related with the standard deviation of population σ through the relations:

$$S.E.(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

This is true when population is infinite which means N is very large or the sampling is done with replacement from finite or infinite population.

$$S.E.(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

This is true when sampling is without replacement from finite population. The above two equations between $\sigma_{\bar{X}}$ and σ are true both for small as well as large sample sizes.

Example 11.1.

Draw all possible samples of size 2 without replacement from a population consisting of 3, 6, 9, 12, 15. Form the sampling distribution of sample means and verify the results:

$$(i) E(\bar{X}) = \mu \quad (ii) \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$$

Solution:

We have population values 3, 6, 9, 12, 15, population size $N = 5$ and sample size $n = 2$. Thus, the number of possible samples which can be drawn without replacement is

$$\binom{N}{n} = \binom{5}{2} = 10.$$

Sample No.	Sample Values	Sample Mean (\bar{X})	Sample No.	Sample Values	Sample Mean (\bar{X})
1	3, 6	4.5	6	6, 12	9.0
2	3, 9	6.0	7	6, 15	10.5
3	3, 12	7.5	8	9, 12	10.5
4	3, 15	9.0	9	9, 15	12.0
5	6, 9	7.5	10	12, 15	13.5

The sampling distribution of the sample mean \bar{X} and its mean and standard deviation are:

\bar{X}	f	$f(\bar{X})$	$\bar{X} f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
4.5	1	1/10	4.5/10	20.25/10
6.0	1	1/10	6.0/10	36.00/10
7.5	2	2/10	15.0/10	112.50/10
9.0	2	2/10	18.0/10	162.00/10
10.5	2	2/10	21.0/10	220.50/10
12.0	1	1/10	12.0/10	144.00/10
13.5	1	1/10	13.5/10	182.25/10
Total	10	1	90/10	877.5/10

$$E(\bar{X}) = \Sigma \bar{X} f(\bar{X}) = \frac{90}{10} = 9$$

$$\text{Var}(\bar{X}) = \Sigma \bar{X}^2 f(\bar{X}) - [\Sigma \bar{X} f(\bar{X})]^2 = \frac{877.5}{10} - \left(\frac{90}{10} \right)^2 = 87.75 - 81 = 6.75$$

The mean and variance of the population are:

X	3	6	9	12	15	$\Sigma X = 45$
X^2	9	36	81	144	225	$\Sigma X^2 = 495$

$$\mu = \frac{\Sigma X}{N} = \frac{45}{5} = 9 \text{ and } \sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{495}{5} - \left(\frac{45}{5}\right)^2 = 99 - 81 = 18$$

Verification:

$$(i) E(\bar{X}) = \mu = 9 \quad (ii) \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right) = \frac{18}{2} \left(\frac{5-2}{5-1} \right) = 6.75$$

Example 11.2

If random samples of size three are drawn without replacement from the population consisting of four numbers 4, 5, 5, 7. Find sample mean \bar{X} for each sample and make sampling distribution of \bar{X} . Calculate the mean and standard deviation of this sampling distribution. Compare your calculations with population parameters.

Solution:

We have population values 4, 5, 5, 7, population size $N = 4$ and sample size $n = 3$. Thus, the number of possible samples which can be drawn without replacement is $\binom{N}{n} = \binom{4}{3} = 4$.

Sample No.	Sample Values	Sample Mean (\bar{X})
1	4, 5, 5	14/3
2	4, 5, 7	16/3
3	4, 5, 7	16/3
4	5, 5, 7	17/3

The sampling distribution of the sample mean \bar{X} and its mean and standard deviation are:

\bar{X}	f	$f(\bar{X})$	$\bar{X} f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
14/3	1	1/4	14/12	196/36
16/3	2	2/4	32/12	512/36
17/3	1	1/4	17/12	289/36
Total	4	1	63/12	997/36

$$\mu_{\bar{X}} = \Sigma \bar{X} f(\bar{X}) = \frac{63}{12} = 5.25$$

$$\sigma_{\bar{X}} = \sqrt{\Sigma \bar{X}^2 f(\bar{X}) - [\Sigma \bar{X} f(\bar{X})]^2} = \sqrt{\frac{997}{36} - \left(\frac{63}{12}\right)^2} = 0.3632$$

The mean and standard deviation of the population are:

X	4	5	5	7	$\Sigma X = 21$
X^2	16	25	25	49	$\Sigma X^2 = 115$

$$\mu = \frac{\Sigma X}{N} = \frac{21}{4} = 5.25 \text{ and } \sigma = \sqrt{\frac{\Sigma X^2}{N}} = \left(\frac{\Sigma X}{N}\right)^2 = \sqrt{\frac{115}{4}} = \left(\frac{21}{4}\right)^2 = 1.0897$$

$$\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{1.0897}{\sqrt{3}} \sqrt{\frac{4-3}{4-1}} = 0.3682$$

$$\text{Hence } \mu_{\bar{X}} = \mu \text{ and } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Example 11.3

Take all possible samples of size two with replacement from the population 2, 2, 8. Show that the population mean is equal to the mean of means of all samples and population variance is twice the variance of sample means.

Solution:

We have population values 2, 2, 8, population size N = 3 and sample size n = 2. Thus, the number of possible samples which can be drawn with replacement is $N^n = 3^2 = 9$.

Sample No.	Sample Values	Sample Mean (\bar{X})	Sample No.	Sample Values	Sample Mean (\bar{X})
1	2, 2	2	6	2, 8	5
2	2, 2	2	7	8, 2	5
3	2, 8	5	8	8, 2	5
4	2, 2	2	9	8, 8	8
5	2, 2	2			

The sampling distribution of the sample mean \bar{X} and its mean and variance are:

\bar{X}	Tally	f	$f(\bar{X})$	$\bar{X} f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
2		4	4/9	8/9	16/9
5		4	4/9	20/9	100/9
8		1	1/9	8/9	64/9
Total		9	1	36/9	180/9

$$E(\bar{X}) = \Sigma \bar{X} f(\bar{X}) = \frac{36}{9} = 4$$

$$\text{Var}(\bar{X}) = \Sigma \bar{X}^2 f(\bar{X}) - [\Sigma \bar{X} f(\bar{X})]^2 = \frac{180}{9} - \left(\frac{36}{9} \right)^2 = 4$$

$$2\text{Var}(\bar{X}) = 2(4) = 8$$

The mean and variance of the population are:

X	2	2	8	$\Sigma X = 12$
X^2	4	4	64	$\Sigma X^2 = 72$

$$\mu = \frac{\Sigma X}{N} = \frac{12}{3} = 4 \text{ and } \sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{72}{3} - \left(\frac{12}{3}\right)^2 = 8$$

Hence $E(\bar{X}) = \mu = 4$ and $\sigma^2 = 2$ $\text{Var}(\bar{X}) = 8$.

Example 11.4.

A population has the values 10, 12, 14, 16, 18 and 20. Draw all possible samples of size 2 without replacement and calculate the sample mean \bar{X} for each sample. Write the sampling distribution of \bar{X} . Find the following probabilities:

- (i) \bar{X} will be greater than 16.
- (ii) \bar{X} will differ from μ by less than 3 units.
- (iii) Sampling error will be less than 2.
- (iv) \bar{X} will be equal to μ .

Solution:

All possible samples of size 2 will be equal to ${}^6C_2 = \frac{6!}{2!4!} = 15$

The samples, their means and necessary calculations are as under:

Sample No.	Sample Values	Sample Mean (\bar{X})	Sample No.	Sample Values	Sample Mean (\bar{X})
1	10, 12	11	9	12, 20	16
2	10, 14	12	10	14, 16	15
3	10, 16	13	11	14, 18	16
4	10, 18	14	12	14, 20	17
5	10, 20	15	13	16, 18	17
6	12, 14	13	14	16, 20	18
7	12, 16	14	15	18, 20	19
8	12, 18	15			

Sampling Distribution of \bar{X}		
\bar{X}	f	$f(\bar{X})$
11	1	1/15
12	1	1/15
13	2	2/15
14	2	2/15
15	3	3/15
16	2	2/15
17	2	2/15
18	1	1/15
19	1	1/15
Total	15	1

$$\text{Population mean } \mu = \frac{10 + 12 + 14 + 16 + 18 + 20}{6} = \frac{90}{6} = 15$$

$$(i) P(\bar{X} > 16) = \frac{2}{15} + \frac{1}{15} + \frac{1}{15} = \frac{4}{15}$$

(ii) \bar{X} will differ from μ by less than 3 units if \bar{X} is greater than 12 and is less than 18.

$$\text{Thus } P[|\bar{X} - \mu| < 3] = P(12 < \bar{X} < 18) = \frac{2}{15} + \frac{2}{15} + \frac{3}{15} + \frac{2}{15} + \frac{2}{15} = \frac{11}{15}$$

(iii) The sampling error will be less than 2 if the random variable \bar{X} is greater than 13 and less than 17. Thus $P(13 < \bar{X} < 17) = P(14 \leq \bar{X} \leq 16) = P[|S.E.| < 2]$
 $= \frac{2}{15} + \frac{3}{15} + \frac{2}{15} = \frac{7}{15}$

$$(iv) P(\bar{X} = \mu) = P(\bar{X} = 15) = \frac{3}{15}$$

Example 11.5

Certain tubes produced by a company have a mean lifetime of 900 hours and a standard deviation of 100 hours. The company sends out 2000 lots of 100 tubes each.

Compute the mean and standard deviation of the sampling distribution of the sample mean \bar{X} if sampling is done: (i) with replacement (ii) without replacement.

Solution:

Here $N = 2000$, $n = 100$, $\mu = 900$, $\sigma = 100$

(i) Sampling with replacement

$$\mu_{\bar{x}} = \mu = 900 \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{100}} = 10$$

(ii) Sampling without replacement

$$\mu_{\bar{x}} = \mu = 900 \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{100}{\sqrt{100}} \sqrt{\frac{2000-100}{2000-1}} = 9.75$$

11.8.3 SAMPLING DISTRIBUTION OF s^2 and S^2

Suppose we draw all possible samples of size n from a finite population and calculate the sample variance $s^2 = \frac{\sum(X - \bar{X})^2}{n-1}$ for each sample. The mean of the sampling distribution of s^2 is denoted by $E(s^2)$ or μ_{s^2} . It can be shown that if sampling is with replacement, then $E(s^2) = \mu_{s^2} = \sigma^2$. Thus s^2 is an unbiased estimator of σ^2 . The sample variance S^2 is defined as: $S^2 = \frac{\sum(X - \bar{X})^2}{n}$. If samples are drawn with replacement, it can be shown that: $E(S^2) = \frac{n}{n-1} \sigma^2$ [$E(S^2) \neq \sigma^2$].

Thus S^2 is a biased estimator of σ^2 . In case of sampling without replacement, we have the following relations:

$$E(s^2) \frac{N-1}{N} = \sigma^2 \quad \text{or} \quad E(s^2) = \left(\frac{N}{N-1}\right) \sigma^2$$

$$E(S^2) \frac{n}{n-1} \cdot \frac{N-1}{N} = \sigma^2 \quad \text{or} \quad E(S^2) = \left(\frac{N}{N-1}\right) \left(\frac{n-1}{n}\right) \sigma^2$$

Example 11.6

A population consists of three numbers 10, 12, 14. Take all possible samples of size two with replacement from this population. Find the mean and the unbiased variance for each sample. Show that $E(s^2) = \sigma^2$ where $s^2 = \sum(X - \bar{X})^2/(n-1)$

Solution:

We have population values 10, 12, 14, population size $N = 3$ and sample size $n = 2$. Thus, the number of possible samples which can be drawn with replacement is $N^n = 3^2 = 9$.

Sample No.	Sample Values	Sample Mean $\bar{X} = \Sigma X/n$	Sample Variance $s^2 = \Sigma(X - \bar{X})^2/(n - 1)$
1	10, 10	$\frac{10 + 10}{2} = 10$	$\frac{(10 - 10)^2 + (10 - 10)^2}{2 - 1} = 0$
2	10, 12	$\frac{10 + 12}{2} = 11$	$\frac{(10 - 11)^2 + (12 - 11)^2}{2 - 1} = 2$
3	10, 14	$\frac{10 + 14}{2} = 12$	$\frac{(10 - 12)^2 + (14 - 12)^2}{2 - 1} = 8$
4	12, 10	$\frac{12 + 10}{2} = 11$	$\frac{(12 - 11)^2 + (10 - 11)^2}{2 - 1} = 2$
5	12, 12	$\frac{12 + 12}{2} = 12$	$\frac{(12 - 12)^2 + (12 - 12)^2}{2 - 1} = 0$
6	12, 14	$\frac{12 + 14}{2} = 13$	$\frac{(12 - 13)^2 + (14 - 13)^2}{2 - 1} = 2$
7	14, 10	$\frac{14 + 10}{2} = 12$	$\frac{(14 - 12)^2 + (10 - 12)^2}{2 - 1} = 8$
8	14, 12	$\frac{14 + 12}{2} = 13$	$\frac{(14 - 13)^2 + (12 - 13)^2}{2 - 1} = 2$
9	14, 14	$\frac{14 + 14}{2} = 14$	$\frac{(14 - 14)^2 + (14 - 14)^2}{2 - 1} = 0$

The sampling distribution of the sample variance s^2 and its mean is:

s^2	Tally	f	$f(s^2)$	$s^2 f(s^2)$
0		3	3/0	0
2		4	4/0	8/0
8		2	2/0	16/0
Total		9	1	24/0

$$\begin{aligned} E(s^2) &= \sum s^2 f(s^2) \\ &= \frac{24}{9} = 2.67 \end{aligned}$$

The variance of the population is:

X	10	12	14	$\Sigma X = 36$
X^2	100	144	196	$\Sigma X^2 = 440$

$$\sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N} \right)^2 = \frac{440}{3} - \left(\frac{36}{3} \right)^2 = 2.67$$

Hence $E(s^2) = \sigma^2 = 2.67$

Example 11.7.

A population consists of five values 4, 6, 8, 10, 12. Take all possible samples of size two without replacement from this population and verify that

$$E(S^2) = \left(\frac{N}{N-1}\right) \left(\frac{n-1}{n}\right) \sigma^2$$

Solution:

We have population values 4, 6, 8, 10, 12, population size $N = 5$ and sample size $n = 2$. Thus, the number of possible samples which can be drawn without replacement is $\binom{N}{n} = \binom{5}{2} = 10$.

Sample No.	Sample Values	Sample Mean $\bar{X} = \frac{\sum X}{n}$	Sample Variance $S^2 = \frac{\sum (X - \bar{X})^2}{n}$
1	4, 6	$\frac{4+6}{2} = 5$	$\frac{(4-5)^2 + (6-5)^2}{2} = 1$
2	4, 8	$\frac{4+8}{2} = 6$	$\frac{(4-6)^2 + (8-6)^2}{2} = 4$
3	4, 10	$\frac{4+10}{2} = 7$	$\frac{(4-7)^2 + (10-7)^2}{2} = 9$
4	4, 12	$\frac{4+12}{2} = 8$	$\frac{(4-8)^2 + (12-8)^2}{2} = 16$
5	6, 8	$\frac{6+8}{2} = 7$	$\frac{(6-7)^2 + (8-7)^2}{2} = 1$
6	6, 10	$\frac{6+10}{2} = 8$	$\frac{(6-8)^2 + (10-8)^2}{2} = 4$
7	6, 12	$\frac{6+12}{2} = 9$	$\frac{(6-9)^2 + (12-9)^2}{2} = 9$
8	8, 10	$\frac{8+10}{2} = 9$	$\frac{(8-9)^2 + (10-9)^2}{2} = 1$
9	8, 12	$\frac{8+12}{2} = 10$	$\frac{(8-10)^2 + (12-10)^2}{2} = 4$
10	10, 12	$\frac{10+12}{2} = 11$	$\frac{(10-11)^2 + (12-11)^2}{2} = 1$

The sampling distribution of the sample variance S^2 and its mean is:

S^2	f	$f(S^2)$	$S^2 f(S^2)$
1	4	4/10	4/10
4	3	3/10	12/10
9	2	2/10	18/10
16	1	1/10	16/10
Total	10	1	50/10

$$E(S^2) = \mu_{S^2} = \sum S^2 f(S^2) = \frac{50}{10} = 5$$

The variance of the population is:

X	4	6	8	10	12	$\Sigma X = 40$
X^2	16	36	64	100	144	$\Sigma X^2 = 360$

$$\sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{360}{5} - \left(\frac{40}{5}\right)^2 = 72 - 64 = 8$$

$$\left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2 = \left(\frac{5}{5-1}\right)\left(\frac{2-1}{2}\right)8 = \left(\frac{5}{4}\right)\left(\frac{8}{2}\right) = \frac{40}{8} = 5$$

$$\text{Hence } E(S^2) = \left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2 = 5$$

Example 11.8

A population of 10 numbers has a mean of 100 and a standard deviation of 10. If samples of size 5 are drawn from this population, find the mean of the sampling distribution of variances when sampling is done

- (i) with replacement (ii) without replacement.

Solution:

Here $N = 10$, $\mu = 100$, $\sigma = 10$, $\sigma^2 = 100$, $n = 5$

- (i) Sampling with replacement

$$E(S^2) = \mu_{S^2} = \left(\frac{n-1}{n}\right)\sigma^2 = \left(\frac{5-1}{5}\right)100 = 80$$

- (ii) Sampling without replacement

$$E(S^2) = \mu_{S^2} = \left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2 = \left(\frac{10}{10-1}\right)\left(\frac{5-1}{5}\right)100 = 88.89$$

11.8.4 SAMPLING DISTRIBUTION OF DIFFERENCE BETWEEN TWO MEANS

Suppose there is a population with mean μ_1 and variance σ_1^2 . Another population has the mean μ_2 and variance σ_2^2 . All possible simple random samples of size n_1 are selected from the first population and the sample means \bar{X}_1 for each sample are calculated. Similarly, all possible simple random samples of size n_2 are selected from the second population and the sample means \bar{X}_2 are calculated. The

difference $(\bar{X}_1 - \bar{X}_2)$ is another random variable and its distribution is called sampling distribution of $\bar{X}_1 - \bar{X}_2$. Some properties of this distribution are:

- (i) The mean of the distribution of $\bar{X}_1 - \bar{X}_2$ is equal to the difference $\mu_1 - \mu_2$. Thus

$$E(\bar{X}_1 - \bar{X}_2) = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

Similarly the distribution of $\bar{X}_2 - \bar{X}_1$ has the mean $\mu_{\bar{X}_2 - \bar{X}_1} = \mu_2 - \mu_1$.

If $\mu_1 = \mu_2$, then $E(\bar{X}_1 - \bar{X}_2) = 0$

The above relations are true for any type of population with any sample size, small or large and the samples may be drawn by without replacement or with replacement.

- (ii) When samples are selected by without replacement from a finite population, the standard error of $\bar{X}_1 - \bar{X}_2$ has the following relation with σ_1^2 and σ_2^2 .

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} \left(\frac{N_1 - n_1}{N_1 - 1} \right) + \frac{\sigma_2^2}{n_2} \left(\frac{N_2 - n_2}{N_2 - 1} \right)}$$

When samples are drawn with replacement or they are drawn from infinite populations (N_1 and N_2 are very large), the relation becomes:

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

It may be noted that in practical life, N_1 and N_2 are usually very large and the fractions $\frac{N_1 - n_1}{N_1 - 1}$ and $\frac{N_2 - n_2}{N_2 - 1}$ are almost equal to unity. Thus in the subsequent

chapter, we shall frequently use the relation $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

- (iii) The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is a normal distribution when $n_1 > 30$ and $n_2 > 30$. The sample sizes n_1 and n_2 may be equal or unequal but both should be large in size. The difference $(\bar{X}_1 - \bar{X}_2)$ is a random variable with normal distribution and the standard normal variable Z can be written as

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The distribution of $\bar{X}_1 - \bar{X}_2$ has the t-distribution when both n_1 and n_2 are small in size.

Example 11.9.

Draw all possible random samples of size $n_1 = 2$ without replacement from the finite population 2, 2, 6. Similarly, draw all possible random samples of size $n_2 = 2$ without replacement from the population 1, 1, 2, 4.

- Find the possible differences between the sample means of the two populations.
- Construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$ and compute its mean and variance.
- Verify that: $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$ and $\text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} \left(\frac{N_1 - n_1}{N_1 - 1} \right) + \frac{\sigma_2^2}{n_2} \left(\frac{N_2 - n_2}{N_2 - 1} \right)$

Solution:

Population I: 2, 2, 6

Population size $N_1 = 3$

Sample size $n_1 = 2$

The number of possible samples which can be drawn without replacement

$$= \binom{N_1}{n_1} = \binom{3}{2} = 3$$

Population II: 1, 1, 2, 4

Population size $N_2 = 4$

Sample size $n_2 = 2$

The number of possible samples which can be drawn without replacement

$$= \binom{N_2}{n_2} = \binom{4}{2} = 6$$

From Population I			From Population II		
Sample No.	Sample Values	Sample Mean (\bar{X}_1)	Sample No.	Sample Values	Sample Mean (\bar{X}_2)
1	2, 2	2	1	1, 1	1.0
2	2, 6	4	2	1, 2	1.5
3	2, 6	4	3	1, 4	2.5
			4	1, 2	1.5
			5	1, 4	2.5
			6	2, 4	3.0

- (i) The 18 possible differences $\bar{X}_1 - \bar{X}_2$ are shown in the following table.

\bar{X}_2	\bar{X}_1		
	2	4	4
1.0	1.0	3.0	3.0
1.5	0.5	2.5	2.5
2.5	-0.5	1.5	1.5
1.5	0.5	2.5	2.5
2.5	-0.5	1.5	1.5
3.0	-1.0	1.0	1.0

- (ii) The sampling distribution of differences between sample means $\bar{X}_1 - \bar{X}_2$ and its mean and variance are computed below.

$\bar{X}_1 - \bar{X}_2 = d$	f	f(d)	$d f(d)$	$d^2 f(d)$
-1.0	1	1/18	-1/18	1.0/18
-0.5	2	2/18	-1/18	0.5/18
0.5	2	2/18	1/18	0.5/18
1.0	3	3/18	3/18	3.0/18
1.5	4	4/18	6/18	9.0/18
2.5	4	4/18	10/18	25.0/18
3.0	2	2/18	6/18	18.0/18
Total	18	1	24/18	57/18

$$E(\bar{X}_1 - \bar{X}_2) = E(d) = \sum d f(d) = \frac{24}{18} = \frac{4}{3}$$

$$\text{Var}(\bar{X}_1 - \bar{X}_2) = \text{Var}(d) = \sum d^2 f(d) = \left[\sum d f(d) \right]^2 = \frac{57}{18} = \left(\frac{4}{3} \right)^2 = \frac{57 - 32}{18} = \frac{25}{18}$$

- (iii) The mean and variance of the first population are:

X_1	2	2	6	$\Sigma X_1 = 10$
X_1^2	4	4	36	$\Sigma X_1^2 = 44$

$$\mu_1 = \frac{\Sigma X_1}{N_1} = \frac{10}{3} \text{ and } \sigma_1^2 = \frac{\Sigma X_1^2}{N_1} - \left(\frac{\Sigma X_1}{N_1} \right)^2 = \frac{44}{9} - \left(\frac{10}{3} \right)^2 = \frac{44}{9} - \frac{100}{9} = \frac{182 - 100}{9} = \frac{32}{9}$$

The mean and variance of the second population are;

X_2	1	1	2	4	$\Sigma X_2 = 8$
X_2^2	1	1	4	16	$\Sigma X_2^2 = 22$

$$\mu_2 = \frac{\Sigma X_2}{N_2} = \frac{8}{4} = 2 \text{ and } \sigma_2^2 = \frac{\Sigma X_2^2}{N_2} - \left(\frac{\Sigma X_2}{N_2}\right)^2 = \frac{22}{4} - \left(\frac{8}{4}\right)^2 = \frac{22}{4} - 4 = \frac{22 - 16}{4} = \frac{3}{2}$$

$$\mu_1 - \mu_2 = \frac{10}{3} - 2 = \frac{10 - 6}{3} = \frac{4}{3}$$

$$\frac{\sigma_1^2}{n_1} \left(\frac{N_1 - n_1}{N_1 - 1} \right) + \frac{\sigma_2^2}{n_2} \left(\frac{N_2 - n_2}{N_2 - 1} \right) = \frac{32}{18} \left(\frac{3 - 2}{3 - 1} \right) + \frac{3}{4} \left(\frac{4 - 2}{4 - 1} \right) = \frac{16}{18} + \frac{1}{2} = \frac{16 + 9}{18} = \frac{25}{18}$$

$$\text{Hence } E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 = \frac{4}{3} \text{ and } \text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} \left(\frac{N_1 - n_1}{N_1 - 1} \right) + \frac{\sigma_2^2}{n_2} \left(\frac{N_2 - n_2}{N_2 - 1} \right) = \frac{25}{18}$$

Example 11.10

Given $N_1 = 800$, $N_2 = 600$, $n_1 = 200$, $n_2 = 124$, $\mu_1 = 1800$, $\mu_2 = 1600$, $\sigma_1 = 200$ and $\sigma_2 = 124$. Compute the mean and standard error of the sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$ if sampling is done (i) with replacement (ii) without replacement.

Solution:

(i) Sampling with replacement

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 = 1800 - 1600 = 200$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(200)^2}{200} + \frac{(124)^2}{124}} = 18$$

(ii) Sampling without replacement

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 = 1800 - 1600 = 200$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} \left(\frac{N_1 - n_1}{N_1 - 1} \right) + \frac{\sigma_2^2}{n_2} \left(\frac{N_2 - n_2}{N_2 - 1} \right)} = \sqrt{\frac{(200)^2}{200} \left(\frac{800 - 200}{800 - 1} \right) + \frac{(124)^2}{124} \left(\frac{600 - 124}{600 - 1} \right)} \\ = 15.77$$

11.8.5 PROPORTION

What is a proportion? Suppose there are 1000 students in a school out of which 600 are male and 400 are female. The ratio of 600 to the total is called the proportion of males and is denoted by p . Thus proportion of males $= p = \frac{600}{1000} = 0.6$

and proportion of females $= q = \frac{400}{1000} = 0.4$

Let us denote male by success and female by a failure. If the male students are assigned the number 1 and females are assigned the number 0, then the population contains 600 ones and 400 zeros. This can be written as below in the form of a distribution called the Bernoulli distribution. Let us calculate the mean of this distribution.

Random Variable (X)	f	f(X)	$\Sigma X f(X)$
0	400	$400/1000 = 0.4$	0
1	600	$600/1000 = 0.6$	0.6
Total	1000	1	0.6

$$E(X) = \text{Mean} = \Sigma X f(X) = 0.6$$

Thus the proportion p of the population called the binomial population is equal to the mean of the population containing 0's and 1's.

11.8.6 SAMPLING DISTRIBUTION OF PROPORTION

Suppose there is a finite population in which the proportion of successes is p and the proportion of failures is q . Suppose we draw all possible samples of size n from the population and calculate the sample proportion \hat{p} for each sample. The sampling distribution of \hat{p} has the following properties.

- The mean of the sampling distribution of \hat{p} is equal to the population proportion p . Thus $E(\hat{p}) = \mu_{\hat{p}} = p$
This relation is true in sampling with replacement and without replacement for any sample size.
- The standard error of \hat{p} is related to the population parameters p and q through the equations:

$$\text{S.E.}(\hat{p}) = \sigma_{\hat{p}} = \sqrt{\frac{pq}{n(N-n)}} \quad (\text{True for sampling without replacement})$$

$$\text{and } \text{S.E.}(\hat{p}) = \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \quad (\text{True for sampling with replacement}) \\ \text{or when } N \text{ is very large}$$

- The shape of the distribution of \hat{p} is normal when $n > 30$. The value of Z can be calculated from \hat{p} , where $Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$.

Warning: It is important to note that when n is small, the distribution of \hat{p} is not the t-distribution.

Example 11.VI

A population consists of five numbers 2, 5, 6, 7, 9. Take all possible samples of size 3 from this population without replacement and compute the proportion of odd numbers for each sample. Verify that: (i) $\mu_{\hat{p}} = p$ (ii) $\sigma^2_{\hat{p}} = \frac{pq}{n(N-n)}$

Solution:

We have population values 2, 5, 6, 7, 9, population size $N = 5$ and sample size $n = 3$. Thus, the number of possible samples which can be drawn without replacement is $\binom{N}{n} = \binom{5}{3} = 10$. Let \hat{p} represent the proportion of odd numbers in the sample.

Sample No.	Sample Values	Sample Proportion (\hat{p})	Sample No.	Sample Values	Sample Proportion (\hat{p})
1	2, 5, 6	1/3	6	2, 7, 9	2/3
2	2, 5, 7	2/3	7	5, 6, 7	2/3
3	2, 5, 9	2/3	8	5, 6, 9	2/3
4	2, 6, 7	1/3	9	5, 7, 9	3/3
5	2, 6, 9	1/3	10	6, 7, 9	2/3

The sampling distribution of the sample proportion \hat{p} and its mean and variance are:

\hat{p}	Tally	f	$f(\hat{p})$	$\hat{p} f(\hat{p})$	$\hat{p}^2 f(\hat{p})$
1/3		3	3/10	3/30	3/90
2/3	+++	6	6/10	12/30	24/90
3/3		1	1/10	3/30	9/90
Total		10	1	18/30	36/90

$$\mu_{\hat{p}} = \sum \hat{p} f(\hat{p}) = \frac{18}{30} = 0.6$$

$$\sigma_{\hat{p}}^2 = \sum \hat{p}^2 f(\hat{p}) = [\sum \hat{p} f(\hat{p})]^2 = \frac{36}{90} = \left(\frac{18}{30}\right)^2 = 0.40 = 0.36 = 0.04$$

$$\text{Population proportion } p = \frac{X}{N} = \frac{3}{5} = 0.6, q = 1 - p = 0.4$$

where X represents the number of odd digits in the population.

$$\frac{pq}{n} \left(\frac{N-n}{N-1} \right) = \frac{(0.6)(0.4)}{3} \left(\frac{5-3}{5-1} \right) = 0.04$$

$$\text{Hence (i) } \mu_{\hat{p}} = p = 0.6 \quad \text{(ii) } \sigma_{\hat{p}}^2 = \frac{pq}{n} \left(\frac{N-n}{N-1} \right) = 0.04$$

Example 11.12.

A finite population contains 4 smokers denoted by S_1, S_2, S_3 and S_4 and 2 non-smokers denoted by N_1 and N_2 . Draw all possible random samples of size 2 without replacement from the population and calculate the proportion of smokers \hat{p} in each.

Ex-I Example. Write the probability distribution (sampling distribution) of \hat{p} and find the following probabilities:

- (i) \hat{p} is more than p (ii) \hat{p} is equal to p (iii) $\hat{p} = \frac{1}{2}$ (iv) that both are smokers.

Solution:

We have population values $S_1, S_2, S_3, S_4, N_1, N_2$, population size $N = 6$ and sample size $n = 2$. Thus, the number of possible samples which can be drawn without replacement is $\binom{N}{n} = \binom{6}{2} = 15$.

Sample No.	Sample Values	Sample proportion (\hat{p})	Sample No.	Sample Values	Sample proportion (\hat{p})
1	S_1, S_2	2/2	9	S_2, N_2	1/2
2	S_1, S_3	2/2	10	S_3, S_4	2/2
3	S_1, S_4	2/2	11	S_3, N_1	1/2
4	S_1, N_1	1/2	12	S_4, N_2	1/2
5	S_1, N_2	1/2	13	S_4, N_1	1/2
6	S_2, S_3	2/2	14	S_4, N_2	1/2
7	S_2, S_4	2/2	15	N_1, N_2	0
8	S_2, N_1	1/2			

The sampling distribution of the sample proportion \hat{p} is:

\hat{p}	f	$f(\hat{p})$
0	1	1/15
1/2	8	8/15
2/2	6	6/15
Total	15	1

$$\text{Population proportion } p = \frac{4}{6} = \frac{2}{3}$$

$$(i) P(\hat{p} > p) = \frac{6}{15} \quad (ii) P(\hat{p} = p) = 0$$

$$(iii) P\left(\hat{p} = \frac{1}{2}\right) = \frac{8}{15} \quad (iv) P(\text{both are smokers}) = \frac{6}{15}$$

Example 11.13

If samples of $n = 200$ observations are to be drawn from a large population $N = 2500$ in which the population proportion is 20 %. Determine the expected mean and standard deviation of the sampling distribution of proportions when sampling is done (i) with replacement (ii) without replacement.

Solution:

Here $N = 2500$, $n = 200$, $p = 0.20$, $q = 1 - p = 1 - 0.20 = 0.80$

(i) Sampling with replacement

$$E(\hat{p}) = p = 0.20 \text{ and } S.E.(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.20)(0.80)}{200}} = 0.0283$$

(ii) Sampling without replacement

$$E(\hat{p}) = p = 0.20 \text{ and } S.E.(\hat{p}) = \sqrt{\frac{pq(N-n)}{n(N-1)}} = \sqrt{\frac{(0.20)(0.80)(2500-200)}{200(2500-1)}} = 0.0271$$

11.8.7 SAMPLING DISTRIBUTION OF DIFFERENCE BETWEEN \hat{p}_1 and \hat{p}_2

Suppose there are two populations with proportions p_1 and p_2 and all possible simple random samples of size n_1 and n_2 are selected from the populations respectively. The sample proportions calculated from the samples are \hat{p}_1 and \hat{p}_2 . The difference $\hat{p}_1 - \hat{p}_2$ is a random variable and its distribution is called the sampling distribution of $\hat{p}_1 - \hat{p}_2$. The properties of this distribution are:

(i) The mean of the distribution of $\hat{p}_1 - \hat{p}_2$ is equal to the difference between p_1 and p_2 . Thus $\mu_{\hat{p}_1 - \hat{p}_2} = E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

This relation is true for any sample size and for sampling with and without replacement.

(ii) The standard error of the distribution of $(\hat{p}_1 - \hat{p}_2)$ has the following relation with population parameters

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1 (N_1 - n_1)}{n_1 (N_1 - 1)} + \frac{p_2 q_2 (N_2 - n_2)}{n_2 (N_2 - 1)}}$$

(True for sampling without replacement)

$$\text{and } S.E.(\hat{p}_1 - \hat{p}_2) = \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

(True for sampling with replacement or when N is very large)

(iii) The distribution of $\hat{p}_1 - \hat{p}_2$ has the normal distribution when both n_1 and n_2 are large in size, when n_1 and n_2 are small in size, the distribution of $\hat{p}_1 - \hat{p}_2$ does not form any standard distribution. The random difference $(\hat{p}_1 - \hat{p}_2)$ can be

transformed into standard normal variable Z where $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$

Example 11.14

Given the data: $N_1 = 6$, $n_1 = 3$, $X_1 = 3$, $N_2 = 5$, $n_2 = 2$, $X_2 = 2$.

Find $E(\hat{p}_1 - \hat{p}_2)$ and $\text{Var}(\hat{p}_1 - \hat{p}_2)$ if sampling is done

- (i) with replacement
- (ii) without replacement

Solution:

$$\text{Here } N_1 = 6, n_1 = 3, X_1 = 3, p_1 = \frac{X_1}{N_1} = \frac{3}{6} = 0.5, q_1 = 1 - p_1 = 0.5$$

$$N_2 = 5, n_2 = 2, X_2 = 2, p_2 = \frac{X_2}{N_2} = \frac{2}{5} = 0.4, q_2 = 1 - p_2 = 0.6$$

- (i) Sampling with replacement

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 = 0.5 - 0.4 = 0.1$$

$$\begin{aligned} \text{Var}(\hat{p}_1 - \hat{p}_2) &= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} = \frac{(0.5)(0.5)}{3} + \frac{(0.4)(0.6)}{2} \\ &= 0.0833 + 0.12 = 0.2083 \end{aligned}$$

- (ii) Sampling without replacement

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 = 0.5 - 0.4 = 0.1$$

$$\begin{aligned} \text{Var}(\hat{p}_1 - \hat{p}_2) &= \frac{p_1 q_1}{n_1} \left(\frac{N_1 - n_1}{N_1 - 1} \right) + \frac{p_2 q_2}{n_2} \left(\frac{N_2 - n_2}{N_2 - 1} \right) \\ &= \frac{(0.5)(0.5)}{3} \left(\frac{6 - 3}{6 - 1} \right) + \frac{(0.4)(0.6)}{2} \left(\frac{5 - 2}{5 - 1} \right) \\ &= 0.05 + 0.09 = 0.14 \end{aligned}$$

SHORT DEFINITIONS

Population whole of aggregate of items is called -

A population is the total set of measurements of interest in a particular problem.
or

The population is a set of data that characterizes some phenomenon.

Finite Population

If a population has finite number of elements, it is called as finite population. For example human population, number of chairs in a college.

Infinite Population

If a population has infinite number of elements, it is called as infinite population. For example number of points on line, number of stars in the sky.

Target Population

A population about which we want to get some information is called target population.

Sampled Population

A population from which a sample is drawn is called sampled population.

Sample

A sample is a subset of data selected from a population.

or

A sample is a subset of the population that contains measurements obtained by an experiment.

Random Sample

A sample obtained by random sampling is called a random sample.

or

If a sample is selected from such a population whose sampling units have known probability that may be equal or unequal, the sample is said to be a random sample.

Sampling

Sampling is the process of drawing sample from the population.

Random Sampling

Any procedure for selecting members from a group on the basis of chance or luck is called a random sampling.

or

A method of selecting samples so that each sample of a given size in a population has an equal or unequal chance of being selected.

Sampling Units

Sampling units are nonoverlapping collections of elements from the population.

or

The basic elements that constitutes a population are known as sampling units.

Simple Random Sample

A simple random sample is one in which every item from a population has the same chance of selection as any other item.

or

A sample selected in such a manner that each possible sample of a specified size has an equal chance of being selected.

119	Short Definitions	15
119	Multiple Choice Questions	15
120	Short Questions	15
120	Exercises.....	160-16
121	Chapter 14	
122	• Regression and Correlation	165-216
123	14.1. Introduction	165
124	14.2. Mathematical Model or Equation.....	165
125	14.3. Non-Linear Model.....	168
126	14.4. Statistical Model	169
127	14.4.1. Independent and Dependent Variables.....	171
128	14.4.2. Cause and Effect Relation.....	172
129	14.5. Regression.....	172
130	14.5.1. Simple Linear Regression	173
131	14.5.2. Purpose of Regression Analysis	173
132	14.5.3. Scatter Diagram.....	173
133	14.6. Fitting a Linear Regression Line—the Method of Least Squares.....	175
134	14.6.1. Properties of the Regression Line	179
135	14.6.2. Regression Equation of X on Y	179
136	14.7. Introduction	184
137	14.8. Correlation.....	184
138	14.8.1. Measurement of Correlation.....	184
139	14.8.2. Perfect Positive Correlation.....	185
140	14.8.3. Perfect Negative Correlation	185
141	14.8.4. No Correlation	185
142	14.8.5. Scatter Diagrams	186
143	14.9. Correlation Coefficient for Sample Data	187
144	14.9.1. Causation in Correlation.....	191
145	14.9.2. Spurious Correlation	191
146	14.9.3. Change of Origin	191
147	14.9.4. Change of Scale	192
148	14.9.5. Change of Origin and Scale	193
149	14.9.6. 'r' in a Linear Regression Relation	193
150	14.9.7. 'r' for Random Variables.....	193

Simple Random Sampling

A procedure for selecting members from a population in such a manner that each drawing gives every available member an equal chance of selection.

or

A method of selecting items from a population so that every possible sample of a specified size has an equal chance of being selected.

Stratified Random Sampling

In a stratified random sampling the population is first divided into subgroups, called strata and a random sample is then taken from each stratum.

or

A stratified random sampling is obtained by partitioning the sampling units in the population into nonoverlapping subpopulations called strata. Random samples are then selected from each stratum.

Parameter

A parameter is a numerical descriptive measure of a population.

or

A parameter is any measure which describes a population.

Statistic

A statistic is quantity calculated from the observations in a sample.

or

A measure computed on the basis of sample data is termed as statistic.

Census

The study of all the data points in a population is called a census.

or

To study all the individual observations of the entire population is called census.

Basic Aims of Sampling

- (i) To get maximum information about a population without examining each and every unit of the population.
- (ii) To find reliability of estimates derived from the sample.

Advantages of Sampling

- (i) Sampling is cheaper than complete count.
- (ii) The data are collected and analyzed more quickly.
- (iii) Sampling saves time.
- (iv) A higher quality of labour with better supervision can be employed due to reduced volume of material.
- (v) A small fraction of population gives sometimes comprehensive and detailed results.

Sampling Design

A sampling design is a definite statistical plan which has all steps taken in the selection of the sample and method of estimation.

or

The sampling design specifies the method of collecting the sample.

- Sampling Frame

A sampling frame is a list of all sampling units in the population.

or

A list of the sampling units for a study is called a sampling frame.

- Probability sampling

A probability sampling is one in which the sampling units are chosen on the basis of known probabilities.

or

When each and every element of the population has known probability of being selected in the sample, then sampling is said to be probability sampling.

- Non-Probability Sampling

The sampling is said to be non-probability sampling when the procedure of selecting the elements from the population is not based on probability but personal judgement is involved in selection.

- Sampling with Replacement

When an object is selected from the population and is replaced before the next object is selected, such a selection is known as sampling with replacement.

or

Sampling is said to be with replacement when we draw a sampling unit from a population and return it to the population before the next unit is drawn. In sampling with replacement, an element can be chosen more than once in a sample.

- Sampling without Replacement

Sampling without replacement is performed when an object is not replaced in the population after it has been selected.

or

Sampling is said to be without replacement when we draw a sampling unit from a population and do not return it to the population before the next unit is drawn. In sampling without replacement an element cannot be chosen more than once in a sample.

- Permutation

A permutation is an arrangement in which the order of the objects selected from a specific pool of objects is important.

or

A permutation is an ordered arrangement of objects.

- Combination

A combination is collection of a group of objects without regard to order.

or

A combination is an arrangement of objects without regard to order.

- Sampling Error

The sampling error is the difference between a population parameter and a sample statistic.

or

The difference between a sample statistic and its corresponding population parameter is called sampling error.

Non-Sampling Error

All types of error other than sampling error, such as measurement error, interviewer error and processing error is called non-sampling error.

or

Non-sampling error is introduced by bias, consciously or unconsciously, on the part of the researcher. This is due to improper sample selection, improper questionnaires, etc.

Bias Unbias: When expectation of any statistic is equal to its parameter. The difference between the mean or expected value of a statistic and the value of the parameter being estimated is called bias.

Bias is not equal to its parameters. This technique is called *Unbiased*.

Bias means a systematic component of error which deprives a statistical result of its representativeness.

Sampling Distribution

The distribution of all possible values that can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population, is called the sampling distribution of that statistic.

or

A probability distribution consisting of all possible values of a sample statistic is known as sampling distribution.

Standard Error

The standard deviation of the sampling distribution for a statistic is called the standard error.

or

The standard deviation of any estimator is called the standard error of the estimator.

Sampling Distribution of the Mean

If we take all possible samples of a given size from a population and determine the mean of each sample, the probability distribution of the sample means is called the sampling distribution of the mean.

or

A probability distribution of all possible sample means of a given sample size is known as sampling distribution of the mean.

Central Limit Theorem

If all samples of a specified size are selected from any population, the sampling distribution of the sample mean is approximately a normal distribution. This approximation improves with larger samples.

or

If the sample size is large, the theoretical sampling distribution of the mean can be approximated closely with a normal distribution.

Population Proportion

The fraction of values in a population which has a specific attribute is called population proportion.

Sample Proportion

A sample proportion is the fraction of items in a sample that has the attribute of interest.

MULTIPLE - CHOICE QUESTIONS

13. For making voters list in Pakistan we need:
- sampling error
 - standard error
 - census
 - simple random sampling
14. Sampling based upon equal probability is called:
- probability sampling
 - systematic sampling
 - simple random sampling
 - stratified random sampling
15. In sampling with replacement, an element can be chosen:
- less than once
 - more than once
 - only once
 - difficult to tell
16. In sampling without replacement, an element can be chosen:
- less than once
 - more than once
 - only once
 - difficult to tell
17. In sampling with replacement, the following is always true:
- $n = N$
 - $n < N$
 - $n > N$
 - all of the above
18. Suppose a finite population has 6 items and 2 items are selected at random without replacement, then all possible samples will be:
- 6
 - 12
 - 15
 - 36
19. Suppose a finite population contains 7 items and 3 items are selected at random without replacement, then all possible samples will be:
- 21
 - 35
 - 14
 - 7
20. A population contains N items and all possible samples of size n are selected without replacement. The possible number of samples will be:
- N
 - n^N
 - ${}^N C_n$
 - N^n
21. Suppose a finite population contains 4 items and 2 items are selected at random with replacement, then all possible samples will be:
- 0
 - 10
 - 8
 - 4
22. A population contains 2 items and 4 items are selected at random with replacement, then all possible samples will be:
- 16
 - 8
 - ${}^4 C_2$
 - 4
23. Suppose a population has N items and n items are selected with replacement. Number of all possible samples will be:
- N^n
 - ${}^N C_n$
 - N
 - n

24. In random sampling, the probability of selecting an item from the population is:
- unknown
 - known
 - un-decided
 - one
25. Random sampling is also called:
- probability sampling
 - non-probability sampling
 - sampling error
 - random error.
26. Non-random sampling is also called:
- biased sampling
 - non-probability sampling
 - random sampling
 - representative sample
27. Sampling error can be reduced by:
- non-random sampling
 - increasing the population
 - decreasing the sample size
 - increasing the sample size.
28. If N is the size of the population and n is the size of the sample, then sampling fraction is:
- n^N
 - N^n
 - $\frac{n}{N}$
 - ${}^N C_n$
29. The finite population correction factor is:
- $\sqrt{\frac{N-1}{N-n}}$
 - $\sqrt{\frac{N+n}{N+1}}$
 - $\sqrt{\frac{N-n}{N-1}}$
 - $\sqrt{\frac{n}{n-1}}$
30. In sampling with replacement, the standard error of \bar{X} is equal to:
- $\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
 - $\frac{\sigma^2}{n}$
 - $\frac{\sigma}{\sqrt{n}}$
 - $\frac{\sigma}{\sqrt{n}} \cdot \frac{n}{N}$
31. If $\sigma_1 = \sigma_2 = \sigma$ and $n_1 \neq n_2$, then S.E. ($\bar{X}_2 - \bar{X}_1$) is equal to:
- $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
 - $\sqrt{\frac{\sigma_1^2}{n_2} + \frac{\sigma_2^2}{n_1}}$
 - $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}$
 - $\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
32. In sampling with replacement, standard error of the sample proportion \hat{p} is equal to:
- $\frac{p(1-p)}{n}$
 - $\sqrt{\frac{p(1-p)}{n}}$
 - $\sqrt{\frac{p + q}{2}}$
 - $\sqrt{\frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}}$

33. If $p_1 = p_2 = p$ and $n_1 \neq n_2$, then S.E ($\hat{p}_1 - \hat{p}_2$) is equal to:

(a) $\frac{p_1 q_1 + p_2 q_2}{n_1 + n_2}$

(b) $\frac{p_1 q_1 - p_2 q_2}{n_1 + n_2}$

(c) $\sqrt{\frac{p_1 q_1 + p_2 q_2}{n_1 + n_2}}$

(d) $\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

34. The selection of cricket team for the world cup is called:

(a) random sampling

(b) systematic sampling

(c) purposive sampling

(d) cluster sampling

35. Random sampling is also called:

(a) probability sampling

(b) judgment sampling

(c) quota sampling

(d) sequential sampling

36. A complete list of all the sampling units is called:

(a) sampling design

(b) sampling frame

(c) population frame

(d) cluster

37. A plan for obtaining a sample from a population is called:

(a) population design

(b) sampling design

(c) sampling frame

(d) sampling distribution

38. If a survey is conducted by a sampling design is called:

(a) sample survey

(b) population survey

(c) systematic survey

(d) none of the above

39. The difference between the expected value of a statistic and the value of the parameter being estimated is called a:

(a) sampling error

(b) non-sampling error

(c) standard error

(d) bias

40. The standard deviation of any sampling distribution is called:

(a) standard error

(b) non sampling error

(c) type - I error

(d) type-II error

41. The standard error increases when sample size is:

(a) increased

(b) decreased

(c) fixed

(d) more than 30

42. The mean of sampling distribution of means is equal to:

(a) \bar{X}

(b) μ

(c) p

(d) none of the above

43. The mean of the sample means is exactly equal to the:

(a) sample mean

(b) population mean

(c) weighted mean

(d) combined mean

44. Sum of all sample means is equal to:
Total number of samples
- (a) $E(\bar{X})$
 - (b) μ
 - (c) both (a) and (b)
 - (d) none of the above
45. A sample which is free from bias is called:
- (a) biased
 - (b) unbiased
 - (c) positively biased
 - (d) negatively biased
46. If $E(\bar{X}) = \mu$ then bias is:
- (a) positive
 - (b) negative
 - (c) zero
 - (d) 100%
47. If $E(\bar{X}) = 10$ and $\mu = 10$ then bias is equal to:
- (a) 0
 - (b) 10
 - (c) 20
 - (d) difficult to tell
48. If $\bar{X} = 10$ and $\mu = 12$ then sampling error is equal to:
- (a) 22
 - (b) 10
 - (c) 12
 - (d) 2
49. The standard deviation of the distribution of sample means is equal to:
- (a) σ^2 / \sqrt{n}
 - (b) \sqrt{n} / σ
 - (c) σ / \sqrt{n}
 - (d) σ / n
50. If $n = 25$, $\sigma^2 = 25$ and $\bar{X} = 25$, then standard error of \bar{X} will be:
- (a) 25
 - (b) 5
 - (c) 1 (d) 0
1. $S^2 = \frac{\sum(X - \bar{X})^2}{n}$ is called:
- (a) unbiased sample variance
 - (b) population variance
 - (c) biased sample variance
 - (d) all of the above
2. $s^2 = \frac{\sum(X - \bar{X})^2}{n-1}$ is called:
- (a) unbiased sample variance
 - (b) true variance
 - (c) biased sample variance
 - (d) variance of means
3. If $E(s^2) = 9$ and $\sigma^2 = 2$ then bias will be:
- (a) 5
 - (b) 9
 - (c) 2
 - (d) 1

54. In sampling without replacement, the standard error of sampling distribution of sample proportion \hat{p} is equal to:

(a) $\hat{p}q \left(\frac{N-n}{N-1} \right)$

(b) $\frac{\hat{p}q}{n} \left(\frac{N-n}{N-1} \right)$

(c) $\sqrt{\frac{\hat{p}q}{n} \left(\frac{N-n}{N-1} \right)}$

(d) $\frac{\sqrt{\hat{p}q}}{n} \left(\frac{N-n}{N-1} \right)$

55. When sampling is done without replacement $\sigma_{\bar{x}}$ is equal to:

(a) $\frac{\sigma^2}{n}$

(b) $\frac{\sigma^2}{\sqrt{n}}$

(c) $\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

(d) $\frac{\sigma}{n} \sqrt{\frac{N-n}{N-1}}$

56. In case of sampling with replacement $\sigma_{\hat{p}_1 - \hat{p}_2}$ is equal to:

(a) $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

(b) $\sqrt{\frac{\hat{p}q}{n}}$

(c) $\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$

(d) $\sqrt{p_1q_1 + p_2q_2}$

57. The distribution of the means of samples of size 4, taken from a population with a standard deviation σ , has a standard deviation of:

(a) σ

(b) $\sigma/4$

(c) $\sigma/2$

(d) $\sigma^2/2$

58. In sampling with replacement, $\sigma_{\bar{x}_1 - \bar{x}_2}$ is equal to:

(a) $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

(b) $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

(c) $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

(d) $\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n_1 + n_2}}$

59. When sampling is done with or without replacement, $E(\hat{p}_1 - \hat{p}_2)$ is equal to:

(a) $\hat{p}_1 - \hat{p}_2$

(b) $p_1 - p_2$

(c) $p_1 + p_2$

(d) p_1p_2

60. In case of sampling with replacement, $E(S^2)$ is equal to:

(a) $\left(\frac{n-1}{n} \right) \sigma^2$

(b) $\left(\frac{n}{n-1} \right) \sigma^2$

(c) $\left(\frac{N}{N-1} \right) \sigma^2$

(d) $\frac{\sigma^2}{n}$

61. In sampling without replacement, the expected value of S^2 is equal to:

(a) $\left(\frac{N-1}{N}\right)\left(\frac{n-1}{n}\right)\sigma^2$

(b) $\left(\frac{N}{N-1}\right)\left(\frac{n}{n-1}\right)\sigma^2$

(c) $\left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2$

(d) $\sqrt{\left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2}$

62. When sampling is done with replacement, then μ_{S^2} is equal to:

(a) σ^2

(b) $\frac{\sigma^2}{n}$

(c) $\sqrt{\frac{\sigma^2}{n}}$

(d) $\left(\frac{n-1}{n}\right)\sigma^2$

63. In sampling without replacement, μ_{S^2} is equal to:

(a) $\left(\frac{N-1}{N}\right)\sigma^2$

(b) $\left(\frac{N}{N-1}\right)\sigma^2$

(c) $\left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2$

(d) $\left(\frac{n-1}{n}\right)\sigma^2$

64. When sampling is done with or without replacement, $\mu_{\bar{X}_1 - \bar{X}_2}$ is equal to:

(a) $\mu_1 - \mu_2$

(b) $\mu_1 + \mu_2$

(c) $\mu_2 - \mu_1$

(d) $\frac{\mu_1}{n_1} - \frac{\mu_2}{n_2}$

65. If X represent the number of units having the specified characteristic and n is the size of the sample, then sample proportion \hat{p} is equal to:

(a) $\frac{n}{X}$

(b) $\frac{X}{n}$

(c) $\frac{X + \sigma}{n}$

(d) $\frac{\sigma}{\sqrt{n}}$

66. If X represent the number of units having the specified characteristic and N is the size of the population, then population proportion p is equal to:

(a) $\frac{X}{n}$

(b) $\frac{N}{X}$

(c) $\frac{X}{N}$

(d) $\frac{\sigma^2}{N}$

Answers

1. (a)	2. (c)	3. (b)	4. (b)	5. (b)	6. (d)	7. (c)	8. (c)
9. (d)	10. (b)	11. (d)	12. (d)	13. (c)	14. (c)	15. (b)	16. (c)
17. (d)	18. (c)	19. (b)	20. (c)	21. (b)	22. (a)	23. (a)	24. (b)
25. (a)	26. (b)	27. (d)	28. (c)	29. (c)	30. (c)	31. (d)	32. (b)
33. (d)	34. (c)	35. (a)	36. (b)	37. (b)	38. (a)	39. (d)	40. (a)
41. (b)	42. (b)	43. (b)	44. (c)	45. (b)	46. (c)	47. (a)	48. (d)
49. (c)	50. (c)	51. (c)	52. (a)	53. (d)	54. (c)	55. (c)	56. (c)
57. (c)	58. (c)	59. (b)	60. (a)	61. (c)	62. (a)	63. (b)	64. (a)
65. (b)	66. (c)						

SHORT QUESTIONS

1. Given $\mu = 6$ and $n = 30$. Find $\mu_{\bar{x}}$.

Ans. 6

2. Given $n = 36$ and $\sigma = 6$. Find $\sigma_{\bar{x}}^2$.

Ans. 1

3. Given $n = 25$ and $\sigma_{\bar{x}} = 5$. Find the value of σ^2 .

Ans. 625

4. Given $\mu_1 = 10$ and $\mu_2 = 6$. Find $\mu_{\bar{x}_1 - \bar{x}_2}$.

Ans. 4

5. Given $n_1 = 30$, $n_2 = 25$, $\sigma_1^2 = 300$ and $\sigma_2^2 = 150$. Find $\sigma_{\bar{x}_1 - \bar{x}_2}^2$.

Ans. 16

6. Given $N = 300$, $n = 100$ and $\sigma^2 = 200$. If sampling is done without replacement, then find the value of $\sigma_{\bar{x}}$.

Ans. 1.16

7. Given $N = 310$, $n = 100$ and $\sigma^2 = 35$. If sampling is done without replacement, then find σ^2 .

Ans. 5150

8. Given $N_1 = 3$, $n_1 = 2$, $N_2 = 4$, $n_2 = 2$, $\sigma_1^2 = 8/3$ and $\sigma_2^2 = 5/4$. If sampling is done without replacement, then find the value of $\sigma_{\bar{x}_1 - \bar{x}_2}^2$.

Ans. 1.08

9. Given $N = 7$, $n = 2$ and $\sigma^2 = 16$. If sampling is done without replacement, then find $E(S^2)$.

Ans. 9.33

10. Given $N = 7$, $n = 2$ and $\sigma^2 = 16$. If sampling is done without replacement, then find μ_{s^2} .

Ans. 18.67

11. Given $\mu = 6$, $n = 2$ and $\sigma^2 = 10.8$. Find $E(S^2)$.

Ans. 5.4

12. Given $\mu = 6$, $n = 2$ and $\sigma^2 = 10.8$. Find $E(s^2)$.

Ans. 10.8

13. Given $N = 7$, $n = 3$, $\mu_p = 3 / 7$. Find the value of population proportion p .

Ans. 3/7

14. Given $N = 7$, $n = 3$ and $\mu_{\hat{p}} = 3/7$. If sampling is done without replacement, find $\sigma_{\hat{p}}^2$.

Ans. 0.0544

15. Given $n = 5$ and $p = 0.5$. Find $\sigma_{\hat{p}}^2$.

Ans. 0.05

16. Given $p_1 = 2/3$, $n_1 = 2$, $p_2 = 1/2$ and $n_2 = 2$. Find $\mu_{\hat{p}_1 - \hat{p}_2}$.

Ans. 1/6

17. Given $p_1 = 2/3$, $n_1 = 2$, $p_2 = 1/2$ and $n_2 = 2$. Find $\sigma_{\hat{p}_1 - \hat{p}_2}^2$.

Ans. 0.24

18. Given $N_1 = 4$, $n_1 = 2$, $N_2 = 4$, $n_2 = 2$, $p_1 = 1/2$ and $p_2 = 1/4$. If sampling is done without replacement, find S.E. ($\hat{p}_1 - \hat{p}_2$).

Ans. 0.3819

19. What is the value of the finite population correction factor when $n = 18$ and $N = 125$.

Ans. 0.93

20. Differentiate between sampling with and without replacement.
21. Distinguish between probability and non-probability sampling.
22. Differentiate between parameter and statistic.
23. Distinguish between population and sample.
24. Distinguish between sampling and non-sampling errors.
25. Differentiate between simple random sampling and stratified random sampling.
26. Explain the term sampling frame.
27. Define the standard error.
28. Distinguish between simple random sample and simple random sampling.
29. Explain the term sampling design.
30. Differentiate between finite and infinite populations.
31. Define the sampling distribution.
32. Differentiate between random sample and simple random sample.
33. Write down the advantages of sampling.
34. Write down the basic aims of sampling.
35. Define the terms sample and sampling.
36. Define the sampling distribution of means.
37. Describe the properties of the sampling distribution of sample means.
38. Define the sampling distribution of sample proportion and describe its properties.
39. What is meant by bias?

EXERCISES

1. A population consists of five numbers 3, 7, 11, 15 and 19. Take all possible samples of size two without replacement from this population. Find the mean and standard deviation of the sampling distribution of means.

Ans: $\mu_{\bar{X}} = 11$, $\sigma_{\bar{X}} = 3.46$

2. Take all possible samples of size 3 without replacement from the population 2, 6, 8, 12 and 14. Form sampling distribution of mean and find its mean and variance. Verify that: $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$

Ans: $\mu_{\bar{X}} = 8.4$, $\sigma_{\bar{X}}^2 = 3.04$, $\mu = 8.4$, $\sigma^2 = 18.24$

3. Draw all possible samples of size two without replacement from the population 16, 18, 20 and 22. Calculate their means and prepare the frequency distribution of sample mean. Compute mean and variance of frequency distribution of mean and compare them with population mean and variance.

Ans: $\mu_{\bar{X}} = 19$, $\sigma_{\bar{X}}^2 = 1.67$, $\mu = 19$, $\sigma^2 = 5$

4. A population consists of four numbers 5, 6, 7 and 8. Take all possible samples of size three without replacement from this population. Calculate mean and variance of sample means and compare them with population mean and variance.

Ans: $\mu_{\bar{X}} = 6.5$, $\sigma_{\bar{X}}^2 = 0.14$, $\mu = 6.5$, $\sigma^2 = 1.25$

5. A population consists of two elements 24 and 35. Take all possible samples of size two with replacement and find their means. Make a sampling distribution of sample means and find its mean and standard deviation. Verify that:

$$(i) \quad \mu_{\bar{X}} = \mu \quad (ii) \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Ans: $\mu_{\bar{X}} = 29.5$, $\sigma_{\bar{X}} = 3.89$, $\mu = 29.5$, $\sigma = 5.5$

6. A population consists of three values 20, 40 and 60.

- (i) Take all possible samples of size 2, which can be drawn with replacement from this population and find means of these samples.
- (ii) Make a frequency distribution of the sample mean and show that the variance of this distribution is equal to the population variance divided by the sample size.

Ans: $\text{Var}(\bar{X}) = 133.33$, $\sigma^2 = 266.67$

7. A population consists of values 6, 8 and 12. Take all possible samples of size 2 using simple random sampling with replacement. Prepare sampling distribution of mean and prove that standard error of mean is the square root of population variance divided by the sample size.

Ans: $S.E.(\bar{X}) = 1.76, \sigma^2 = 6.2222$

8. Draw all possible samples of size 3 with replacement from 2, 4. Then show that: (i) population mean = mean of sample means
(ii) standard error = population S.D. \sqrt{n} .

Ans. $\mu_{\bar{X}} = 3, S.E.(\bar{X}) = 0.5773, \mu = 3, \sigma = 1$

9. If mean and variance of a population are 5 and 2.15 respectively. What would be the standard error of mean if samples of size 4 are drawn with replacement.

Ans. $S.E.(\bar{X}) = 0.7331$

10. If mean and variance of population are 7 and 3.15 respectively. What would be standard error of mean if samples are drawn without replacement of size 6 from a population of size 10.

Ans. $S.E.(\bar{X}) = 0.4830$

11. What will be the mean and variance of sample means if (i) Sample of 36 is drawn with replacement from the population 1, 2, 3, 4, 4, 4, 5, 6, 6, 7. (ii) Sample of 4 is drawn without replacement from the population given in (i).

Ans. (i) $\mu_{\bar{X}} = 4.2, \text{Var}(\bar{X}) = 0.088$ (ii) $\mu_{\bar{X}} = 4.2, \text{Var}(\bar{X}) = 0.527$

12. Assume that the masses of 2000 male students at a college are normally distributed with mean 65.0 kg and standard deviation 3.0 kg. If 100 samples consisting of 36 students each are obtained, what would be the expected mean and standard deviation of the resulting sampling distribution of means if sampling were done (i) with replacement (ii) without replacement?

Ans. (i) $\mu_{\bar{X}} = 65, \sigma_{\bar{X}} = 0.5$ (ii) $\mu_{\bar{X}} = 65, \sigma_{\bar{X}} = 0.5$

13. The heights of 1500 students are normally distributed with a mean of 68.0 inches and a standard deviation of 2.5 inches. If 300 random samples of size 25 are drawn from this population, determine the expected mean and standard deviation of the sampling distribution of means if sampling is done (i) with replacement (ii) without replacement.

Ans. (i) $\mu_{\bar{X}} = 68, \sigma_{\bar{X}} = 0.5$ (ii) $\mu_{\bar{X}} = 68, \sigma_{\bar{X}} = 0.5$

14. The Federal Bureau of Statistics collects data on earnings of industry workers. The average weekly earnings of workers in the industry is Rs. 1500. Suppose that n such workers are to be selected at random. Let \bar{X} denote the mean

$$\mu = \mu_{\bar{X}} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \left(\frac{\sqrt{N-n}}{N-1} \right) \text{ for (W.O.R)}$$

[Chapter 11] Sampling and Sampling Distributions

75

weekly salary of the workers chosen. Assuming a population standard deviation of Rs. 200, find the mean and standard deviation of \bar{X} if:

- (i) $n = 25$ (ii) $n = 100$ (iii) $n = 400$

Ans: (i) $\mu_{\bar{X}} = 1500, \sigma_{\bar{X}} = 40$ (ii) $\mu_{\bar{X}} = 1500, \sigma_{\bar{X}} = 20$ (iii) $\mu_{\bar{X}} = 1500, \sigma_{\bar{X}} = 10$

15. Suppose a random sample of size n is taken from a population of size N .

- (a) Assume $n = 1$. Compute $\sigma_{\bar{X}}$, if the sampling is done

(i) with replacement (ii) without replacement.

- (b) Assume $n = N$. Compute $\sigma_{\bar{X}}$, when sampling is done without replacement.

Ans: (a)(i) $\sigma_{\bar{X}} = \sigma$ (ii) $\sigma_{\bar{X}} = \sigma$ (b) $\sigma_{\bar{X}} = 0$

16. A population consists of the three numbers 2, 4, 6. Consider all possible samples of size two which can be drawn with replacement from this population. Find the mean of the sampling distribution of variances.

Ans. $\mu_{S^2} = 1.3333$

17. A population of 7 numbers has a mean of 40 and a standard deviation of 3. If

samples of size 5 are drawn from this population and the variance $S^2 = \frac{\sum (X - \bar{X})^2}{n}$ of each sample is computed, find the mean of the sampling distribution of variances if sampling is: (i) with replacement (ii) without replacement.

Ans. (i) 7.2 (ii) 8.4

18. A population consists of four values 4, 10, 14, 20. Take all possible samples of size two without replacement from this population and verify that

$$\mu_{S^2} = \left(\frac{N}{N-1} \right) \left(\frac{n-1}{n} \right) \sigma^2$$

Ans. $\mu_{S^2} = 22.67, \sigma^2 = 34$

19. A population consists of three numbers 4, 6, 8. Take all possible samples of size two with replacement from this population. Find the mean and the unbiased variance for each sample. Show that (i) $\mu_{\bar{X}} = \mu$ and (ii) $\mu_{S^2} = \sigma^2$.

Ans. $\mu_{\bar{X}} = 6, \mu_{S^2} = 2.67, \mu = 6, \sigma^2 = 2.67$

20. A population consists of four values 4, 6, 8, 10. Take all possible samples of size two without replacement from this population and verify that $E(S^2) = \left(\frac{N}{N-1} \right) \sigma^2$

where $s^2 = \sum (X - \bar{X})^2 / (n - 1)$.

Ans: $E(S^2) = 6.67, \sigma^2 = 5$.

21. Let \bar{X}_1 represent the mean of a random sample of size $n_1 = 2$ with replacement from a finite population consisting of values 4, 8. Similarly, let \bar{X}_2 represent the mean of a random sample of size $n_2 = 2$ with replacement from another finite population consisting of values 2, 4. Form a sampling distribution of $\bar{X}_1 - \bar{X}_2$.

$$\text{Verify that: (i) } E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 \text{ (ii) } \text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\text{Ans. } E(\bar{X}_1 - \bar{X}_2) = 3, \text{ Var}(\bar{X}_1 - \bar{X}_2) = 2.5, \mu_1 = 6, \sigma_1^2 = 4, \mu_2 = 3, \sigma_2^2 = 1$$

22. Draw all possible random samples of size $n_1 = 2$ without replacement from a finite population consisting of 3, 6, 9. Similarly draw all possible random samples of size $n_2 = 2$ without replacement from another finite population consisting of 2, 4, 6.

(a) Find the possible differences between the sample means of the two populations.

(b) Construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$ and compute its mean and variance.

$$(c) \text{ Verify that: (i) } \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ (ii) } \sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} \left(\frac{N_1 - n_1}{N_1 - 1} \right) + \frac{\sigma_2^2}{n_2} \left(\frac{N_2 - n_2}{N_2 - 1} \right)$$

$$\text{Ans. } \mu_{\bar{X}_1 - \bar{X}_2} = 2, \sigma_{\bar{X}_1 - \bar{X}_2}^2 = 2.1667, \mu_1 = 6, \sigma_1^2 = 6, \mu_2 = 4, \sigma_2^2 = \frac{8}{3}$$

23. Given $\mu_1 = 4500, \mu_2 = 4000, \sigma_1 = 200, \sigma_2 = 250, N_1 = 400, N_2 = 300, n_1 = 100$ and $n_2 = 50$. Determine the expected mean and standard deviation of the sampling distribution of difference of the means if sampling is done

(i) with replacement (ii) without replacement

$$\text{Ans: (i) 500 and 40.62 (ii) 500 and 36.69}$$

24. Given $N_1 = 125, n_1 = 30, \mu_1 = 78, \sigma_1^2 = 150, N_2 = 200, n_2 = 50, \mu_2 = 85$ and $\sigma_2^2 = 200$. Compute $E(\bar{X}_2 - \bar{X}_1)$ and $\text{Var}(\bar{X}_2 - \bar{X}_1)$ when sampling is done

(i) with replacement (ii) without replacement

$$\text{Ans: (i) 7 and 9 (ii) 7 and 6.85}$$

25. There are five digits in a population i.e; 2, 4, 5, 7 and 10. Draw all possible samples of size 3 without replacement and find the sample proportion (\hat{p}) of even digits in each sample. Verify that:

$$(i) E(\hat{p}) = p(\text{population proportion}) \quad (ii) S.E.(\hat{p}) = \sqrt{\frac{pq}{n} \left(\frac{N-n}{N-1} \right)}$$

$$\text{Ans: } E(\hat{p}) = 0.6, S.E.(\hat{p}) = 0.2, p = 0.6$$

26. Draw all possible samples of size three without replacement from the population 3, 4, 5 and 7. Calculate proportion of odd numbers in each sample and verify that: (i) $E(\hat{p}) = p$ (ii) $\text{Var}(\hat{p}) = \frac{pq}{n} \left(\frac{N-n}{N-1} \right)$

Where \hat{p} and p are proportions of odd numbers in sample and population, respectively.

Ans: $E(\hat{p}) = 0.75$, $\text{Var}(\hat{p}) = 0.0208$, $p = 0.75$

27. In a private lodge, there live five friends and their marital status is U, M, M, U, M where U and M stand for unmarried and married respectively. Find the proportion of married friends in the population. Take all possible samples of two friends without replacement from this population and find the proportion of married friends in each sample. Make the sampling distribution of the sample proportion and verify that: (i) $\mu_{\hat{p}} = p$ (ii) $\sigma_{\hat{p}}^2 = \frac{pq}{n} \left(\frac{N-n}{N-1} \right)$

Ans: $\mu_{\hat{p}} = 0.6$, $\sigma_{\hat{p}}^2 = 0.09$, $p = 0.6$

28. (a) Draw all possible samples of two letters each, with replacement from the letters of the word "NEW".
 (b) Find proportion of letter "E" in each sample.
 (c) Make sampling distribution of proportions obtained in part (b).
 (d) Find mean and variance of the distribution.
 (e) Verify that: (i) $\mu_{\hat{p}} = p$ (ii) $\sigma_{\hat{p}}^2 = \frac{pq}{n}$

Ans. $\mu_{\hat{p}} = 1/3$, $\sigma_{\hat{p}}^2 = 1/9$, $p = 1/3$

29. Suppose a random sample of size $n = 80$ is taken from a population of size $N = 500$. The proportion in the population being calculated is 72.3 %. Find the mean and standard deviation of the distribution of sample proportions when sampling is done (i) with replacement (ii) without replacement.

Ans: (i) $\mu_{\hat{p}} = 0.723$, $\sigma_{\hat{p}} = 0.050$ (ii) $\mu_{\hat{p}} = 0.723$, $\sigma_{\hat{p}} = 0.046$

30. Find the mean and standard deviation of the sampling distribution of proportions for $n = 100$ and a population proportion of

(i) 20 % (ii) 40 % (iii) 50 % (iv) 90 %

Ans: (i) $\mu_{\hat{p}} = 0.20$, $\sigma_{\hat{p}} = 0.04$ (ii) $\mu_{\hat{p}} = 0.40$, $\sigma_{\hat{p}} = 0.05$

(iii) $\mu_{\hat{p}} = 0.50$, $\sigma_{\hat{p}} = 0.05$ (iv) $\mu_{\hat{p}} = 0.90$, $\sigma_{\hat{p}} = 0.03$

31. Let \hat{p}_1 represent the proportion of even numbers in a random sample of size $n_1 = 2$ without replacement from a finite population consisting of values 4, 6, 9. Similarly, let \hat{p}_2 represent the proportion of even numbers in a random sample of

size $n_2 = 2$ without replacement from another finite population consisting of values 2, 2, 5. Form a sampling distribution of $(\hat{p}_1 - \hat{p}_2)$. Verify that:

$$(i) E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 \quad (ii) \text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} \left(\frac{N_1 - n_1}{N_1 - 1} \right) + \frac{p_2 q_2}{n_2} \left(\frac{N_2 - n_2}{N_2 - 1} \right)$$

Ans. $E(\hat{p}_1 - \hat{p}_2) = 0$, $\text{Var}(\hat{p}_1 - \hat{p}_2) = 1/9$, $p_1 = 2/3$, $p_2 = 2/3$

32. Solve directly:

(i) Given $N = 1500$, $n = 36$, $\mu = 22.4$. Find $\mu_{\bar{X}}$.

(ii) Given $n = 25$, $\sigma_{\bar{X}} = 2.5$. Find σ^2 .

(iii) Given $N = 310$, $n = 100$, $\mu_{\bar{X}} = 24000$, $\sigma^2 = 5000$. Find $\sigma_{\bar{X}}^2$ (without replacement).

(iv) Given $\mu = 6$, $\sigma^2 = 8$, $n = 2$. Find $\mu_{\bar{X}}$ and μ_{S^2} when sampling is done with replacement and $S^2 = \sum(X - \bar{X})^2/n$.

(v) Given $\mu_{\bar{X}} = 5$, $\mu_{S^2} = 4$, $n = 2$. Find μ and σ^2 when sampling is done with replacement and $S^2 = \sum(X - \bar{X})^2/n$.

(vi) Given $N = 5$, $n = 2$, $\mu = 6$, $\sigma^2 = 8$. Find $\mu_{\bar{X}}$ and μ_{S^2} when sampling is done without replacement and $S^2 = \sum(X - \bar{X})^2/n$.

(vii) Given $n_1 = 2$, $n_2 = 2$, $\mu_1 = 6$, $\mu_2 = 2$, $\sigma_1^2 = 2.67$, $\sigma_2^2 = 0.67$. Find $\mu_{\bar{X}_1 - \bar{X}_2}$ and $\sigma_{\bar{X}_1 - \bar{X}_2}^2$.

(viii) Given $n_1 = 49$, $n_2 = 36$, $\mu_{\bar{X}_1 - \bar{X}_2} = 0.3$, $\sigma_1^2 = 0.36$, $\sigma_2^2 = 0.16$. Find $\mu_1 - \mu_2$ and $\sigma_{\bar{X}_1 - \bar{X}_2}$.

(ix) Given $\mu_{\bar{X}_1 - \bar{X}_2} = 4$, $\mu_2 = 6$, $\sigma_1 = 2.25$, $N_1 = 30$, $N_2 = 25$, $n_1 = 4$, $n_2 = 4$, $\sigma_{\bar{X}_1 - \bar{X}_2} = 6.25$. Find μ_1 and σ_2 when sampling is done without replacement.

(x) Given $N = 5$, $n = 2$, $\mu_{\hat{p}} = 2/5$. Find p and $\sigma_{\hat{p}}^2$ when sampling is done without replacement.

(xi) Given $p_1 = 1/2$, $p_2 = 1/3$, $N_1 = 3$, $n_1 = 2$, $N_2 = 3$, $n_2 = 2$. Find $E(\hat{p}_1 - \hat{p}_2)$ and S.E. ($\hat{p}_1 - \hat{p}_2$) when sampling is done without replacement.

Ans. (i) 22.4 (ii) 156.25 (iii) 33.98 (iv) 6, 4 (v) 5, 8 (vi) 6, 5

(vii) 4, 1.67 (viii) 0.3, 0.1082 (ix) 10, 13.166 (x) 2/5, 0.09 (xi) 1/6, 0.3436.

Chapter

12

STATISTICAL INFERENCE ESTIMATION

12.1 INTRODUCTION

A person selected at random from a certain place shows the colour, habits and language of the people of his area. He gives some information about all the people. We say that he is the representative of all the people of his area. Whatever we gain from this person is the inference about the lot of people among whom he has been selected. From a particular person, we gain information about the main group of people. The information travels from the particular individual to the general group of people. To get better information about the people, we may select more than one person out of them. We say that the sample size has been increased. A large sample usually contains greater information about the population. In our daily life, if an individual says something, we may not accept it. We may have some doubts about it. But if many people say something, we feel like accepting whatever they say. The same logic works in statistical reasoning. Every individual from a statistical population speaks about the properties of the population. Every drop of water from an ocean has certain characteristics which lead us to some conclusions about the water in the ocean. A single drop of water or oil is sometimes sufficient to give clear picture about these liquids in the big containers. In statistical studies, if the population contains individuals whose characteristics are similar, then a small simple random sample is sufficient to give the required information about the properties of the population. We decide about the sample size according to the nature of population and the nature of the study. The conclusions based on a sample data are related to probability.

12.2 STATISTICAL INFERENCE

The information gained from the sample data is used to reach some conclusions about the characteristics of the population. This process is called *statistical inference*. *Statistical inference* is based on the principles of the sampling theory. There are two approaches of statistical inference namely,

- (i) Estimation of parameters (ii) Hypothesis testing or testing of hypothesis.

In this chapter we discuss *estimation* and hypothesis testing will be discussed in the next chapter.

12.2.1 APPROACHES OF STATISTICAL INFERENCE

There are two approaches of statistical inference called *estimation* and testing of hypothesis. These two approaches are completely different as far as their

conclusions are concerned but both are based on the same theory of probability and the principles of sampling theory. They start with a simple random sample from the population but later on they move on separate paths. Suppose we are interested to know the percentage of smokers in our country. This is a problem which comes under *estimation*. Suppose a manufacturer of copper wires of some size claims that the breaking strength of his product is 10 kg. His claim is to be tested with the help of some tests. This is a problem which comes under hypothesis testing.

12.3 ESTIMATION

Statistical inference about the unknown values of the population parameters is called *estimation* of parameters. Suppose we are interested to know the average life of tires of a certain firm. This means we want an estimate of something which is not known to us. It is a problem of estimation. The minimum and maximum cholesterol level of persons is also a problem of estimation. These estimates are provided by the sample observations. Estimation of parameters is done by two methods which are:

- (i) Point estimation
- (ii) Interval estimation.

12.3.1 POINT ESTIMATOR AND POINT ESTIMATE

Point estimate is a value calculated from the sample data. A single value calculated from a sample or samples is called point estimate. Consider a simple random sample of size 4 with values as 10, 15, 20 and 25. The mean of this sample is $(10 + 15 + 20 + 25)/4 = 17.5$. Thus the sample mean 17.5 is a point estimate of the unknown population parameter μ . We are interested to know the percentage of children under 5 years who take tea regularly with the breakfast. We have taken a simple random sample of 100 children and 60 are found to be habitual of taking tea. The sample proportion $60/100 = 0.6$ is called the point estimate of the unknown population proportion p . If the parameter is θ , the specific value calculated from the sample is called a point estimate of $\hat{\theta}$. Suppose we have two samples of sizes n_1 and n_2 and we have calculated their proportions \hat{p}_1 and \hat{p}_2 . The difference $(\hat{p}_1 - \hat{p}_2)$ is also a point estimate of the actual difference between the population parameters. This unknown parameter may be denoted by $(p_1 - p_2)$.

The word *estimator* is used in general for a statistic. *Estimator* is based on the sample and in general differs from sample to sample. It is a random variable with a probability distribution called the sampling distribution. If the unknown population parameter is denoted by θ , its *estimator* is generally denoted by $\hat{\theta}$ (theta hat). If the

parameter is the population mean μ , the sample mean $\bar{X} = \Sigma X/n$ (with any value) is the *estimator* of μ . Median and mode are also *estimators* of μ . The sample proportion

\hat{p} is an *estimator* of population proportion p . The sample variances $s^2 = \frac{\Sigma(X - \bar{X})^2}{n-1}$ and

$S^2 = \frac{\Sigma(X - \bar{X})^2}{n}$ are *estimators* of population variance σ^2 . We have already learnt in chapter 11 that $E(s^2) = \sigma^2$ and $E(S^2) \neq \sigma^2$. At some later stage we shall call s^2 an unbiased *estimator* of σ^2 and S^2 a biased estimator of σ^2 .

12.3.2 POINT ESTIMATION

Point estimation is a process of getting a single value from the sample as an *estimate* of the unknown population parameter. A *point estimate*, in general, is not equal to the population parameter. *Point estimation* is of great importance in practical life. In our daily life it is quite common that we make use of the *point estimates* without referring to the idea of statistical inference. The percentage of people living in rented houses is a problem of *point estimation*. The percentage of bottles properly filled is provided by a *point estimate*. The percentage of babies who are born with physical defects and the percentage of children who do not get admission in the schools are the areas of *point estimation*. A serious drawback in *point estimation* is that the amount of error cannot be calculated in *point estimation*.

12.3.3 UNBIASEDNESS

When a large number of random samples of a given size are taken and the value of the estimator is calculated for each sample, the average of these values may be equal to the population parameter. An estimator having this property is called *unbiased estimator*. In other words, the estimator $\hat{\theta}$ is called unbiased estimator of the parameter θ if $E(\hat{\theta}) = \theta$. The estimator is a statistic with a probability distribution which is called the sampling distribution of the statistic. An estimator is called an unbiased estimator of the population parameter if the mean of its sampling distribution is equal to the parameter. The sample mean \bar{X} is an *unbiased estimator* of the population mean μ . It means $E(\bar{X}) = \mu$. The sample proportion \hat{p} is also *unbiased estimator* of the population proportion p because $E(\hat{p}) = p$. The sample

variance $s^2 = \frac{\sum(X - \bar{X})^2}{n-1}$ is *unbiased estimator* of the population variance σ^2 but

$S^2 = \frac{\sum(X - \bar{X})^2}{n}$ is a biased estimator of σ^2 and the bias is equal to the difference $E(S^2) - \sigma^2$.

Unbiasedness is one of the important properties of good point estimators. Other properties of good point estimators are *consistency*, *efficiency* and *sufficiency*. These properties will not be discussed in this book.

12.3.4 IMPORTANCE OF UNBIASEDNESS

Unbiasedness plays a major role in statistical inference. The next chapter is about the testing of hypothesis wherein we shall learn that the hypothesis about the population parameter is in fact the hypothesis about the sampling distribution of the estimator of that parameter. If we infer that the mean of the sampling distribution of \bar{X} is, say 150, it means that the mean of the population is also 150. This is due to the unbiasedness of \bar{X} . If \bar{X} were a biased estimator, some very important tests of hypotheses about μ would not have been possible.

12.4 INTERVAL ESTIMATION

If a random interval is calculated so that it contains the unknown parameter with a known probability, then the interval is called *confidence interval estimate* or simply confidence interval for the parameter. The process of finding such intervals is called *interval estimation*. The interval estimation has gained a lot of importance in statistical inference. It is based on random sampling. Thus the confidence interval constructed is a random term because it is based on the sample data.

A drawback in point estimation is that it does not provide the estimate of error. No assurance is attached to the point estimate. Point estimate is a single value and it is wrong to think that a single value will be equal to the value of the unknown parameter. The interval estimate has some assurance of containing the population parameter. Confidence interval tells us with a known degree of confidence as to where the population parameter actually lies.

12.4.1 CONFIDENCE COEFFICIENT

The probability attached to the confidence interval is called *confidence coefficient* or level of confidence. It is denoted by $1 - \alpha$. If α is specified as 0.05, then $1 - \alpha = 1 - 0.05 = 0.95$ or 95 %. We can speak of confidence coefficient in terms of unity or in terms of percentage. The confidence coefficients which are commonly used are 90 %, 95 %, 98 % and 99 %.

12.5 CONSTRUCTION OF CONFIDENCE INTERVAL

To make a confidence interval estimate of the parameter θ , we adopt the following procedure.

- (i) We take a random sample of size n with observations $X_1, X_2, X_3, \dots, X_n$ from a population with unknown parameter θ .
- (ii) The point estimator of θ is decided. Let it be $\hat{\theta}$. The point estimate denoted by $\hat{\theta}_p$ is calculated from the sample.
- (iii) The confidence coefficient is decided. Let it be $(1 - \alpha)$.
- (iv) Let the interval be denoted by (L, U) where L is the lower limit and U is the upper limit.
- (v) A certain procedure of calculating L and U is adopted such that probability is $(1 - \alpha)$ that the interval (L, U) contains the parameter θ . In symbols, we may write $P[L < \theta < U] = 1 - \alpha$ where α lies between 0 and 1 but it is usually small. The extreme ends of the interval are L and U which are called the *lower and upper confidence limits* of the parameter θ . L and U are random terms based on the sample data.

The lower limit L and the upper limit U are calculated from the point estimate $\hat{\theta}_p$. Thus, as a general rule,

$$L = \hat{\theta}_p - k \text{ (Standard error of } \hat{\theta}) \text{ and } U = \hat{\theta}_p + k \text{ (Standard error of } \hat{\theta})$$

where k depends upon the shape of the sampling distribution of $\hat{\theta}$ and the confidence coefficient $1 - \alpha$. The estimator $\hat{\theta}$ may be sample mean \bar{X} , sample proportion \hat{p} , the difference between means $(\bar{X}_1 - \bar{X}_2)$ or the difference between proportions $(\hat{p}_1 - \hat{p}_2)$.

12.5.1 SELECTION OF PROPER CONFIDENCE INTERVAL

For making the confidence interval estimate for some parameter, we have to use the appropriate formula. Some intervals are based on the normal distribution and some are based on the t-distribution. It is in fact the sampling distribution of the statistic which decides the formula. If the sampling distribution of the statistic is a normal distribution, then the standard normal variate Z is used in the interval and if the distribution is 't', then the random variable 't', is used in the formula. It is important to note that it is the sampling distribution which decides the proper formula. It is not the parent population which decides the interval, though the shape of the population distribution also plays its role in determining the proper interval. We have to examine the following points for making the confidence interval for the population mean μ .

(i) Parent Population

What is the shape of the population which is sampled? Is it normal, approximately normal for practical purposes or known to be non-normal?

(ii) Sample Size

The sample size n plays an important role in the statistical inference about μ . When $n > 30$, it is called large sample size. According to the Central limit theorem, the sampling distribution of \bar{X} tends to normality by increasing the sample size.

(iii) σ is Known or Unknown

If σ is known, the distribution of \bar{X} can be assumed normal even if $n \leq 30$, provided the population is normal. If σ is unknown and $n \leq 30$, the distribution of \bar{X} is not assumed to be normal.

t - DISTRIBUTION

The sampling distribution of \bar{X} forms t-distribution under the following conditions.

- The simple random sample of small size is drawn from a normal population with mean μ . This assumption is very important for any inference about \bar{X} .
- The sample $X_1, X_2, X_3, \dots, X_n$ is selected at random.
- If there are two populations under consideration, both are normal with equal variances.

If $X_1, X_2, X_3, \dots, X_n$ is a random sample of size n from a normal population with mean μ and variance σ^2 and \bar{X} is the sample mean, then the random variable 't' is

defined as $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ where s is the sample standard deviation defined by

$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}}$. The random variable 't' forms the t-distribution with $(n-1)$ degrees of freedom (d.f.)

The t-distribution is symmetrical about its mean zero like the normal distribution. The shape of the t-distribution changes by increasing the sample size. When sample size is sufficiently large ($n > 30$), the t-distribution tends to the normal distribution.

Tables are available from which we can read the t-values for given values of α . If $\alpha = 0.05$ and degrees of freedom is 9, then from the t-table, we read under column 0.05 and against 9 degrees of freedom. We get 1.833. It is written as $t_{0.05}(9) = 1.833$. Similarly $t_{0.025}(8) = 2.306$.

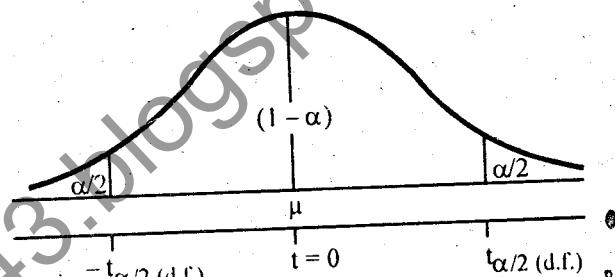


Figure 12.1

12.6 CONFIDENCE INTERVAL ESTIMATE OF POPULATION MEAN μ (LARGE SAMPLE)

σ -Known

Let us consider a population (normal or non-normal) with mean μ which is unknown and the variance σ^2 which is assumed to be known. A simple random sample of size n is selected from the population and the sample mean \bar{X} is calculated. When the sample size is large ($n > 30$), the sampling distribution of \bar{X} is a normal distribution with mean $\mu_{\bar{X}} = \mu$ and standard error $\sigma_{\bar{X}}$ where $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. It is assumed here that the population is infinite or it is very large. The population may or may not be normal. The normal distribution of \bar{X} is shown below:

The random variable \bar{X} can be transformed into standard normal variable Z

where $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. The random variable Z can take any value between $-\infty$ to $+\infty$. Let us mark two points $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$ on Z-scale, where α lies between 0 and 1. $-Z_{\frac{\alpha}{2}}$ is a

point on the left of which the area under normal distribution of Z is $\frac{\alpha}{2}$ and $Z_{\frac{\alpha}{2}}$ cuts off an area $\frac{\alpha}{2}$ to the right. Thus the area of the normal curve between $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$ is $1 - \alpha$, the total area under the normal curve being unity. Out of all possible values of Z , $100(1 - \alpha)\%$ of the values occupy the space marked $1 - \alpha$. Thus the probability is $(1 - \alpha)$ that the random variable Z will take a value between $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$. This probability statement can be written in symbols as $P[-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}] = 1 - \alpha$. Putting $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$, we get

$$P\left[-Z_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < Z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

Without proof, we write the confidence interval for μ which is

$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

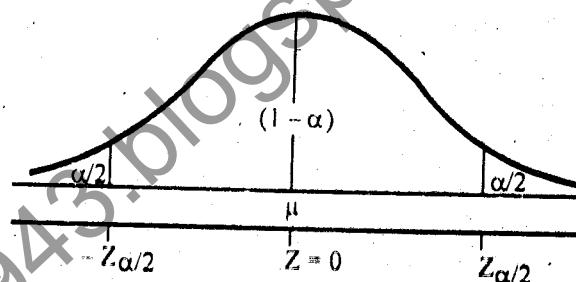


Figure 12.2

It is called $100(1 - \alpha)\%$ confidence interval for μ . The lower confidence limit is

$$L = \bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$
 and the upper confidence limit is $U = \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

The interval can also be written as $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

For 95% confidence interval for μ ,

$\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$. From the area

table of normal distribution

$$Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96.$$

Thus 95% confidence interval for μ is

$$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$

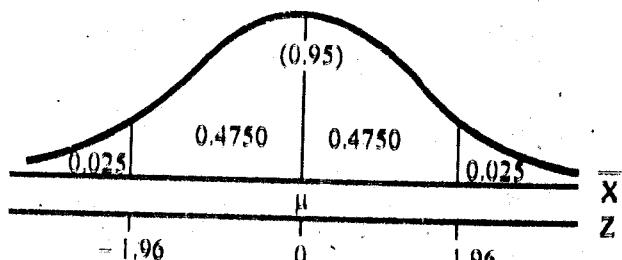


Figure 12.3

For 99% confidence interval for μ , $\alpha = 0.01$, $\frac{\alpha}{2} = 0.005$ and $Z_{0.005} = 2.58$ (From the area table of normal distribution)

$$\bar{X} - 2.58 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2.58 \frac{\sigma}{\sqrt{n}}$$

This interval is wider than the above 95 % interval. A very wide interval, which gives the most probable confidence limits for μ is

$$\bar{X} - 3 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 3 \frac{\sigma}{\sqrt{n}}$$

This interval is almost certain to contain the true value of μ .

Finite Population

When the population is finite or sampling is done without replacement, the standard error of \bar{X} is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$. When N is given, the confidence interval for μ would become.

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

12.6.1 MEANING OF THE CONFIDENCE INTERVAL

Let us consider 95 % confidence interval for μ . The interval has $L = \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}$ and

$U = \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$. The probability is 0.95 that the random interval (L, U) contains the unknown parameter μ . This means that if samples of size n were repeatedly taken from the population and if the random confidence interval $(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}})$ were computed for each sample, then 95 out of 100 such

intervals would in the long run contain the unknown parameter μ .

When we construct 95 % confidence interval, then there are 5 % chances that

we shall get an interval which will not cover the value of μ . Suppose $\bar{X} = 80$, $\sigma = 24$ and $n = 36$. Let us put these values in 95 % confidence interval, we get

$$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} = 80 - 1.96 \frac{24}{\sqrt{36}} = 80 - 1.96 \times 7.84 = 64.63$$

$$\text{and } \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} = 80 + 1.96 \frac{24}{\sqrt{36}} = 80 + 1.96 \times 7.84 = 95.37$$

For a specified random sample, the 95 % confidence interval for μ is (64.63 to 95.37). We call it 95 % confidence interval. At this stage we cannot say that probability is 0.95 that the interval (64.63, 95.37) contains the value of μ . Before tossing a die, we say that probability is $1/6$ that 4 will come on the die. But when a die has been tossed and the face 4 has been observed or it has not been observed then we are not in a situation of probability. Now we cannot say that probability of getting 4 on the die is $1/6$. If 2% of the drivers on the average violate the traffic rules, then the probability is $2/100 = 0.02$ that a driver will violate the rules. But when a driver has made a mistake and he has been arrested, now he has no concern with the probability of 0.02. Probability is for uncertain situations. When something

has happened or it has not happened, it is now ridiculous to talk about the probability of its happening. When Mr. A has died, we do not say that probability of his survival is, say 0.90 when a confidence interval has been constructed from the sample data, it is now something which has happened or which has been determined. The different possibilities are not involved now. The calculated interval is now not a random variable. It is the realized value of the random interval.

Example 12.1.

- An electrical firm manufactures light bulbs that have a length of life with mean μ and a standard deviation of 40 hours. If a sample of 100 bulbs has an average life of 780 hours, find a 95 % confidence interval for the population mean of all bulbs produced by this firm.
- A random sample of size $n = 400$, selected without replacement from a population of size $N = 2000$ with $\sigma = 4$, the sample mean is found to be $\bar{X} = 80$. Construct a 90 % confidence interval for the true mean of the population.

Solution:

- A 100 $(1 - \alpha)$ % confidence interval for μ is

$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Here $\bar{X} = 780$, $\sigma = 40$, $n = 100$, $1 - \alpha = 0.95$ or $\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$

From the area table of normal distribution, we have $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$

Hence the 95 % confidence interval for μ is

$$780 - 1.96 \left(\frac{40}{\sqrt{100}} \right) < \mu < 780 + 1.96 \left(\frac{40}{\sqrt{100}} \right)$$

$$780 - 7.84 < \mu < 780 + 7.84$$

$$772.16 < \mu < 787.84$$

- A 100 $(1 - \alpha)$ % confidence interval for μ is

$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Here $n = 400$, $N = 2000$, $\sigma = 4$, $\bar{X} = 80$, $1 - \alpha = 0.90$ or $\alpha = 0.10$ and $\frac{\alpha}{2} = 0.05$

From the area table of normal distribution, we have $Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$

Hence the 90 % confidence interval for μ is

$$80 - 1.645 \left(\frac{4}{\sqrt{400}} \right) \sqrt{\frac{2000-400}{2000-1}} < \mu < 80 + 1.645 \left(\frac{4}{\sqrt{400}} \right) \sqrt{\frac{2000-400}{2000-1}}$$

$$80 - 0.294 < \mu < 80 + 0.294$$

$$79.706 < \mu < 80.294$$

σ-Unknown

In the previous article it was assumed that σ is known. In practical situations, σ is usually not known. When σ is not known, we can replace it by the sample standard deviation S . In this case the confidence interval for μ is

$$\bar{X} - \frac{Z_{\alpha/2}}{2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + \frac{Z_{\alpha/2}}{2} \frac{S}{\sqrt{n}}$$

It is important to note that this interval estimate can be used only when n is large. But the population may or may not be normal. When the population is finite, the interval for μ would become

$$\bar{X} - \frac{Z_{\alpha/2}}{2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + \frac{Z_{\alpha/2}}{2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

This interval can be calculated when N is given.

Example 12.2.

The heights of a random sample of 50 college students showed a mean of 174.5 centimeters and a standard deviation of 6.9 centimeters. Construct a 98 % confidence interval for the mean height of all college students.

Solution:

A 100 $(1 - \alpha)$ % confidence interval for μ is

$$\bar{X} - \frac{Z_{\alpha/2}}{2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + \frac{Z_{\alpha/2}}{2} \frac{S}{\sqrt{n}}$$

Here $\bar{X} = 174.5$, $S = 6.9$, $n = 50$, $1 - \alpha = 0.98$ or $\alpha = 0.02$ and $\frac{\alpha}{2} = 0.01$

From the area table of normal distribution, we have $Z_{\alpha/2} = Z_{0.01} = 2.326$

Hence the 98 % confidence interval for μ is

$$174.5 - 2.326 \left(\frac{6.9}{\sqrt{50}} \right) < \mu < 174.5 + 2.326 \left(\frac{6.9}{\sqrt{50}} \right)$$

$$174.5 - 2.27 < \mu < 174.5 + 2.27$$

$$172.23 < \mu < 176.77$$

Example 12.3.

The systolic blood pressure of 90 men has a mean of 128.9 mm of mercury and a standard deviation of 17 mm of mercury. Assuming that these are a random sample of blood pressures, calculate a 99 % confidence interval for the mean blood pressure in the population.

Solution: A 100 $(1 - \alpha)$ % confidence interval for μ is

$$\bar{X} - \frac{Z_{\alpha/2}}{2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + \frac{Z_{\alpha/2}}{2} \frac{S}{\sqrt{n}}$$

Here $\bar{X} = 128.9$, $S = 17$, $n = 90$, $1 - \alpha = 0.99$ or $\alpha = 0.01$ and $\frac{\alpha}{2} = 0.005$

From the area table of normal distribution, we have $Z_{\frac{\alpha}{2}} = Z_{0.005} = 2.575$

Hence the 99 % confidence interval for μ is

$$128.9 - 2.575 \left(\frac{17}{\sqrt{90}} \right) < \mu < 128.9 + 2.575 \left(\frac{17}{\sqrt{90}} \right)$$

$$128.9 - 4.61 < \mu < 128.9 + 4.61$$

$$124.29 < \mu < 133.51$$

12.7 CONFIDENCE INTERVAL ESTIMATE FOR POPULATION MEAN μ - POPULATION NORMAL (SMALL SAMPLE)

σ -Known, Population Normal

Here we are stressing that the population is normal. If n is small, σ is known, then the random variable Z can be used in the interval only when the population is normal. The confidence interval for μ is the same as for the large sample. For the convenience of students, the $100(1 - \alpha)\%$ confidence interval for μ is reproduced here i.e.,

$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

σ Not Known, Population Normal

This is an important case in which the random variable Z cannot be used. When n is small, population is normal with unknown σ , the random variable

$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$ has the t -distribution with $(n - 1)$ degrees of freedom.

Let us mark two points $-t_{\frac{\alpha}{2}(n-1)}$

and $t_{\frac{\alpha}{2}(n-1)}$ on the t -scale in the figure.

Using tables of the t -distribution, we can find $-t_{\frac{\alpha}{2}(n-1)}$ and $t_{\frac{\alpha}{2}(n-1)}$. The area

of the t -distribution between $-t_{\frac{\alpha}{2}(n-1)}$ and $t_{\frac{\alpha}{2}(n-1)}$ is $(1-\alpha)$.

Thus the

probability is $(1-\alpha)$ that the random

variable ' t ' will fall between $-t_{\frac{\alpha}{2}(n-1)}$ and $t_{\frac{\alpha}{2}(n-1)}$. We can write the probability

statement as: $P \left[-t_{\frac{\alpha}{2}(n-1)} < t < t_{\frac{\alpha}{2}(n-1)} \right] = 1 - \alpha$

Putting the value of ' t ', we have

$$P \left[-t_{\frac{\alpha}{2}(n-1)} < \frac{\bar{X} - \mu}{s / \sqrt{n}} < t_{\frac{\alpha}{2}(n-1)} \right] = 1 - \alpha$$

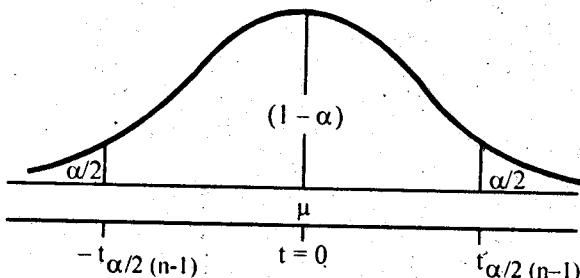


Figure 12.4

From this inequality within the brackets we can get the confidence interval for μ . Thus $100(1 - \alpha)$ % confidence interval for μ is

$$\bar{X} - t_{\frac{\alpha}{2}(n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}(n-1)} \frac{s}{\sqrt{n}}$$

Use of Z or t

Table 12.1. can be used to decide whether the random variable 'Z' or 't' is to be used in making the confidence interval for μ .

Table 12.1.

	n - Large	n - Small
σ - known	Z	Z (normal population)
σ - unknown	Z	t (normal population)

Example 12.4.

A random sample of $n = 20$ from a normal population gives the sample mean 140 and the sample standard deviation, $s = 8$. Construct a 98 % confidence interval for the population mean.

Solution:

Here n is small, therefore the confidence interval based on t-distribution is used. A $100(1 - \alpha)$ % confidence interval for μ is

$$\bar{X} - t_{\frac{\alpha}{2}(n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}(n-1)} \frac{s}{\sqrt{n}}$$

Here $\bar{X} = 140$, $s = 8$, $n = 20$, $1 - \alpha = 0.98$ or $\alpha = 0.02$ and $\frac{\alpha}{2} = 0.01$

From the t-table, we have $t_{\frac{\alpha}{2}(n-1)} = t_{0.01(19)} = 2.539$

Hence the 98 % confidence interval for μ is

$$140 - 2.539 \left(\frac{8}{\sqrt{20}} \right) < \mu < 140 + 2.539 \left(\frac{8}{\sqrt{20}} \right)$$

$$140 - 4.54 < \mu < 140 + 4.54$$

$$135.46 < \mu < 144.54$$

Example 12.5.

A machine is producing metal pieces that are cylindrical in shape. A sample of pieces is taken and their diameters are 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01 and 1.03 centimeters. Find a 99 % confidence interval for the mean diameter of pieces produced by this machine, assuming an approximate normal population.

Solution:

A $100(1 - \alpha)$ % confidence interval for μ is

$$\bar{X} - t_{\frac{\alpha}{2}(n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}(n-1)} \frac{s}{\sqrt{n}}$$

Here $\Sigma X = 9.05$, $\Sigma X^2 = 9.1051$, $n = 9$, $\bar{X} = \frac{\Sigma X}{n} = \frac{9.05}{9} = 1.0056$,

$$s^2 = \frac{1}{n-1} \left[\Sigma X^2 - \frac{(\Sigma X)^2}{n} \right] = \frac{1}{8} \left[9.1051 - \frac{(9.05)^2}{9} \right] = 0.0006, s = 0.0245,$$

$$1 - \alpha = 0.99 \text{ or } \alpha = 0.01 \text{ and } \frac{\alpha}{2} = 0.005$$

From the t-table, we have $t_{\frac{\alpha}{2}(n-1)} = t_{0.005(8)} = 3.355$

Hence the 99 % confidence interval for μ is

$$1.0056 - 3.355 \left(\frac{0.0245}{\sqrt{9}} \right) < \mu < 1.0056 + 3.355 \left(\frac{0.0245}{\sqrt{9}} \right)$$

$$1.0056 - 0.0274 < \mu < 1.0056 + 0.0274$$

$$0.9782 < \mu < 1.0330$$

12.8 CONFIDENCE INTERVAL ESTIMATE FOR THE DIFFERENCE BETWEEN TWO POPULATION MEANS (LARGE SAMPLES)

σ_1^2 and σ_2^2 known

Consider two large populations with means μ_1 and μ_2 which are unknown and variances σ_1^2 and σ_2^2 which are assumed to be known. The populations may or may not be normal. Two independent random samples of sizes n_1 and n_2 are selected from the populations and sample means \bar{X}_1 and \bar{X}_2 are calculated. The point estimator of the difference between μ_1 and μ_2 is given by the statistic $\bar{X}_1 - \bar{X}_2$. The statistic $\bar{X}_1 - \bar{X}_2$ is an unbiased estimator of $\mu_1 - \mu_2$ and has the normal distribution with mean $\mu_1 - \mu_2$ and standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. The standard normal variable of $(\bar{X}_1 - \bar{X}_2)$ is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The probability is $(1 - \alpha)$ that the value of random variable Z will fall between two selected points $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$. We can write the probability statement as

$$P \left[-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}} \right] = 1 - \alpha$$

$$\text{or } P \left[-Z_{\frac{\alpha}{2}} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < Z_{\frac{\alpha}{2}} \right] = 1 - \alpha$$

We can simplify this inequality to get the $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ which is

$$(\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

This interval estimate can be written as

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If we want to get the confidence interval of $\mu_2 - \mu_1$, we shall use the interval in which the difference $(\bar{X}_2 - \bar{X}_1)$ is used. Thus the confidence interval for $\mu_2 - \mu_1$ is

$$(\bar{X}_2 - \bar{X}_1) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

In the numerical questions, the values of \bar{X}_1 and \bar{X}_2 are usually positive but the difference $(\bar{X}_1 - \bar{X}_2)$ or $(\bar{X}_2 - \bar{X}_1)$ may be positive or negative. The confidence limits of $(\mu_1 - \mu_2)$ or $(\mu_2 - \mu_1)$ are sometimes negative.

Example 12.6.

A random sample of size $n_1 = 25$ taken from a normal population with a standard deviation $\sigma_1 = 5$ has a mean $\bar{X}_1 = 80$. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3$, has a mean $\bar{X}_2 = 75$. Find a 94 % confidence interval for $\mu_1 - \mu_2$.

Solution:

A $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Here $n_1 = 25$, $\sigma_1 = 5$, $\sigma_1^2 = 25$, $\bar{X}_1 = 80$, $n_2 = 36$, $\sigma_2 = 3$, $\sigma_2^2 = 9$, $\bar{X}_2 = 75$,

$1 - \alpha = 0.94$ or $\alpha = 0.06$ and $\alpha/2 = 0.03$

From the area table of normal distribution, we have $Z_{\frac{\alpha}{2}} = Z_{0.03} = 1.88$

Hence the 94 % confidence interval for $\mu_1 - \mu_2$ is

$$(80 - 75) - 1.88 \sqrt{\frac{25}{25} + \frac{9}{36}} < \mu_1 - \mu_2 < (80 - 75) + 1.88 \sqrt{\frac{25}{25} + \frac{9}{36}}$$

$$5 - 2.1 < \mu_1 - \mu_2 < 5 + 2.1$$

$$2.9 < \mu_1 - \mu_2 < 7.1$$

σ_1^2 and σ_2^2 are Unknown

When the population variances σ_1^2 and σ_2^2 are not given, they are estimated by the sample variances S_1^2 and S_2^2 . The populations may or may not be normal. The confidence interval for $(\mu_1 - \mu_2)$ becomes

$$(\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Example 12.7.

Construct a 95 % confidence interval for the true difference between the average time in breakdowns of two kinds of devices, given that a random sample of 40 devices of type A on the average lasted 208 hours of continuous use between breakdowns with a standard deviation of 26 hours, and that a random sample of 50 devices of type B lasted on the average 192 hours with a standard deviation of 22 hours.

Solution:

A 100 $(1 - \alpha)$ % confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Here $n_1 = 40$, $\bar{X}_1 = 208$, $S_1 = 26$, $S_1^2 = 676$,

$n_2 = 50$, $\bar{X}_2 = 192$, $S_2 = 22$, $S_2^2 = 484$

$1 - \alpha = 0.95$ or $\alpha = 0.05$ and $\alpha/2 = 0.025$

From the area table of normal distribution, we have $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$

Hence the 95 % confidence interval for $\mu_1 - \mu_2$ is

$$(208 - 192) - 1.96 \sqrt{\frac{676}{40} + \frac{484}{50}} < \mu_1 - \mu_2 < (208 - 192) + 1.96 \sqrt{\frac{676}{40} + \frac{484}{50}}$$

$$16 - 10.1 < \mu_1 - \mu_2 < 16 + 10.1$$

$$\therefore 5.9 < \mu_1 - \mu_2 < 26.1$$

12.9 CONFIDENCE INTERVAL ESTIMATE FOR THE DIFFERENCE BETWEEN TWO POPULATION MEANS-POPULATIONS NORMAL (SMALL SAMPLES)

σ_1^2 and σ_2^2 Known

When the populations are normal and their variances are known, the formula for confidence interval for $(\mu_1 - \mu_2)$ for small samples is the same as for large

samples. To keep the different intervals in their proper order, the interval is highlighted in this section and is written here again. Thus $100(1 - \alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

σ_1^2 and σ_2^2 Unknown, $\sigma_1^2 = \sigma_2^2 = \sigma^2$

When σ_1^2 and σ_2^2 are not known, the random variable Z cannot be used to find a confidence interval for $\mu_1 - \mu_2$. For normal populations, with small sample sizes, the statistic $(\bar{X}_1 - \bar{X}_2)$ has the t-distribution with $(n_1 + n_2 - 2)$ degree of freedom. But we have to make another assumption that the variances of the populations are equal i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say). The population variance σ^2 which is common for both the populations can be estimated by a pooled estimator s_p^2 where,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\sum(X_1 - \bar{X}_1)^2 + \sum(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

s_1^2 and s_2^2 are the unbiased sample variances. Thus, if independent samples of small sizes are drawn from the normal populations with $\sigma_1^2 = \sigma_2^2$ the statistic $(\bar{X}_1 - \bar{X}_2)$ has the t-distribution with $(n_1 + n_2 - 2)$ degree of freedom, where

$$\begin{aligned} t &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \\ &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \end{aligned}$$

when $n_1 = n_2 = n$, we can write,

$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{2}{n}}}$. The probability is $(1 - \alpha)$ that the random variable t will fall

between $-\frac{t_{\alpha/2}(n_1+n_2-2)}{2}$ and $\frac{t_{\alpha/2}(n_1+n_2-2)}{2}$. The probability statement for the random variable t is:

$$P \left[-\frac{t_{\alpha/2}(n_1+n_2-2)}{2} < t < \frac{t_{\alpha/2}(n_1+n_2-2)}{2} \right] = 1 - \alpha$$

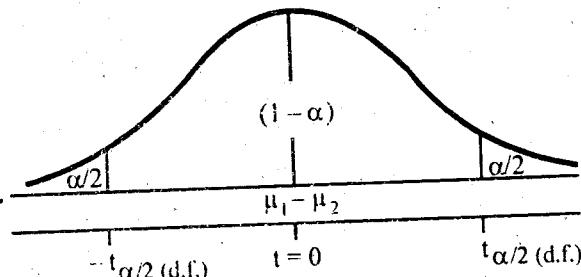


Figure 12.5

Putting the value of t, we get

$$P \left[-\frac{t_{\alpha/2}(n_1 + n_2 - 2)}{s_p} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < \frac{t_{\alpha/2}(n_1 + n_2 - 2)}{s_p} \right] = 1 - \alpha$$

Now, we directly write the $100(1 - \alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ which is

$$(\bar{X}_1 - \bar{X}_2) \pm \frac{t_{\alpha/2}(n_1 + n_2 - 2)}{s_p} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example 12.8.

The following summary statistics are recorded about the strength of two types of synthetic rubber.

Type I	$n_1 = 16$	$\bar{X}_1 = 15.3$	$s_1 = 4.4$
Type II	$n_2 = 9$	$\bar{X}_2 = 13.8$	$s_2 = 3.9$

Assume that the distribution of strengths for the two types of rubber are normal with equal variances. Compute a 99 % confidence interval for the difference $\mu_1 - \mu_2$.

Solution:

A $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) - \frac{t_{\alpha/2}(v)}{s_p} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + \frac{t_{\alpha/2}(v)}{s_p} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Here $n_1 = 16$, $\bar{X}_1 = 15.3$, $s_1 = 4.4$, $s_1^2 = 19.36$

$n_2 = 9$, $\bar{X}_2 = 13.8$, $s_2 = 3.9$, $s_2^2 = 15.21$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(16 - 1)19.36 + (9 - 1)15.21}{16 + 9 - 2} = \frac{290.4 + 121.68}{23}$$

$$= \frac{412.08}{23} = 17.916, s_p = \sqrt{17.916} = 4.233$$

$$1 - \alpha = 0.99 \quad \text{or} \quad \alpha = 0.01 \quad \text{and} \quad \frac{\alpha}{2} = 0.005, v = n_1 + n_2 - 2 = 16 + 9 - 2 = 23$$

From the t-table, we have $t_{\alpha/2(v)} = t_{0.005(23)} = 2.807$

Hence the 99 % confidence interval for $\mu_1 - \mu_2$ is

$$(15.3 - 13.8) - (2.807)(4.233) \sqrt{\frac{1}{16} + \frac{1}{9}} < \mu_1 - \mu_2 < (15.3 - 13.8) + (2.807)(4.233) \sqrt{\frac{1}{16} + \frac{1}{9}}$$

$$1.5 - 4.95 < \mu_1 - \mu_2 < 1.5 + 4.95$$

$$-3.45 < \mu_1 - \mu_2 < 6.45$$

12.10 CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO POPULATION MEANS-DEPENDENT SAMPLES (PAIRED OBSERVATIONS)

Suppose we give a test to a sample of students and the marks obtained by them are denoted by X where X takes the values $X_1, X_2, X_3, \dots, X_n$. The students are given some extra coaching and again they are given the test of the same difficulty and the marks obtained by them are denoted by Y where Y takes the values $Y_1, Y_2, Y_3, \dots, Y_n$. The marks obtained in the first test are called 'before' and the marks obtained in the second test are called 'after' observations. These two sets of marks are in pairs like $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots (X_n, Y_n)$ and are called paired observations. Obviously the Y values depend upon the X values, hence the samples are dependent. Let us write the paired observations in the following form and calculate the difference d_i for each pair.

X_i	Y_i	$d_i = X_i - Y_i$
X_1	Y_1	$X_1 - Y_1 = d_1$
X_2	Y_2	$X_2 - Y_2 = d_2$
X_3	Y_3	$X_3 - Y_3 = d_3$
\vdots	\vdots	\vdots
X_n	Y_n	$X_n - Y_n = d_n$

The mean of 'd' values is denoted by \bar{d} where $\bar{d} = \frac{\sum d_i}{n}$. We can think of a population of X_i and Y_i observations with means μ_1 and μ_2 and the population of random differences d_i with mean μ_d and standard error σ_d . It is required to calculate the confidence interval for the mean μ_d . The distribution of \bar{d} has the t-distribution with $(n - 1)$ degrees of freedom having mean μ_d and standard error $\frac{\sigma_d}{\sqrt{n}}$. The random

variable \bar{d} can be transformed into random variable t where $t = \frac{\bar{d} - \mu_d}{\sigma_d / \sqrt{n}}$.

The standard deviation σ_d is unknown and is replaced by its sample estimate s_d

where $s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}}$. Thus $t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$. The random variable 't' lies between $-\frac{t_{\alpha/2}(n-1)}{2}$ and $\frac{t_{\alpha/2}(n-1)}{2}$ with a probability of $(1-\alpha)$. We can write the probability statement

$$P \left[-\frac{t_{\alpha/2}(n-1)}{2} < t < \frac{t_{\alpha/2}(n-1)}{2} \right] = 1 - \alpha \quad \text{or} \quad P \left[-\frac{t_{\alpha/2}(n-1)}{2} < \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} < \frac{t_{\alpha/2}(n-1)}{2} \right] = 1 - \alpha$$

The terms within the brackets can be written as:

$$P\left[\bar{d} - \frac{t_{\alpha/2}(n-1)}{\sqrt{n}} s_d < \mu_d < \bar{d} + \frac{t_{\alpha/2}(n-1)}{\sqrt{n}} s_d\right] = 1 - \alpha$$

Thus $100(1 - \alpha)\%$ confidence interval for μ_d is

$$\bar{d} - \frac{t_{\alpha/2}(n-1)}{\sqrt{n}} s_d < \mu_d < \bar{d} + \frac{t_{\alpha/2}(n-1)}{\sqrt{n}} s_d$$

Example 12.9.

The following data give paired yields of two varieties of wheat. Each pair was planted in a different locality.

Locality	1	2	3	4	5
Variety I	40	25	37	43	46
Variety II	47	27	33	40	52

Compute a 95 % confidence interval for the mean difference between the yields of the two varieties, assuming the differences of yields to be approximately normally distributed.

Solution:

A $100(1 - \alpha)\%$ confidence interval for $\mu_d = \mu_1 - \mu_2$ is

$$\bar{d} - \frac{t_{\alpha/2}(n-1)}{\sqrt{n}} s_d < \mu_d < \bar{d} + \frac{t_{\alpha/2}(n-1)}{\sqrt{n}} s_d$$

The necessary calculations are given below:

X_1	40	25	37	43	46	
X_2	47	27	33	40	52	
$d_i = X_1 - X_2$	-7	-2	4	3	-6	$\sum d_i = -8$
d_i^2	49	4	16	9	36	$\sum d_i^2 = 114$

$$\bar{d} = \frac{\sum d_i}{n} = -\frac{8}{5} = -1.6$$

$$s_d^2 = \frac{1}{n-1} \left[\sum d_i^2 - \frac{(\sum d_i)^2}{n} \right] = \frac{1}{5-1} \left[114 - \frac{(-8)^2}{5} \right] = \frac{1}{4} [101.2] = 25.3,$$

$$s_d = 5.03, 1 - \alpha = 0.95 \text{ or } \alpha = 0.05 \text{ and } \frac{\alpha}{2} = 0.025$$

From the t-table, we have $t_{\alpha/2(n-1)} = t_{0.025(4)} = 2.776$

Hence the 95 % confidence interval for $\mu_d = \mu_1 - \mu_2$ is

$$\begin{aligned} -1.6 - 2.776 \frac{5.03}{\sqrt{5}} &< \mu_d < -1.6 + 2.776 \frac{5.03}{\sqrt{5}} \\ -1.6 - 6.24 &< \mu_d < -1.6 + 6.24 \\ -7.84 &< \mu_d < 4.64 \end{aligned}$$

12.11 PROPORTION

Suppose a population is divided into two groups. The observations in the first group are called 'successes' and the observations of the second group are called 'failures'. For example the people may be divided into literates and illiterates. The proportion of successes in the population is defined as

$$\frac{\text{number of successes}}{\text{Total number of observations in the population}}$$

This proportion is denoted by p . The proportion of 'failures' is denoted by q and $q = 1 - p$ or $q + p = 1$. Let us see how $q + p = 1$. Let N denote the total number of observations in the population. We have,

$N = \text{number of successes} + \text{number of failures}$. Divide both sides by N

$$\begin{aligned}\frac{N}{N} &= \frac{\text{number of successes} + \text{number of failures}}{N} \\ &= \frac{\text{number of successes}}{N} + \frac{\text{number of failures}}{N}\end{aligned}$$

$$1 = p + q$$

Suppose a random sample of size n is selected from the population. Let there be X successes in the sample. The ratio X/n is the sample proportion and is denoted by \hat{p} . Thus $\hat{p} = X/n$, where \hat{p} is random variable and X is also random variable.

POINT ESTIMATE

The sample proportion \hat{p} calculated from a sample is the point estimate of the population proportion p . The statistic \hat{p} is unbiased estimator of p . Hence $E(\hat{p}) = p$.

12.12 CONFIDENCE INTERVAL ESTIMATE FOR POPULATION PROPORTION p (LARGE SAMPLE)

Suppose a population proportion is p which is unknown. A random sample of size n ($n > 30$) is selected from the population and sample proportion \hat{p} is calculated. The statistic \hat{p} is the estimator of p . The distribution of \hat{p} is normal with mean $\mu_{\hat{p}} = p$ and standard error $\sqrt{\frac{pq}{n}}$. Thus the random variable \hat{p} can be transformed

into random variable Z , where $Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$. When n is large, the terms p and q in

the denominator can be replaced by their sample estimates \hat{p} and \hat{q} . Thus

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}}$$

We take two points on Z - scale. These are $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$. The area of the normal curve between $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$ is $(1 - \alpha)$. The random variable Z will fall between $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$ with a probability of $(1 - \alpha)$. This statement can be expressed in the following form:

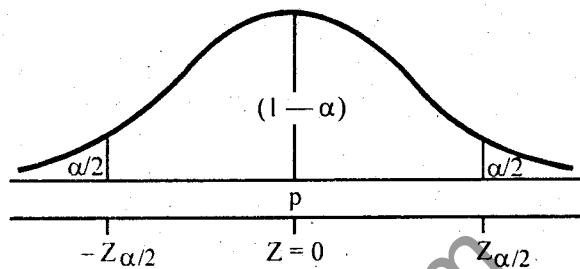


Figure 12.6

$$P\left[-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right] = 1 - \alpha \text{ or } P\left[-Z_{\frac{\alpha}{2}} < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} < Z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

The terms within the brackets can be written as

$$P\left[\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right] = 1 - \alpha$$

Thus $100(1 - \alpha)\%$ confidence interval estimate for p is

$$\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \text{ and } \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

For 95 % confidence interval we have $\alpha = 0.05$, $\alpha/2 = 0.025$ and $Z_{0.025} = 1.96$,

$-Z_{0.025} = -1.96$. Thus 95 % confidence interval for p is $\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$ and

$\hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$. For most probable confidence limits we take $Z = 3$.

Example 12.10.

A random sample of 200 persons from a city was interviewed and 50 of them were found to be literate. Calculate a 90 % confidence interval for the proportion of literate persons in the city. Also calculate a confidence interval for the proportion of illiterate persons in the city.

Solution:

A $100(1 - \alpha)\%$ confidence interval for p (literate persons) is

$$\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Here $n = 200$, $X = 50$ (number of literate persons), $\hat{p} = \frac{X}{n} = \frac{50}{200} = 0.25$,

$$\hat{q} = 1 - \hat{p} = 0.75, 1 - \alpha = 0.90 \text{ or } \alpha = 0.10 \text{ and } \alpha/2 = 0.05$$

From the area table of normal distribution, we have $Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$

Hence the 90 % confidence interval for p (literate persons) is

$$0.25 - 1.645 \sqrt{\frac{(0.25)(0.75)}{200}} < p < 0.25 + 1.645 \sqrt{\frac{(0.25)(0.75)}{200}}$$

$$0.25 - 0.05 < p < 0.25 + 0.05$$

$$0.2 < p < 0.3$$

Also

A $100(1 - \alpha)\%$ confidence interval for p (illiterate persons) is

$$\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Here $n = 200$, $X = 150$ (number of illiterate persons)

$$\hat{p} = \frac{X}{n} = \frac{150}{200} = 0.75, \hat{q} = 1 - \hat{p} = 0.25$$

Hence the 90 % confidence interval for p (illiterate persons) is

$$0.75 - 1.645 \sqrt{\frac{(0.75)(0.25)}{200}} < p < 0.75 + 1.645 \sqrt{\frac{(0.75)(0.25)}{200}}$$

$$0.75 - 0.05 < p < 0.75 + 0.05$$

$$0.7 < p < 0.8$$

12.13 CONFIDENCE INTERVAL ESTIMATE FOR THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS (LARGE SAMPLES)

Suppose there are two populations having proportions p_1 and p_2 which are unknown. It is required to calculate the confidence interval for the difference $(p_1 - p_2)$. Two independent random samples of size n_1 and n_2 are selected from the populations and the sample proportions are calculated which are \hat{p}_1 and \hat{p}_2 respectively. The statistic $(\hat{p}_1 - \hat{p}_2)$ is estimator of the parameter $(p_1 - p_2)$. When n_1 and n_2 are large, the random variable $(\hat{p}_1 - \hat{p}_2)$ has the normal distribution with

mean $(p_1 - p_2)$ and standard error $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$. The standard normal random

variable Z is written as $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$. The probability is $(1 - \alpha)$ that the

random variable Z will take on a value between $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$. This statement can be written as below:

$$P\left[-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right] = 1 - \alpha \text{ or } P\left[-Z_{\frac{\alpha}{2}} < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} < Z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

The terms within the brackets can be written as:

$$P\left[\left(\hat{p}_1 - \hat{p}_2\right) - Z_{\frac{\alpha}{2}} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} < p_1 - p_2 < \left(\hat{p}_1 - \hat{p}_2\right) + Z_{\frac{\alpha}{2}} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right] = 1 - \alpha$$

Thus $100(1 - \alpha)\%$ confidence interval estimate for $(p_1 - p_2)$ is

$$\left(\hat{p}_1 - \hat{p}_2\right) - Z_{\frac{\alpha}{2}} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} < p_1 - p_2 < \left(\hat{p}_1 - \hat{p}_2\right) + Z_{\frac{\alpha}{2}} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

But the terms p_1 , q_1 , p_2 and q_2 are for the populations and are unknown. For large sample sizes, they can be estimated by their sample estimates which are \hat{p}_1 , \hat{q}_1 , \hat{p}_2 and \hat{q}_2 respectively. Thus the confidence limits for $(p_1 - p_2)$ are:

$$\text{Lower Limit} = \left(\hat{p}_1 - \hat{p}_2\right) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$\text{Upper Limit} = \left(\hat{p}_1 - \hat{p}_2\right) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Example 12.11.

Consider two pain relieving drugs compared on two independent samples of 1000 individuals each. Suppose 750 of those individuals receiving drug I and 800 of those receiving drug II reported some pain relief. Construct a 90 % confidence interval for the difference between population proportions.

Solution:

A $100(1 - \alpha)\%$ confidence interval for $p_1 - p_2$ is

$$\left(\hat{p}_1 - \hat{p}_2\right) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < \left(\hat{p}_1 - \hat{p}_2\right) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$\text{Hence } n_1 = 1000, \quad X_1 = 750, \quad \hat{p}_1 = \frac{X_1}{n_1} = \frac{750}{1000} = 0.75, \quad \hat{q}_1 = 1 - \hat{p}_1 = 0.25,$$

$$n_2 = 1000, \quad X_2 = 800, \quad \hat{p}_2 = \frac{X_2}{n_2} = \frac{800}{1000} = 0.80, \quad \hat{q}_2 = 1 - \hat{p}_2 = 0.20,$$

$$1 - \alpha = 0.90 \text{ or } \alpha = 0.10 \text{ and } \alpha/2 = 0.05$$

From the area table of normal distribution, we have $Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$

Hence the 90 % confidence interval for $p_1 - p_2$ is

$$(0.75 - 0.8) - 1.645 \sqrt{\frac{(0.75)(0.25)}{1000} + \frac{(0.8)(0.2)}{1000}} < p_1 - p_2 < (0.75 - 0.8) + 1.645 \sqrt{\frac{(0.75)(0.25)}{1000} + \frac{(0.8)(0.2)}{1000}}$$

$$- 0.05 - 0.03 < p_1 - p_2 < - 0.05 + 0.03$$

$$- 0.08 < p_1 - p_2 < - 0.02$$

SHORT DEFINITIONS**Statistical Inference**

The process by which decision makers reach conclusions about a population based on sample information collected from the population.

or

A statistical inference is a decision, estimate, prediction, or generalization about the population based on information contained in a sample.

Estimation

Estimation is the process by which we attempt to determine the value of a population parameter from sample information.

or

Estimation is a process by which we get information about unknown value of population parameter by using sample values.

Estimate

An estimate is the numerical value calculated from sample data.

or

An estimate is the numerical value of the estimator.

Estimator

An estimator is a rule that tells how to calculate an estimate based on the measurements contained in a sample.

or

An estimator is a statistic that specifies how to use the sample data to estimate an unknown parameter of the population.

Point Estimate

A point estimate is a number representing an estimate of the population parameter based on a sample.

or

A point estimate consists of a single sample statistic that is used to estimate the true population parameter.

Interval Estimate

An interval estimate is the range of values within which the value of the parameter is expected to lie.

or

An estimate expressed by a range of values within which the true value of the population parameter is believed to lie, is referred to as an interval estimate.

Error of Estimation

The distance between an estimate and the estimated parameter is called the error of estimation.

Unbiased Estimator

An estimator is unbiased if its expected value is equal to the population parameter being estimated.

or

An estimator of a population parameter is said to be unbiased if the mean of its sampling distribution is equal to the parameter.

Biased Estimator

If the mean of the estimator is not equal to the population parameter, the estimator is said to be biased.

or

An estimator $\hat{\theta}$ is said to be biased if the expected value of the estimator is not equal to the population parameter being estimated i.e; $E(\hat{\theta}) \neq \theta$.

Confidence Interval

A confidence interval is a range of values within which the population parameter is expected to occur.

or

An interval that estimates a population parameter within a range of possible values with a specified probability.

Confidence Limits

The two endpoints of a confidence interval are called confidence limits.

Level of Confidence

The probability that the population parameter is included within the confidence interval is called the level of confidence.

or

The probability of correctly accepting the null hypothesis $(1 - \alpha)$, is called the level of confidence.

Degrees of Freedom

Degrees of freedom is the number of values that are free to vary after we have placed certain restrictions upon the data.

MULTIPLE - CHOICE QUESTIONS

1. The process of making estimates about the population parameter from a sample is called:
 - (a) statistical independence
 - (b) statistical inference
 - (c) statistical hypothesis
 - (d) statistical decision
2. Statistical inference has two branches namely:
 - (a) level of confidence and degrees of freedom
 - (b) biased estimator and unbiased estimator
 - (c) point estimate and interval estimate
 - (d) estimation of parameter and testing of hypothesis

3. Estimation is of two types:
- one sided and two sided
 - type I and type II
 - point estimation and interval estimation
 - biased and unbiased
4. A formula or rule used for estimating the parameter is called:
- estimation
 - estimate
 - estimator
 - interval estimate
5. Statistic is an estimator and its calculated value is called:
- biased estimate
 - estimation
 - interval estimate
 - estimate
6. Estimate is the observed value of an:
- unbiased estimator
 - estimator
 - estimation
 - interval estimation
7. The process of using sample data to estimate the values of unknown population parameters is called:
- estimate
 - estimator
 - estimation
 - interval estimate
8. The numerical value which we determine from the sample for population parameter is called:
- estimation
 - estimate
 - estimator
 - confidence coefficient
9. A single value used to estimate a population value is called:
- interval estimate
 - point estimate
 - confidence interval
 - level of confidence
10. An interval calculated from the sample data and it is likely to contain the value of parameter with some probability is called:
- interval estimate
 - point estimate
 - level of confidence
 - degrees of freedom
11. A range of values within which the population parameter is expected to occur is called:
- confidence coefficient
 - confidence interval
 - confidence limits
 - level of significance
12. The end points of a confidence interval are called:
- confidence coefficient
 - confidence limits
 - error of estimation
 - parameters

13. The probability associated with confidence interval is called:
- (a) level of confidence
 - (b) confidence coefficient
 - (c) both (a) and (b)
 - (d) confidence limits
14. If the mean of the estimator is not equal to the population parameter, the estimator is said to be:
- (a) unbiased
 - (b) biased
 - (c) positively biased
 - (d) negatively biased
15. If $\hat{\theta}$ is the estimator of the parameter θ , then $\hat{\theta}$ is called unbiased if:
- (a) $E(\hat{\theta}) > \theta$
 - (b) $E(\hat{\theta}) < \theta$
 - (c) $E(\hat{\theta}) \neq \theta$
 - (d) $E(\hat{\theta}) = \theta$
16. Estimates given in the form of confidence intervals are called:
- (a) point estimates
 - (b) interval estimates
 - (c) confidence limits
 - (d) degrees of freedom
17. $(1 - \alpha)$ is called:
- (a) critical value
 - (b) level of significance
 - (c) level of confidence
 - (d) interval estimate
18. If $(1 - \alpha)$ is increased, the width of a confidence interval is:
- (a) decreased
 - (b) increased
 - (c) constant
 - (d) same
19. By decreasing the sample size, the confidence interval becomes:
- (a) narrower
 - (b) wider
 - (c) fixed
 - (d) all of the above
20. Confidence interval becomes narrow by increasing the:
- (a) sample size
 - (b) population size
 - (c) level of confidence
 - (d) degrees of freedom
21. By increasing the sample size, the precision of confidence interval is:
- (a) increased
 - (b) decreased
 - (c) same
 - (d) unchanged
22. The distance between an estimate and the estimated parameter is called:
- (a) sampling error
 - (b) error of estimation
 - (c) bias
 - (d) standard error
23. The number of values that are free to vary after we have placed certain restrictions upon the data is called:
- (a) degrees of freedom
 - (b) confidence coefficient
 - (c) number of parameters
 - (d) number of samples

24. A 95 % confidence interval for the mean of a population is such that:
- It contains 95 % of the values in the population
 - There is a 95 % chance that it contains all the values in the population
 - There is a 95 % chance that it contains the mean of the population
 - There is a 95 % chance that it contains the standard deviation of the population.
25. A confidence interval will be widened if:
- The confidence level is increased and the sample size is reduced
 - The confidence level is increased and the sample size is increased
 - The confidence level is decreased and the sample size is increased
 - The confidence level is decreased and the sample size is decreased.
26. A statistician calculates a 95 % confidence interval for μ when σ is known. The confidence interval is Rs. 18000 to Rs. 22000, the amount of the sample mean \bar{X} is:
- | | |
|---------------|---------------|
| (a) Rs. 18000 | (b) Rs. 20000 |
| (c) Rs. 22000 | (d) Rs. 40000 |
27. If the population standard deviation σ is known, the confidence interval for the population mean μ is based on:
- | | |
|--------------------------------|-----------------------------|
| (a) the poisson distribution | (b) the t-distribution |
| (c) the χ^2 -distribution | (d) the normal distribution |
28. If the population standard deviation σ is unknown, and the sample size is small i.e; $n \leq 30$, the confidence interval for the population mean μ is based on:
- | | |
|-------------------------------|-------------------------------------|
| (a) the t-distribution | (b) the normal distribution |
| (c) the binomial distribution | (d) the hypergeometric distribution |
29. The shape of the t-distribution depends upon the:
- | | |
|-----------------|------------------------|
| (a) sample size | (b) population size |
| (c) parameters | (d) degrees of freedom |
30. If the population standard deviation σ is doubles, the width of the confidence interval for the population mean μ (i.e; the upper limit of the confidence interval – lower limit of the confidence interval) will be:
- | | |
|------------------|------------------------------|
| (a) divided by 2 | (b) multiplied by $\sqrt{2}$ |
| (c) doubled | (d) decrease |
31. The following statistics are unbiased estimators:
- | | |
|---------------------------|---|
| (a) the sample mean | (b) the sample variance $s^2 = \frac{\sum(X - \bar{X})^2}{n - 1}$ |
| (c) the sample proportion | (d) all the above |

32. A statistic is an unbiased estimator of a parameter if:
- $E(\text{statistic}) = \text{parameter}$
 - $E(\text{mean}) = \text{variance}$
 - $E(\text{variance}) = \text{mean}$
 - $E(\text{sample mean}) = \text{proportion}$
33. Which of the following is biased estimator?
- $\bar{X} = \frac{\sum X}{n}$
 - $\hat{p} = \frac{X}{n}$
 - $s^2 = \frac{\sum (X - \bar{X})^2}{n-1}$
 - $S^2 = \frac{\sum (X - \bar{X})^2}{n}$
34. If the observations are paired and the number of pairs is n , then degrees of freedom is equal to:
- n
 - $n-1$
 - $n_1 + n_2 - 2$
 - $n/2$
35. If $\alpha = 0.10$ and $n = 15$; $t_{\frac{\alpha}{2}}$ equals:
- 1.761
 - 1.753
 - 1.771
 - 2.145
36. If $n_1 = 16$, $n_2 = 9$ and $\alpha = 0.01$; $t_{\frac{\alpha}{2}}$ equals:
- 2.787
 - 2.807
 - 2.797
 - 3.767
37. In t-distribution for two independent samples $n_1 = n_2 = n$, then the degrees of freedom is equal to:
- $2n - 1$
 - $2n - 2$
 - $2n + 1$
 - $n - 1$
38. If $1 - \alpha = 0.90$, then value of $Z_{\frac{\alpha}{2}}$ is:
- 1.96
 - 2.575
 - 1.645
 - 2.326
39. If the population standard deviation σ is known and the sample size n is less than or equal to or more than 30, the confidence interval for the population mean μ is:
- $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
 - $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$
 - $\bar{X} \pm \frac{t_{\frac{\alpha}{2}}}{2} \frac{s}{\sqrt{n}}$
 - $\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$
40. If the population standard deviation σ is unknown and the sample size n is greater than 30, the confidence interval for the population mean μ is:
- $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
 - $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$
 - $\bar{X} \pm \frac{t_{\frac{\alpha}{2}}}{2} \frac{s}{\sqrt{n}}$
 - $\bar{X} \pm \frac{t_{\frac{\alpha}{2}}}{2} \frac{s_d}{\sqrt{n}}$

41. If the population standard deviation σ is unknown and the sample size n is less than or equal to 30, the confidence interval for the population mean μ is:
- (a) $\bar{X} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}$ (b) $\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$
 (c) $\bar{X} \pm t_{\alpha/2} \frac{sd}{\sqrt{n}}$ (d) $b \pm t_{\alpha/2} \frac{s}{\sqrt{n-2}}$
42. A student calculates a 90 % confidence interval for population mean μ when population standard deviation σ is unknown and $n = 9$. The confidence interval is -- 24.3 cents to 64.3 cents, the sample mean \bar{X} is:
- (a) 40 (b) - 24.3
 (c) 64.3 (d) 20
43. A 95 % confidence interval for population proportion p is 32.4 % to 47.6 %, the value of sample proportion \hat{p} is:
- (a) 40 % (b) 32.4 %
 (c) 47.6 % (d) 80 %
44. If we have normal populations with known population standard deviations σ_1 and σ_2 , the confidence interval estimate for the difference between two population means ($\mu_1 - \mu_2$) is:
- (a) $(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ (b) $(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
 (c) $(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n_1 + n_2}}$ (d) $(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2 \sigma_2^2}{n_1 n_2}}$
45. If the population standard deviations σ_1 and σ_2 are unknown and sample sizes $n_1, n_2 > 30$, the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:
- (a) $(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{S_1^2 + S_2^2}{n_1 + n_2}}$ (b) $(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{S_1^2 S_2^2}{n_1 n_2}}$
 (c) $(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ (d) $(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
46. If the sample size is large, the confidence interval estimate of a population proportion p is:
- (a) $\bar{X} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}$ (b) $\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$
 (c) $\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p} + \hat{q}}{n}}$ (d) $\hat{p} \pm Z_{\alpha/2} \sqrt{n\hat{p}\hat{q}}$

47. If $n_1, n_2 \leq 30$, the confidence interval estimate for the difference of two population means ($\mu_1 - \mu_2$) when population standard deviations σ_1, σ_2 are unknown but equal in case of pooled variates is:

- (a) $(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ (b) $(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}(v)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
 (c) $(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}(v)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ (d) $(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n_1 + n_2}}$

48. The confidence interval estimate for the difference of two population means $\mu_1 - \mu_2 = \mu_d$ in case of paired observations small sample ($n \leq 30$) is:

- (a) $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ (b) $\bar{d} \pm t_{\frac{\alpha}{2}(n-1)} \frac{s_d}{\sqrt{n}}$
 (c) $\bar{d} \pm t_{\frac{\alpha}{2}(n-1)} \sqrt{\frac{s_1^2 + s_2^2}{n}}$ (d) $b \pm t_{\frac{\alpha}{2}(n-2)} s_b$

49. If the sample size is large, the confidence interval estimate for the difference between two population proportions $p_1 - p_2$ is:

- (a) $(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$ (b) $(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}{n_1 + n_2}}$
 (c) $(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}$ (d) $(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}{n_1 n_2}}$

Answers

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

SHORT QUESTIONS

1. Given $n = 64$, $\bar{X} = 42.7$, $\sigma = 8$ and $Z_{\alpha/2} = 1.645$. Find the confidence interval for μ .

Ans. $41.1 < \mu < 44.3$

2. Given $n = 16$, $\bar{X} = 52.5$, $\sigma = 10$ and $1 - \alpha = 0.90$. Compute the 90% confidence interval for μ .

Ans. $48.4 < \mu < 56.6$

3. Given $N = 500$, $n = 100$, $\bar{X} = 60$, $\sigma = 5$ and $Z_{0.025} = 1.96$. Find the 95% confidence interval for μ .

Ans. $59.1 < \mu < 60.9$

4. Given $n = 144$, $\bar{X} = 750$, $S = 6$ and $Z_{0.01} = 2.326$. Compute the 98% confidence interval for μ .

Ans. $748.837 < \mu < 751.163$

5. Given $n = 16$, $\bar{X} = 80$, $s = 3$ and $t_{0.01(15)} = 2.602$. Construct the 98% confidence interval for population mean μ .

Ans. $78.05 < \mu < 81.95$

6. Given $n_1 = 32$, $n_2 = 50$, $\bar{X}_1 = 125$, $\bar{X}_2 = 100$, $\sigma_1^2 = 16$, $\sigma_2^2 = 25$ and $Z_{0.005} = 2.575$.

Find the 99% confidence interval for the difference between population means $\mu_1 - \mu_2$.

Ans. $22.425 < \mu_1 - \mu_2 < 27.575$

7. Given $n_1 = 48$, $\bar{X}_1 = 90$, $S_1^2 = 12$, $n_2 = 72$, $\bar{X}_2 = 85$, $S_2^2 = 18$ and $Z_{0.025} = 1.96$. Find the 95% confidence interval for $\mu_1 - \mu_2$.

Ans. $3.6 < \mu_1 - \mu_2 < 6.4$

8. Given the following summary statistics for independent random samples from two populations:

$n_1 = 16$, $\bar{X}_1 = 60$, $s_1^2 = 36$, $n_2 = 9$, $\bar{X}_2 = 50$, $s_2^2 = 25$, $s_p = 5.67$ and $t_{0.05(23)} = 1.714$.

Find the 90% confidence interval for $\mu_1 - \mu_2$.

Ans. $5.95 < \mu_1 - \mu_2 < 14.05$

9. Given $\bar{d} = 3$, $n = 9$, $s_d = 3$ and $t_{0.025}(8) = 2.306$. Find the 95% confidence interval for $\mu_d = \mu_1 - \mu_2$.

Ans. $0.694 < \mu_d < 5.306$

10. Given $n = 500$, $\hat{p} = 0.30$ and $Z_{0.03} = 1.88$. Find the 94% confidence interval for the population proportion p .

Ans. $0.26 < p < 0.34$

11. Given $n_1 = 400$, $\hat{p}_1 = 0.8$, $n_2 = 300$, $\hat{p}_2 = 0.6$ and $Z_{0.04} = 1.75$. Find the 92% confidence interval for the difference between population proportions.

Ans. $0.14 < p_1 - p_2 < 0.26$

12. Determine the critical value of Z in each of the following circumstances:

- | | | |
|-------------------------|-------------------------|-------------------------|
| (a) $1 - \alpha = 0.90$ | (b) $1 - \alpha = 0.92$ | (c) $1 - \alpha = 0.94$ |
| (d) $1 - \alpha = 0.96$ | (e) $1 - \alpha = 0.98$ | (f) $1 - \alpha = 0.99$ |

Ans. (a) 1.645 (b) 1.75 (c) 1.88 (d) 2.054 (e) 2.326 (f) 2.575

13. State the formula which is used to calculate a 95 % confidence interval for the population mean μ , when the population standard deviation σ is known.

$$\text{Ans. } \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$

14. State the formula which is used to calculate a 90 % confidence interval for the population mean μ , when the population standard deviation σ is unknown and $n = 10$.

$$\text{Ans. } \bar{X} - 1.833 \frac{s}{\sqrt{10}} < \mu < \bar{X} + 1.833 \frac{s}{\sqrt{10}}$$

15. State the formula which is used to calculate a 92 % confidence interval for the population proportion p , when $n > 30$.

$$\text{Ans. } \hat{p} - 1.75 \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + 1.75 \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

16. Determine the critical value of 't' in each of the following circumstances:

- | | |
|------------------------------------|---|
| (a) $1 - \alpha = 0.95$, $n = 12$ | (b) $1 - \alpha = 0.98$, $n_1 = 10$, $n_2 = 12$ |
| (c) $1 - \alpha = 0.90$, $n = 16$ | (d) $1 - \alpha = 0.99$, $n_1 = 6$, $n_2 = 6$ |

Ans. (a) 2.201 (b) 2.528 (c) 1.753 (d) 3.169

17. If $\bar{X} = 100$, $\sigma = 8$ and $n = 64$, set up a 95 % confidence interval estimate of the population mean μ .

Ans. $98.04 < \mu < 101.96$

18. State the formula which is used to calculate a 98 % confidence interval for the difference between two population means $\mu_1 - \mu_2$, when population variances are known with any sample size.

$$\text{Ans. } (\bar{X}_1 - \bar{X}_2) - 2.326 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + 2.326 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

19. State the formula which is used to calculate a 95 % confidence interval for the mean of a population of paired differences for n = 9.

Ans. $\bar{d} - 2.306 \frac{s_d}{\sqrt{3}} < \mu_d < \bar{d} + 2.306 \frac{s_d}{\sqrt{3}}$

20. Distinguish between point estimate and interval estimate.
21. Explain the terms estimate and estimator.
22. What is meant by estimation?
23. Explain what is meant by confidence interval.
24. Explain what is meant by unbiased estimator.
25. What do you mean by unbiased estimator? Give at least two examples.
26. Define the terms point estimate and interval estimate.
27. Write all the confidence intervals for population mean with small and large samples, population standard deviation being known and unknown.
28. What do you know about statistical inference?
29. Differentiate between biased estimator and unbiased estimator.
30. What is meant by unbiasedness?
31. What is the procedure followed in the construction of confidence interval?
32. Why is a confidence interval estimate of a parameter is more useful than a point estimate?

EXERCISES

1. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 42 hours. If a random sample of 49 bulbs has an average life of 800 hours, find a 95 % confidence interval for the population mean of all bulbs produced by this firm.

Ans. $788.24 < \mu < 811.76$

2. Find a 90 % confidence interval for the mean of a normal distribution if $\sigma = 2$ and a sample of size 8 gave the values 9, 14, 10, 12, 7, 13, 11, 12.

Ans. $9.84 < \mu < 12.16$

3. An advertising agency wants to estimate the average income of families located in a particular area of a low-income section of Karachi. There are 1000 families in this area, and the agency chooses a random sample of 100. The mean income of these families is Rs. 1800 per month. Compute a 95 % confidence interval for the population mean, if the population standard deviation is known to be Rs. 200.

Ans. $1762.79 < \mu < 1837.21$

4. The population of scores of 10-year old children in a psychological performance test is known to have a standard deviation 5.2. If a random sample of size 20 shows a mean of 16.9, find a 95 % confidence interval for the mean score of the population, assuming that the population is normal.

Ans. $14.62 < \mu < 19.18$

5. A tire manufacturer wants to estimate the mean weight of the tires produced by one of its plants. He takes a random sample of 100 tires produced at this plant and finds that the sample mean is 48.1 lbs. and the sample standard deviation is 0.12 lbs. Calculate a 95 % confidence interval for the population mean.

Ans. $48.076 < \mu < 48.124$

6. Compute a 90 % confidence interval for the population mean, if $n = 36$, $\Sigma X = 5400$ and $\Sigma(X - \bar{X})^2 = 1296$.

Ans. $148.355 < \mu < 151.645$

7. The heights of a random sample of 64 college students showed a mean of 172 centimeters and a standard deviation of 6.5 centimeters. Find a 92 % confidence interval for the mean height of all college students.

Ans. $170.578 < \mu < 173.422$

8. The hourly wages of 144 workers of a large factory were recorded, and the sample mean and standard deviation were found to be Rs. 23.52 and Rs. 6.71 respectively. Find a 99 % confidence interval for the mean wages of factory workers.

Ans. $22.08 < \mu < 24.96$

9. A random sample of 9 cigarettes of a certain brand has an average nicotine content of 3.6 milligrams and a standard deviation of 0.9 milligrams. Construct a 95 % confidence interval for the true average nicotine content of this particular brand of cigarettes, assuming an approximate normal distribution.

Ans. $2.91 < \mu < 4.29$

10. A certain machine is used to produce items whose weights are assumed to be normally distributed. Suppose that the variability in the weight of the output of the machine is unknown. For a random sample of size 16, \bar{X} is found to be 282 grams and s is 8 grams. Find a 95 % confidence interval for the true population mean.

Ans. $277.738 < \mu < 286.262$

11. A firm wants to estimate the mean lifetime of a particular kind of tool. From the previous experience, it is known that the lifetimes are normally distributed. It draws a random sample of four of these tools and finds that their lifetimes are 7.9, 9.3, 10.8 and 11.4 years. Calculate a 95 % confidence interval for the mean lifetime of this kind of tool.

Ans. $7.35 < \mu < 12.35$

12. A random sample of size $n_1 = 36$ taken from a normal population with a variance $\sigma_1^2 = 9$, has a mean $\bar{X}_1 = 75$. A second random sample of size $n_2 = 25$, taken from a different normal population with a variance $\sigma_2^2 = 25$, has a mean $\bar{X}_2 = 70$. Find a 98 % confidence interval for $\mu_1 - \mu_2$.

Ans. $2.4 < \mu_1 - \mu_2 < 7.6$

13. Two independent samples of 100 machinists and 100 carpenters are taken to estimate the difference between the weekly wages of the two categories of workers. The relevant data are given below:

	Sample mean wage	Population variance
Machinists	345	196
Carpenters	340	204

Determine the 95 % and 99 % confidence limits for the true difference between the average wages for machinists and carpenters. Which interval is wider?

Ans. 95 % confidence interval: $1.08 < \mu_1 - \mu_2 < 8.92$

99 % confidence interval: $-0.15 < \mu_1 - \mu_2 < 10.15$. Therefore 99% interval is wider.

14. A standardized statistics test was given to 75 boys and 50 girls. The boys made an average grade of 82 with a variance of 64, while the girls made an average grade of 76 with a variance of 36. Find a 99% confidence interval for $\mu_1 - \mu_2$, where μ_1 stands for mean score of all boys and μ_2 stands for mean score of all girls.

Ans. $2.77 < \mu_1 - \mu_2 < 9.23$

15. Two independent random samples of the diameters of tires are drawn, one from a batch of tires produced at plant A, and another from a batch of tires produced at plant B. The results are as follows:

Sample	Sample size	Sample mean (inches)	Sample variance (inches) ²
Plant A	100	50.7	0.09
Plant B	100	50.3	0.04

Calculate a 95 % confidence interval for the difference between the mean diameter of the entire batch produced at plant A and the mean diameter of the entire batch produced at plant B.

Ans. $0.33 < \mu_1 - \mu_2 < 0.47$

16. Suppose that for a random sample of size 8 from population I the sample mean and standard deviation were 14.9 and 4.17, respectively while a random sample of size 5 from population II yielded a sample mean of 10.6 and a sample standard deviation of 3.62 respectively. Assuming that the populations are normally distributed with equal variances. Compute a 90 % confidence interval for the difference between the population means.

Ans. $0.22 < \mu_1 - \mu_2 < 8.38$

17. The following summary statistics are recorded for independent random samples from two populations:

Sample I	$n_1 = 9$	$\bar{X}_1 = 16.18$	$s_1 = 1.54$
Sample II	$n_2 = 6$	$\bar{X}_2 = 4.22$	$s_2 = 1.37$

Assume that populations are normal with the identical standard deviations. Calculate a 95 % confidence interval for $\mu_1 - \mu_2$.

Ans. $10.28 < \mu_1 - \mu_2 < 13.64$

18. It is claimed that a new diet will reduce a person's weight by 5 kilograms on the average in a period of 8 weeks. The weights of 4 men who were given this diet were recorded before and after a 8 week period:

Men	1	2	3	4
Weights before	63	64	66	73
Weights after	64	58	62	66

Compute a 90 % confidence interval for the mean difference in the weights. Assume the distribution of weights to be approximately normal.

Ans. $-0.19 < \mu_d < 8.19$

19. To compare two treatments, a matched-pair experiment was conducted with 12 pairs of subjects, and the paired differences of the response to treatment B from the response to treatment A were recorded: 2, 5, 6, 8, - 6, 4, 18, - 12, 17, - 7, 16, 12. Construct a 95 % confidence interval for the mean difference of the responses to the two treatments.

Ans. $-0.98 < \mu_d < 11.48$

20. A random sample of 300 cigarette smokers is selected and 75 are found to have a preference for Gold leaf. Find a 98 % confidence interval for the fraction of the population of cigarette smokers who prefer Gold leaf.

Ans. $0.19 < p < 0.31$

21. A tire manufacturer draws a random sample of 160 tires produced by a new process and finds that 40 percent wear better than required specifications. Construct a 90 % confidence interval for the population proportion of the tires produced by the new process that will wear better than required specifications.

Ans. $0.34 < p < 0.46$

22. Find a 95 % confidence interval for 'p' if 24 heads are obtained in 40 tosses of a coin.

Ans. $0.45 < p < 0.75$

23. In a random sample of 1000 homes in a certain city, it is found that 228 are heated by oil. Find a 99 % confidence interval for the proportion of homes in this city that are heated by oil.

Ans. $0.194 < p < 0.262$

24. A firm has factories in Karachi and Lahore. It picks a random sample of 100 workers from each factory. In Karachi, 32 percent say that they buy the firm's product; in Lahore, 27 percent. Construct a 95 % confidence interval for the difference between the proportion of workers in Karachi and Lahore who say they buy the firm's product.

Ans. $-0.08 < p_1 - p_2 < 0.18$

25. A poll is taken among the residents of a city and the surrounding semi-urban area to determine the feasibility of a proposal to construct a civic center. If 240 of 500 city residents favour the proposal and 120 of 200 semi-urbans favour it, find a 95 % confidence interval for the true difference in the fractions favouring the proposal to construct the civic centre.

Ans. $-0.20 < p_1 - p_2 < -0.04$

Econ H12 10

Chapter

13

STATISTICAL INFERENCE TESTING OF HYPOTHESES

13.1 INTRODUCTION

Statistical inference consists of estimation of parameters and testing of hypotheses. Estimation has already been discussed in the previous chapter and in this chapter our lesson is about the testing of hypotheses. Point estimation and interval estimation as discussed earlier have their own fields of application. Sometimes there is a situation in which the point estimation and the interval estimation are either not required or the estimation of parameters does not provide any inference. For example, the following situations require inference which is not possible by methods of estimation.

- (i) The contents of a medicine have been changed to improve the effectiveness of the medicine. In this situation both the point estimation and the interval estimation fail to answer the question about the improvement of the medicine. In this case we have to take help from the sample data to decide whether or not the medicine has been improved.
- (ii) A manufacturer of tires claims that the average life of his tires is at least 15000 kilometers. The life of tires is an important factor to settle the price of the tires. It is a big information if we prove with reasonable amount of confidence that the life of the tires is not more than 15000 kilometers. The answer is not provided by a point estimate or by an interval estimate of the life of the tires. What we shall have to do is that we shall examine the claim of the manufacturer on the basis of the experiment conducted on the sample of tires. A certain procedure will be adopted to reach some conclusion. This is what we shall call the test of hypothesis about the life of tires.

13.2 STATISTICAL HYPOTHESES

Any opinion or idea may be formed about the population under study. Consider the following statements: Average consumption of sugar per month for a consumer is 1 kg; Intelligent parents have intelligent children, tall fathers have tall sons, average life of the people of Pakistan is higher than that of India, proper greasing increases the life of ceiling fans, use of coffee increases chances of heart attack, one variety of seed is better than the other, a medicine of allergy gives relief to at least 80 % of the people, more than 25 % people are literate in our country, only 60 % people will go to the polling stations for voting. These statements are the questions

in different fields of life and these questions are to be answered after proper experimentation. These questions have come up in the process of investigations. This is how the hypotheses are generated during various studies. When an assumption is explained in the form of a statement about the distribution of a population or populations, it is called a statistical hypothesis. In simple words, a statistical hypothesis is a statement about the unknown value of the population parameter. The statement may be true or false.

13.2.1 NULL HYPOTHESIS

The hypothesis which is to be tested is called *null hypothesis*. It is denoted by H_0 . It is a starting point in the investigations. A statement which we hope will be rejected is taken as a hypothesis. Modern approach is different. Today any hypothesis we wish to test is called null hypothesis and is denoted by H_0 . In this book we shall follow the old convention. Any hypothesis will be called null hypothesis only when we hope to reject it. Thus the null hypothesis is framed for possible rejection. Tall fathers have tall children. We shall assume that tall fathers do not have tall children. This will be considered as null hypothesis and will be denoted by H_0 . We are hoping that H_0 will be rejected on the basis of sample data. Use of coffee increases chances of heart attack. To start with we shall assume that heart attack has no link with the use of coffee. This will be taken as H_0 and we hope it will be rejected by the sample data.

13.2.2 ALTERNATIVE HYPOTHESIS

The hypothesis which is accepted when the null hypothesis has been rejected is called the alternative hypothesis. It is denoted by H_1 or H_A . Whatever we are expecting from the sample data is taken as the alternate hypothesis. "More than 25% people are literate in our country". We are hoping to get this result from the sample. It will be taken as an alternate hypothesis H_1 and null hypothesis H_0 will be that 25 % or less than that are literate. To be more specific, H_0 will be 25 % or less are literate and H_1 will be more than 25 % are literate. It is written as:

$$H_0: p \leq 0.25 \quad (\text{25 \% or less}) \quad H_1: p > 0.25 \quad (\text{more than 25 \%})$$

To keep the things simple, we can write H_0 in the form of equality as $H_0: p = 0.25$ but it is important to write H_1 with proper direction of inequality. Thus we write $H_1: p > 0.25$.

In this case the H_1 contains the inequality *more than* ($>$). We shall explain later that H_1 may be written with inequality *less than* ($<$) or *not equal* (\neq). In general, if the hypothesis about the population parameter θ is θ_0 , then H_1 can be written in three different ways.

$$\text{For } H_0: \theta = \theta_0, \quad H_1: \theta \neq \theta_0 \quad H_1: \theta > \theta_0 \quad H_1: \theta < \theta_0$$

But this is the simple approach which is allowed for the students. Another way of writing the above hypotheses H_0 and H_1 is

$$(a) H_0: \theta = \theta_0, H_1: \theta \neq \theta_0 \quad (b) H_0: \theta \leq \theta_0, H_1: \theta > \theta_0 \quad (c) H_0: \theta \geq \theta_0, H_1: \theta < \theta_0$$

The alternative hypothesis H_1 never contains the sign of equality. Thus H_1 will not contain '=' or ' \geq ' signs. The equality sign '=' and inequalities like ' \leq ' and ' \geq ' are used for writing H_0 .

13.2.3 SIMPLE HYPOTHESIS

If a hypothesis has a single value for the population parameter, it is called *simple hypothesis*. The breaking strength of copper wire is 10 kg. Here $H_0: \mu = 10$ kg has a single specified value. H_0 is simple hypothesis, similarly $\mu_1 - \mu_2 = 10$ and $p = 0.6$ are simple hypotheses.

13.2.4 COMPOSITE HYPOTHESIS

The hypothesis is called *composite* if it specifies a *range* of values for the parameter. The hypothesis $\mu \geq 10$ is a composite hypothesis. Similarly the hypotheses $(\mu_1 - \mu_2) \geq 10$ and $p \leq 0.6$ are composite.

13.2.5 ACCEPTANCE AND REJECTION OF NULL HYPOTHESIS

The given hypothesis is tested with the help of the sample data. A simple random sample has the full freedom of giving any value to its statistic. The sample is not aware of our plans. We decide about our hypothesis on the basis of the sample statistic. If the sample does not support the null hypothesis, we reject it on probability basis and accept the alternative hypothesis. If the sample does not oppose the hypothesis, the hypothesis is accepted. But here '*accept*' does not mean the *acceptance* of null hypothesis but only means that the sample has not strongly opposed it. "*Not opposed*" does not mean that the sample has strongly supported the hypothesis. The support of the sample in favour of the hypothesis cannot be established. When the hypothesis is *rejected*, it is rejected with a high probability. Thus *rejection* of H_0 is a strong decision and it leads us to the acceptance of H_1 . But acceptance of H_1 is not like the acceptance of H_0 . The *acceptance* of null hypothesis does not give us a certain strong decision. It is a situation which may require some further investigations. At this stage, many factors are to be taken into account. The sample size and certain other things not yet discussed help us to do something more about the null hypothesis before it is finally accepted. Thus *rejection* is a decision but not necessarily true and *acceptance* is not a decision in any sense of the word.

There is a modern approach in which the terms *rejection* and *acceptance* are not used. This modern approach is beyond the level of this book. But it remains true in its place that *acceptance* of a null hypothesis is a weak decision whereas *rejection* is a strong evidence of the sample against the null hypothesis. When the null hypothesis is rejected, it means the sample has done some statistical work but when the null hypothesis is accepted, it means the sample is almost silent. This behaviour of the sample should not be used in favour of the null hypothesis.

13.2.6 TEST STATISTIC

A statistic is calculated from the sample. To begin with we assume that the hypothesis about the population parameter is true. We compare the value of the statistic with the hypothetical value of the parameter. If the difference between

them is small, the hypothesis is accepted and if the difference between them is large, the hypothesis is rejected. A statistic on which the decision can be based whether to accept or reject a hypothesis is called *test statistic*. Some of the test statistics to be discussed in this book are 'Z', 't' and χ^2 (Chi-square)

13.2.7 ACCEPTANCE AND REJECTION REGIONS

The values of the test statistic which we think do not agree with the given hypothesis are called the critical region or rejection region. The values of the test statistic which support the hypothesis form the acceptance region. The rejection region is equal to α and the acceptance region is denoted by $(1 - \alpha)$. These two regions are separate from each other and both regions combined together make the complete sampling distribution of the statistic. These regions are separated by a value (or values), which is called critical value (or values).

13.2.8 TWO-TAILED TEST

When the rejection region is taken on both ends of the sampling distribution, the test is called *two-sided test* or *two-tailed test*. When we are using a two-sided test, half of the rejection region equal to $\alpha/2$ is taken on the right side and the other half equal to $\alpha/2$ is taken on the left side of the sampling distribution. Suppose the sampling distribution of the statistic is a normal distribution and we have to test the hypothesis $H_0: \theta = \theta_0$ against the alternative hypothesis $H_1: \theta \neq \theta_0$ which is two-sided. H_0 is rejected when the calculated value of Z is greater than $Z_{\alpha/2}$ or it is less than $-Z_{\alpha/2}$. Thus the critical region is $Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$, it can also be written as $-Z_{\alpha/2} < Z < Z_{\alpha/2}$.

When H_0 is rejected, then H_1 is accepted. Two-sided test is shown in Fig.13.1.

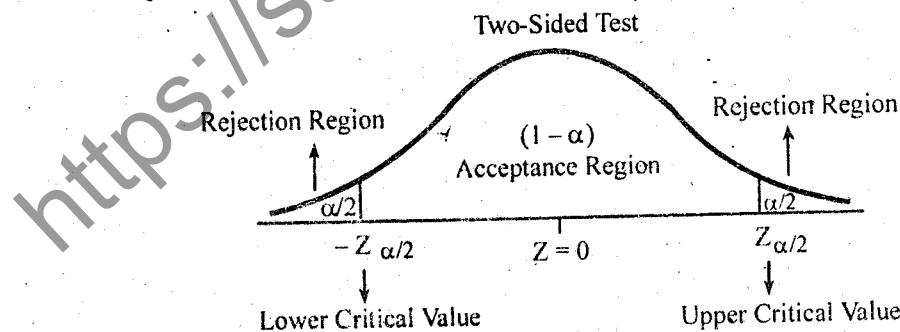


Figure 13.1

13.2.9 ONE-TAILED TEST

When the alternative hypothesis H_1 is one-sided like $\theta > \theta_0$ or $\theta < \theta_0$, then the rejection region is taken only on one side of the sampling distribution. It is called *one-tailed test* or *one-sided test*. When H_1 is one-sided to the right like $\theta > \theta_0$, the entire rejection region equal to α is taken in the right end of the sampling distribution.

The test is called *one-sided* to the right. The hypothesis H_0 is rejected if the calculated value of a statistic, say Z falls in the rejection region. The critical value is Z_α which has the area equal to α to its right. The rejection region and acceptance region are shown in Fig.13.2. The null hypothesis H_0 is rejected when $Z(\text{calculated}) > Z_\alpha$.

If the alternative hypothesis is one-sided to the left like $\theta < \theta_0$, the entire rejection region equal to α is taken on the left tail of the sampling distribution. The test is called one-sided or one-tailed to the left. The critical value is $-Z_\alpha$ which cuts off the area equal to α to its left. The critical region is $Z < -Z_\alpha$ and is shown in Fig.13.3.

For some important values of α , the critical values of Z for two-tailed and one tailed tests are given below:

Critical values of Z

α	Two-sided test	One-sided to the right	One-sided to the left
0.10 (10 %)	-1.645 and +1.645	+1.282	-1.282
0.05 (5 %)	-1.96 and +1.96	+1.645	-1.645
0.02 (2 %)	-2.326 and +2.326	+2.054	-2.054
0.01 (1 %)	-2.575 and +2.575	+2.326	-2.326

13.3 ERRORS IN TESTING OF HYPOTHESIS

The null hypothesis H_0 is accepted or rejected on the basis of the value of the test-statistic which is a function of the sample. The test statistic may land in acceptance region or rejection region. If the calculated value of test-statistic, say Z , is small (insignificant) i.e., Z is close to zero or we can say Z lies between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ is a two-sided alternative test ($H_1: \theta \neq \theta_0$), the hypothesis is accepted. If the calculated value of the test-statistic Z is large (significant), H_0 is rejected and H_1 is accepted. In this rejection plan or acceptance plan, there is the possibility of making any one of the two errors which are called Type I and Type II-errors.

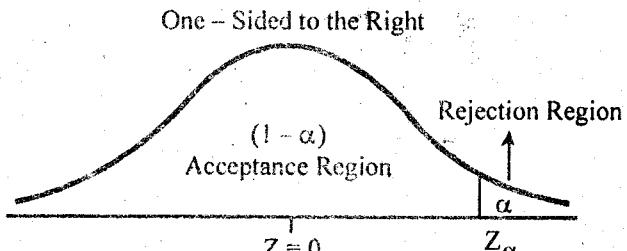


Figure 13.2

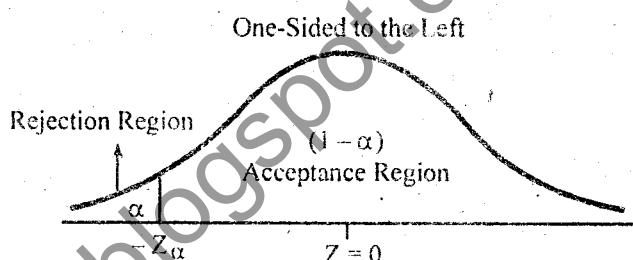


Figure 13.3

13.3.1 TYPE I ERROR

The null hypothesis H_0 may be true but it may be rejected. This is an error and is called *Type I error*. When H_0 is true, the test-statistic, say Z , can take any value between $-\infty$ to $+\infty$. But we reject H_0 when Z lies in the rejection region while the rejection region is also included in the interval $-\infty$ to ∞ . In a two-sided H_1 (like $\theta \neq \theta_0$), the hypothesis is rejected when Z is less than $-Z_{\alpha/2}$ or Z is greater than $Z_{\alpha/2}$. When H_0 is true, Z can fall in the rejection region with a probability equal to the rejection region α . Thus it is possible that H_0 is rejected while H_0 is true. This is called *Type I error*. The probability is $(1 - \alpha)$ that H_0 is accepted when H_0 is true. It is called correct decision. We can say that *Type I error* has been committed when:

- (i) an intelligent student is not promoted to the next class.
- (ii) a good player is not allowed to play the match.
- (iii) an innocent person is punished.
- (iv) a driver is punished for no fault of him.
- (v) a good worker is not paid his salary in time.

These are the examples from practical life. These examples are quoted to make a point clear to the students.

α (ALPHA)

The probability of making *Type I error* is denoted by α (alpha). When a null hypothesis is rejected, we may be wrong in rejecting it or we may be right in rejecting it. We do not know that H_0 is true or false. Whatever our decision will be, it will have the support of probability. A true hypothesis has some probability of rejection and this probability is denoted by α . This probability is also called the size of *Type I error* and is denoted by α .

13.3.2 TYPE II ERROR

The null hypothesis H_0 may be false but it may be accepted. It is an error and is called *Type II error*. The value of the test-statistic may fall in the acceptance region when H_0 is in fact false. Suppose the hypothesis being tested is $H_0: \theta = \theta_0$ and H_0 is false and true value of θ is θ_1 or θ_{true} . If the difference between θ_0 and θ_1 is very large then the chance is very small that θ_0 (wrong) will be accepted. In this case the true sampling distribution of the statistic will be quite away from the sampling distribution under H_0 . There will be hardly any test-statistic which will fall in the acceptance region of H_0 . When the true distribution of the test-statistic overlaps the acceptance region of H_0 , then H_0 is accepted though H_0 is false. If the difference between θ_0 and θ_1 is small, then there is a high chance of accepting H_0 . This action will be an error of *Type II*.

β (BETTA)

The probability of making *Type II error* is denoted by β . Type II error is committed when H_0 is accepted while H_1 is true. The value of β can be calculated only when we happen to know the true value of the population parameter being tested.

13.3.3 RELATION BETWEEN α AND β

Suppose we have to test $H_0: \mu = \mu_0$ against the alternative $H_1: \mu > \mu_0$. A random sample of size n is selected from the population and the sample mean \bar{X} is calculated. The sample size n is large and therefore the sampling distribution of \bar{X} is normal with mean μ . To start with we assume that $H_0: \mu = \mu_0$ is true and \bar{X} has the distribution as shown on left side of the fig. 13.4.

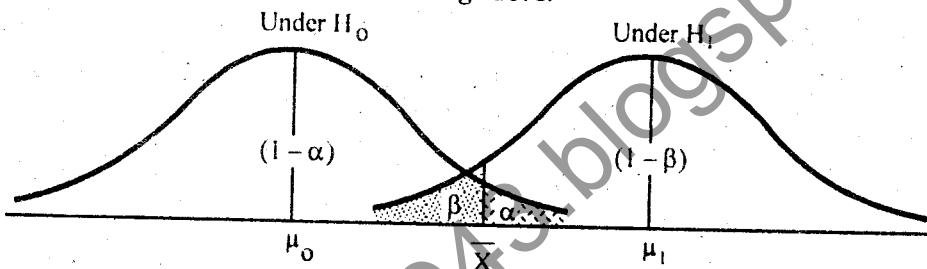


Figure 13.4

Fig.13.4. has two sampling distributions one is on the left side and the other is on the right side. When the null hypothesis $H_0: \mu = \mu_0$ is being tested, there are the following four possibilities.

- H_0 is true and \bar{X} falls in the area marked $(1 - \alpha)$ in the Fig.13.4. The hypothesis H_0 is accepted and this is called correct decision. Probability of this correct decision is $(1 - \alpha)$. We may or may not make this decision.
- H_0 is true and \bar{X} falls in the area marked α . This is the area of the distribution on the left side. Now H_0 is true but it will be rejected because \bar{X} falls in the rejection region. This is an error of Type I and this error will be committed with the probability of α . We do not know whether we have committed α error or not.
- H_0 is false. The true value of μ is say μ_1 and the true distribution of \bar{X} is the distribution on the right side in Fig. 13.4. Now suppose \bar{X} falls in the area marked $(1 - \beta)$. This is outside the acceptance region of the distribution on the left side. Thus $H_0: \mu = \mu_0$ is rejected and the probability of this action is $(1 - \beta)$. It is called correct decision when H_0 is false. In fact, \bar{X} belongs to some distribution. When we take a hypothesis H_0 , this is an assumption about the

mean of the distribution of \bar{X} . If true distribution of \bar{X} is on the right side, then some area of this distribution is falling on the acceptance region of the hypothetical distribution on the left side. This area is marked as β .

- (iv) H_0 is false and the value of \bar{X} falls in the area marked β . In this case H_0 is accepted because \bar{X} has fallen in the acceptance region of the first distribution. Thus H_0 being false may be accepted with probability of β .

If the distribution on the right side is shifted to the right, β will decrease and if this distribution is shifted to the left, β will increase. Thus the value of β depends upon the true value of population mean μ . In a certain given situation when n is fixed the value of β increases when α is decreased. Thus if we want to decrease α , we shall do it at the risk of increasing β . α -error and β -error are also called α -risk and β -risk respectively. Which risk do we want to keep at minimum level? This depends upon the costs of committing α -error and β -error. Suppose we are hesitant of rejecting H_0 when it is true, then we shall take α at a small level. In most of the tests, α is fixed at a small level like 0.01 (1 %) or 0.05 (5 %).

The following table shows four possible decisions in a certain test of hypothesis.

	H_0 is True	H_0 is False
H_0 is Accepted	Correct decision	Type II error
H_0 is Rejected	Type I error	Correct decision

When we are testing a hypothesis, our decision will fall in any one of the above four boxes. The four possible decisions in terms of probabilities are shown below in a tabular form.

	H_0 True	H_0 False
H_0 is Accepted	$(1 - \alpha)$	β
H_0 is Rejected	α	$(1 - \beta)$

It may be noted that α is an area in the right tail of the distribution under H_0 and β is the area in the left tail of the distribution under H_1 . Thus $\alpha + \beta \neq 1$ in general. In some special case and that too very rarely, $\alpha + \beta$ may be equal to 1. Level of α is usually small. Thus probability is small that our decision will fall in the box marked α . But when our decision has fallen in the box marked α , it is a powerful decision against H_0 .

13.4 LEVEL OF SIGNIFICANCE

The α -risk is the probability of rejecting a true null hypothesis. It is also called the significance level or level of significance of the test. It is denoted by α and its level is usually 1 % or 5 %. The value of α is usually decided before the selection of the sample.

13.5 FARMULATING H_0 AND H_1 AND MAKING CRITICAL REGION

Now, when we have discussed different terms used in the testing of hypothesis, we are in a position to discuss a point which is quite confusing sometimes. The question is how to formulate the null hypothesis H_0 and the alternative hypothesis H_1 . We elaborate this point here and we shall repeat here certain points already discussed in this chapter about framing of H_0 and H_1 . Let us consider some cases.

- (i) A machine has been producing components with mean length of 3 cm. which is the required standard. A new machinery has been installed and it is required to test the hypothesis that the mean length of the components is the same. It is obvious that in this case the H_0 and H_1 will be:

$$H_0 : \mu = 3 \text{ cm.} \quad H_1 : \mu \neq 3 \text{ cm.}$$

H_1 contains the inequality ' \neq ' which means that the rejection region is taken in both ends of the sampling distribution.

The test-statistic used is $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$.

The null hypothesis H_0 is rejected if $Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$. It is called *two-tailed test* with rejection region on both sides. H_0 is rejected when

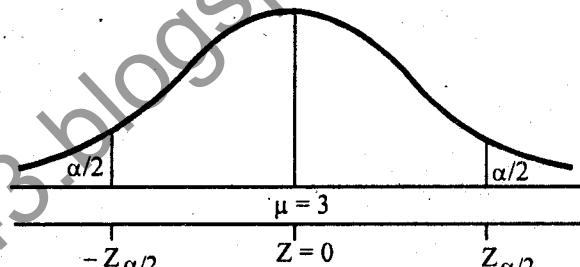


Figure 13.5

sample mean \bar{X} is sufficiently larger than 3 cm. or sufficiently smaller than 3.

- (ii) Suppose that we want to test whether the mean μ of a normal distribution exceeds a specified value μ_0 . We set up the null and alternative hypotheses as follows: $H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$

The null hypothesis H_0 and the alternative hypothesis H_1 in this case can also be written as $H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$

H_1 is complement of H_0 and the area of the distribution under H_0 and H_1 makes the complete distribution. In this case, the region of rejection is taken in the right tail of the distribution.

The test-statistic is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}. \text{ The null hypothesis}$$

H_0 is rejected when the calculated value of Z is greater than the critical value Z_α .

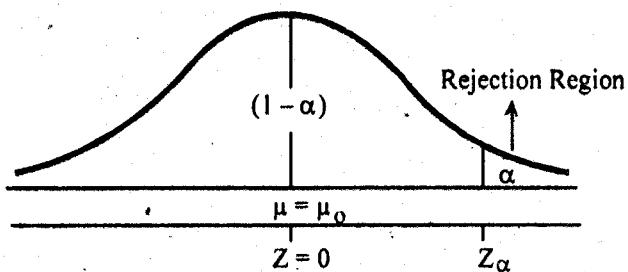


Figure 13.6

- (iii) At least 60 % of the people are in favour of English as medium of instructions.
The sampling distribution of proportion \hat{p} is divided into two parts (1) at least 60 % (2) less than 60 %.

We have a serious doubt about the statement and we hope to disprove it. The proportion of the people $p \geq 0.6$ is to be tested. The idea or suggestion of at least 60 % ($p \geq 0.6$) will be rejected if the sample gives the result well below 60 %. The rejection region is decided by H_1 , which is one-sided to the left. Thus we frame H_0 and H_1 as: $H_0: p \geq 0.6$ $H_1: p < 0.6$

In this case the entire critical region lies in the left tail. If $H_1: p < 0.6$ is true then the sample proportion \hat{p} should lie in the rejection region.

The test statistic used here is

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}. \text{ The hypothesis } H_0$$

is rejected if $Z < -Z_\alpha$.

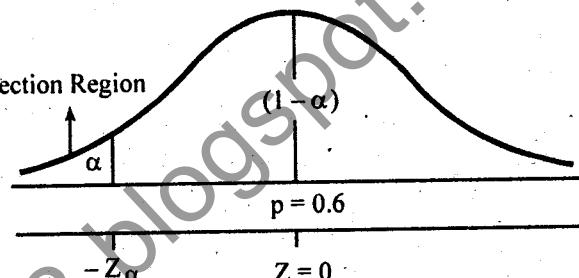


Figure 13.7

Example 13.1.

Indicate the type of errors committed in the following cases:

- (i) $H_0: \mu = 500$, $H_1: \mu \neq 500$. H_0 is rejected while H_0 is true.
(ii) $H_0: \mu = 500$, $H_1: \mu < 500$. H_0 is accepted while true value of $\mu = 600$.

Answer:

- (i) The hypothesis $\mu = 500$ is true and it has been rejected. Type I error has been committed.
(ii) H_0 is false and has been accepted. Type II error has been committed.

13.6 GENERAL PROCEDURE FOR TESTING OF HYPOTHESIS

Following are the main steps involved in the testing of a hypothesis about the population parameter.

1. Formulating Null hypothesis H_0 :

First of all we have to identify the problem and then we frame the hypothesis which we think shall be rejected. Suppose the population parameter is θ about which we have to frame the hypothesis. We specify a value θ_0 for the unknown parameter.

The null hypothesis H_0 can be written in three ways as shown below:

- (i) $H_0: \theta = \theta_0$ (ii) $H_0: \theta \leq \theta_0$ (iii) $H_0: \theta \geq \theta_0$

In some particular situation any one of the above three forms of H_0 is taken. The important thing about H_0 is that H_0 always contains some form of an equality sign such as '=' , '≥' , or '≤'. As H_0 always contains sign of equality of some type, some people always write H_0 as $H_0: \theta = \theta_0$ and they do not write the inequality contained in H_0 .

Alternative hypothesis H_1 :

The alternative hypothesis H_1 is the opposite or complement of H_0 . H_0 and H_1 combined together make the entire sampling distribution. Both H_0 and H_1 are equally important and they are to be defined properly and clearly. As H_1 is complement of H_0 , therefore H_1 stands decided when H_0 has been fixed. For example, for each value of H_0 , the corresponding value of H_1 is given below:

- (i) If $H_0: \theta = \theta_0$ then $H_1: \theta \neq \theta_0$
- (ii) If $H_0: \theta \leq \theta_0$ then $H_1: \theta > \theta_0$
- (iii) If $H_0: \theta \geq \theta_0$ then $H_1: \theta < \theta_0$

2. Level of significance α :

It is the probability of rejecting H_0 when H_0 is true. It is denoted by α . It makes the size of the critical region.

3. Test-statistic:

The *test statistic* depends upon the shape of the sampling distribution of the statistic. If the sampling distribution is a normal distribution, the test-statistic to be used is Z and if it is a t-distribution, the test-statistic to be used is t. Other test statistics are F and χ^2 (chi-square).

4. Critical region:

Critical region or rejection region is decided by H_1 . The size of critical region is equal to α .

- (i) If the alternative hypothesis is $H_1: \theta \neq \theta_0$ the rejection region is taken in both ends of the sampling distribution. Each side has rejection region equal to $\alpha/2$. It is called two-sided rejection region. The rejection regions are separated by the two critical values.
- (ii) When H_1 is $\theta > \theta_0$, then rejection region of size α is taken only in the right side. It is called *one-sided to the right*. The rejection region is separated from the acceptance region by a critical value of test-statistic.
- (iii) When H_1 is $\theta < \theta_0$, the rejection region of size α is taken only on the left side. It is called *one-sided to the left*.

5. Computations:

The relevant test-statistic is calculated from the sample data. The calculated value is to be compared with the tabulated value.

6. Conclusion:

If the calculated value of test-statistic lies in the rejection region, the null hypothesis H_0 is rejected and H_1 is accepted. If the calculated value of the test-statistic falls in the acceptance region, we say that H_0 is accepted but it is not acceptance in the real sense of the word. The word acceptance only means that the sample has not provided sufficient information against the null hypothesis.

13.7 HYPOTHESIS TESTING - POPULATION MEAN μ , σ KNOWN

(LARGE SAMPLE)

Suppose a population has the mean μ which is unknown and the standard deviation σ which is known. A large sample of size n is selected from the population and sample mean \bar{X} is calculated. We are required to test a hypothesis that the population mean μ has the specified value μ_0 . The steps of the procedure are listed below:

1. We frame the null hypothesis H_0 and the alternative hypothesis H_1 . Three different forms of H_0 and H_1 are possible which are:
 - $H_0: \mu = \mu_0$ and $H_1: \mu \neq \mu_0$
 - $H_0: \mu \leq \mu_0$ and $H_1: \mu > \mu_0$
 - $H_0: \mu \geq \mu_0$ and $H_1: \mu < \mu_0$
2. Level of significance α is decided.
3. Test-statistic:

When sample size is large, the sampling distribution of \bar{X} has the normal distribution with mean μ and the standard error σ/\sqrt{n} . The population may or

may not be normal. The test-statistic to be used is Z where $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

4. Critical region:

The critical region depends upon the alternative hypothesis. There are three possible rejection plans. We discuss all the three turn by turn.

- (a) When H_1 is $\mu \neq \mu_0$, the rejection region equal to $\alpha/2$ in size is taken on both ends of the sampling distribution as shown in Fig. 13.8. The critical values of Z which separates the

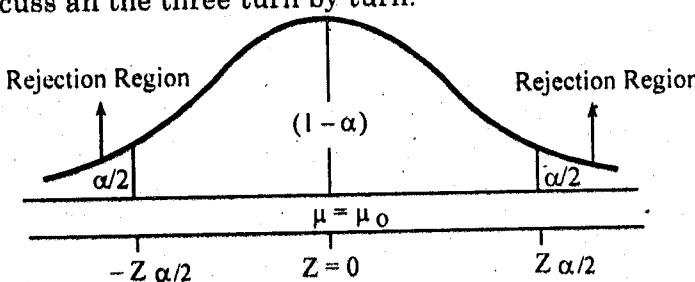


Figure 13.8

critical regions from the central acceptance region are $-Z_{\alpha/2}$ and $Z_{\alpha/2}$. The critical value $-Z_{\alpha/2}$ has the area on its left equal to $\alpha/2$ and the critical value $+Z_{\alpha/2}$ has area on its right equal to $\alpha/2$. H_0 is rejected if the calculated value of Z lies in rejection region. The rejection region is $Z < -Z_{\alpha/2}$ and $Z > Z_{\alpha/2}$. When $\alpha = 0.05$, then $-Z_{\alpha/2} = -Z_{0.025} = -1.96$ and $Z_{0.025} = 1.96$.

- (b) When H_1 is $\mu > \mu_0$, the rejection region equal to α is taken in the right end of the distribution as shown in Fig. 13.9. The test plan is called one-tailed to the right.

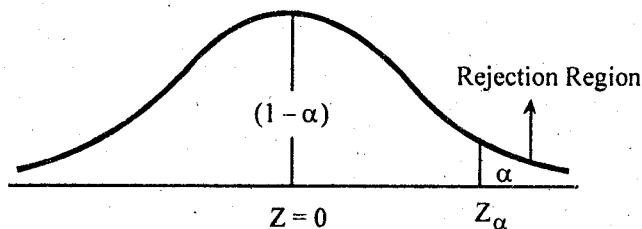


Figure 13.9

The hypothesis is rejected when the calculated value of Z is greater than Z_α , where Z_α is the critical point on the right of which the area is equal to α .

- (c) When H_1 is $\mu < \mu_0$, the rejection region equal to α is taken in the left end of the distribution as shown in Fig. 13.10. The rejection plan is called one-tailed to the left.

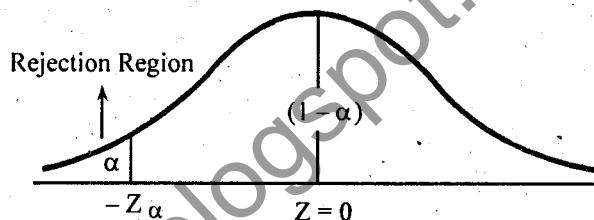


Figure 13.10

The hypothesis is rejected when the calculated value of Z is less than the critical value $-Z_\alpha$ where $-Z_\alpha$ is a critical point on the left of which the area is α . The rejection region is $Z < -Z_\alpha$. Corresponding to each null hypothesis, the alternate hypothesis and the rejection regions are given below:

Null hypothesis	Alternative hypothesis	Rejection region
(a) $H_0: \mu = \mu_0$	$H_1: \mu \neq \mu_0$ (two-sided)	$Z < -Z_{\alpha/2}$ and $Z > Z_{\alpha/2}$
(b) $H_0: \mu \leq \mu_0$	$H_1: \mu > \mu_0$ (one-sided)	$Z > Z_\alpha$
(c) $H_0: \mu \geq \mu_0$	$H_1: \mu < \mu_0$ (one-sided)	$Z < -Z_\alpha$

5. Computations:

The value of Z is calculated by using the formula: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

6. Conclusion:

If the value of Z lies in the acceptance region, the hypothesis is accepted. But acceptance is just an indication that the sample data has failed to provide evidence against the null hypothesis. If the value of Z lies in rejection region the hypothesis is rejected. When H_0 is rejected, there is only 100α % chance that the null hypothesis is true.

Example 13.2.

Past records show that the average score of students in statistics is 57 with standard deviation 10. A new method of teaching is employed and a random sample of 70 students is selected. The sample average is 60. Can we conclude on the basis of these results, at 5 % level of significance, that the average score has increased?

Solution:

1. Null hypothesis: $H_0: \mu = 57$ Alternative hypothesis: $H_1: \mu > 57$

2. Level of significance: $\alpha = 0.05$

3. Test - statistic: $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$

4. Critical region: $Z > 1.645$. Here we use one-sided test to the right. The hypothesis $H_0: \mu = 57$ will be rejected if Z lies in rejection region.

(From the area table of normal distribution, we have $Z_\alpha = Z_{0.05} = 1.645$)

5. Computations: Here $n = 70$, $\bar{X} = 60$, $\sigma = 10$, and hence

$$Z = \frac{60 - 57}{10 / \sqrt{70}} = \frac{3}{10 / \sqrt{70}} = 2.51$$

6. Conclusion: Since the calculated value of $Z = 2.51$ falls in the critical region, so we reject our null hypothesis $H_0: \mu = 57$ at 5 % level of significance and we may conclude that the average score has increased.

Example 13.3.

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a mean of 812 hours and a standard deviation of 40 hours. Test the hypothesis that $\mu = 812$ hours against the alternative $\mu \neq 812$ hours if a random sample of 36 bulbs has an average life of 800 hours. Use a 5 % level of significance.

Solution:

1. Null hypothesis: $H_0: \mu = 812$ Alternative hypothesis: $H_1: \mu \neq 812$

2. Level of significance: $\alpha = 0.05$

3. Test - statistic: $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$

4. Critical region: $|Z| > 1.96$ ($Z < -1.96$ and $Z > 1.96$)

(From the area table of normal distribution, we have $Z_{\alpha/2} = Z_{0.025} = 1.96$)

- 5. Computations:** Here $n = 36$, $\bar{X} = 800$, $\sigma = 40$, and hence

$$Z = \frac{800 - 812}{40 / \sqrt{36}} = -\frac{12}{40} (6) = -1.8$$

- 6. Conclusion:** Since the calculated value of $Z = -1.8$ falls in the acceptance region. Thus $H_0: \mu = 812$ is not rejected.

13.8 HYPOTHESIS TESTING — POPULATION MEAN μ — σ NOT KNOWN (LARGE SAMPLE)

This is an important case in which σ is not known. When sample size n is large, the population may be normal or not, the sampling distribution of \bar{X} has the normal distribution with mean μ and standard error σ/\sqrt{n} . But when σ is unknown, it is estimated by the sample standard deviation S and the estimated standard error is

S/\sqrt{n} . The Z-statistic becomes $Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ where $S^2 = \frac{\sum(X - \bar{X})^2}{n}$. The remaining procedure is exactly the same as discussed earlier. The only difference is that S is used in place of σ in the calculation of Z .

Example 13.4.

A home heating oil delivery company would like to estimate the annual usage for its customers who live in single-family homes. A sample of 100 customers indicated an average annual usage of 1103 gallons and a sample standard deviation of 327.8 gallons. At the 1 % level of significance, is there evidence that the average annual usage exceeds 1000 gallons per year?

Solution:

1. Null hypothesis: $H_0: \mu \leq 1000$ Alternative hypothesis: $H_1: \mu > 1000$

2. Level of significance: $\alpha = 0.01$

3. Test statistic: $Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$

4. Critical region: $Z > 2.326$

(From the area table of normal distribution, we have $Z_\alpha = Z_{0.01} = 2.326$)

5. Computations: Here $n = 100$, $\bar{X} = 1103$, $S = 327.8$, and hence

$$Z = \frac{1103 - 1000}{327.8 / \sqrt{100}} = \frac{103}{327.8} (10) = 3.14$$

6. Conclusion: Since the calculated value of $Z = 3.14$ falls in the critical region, so we reject our null hypothesis $H_0: \mu \leq 1000$ at 1 % level of significance and we may conclude that the average annual usage exceeds 1000 gallons per year.

Example 13.5.

A sample of 42 measurements was taken in order to test the null hypothesis that the population mean equals 8.5 against the alternative that it is different from 8.5. The sample mean and standard deviation were found to be 8.79 and 1.27, respectively. Perform the hypothesis test using 0.01 as the level of significance.

Solution:

1. Null hypothesis: $H_0: \mu = 8.5$ Alternative hypothesis: $H_1: \mu \neq 8.5$
2. Level of significance: $\alpha = 0.01$

3. Test - statistic: $Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$
4. Critical region: $|Z| > 2.575$ ($Z < -2.575$ and $Z > 2.575$)
(From the area table of normal distribution, we have $Z_{\frac{\alpha}{2}} = Z_{0.005} = 2.575$)

5. Computations: Here $n = 42$, $\bar{X} = 8.79$, $S = 1.27$, and hence

$$Z = \frac{8.79 - 8.5}{1.27 / \sqrt{42}} = \frac{0.29}{1.27} \sqrt{42} = 1.48$$

6. Conclusion: Since the calculated value of $Z = 1.48$ falls in the acceptance region, so we accept our null hypothesis $H_0: \mu = 8.5$ at 1 % level of significance.

13.9 HYPOTHESIS TESTING – POPULATION MEAN μ , σ KNOWN – NORMAL POPULATION (SMALL SAMPLE)

Sometimes the hypothesis about the population which is normal and its standard deviation σ is known. In this case Z-test is used both for small and large

sample size. Thus $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$. The procedure for testing of population mean μ is the same as discussed earlier.

13.10 HYPOTHESIS TESTING – POPULATION MEAN μ , σ UNKNOWN – NORMAL POPULATION (SMALL SAMPLE)

When the standard deviation of the population is not known, it is estimated by the sample standard deviation 's' where $s = \sqrt{\frac{1}{n-1} \sum (X - \bar{X})^2}$. The procedure runs as follows:

The different forms of hypotheses are

1. (a) $H_0: \mu = \mu_0$ and $H_1: \mu \neq \mu_0$
 (b) $H_0: \mu \leq \mu_0$ and $H_1: \mu > \mu_0$
 (c) $H_0: \mu \geq \mu_0$ and $H_1: \mu < \mu_0$
2. Level of significance α is decided.

3. Test - statistic:

When population is normal and sample size n is small, the sampling distribution of \bar{X} has the t-distribution with $(n - 1)$ degrees of freedom. The test-statistic is $t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$.

4. Critical region:

The critical region is based on the alternative hypothesis.

- (a) For the alternative hypothesis $H_1 : \mu \neq \mu_0$, the rejection region is two-sided as shown in Fig. 13.11. The two critical values $-t_{\alpha/2}(n-1)$ and $t_{\alpha/2}(n-1)$ are seen from the t-table below $\alpha/2$ and against $(n - 1)$ degrees of freedom. The critical region is $t > t_{\alpha/2}(n-1)$ or $t < -t_{\alpha/2}(n-1)$ as shown in Fig. 13.11.

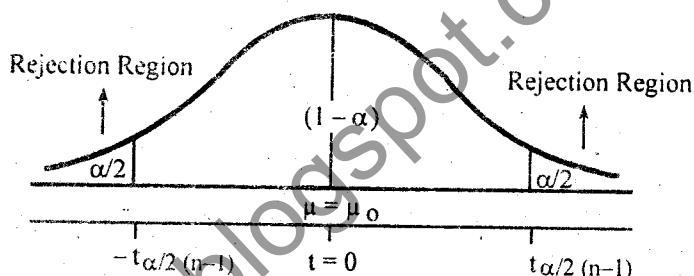


Figure 13.11

- (b) When H_1 is $\mu > \mu_0$, the rejection region is taken on the extreme right side of the sampling distribution as shown in Fig. 13.12. The critical value $t_\alpha(n-1)$ is seen from the t-table below α and against $(n - 1)$ degrees of freedom. The critical region is $t > t_\alpha(n-1)$.

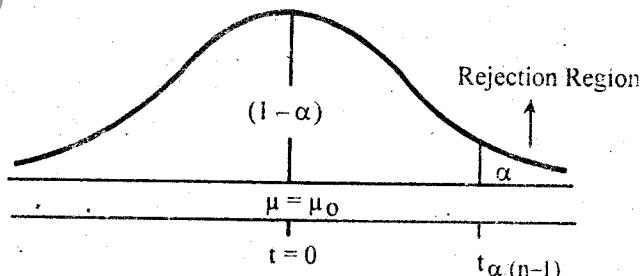


Figure 13.12

- (c) When H_1 is $\mu < \mu_0$, the entire rejection region is taken on the left side of the sampling distribution as shown in Fig. 13.13. The critical value $-t_{\alpha}(n-1)$ is seen from the t-table below α and against $(n - 1)$ degrees of freedom. The critical region is $t < -t_{\alpha}(n-1)$.

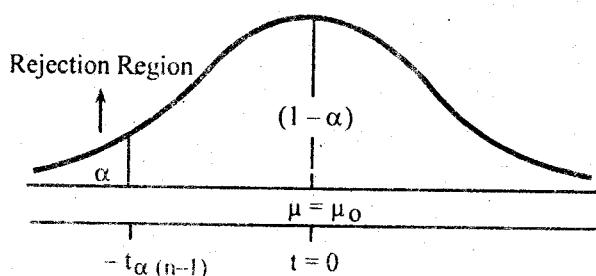


Figure 13.13

5. Computations:

The test-statistic 't' is calculated from the sample data where $t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$

6. Conclusion:

The null hypothesis H_0 is rejected in favour of H_1 when the value of t lies in the rejection region. H_0 is accepted when the value of t lies in acceptance region.

Example 13.6.

A manufacturing company making automobile tires claims that the average life of its product is 35000 miles. A random sample of 16 tires was selected; and it was found that the mean life was 34000 miles with a standard deviation $s = 2000$ miles. Test hypothesis $H_0: \mu = 35000$ against the alternative $H_1: \mu < 35000$ at $\alpha = 0.05$.

Solution:

1. Null hypothesis: $H_0: \mu = 35000$ Alternative hypothesis: $H_1: \mu < 35000$

2. Level of significance: $\alpha = 0.05$

3. Test - statistic: $t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$

4. Critical region: $t < -1.753$

(From the t-table, we have $-t_{\alpha(n-1)} = -t_{0.05(15)} = -1.753$)

5. Computations: Here $n = 16$, $\bar{X} = 34000$, $s = 2000$, and hence

$$t = \frac{34000 - 35000}{2000 / \sqrt{16}} = \frac{-1000}{2000} (4) = -2$$

6. Conclusion:

Since the calculated value of $t = -2$ falls in the critical region, so we reject our null hypothesis $H_0: \mu = 35000$ at 5 % level of significance.

Example 13.7.

A random sample of 8 cigarettes of a certain brand has an average nicotine content of 4.2 milligrams and a standard deviation of 1.4 milligrams. Is this in line with the manufacturer's claim that the average nicotine content does not exceed 3.5 milligrams? Use 1 % level of significance and assume the distribution of nicotine contents to be normal.

Solution:

1. Null hypothesis: $H_0: \mu \leq 3.5$ Alternative hypothesis: $H_1: \mu > 3.5$

2. Level of significance: $\alpha = 0.01$

3. Test - statistic: $t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$

4. **Critical region:** $t > 2.998$

(From the t-table, we have $t_{\alpha(n-1)} = t_{0.01(7)} = 2.998$)

5. **Computations:** Here $n = 8$, $\bar{X} = 4.2$, $s = 1.4$, and hence

$$t = \frac{4.2 - 3.5}{1.4 / \sqrt{8}} = \frac{0.7}{1.4} \sqrt{8} = 1.414$$

6. **Conclusion:** Since the calculated value of $t = 1.414$ falls in the acceptance region, so we accept our null hypothesis $H_0: \mu = 35$ at 1 % level of significance.

13.11 HYPOTHESIS TESTING - DIFFERENCE BETWEEN TWO POPULATION MEANS $\mu_1 - \mu_2$, σ_1^2 AND σ_2^2 KNOWN

(LARGE SAMPLES)

Suppose there are two populations (normal or non-normal) with means μ_1 and μ_2 which are unknown and the variances σ_1^2 and σ_2^2 which are known. Two large random samples of sizes n_1 and n_2 are selected from the populations and the sample means \bar{X}_1 and \bar{X}_2 are calculated. The difference $(\bar{X}_1 - \bar{X}_2)$ is a random variable and its distribution is normal with mean $\mu_1 - \mu_2$ and standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

The procedure for testing the hypothesis $\mu_1 - \mu_2 = 0$ is explained below.

1. The null and the alternative hypotheses which are possible are

(a) $H_0: \mu_1 - \mu_2 = 0$ (or $\mu_1 = \mu_2$) and $H_1: \mu_1 - \mu_2 \neq 0$ (or $\mu_1 \neq \mu_2$)

(b) $H_0: \mu_1 - \mu_2 \leq 0$ (or $\mu_1 \leq \mu_2$) and $H_1: \mu_1 - \mu_2 > 0$ (or $\mu_1 > \mu_2$)

(c) $H_0: \mu_1 - \mu_2 \geq 0$ (or $\mu_1 \geq \mu_2$) and $H_1: \mu_1 - \mu_2 < 0$ (or $\mu_1 < \mu_2$)

2. Level of significance α is decided.

3. **Test - statistic:**

The distribution of $(\bar{X}_1 - \bar{X}_2)$ is normal, therefore the test-statistic to be used is Z ,

$$\text{where } Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

4. **Critical region:**

For each alternate hypothesis H_1 , there is a rejection plan as explained earlier.

5. Computations:

The Z-statistic is calculated using the sample data where,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Sometimes the null hypothesis states some difference between μ_1 and μ_2 and the difference is denoted by Δ . In that case H_0 is $\mu_1 - \mu_2 = \Delta$ (say) and

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

6. Conclusion:

The hypothesis is rejected if the calculated value of Z lies in rejection region. If Z lies in acceptance region, the hypothesis is accepted.

Example 13.8.

Suppose you wish to estimate the effects of a certain sleeping pill on men and women. Two samples are independently taken, and the relevant data are shown below:

	Men	Women
Sample size	$n_1 = 36$	$n_2 = 64$
Sample mean	$\bar{X}_1 = 8.75$	$\bar{X}_2 = 7.25$
Population variance	$\sigma_1^2 = 9$	$\sigma_2^2 = 4$

Test the null hypothesis $H_0: \mu_1 = \mu_2$ against the alternative hypothesis $H_1: \mu_1 > \mu_2$ at $\alpha = 0.05$.

Solution:

1. Null hypothesis: $H_0: \mu_1 = \mu_2$ Alternative hypothesis: $H_1: \mu_1 > \mu_2$

2. Level of significance: $\alpha = 0.05$

3. Test - statistic: $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

4. Critical region: $Z > 1.645$

(From the area table of normal distribution, we have $Z_\alpha = Z_{0.05} = 1.645$)

5. Computations: Here $n_1 = 36$, $\bar{X}_1 = 8.75$, $\sigma_1^2 = 9$, $n_2 = 64$, $\bar{X}_2 = 7.25$, $\sigma_2^2 = 4$,

$$\text{and hence } Z = \frac{(8.75 - 7.25) - 0}{\sqrt{\frac{9}{36} + \frac{4}{64}}} = \frac{1.5}{0.5590} = 2.683$$

- 6. Conclusion:** Since the calculated value of $Z = 2.683$ falls in the critical region, so we reject our null hypothesis $H_0: \mu_1 = \mu_2$ at 5 % level of significance.

Example 13.9.

Two astronomers recorded observations on a certain star. The mean of 30 observations obtained by first astronomer is 8.85 and mean of 40 observations made by second astronomer is 8.20. Past experience shows that each astronomer obtained readings with variance of 1.2. Using $\alpha = 0.01$, can we say that the difference between two results is significant.

Solution:

1. Null hypothesis: $H_0: \mu_1 = \mu_2$ Alternative hypothesis: $H_1: \mu_1 \neq \mu_2$
2. Level of significance: $\alpha = 0.01$

3. Test - statistic: $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$$(\because \sigma_1^2 = \sigma_2^2 = \sigma^2)$$

4. Critical region: $|Z| > 2.575$ ($Z < -2.575$ and $Z > 2.575$)
(From the area table of normal distribution, we have $Z_{\frac{\alpha}{2}} = Z_{0.005} = 2.575$)

5. Computations: Here $n_1 = 30$, $\bar{X}_1 = 8.85$, $n_2 = 40$, $\bar{X}_2 = 8.20$, $\sigma^2 = 1.2$, $\sigma = 1.10$
and hence $Z = \frac{(8.85 - 8.20) - 0}{1.10 \sqrt{\frac{1}{30} + \frac{1}{40}}} = \frac{0.65}{0.27} = 2.407$

6. Conclusion: Since the calculated value of $Z = 2.407$ falls in the acceptance region, so we accept our null hypothesis $H_0: \mu_1 = \mu_2$ at 1 % level of significance. We may conclude that the difference between two results is insignificant.

13.12 HYPOTHESIS TESTING - DIFFERENCE BETWEEN TWO POPULATION MEANS $\mu_1 - \mu_2$, σ_1^2 AND σ_2^2 UNKNOWN (LARGE SAMPLES)

When the population variances σ_1^2 and σ_2^2 are unknown, they are estimated by their sample variances S_1^2 and S_2^2 and the test-statistic to be used becomes,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

This formula is used only for large sample sizes but the populations may or may not be normal. The procedure for testing H_0 is the same as explained earlier.

Example 13.10.

Suppose that two randomly selected samples yield the following information:

	Sample I	Sample II
Size	$n_1 = 82$	$n_2 = 41$
Mean	$\bar{X}_1 = 50$	$\bar{X}_2 = 55$
Variance	$S_1^2 = 405$	$S_2^2 = 324$

Test the null hypothesis that the two population means are equal that is, $H_0: \mu_1 = \mu_2$ against the alternative hypothesis $H_1: \mu_1 < \mu_2$ at $\alpha = 0.01$.

Solution:

1. Null hypothesis: $H_0: \mu_1 = \mu_2$ Alternative hypothesis: $H_1: \mu_1 < \mu_2$

2. Level of significance: $\alpha = 0.01$

3. Test - statistic: $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

4. Critical region: $Z < -2.326$

(From the area table of normal distribution, we have $-Z_{\alpha} = -Z_{0.01} = -2.326$)

5. Computations: Here $n_1 = 82$, $\bar{X}_1 = 50$, $S_1^2 = 405$, $n_2 = 41$, $\bar{X}_2 = 55$, $S_2^2 = 324$,

$$\text{and hence } Z = \frac{(50 - 55) - 0}{\sqrt{\frac{405}{82} + \frac{324}{41}}} = \frac{-5}{3.58} = -1.40$$

6. Conclusion:

Since the calculated value of $Z = -1.40$ falls in the acceptance region, so we accept our null hypothesis $H_0: \mu_1 = \mu_2$ at 1 % level of significance.

13.13 TEST ABOUT $\mu_1 - \mu_2$, σ_1^2 AND σ_2^2 KNOWN, POPULATIONS NORMAL (SMALL SAMPLES)

In case of small sample sizes, we can use Z-test for testing the difference between μ_1 and μ_2 when σ_1^2 and σ_2^2 are known and the populations are necessarily

normal. The Z-test used is $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

13.14 TEST ABOUT $\mu_1 - \mu_2$, σ_1^2 AND σ_2^2 NOT KNOWN, POPULATIONS NORMAL (SMALL SAMPLES)

This is a case which is different from the previous three cases. Here the conditions are that:

- (i) the populations are normal
- (ii) σ_1^2 and σ_2^2 are unknown but assumed to be equal.

- (iii) the sample sizes n_1 and n_2 are small and are selected independently.

The variances σ_1^2 and σ_2^2 are unknown but $\sigma_1^2 = \sigma_2^2 = \sigma^2$. The parameter σ^2 is estimated by the sample variances. The sample estimator of σ^2 is s_p^2 , where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\sum(X_1 - \bar{X}_1)^2 + \sum(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

$$\text{and } s_p = \sqrt{\frac{\sum(X_1 - \bar{X}_1)^2 + \sum(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

s_p^2 is called pooled estimator of the common population variance σ^2 . The difference $(\bar{X}_1 - \bar{X}_2)$ has the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom where

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The tabulated value of 't' for $n_1 + n_2 - 2$ degrees of freedom is seen from the t-table.

For $H_1: \mu_1 \neq \mu_2$ the critical values are $-t_{\alpha/2(n_1+n_2-2)}$ and $t_{\alpha/2(n_1+n_2-2)}$

For $H_1: \mu_1 > \mu_2$ the critical value is $t_{\alpha(n_1+n_2-2)}$

and For $H_1: \mu_1 < \mu_2$ the critical value is $-t_{\alpha(n_1+n_2-2)}$

The null hypothesis H_0 is rejected when the calculated value of t lies in rejection region.

Example 13.11.

Two samples are randomly selected from two classes of students who have been taught by different methods. An examination is given and the results are shown as follows:

	Class I	Class II
Sample Size	$n_1 = 8$	$n_2 = 10$
Mean	$\bar{X}_1 = 95$	$\bar{X}_2 = 97$
Variance	$s_1^2 = 47$	$s_2^2 = 30$

On the assumption that the test scores of the two classes of students have identical variances, determine whether the two different methods of teaching are equally effective at $\alpha = 0.01$.

Solution:

1. Null hypothesis: $H_0: \mu_1 = \mu_2$ Alternative hypothesis: $H_1: \mu_1 \neq \mu_2$
2. Level of significance: $\alpha = 0.01$

3. Test - statistic: $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

4. Critical region: $|t| > 2.921$ ($t < -2.921$ and $t > 2.921$)

(From the t-table, we have $t_{\frac{\alpha}{2}(n_1+n_2-2)} = t_{0.005(16)} = 2.921$)

5. Computations: Here $n_1 = 8$, $\bar{X}_1 = 95$, $s_1^2 = 47$, $n_2 = 10$, $\bar{X}_2 = 97$, $s_2^2 = 30$,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(8 - 1)47 + (10 - 1)30}{8 + 10 - 2} = \frac{599}{16} = 37.4375,$$

$$s_p = \sqrt{37.4375} = 6.12, \text{ and hence } t = \frac{(95 - 97) - 0}{6.12 \sqrt{\frac{1}{8} + \frac{1}{10}}} = \frac{-2}{2.9030} = -0.689$$

6. Conclusion:

Since the calculated value of $t = -0.689$ falls in the acceptance region, so we accept our null hypothesis $H_0: \mu_1 = \mu_2$ at 1 % level of significance. On the basis of the evidence, we may conclude that the two different methods of teaching are equally effective.

13.15 TEST ABOUT $\mu_1 - \mu_2$, DEPENDENT SAMPLES, POPULATIONS NORMAL

Suppose there are two populations with mean μ_1 and μ_2 which are unknown. Two random samples of sizes n_1 and n_2 are selected. It is further assumed that the samples are dependent. Suppose we record blood pressures of a sample of some patients. The patients are given a treatment for some period and again their blood pressures are recorded. These two sets of observations are called dependent samples. The first set of observations is called 'before' and the second set of observations is called 'after' observations. These observations are in pairs. If $X_1, X_2, X_3, \dots, X_n$ are the 'before' observations and $Y_1, Y_2, Y_3, \dots, Y_n$ are the 'after' observations, then the paired observations are $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n)$. Let us find the difference between the paired values. Let difference $d_1 = X_1 - Y_1, d_2 = X_2 - Y_2, d_3 = X_3 - Y_3, \dots, d_n = X_n - Y_n$.

The mean of the sample 'd' values is denoted by \bar{d} . Suppose the corresponding parameter of the difference between paired observations in the populations is denoted by μ_D . The various steps of the procedure are:

- Three different forms of null and alternative hypotheses are
 - $H_0: \mu_D = 0$ (or $\mu_1 = \mu_2$) and $H_1: \mu_D \neq 0$ (or $\mu_1 \neq \mu_2$)
 - $H_0: \mu_D \leq 0$ (or $\mu_1 \leq \mu_2$) and $H_1: \mu_D > 0$ (or $\mu_1 > \mu_2$)
 - $H_0: \mu_D \geq 0$ (or $\mu_1 \geq \mu_2$) and $H_1: \mu_D \leq 0$ (or $\mu_1 < \mu_2$)

Sometimes we have to examine that the differences of the paired observations in the population have some specified value say Δ . In that case $\mu_D = \Delta$.

- Level of significance α is decided.
- Test-statistic:

\bar{d} has the t-distribution with $(n - 1)$ degrees of freedom.

$$t = \frac{\bar{d} - d_0}{s_d / \sqrt{n}} \text{ where } s_d = \sqrt{\frac{\sum(d - \bar{d})^2}{n-1}} = \sqrt{\frac{1}{n-1} \left[\sum d_i^2 - \frac{(\sum d_i)^2}{n} \right]}$$

- Critical region:

Corresponding to each H_1 , there is a critical region.

- Computations: The test-statistic t is calculated where $t = \frac{\bar{d} - d_0}{s_d / \sqrt{n}}$

$$\text{When } H_0 \text{ is } \mu_D = 0, \text{ then } t = \frac{\bar{d} - 0}{s_d / \sqrt{n}} = \frac{\bar{d} \sqrt{n}}{s_d}$$

- Conclusion:

The hypothesis $\mu_D = 0$ is rejected if the calculated value of 't' lies in the rejection region.

Example 13.12.

Suppose that a shoe company wanted to test material for the sales of shoes. For each pair of shoes the new material was placed on one shoe and the old material was placed on the other shoe. After a given period of time a random sample of ten pairs of shoes was selected and the wear was measured on a ten-point scale with the following results:

Pair number	1	2	3	4	5	6	7	8	9	10
New material	2	4	5	7	7	5	9	8	8	7
Old material	4	5	3	8	9	4	7	8	5	6
Differences	-2	-1	+2	-1	-2	+1	+2	0	+3	+1

At the 0.05 level of significance, is there evidence that the average wear is higher for the new material than the old material?

Solution:

1. Null hypothesis: $H_0: \mu_{\text{new}} \leq \mu_{\text{old}} \text{ or } \mu_D = \mu_{\text{new}} - \mu_{\text{old}} \leq 0$

Alternative hypothesis: $H_1: \mu_{\text{new}} > \mu_{\text{old}} \text{ or } \mu_D = \mu_{\text{new}} - \mu_{\text{old}} > 0$

2. Level of significance: $\alpha = 0.05$

3. Test - statistic: $t = \frac{\bar{d} - d_0}{s_d / \sqrt{n}}$

4. Critical region: $t > 1.833$

(From the t-table, we have $t_{\alpha(n-1)} = t_{0.05(9)} = 1.833$)

5. Computations: Let X_1 = new material and X_2 = old material.

The necessary calculations are given below:

X_1	2	4	5	7	7	5	9	8	8	7
X_2	4	5	3	8	9	4	7	8	5	6
$d = X_1 - X_2$	-2	-1	+2	-1	-2	+1	+2	0	+3	+1
d^2	4	1	4	1	4	1	4	0	9	1

Here $n = 10$, $\sum d = 3$, $\sum d^2 = 29$, $\bar{d} = \frac{\sum d}{n} = \frac{3}{10} = 0.3$,

$$s_d^2 = \frac{1}{n-1} \left[\sum d^2 - \frac{(\sum d)^2}{n} \right] = \frac{1}{10-1} \left[29 - \frac{(3)^2}{10} \right] \\ = 3.1222, \quad s_d = 1.77, \text{ and hence}$$

$$t = \frac{0.3 - 0}{1.77 / \sqrt{10}} = \frac{0.3}{1.77} \sqrt{10} = 0.536$$

6. Conclusion:

Since the calculated value of $t = 0.536$ falls in the acceptance region, so we accept our null hypothesis $H_0: \mu_{\text{new}} \leq \mu_{\text{old}}$ at 5 % level of significance. On the basis of the evidence, we may conclude that the average wear is not higher for the new material than the old material.

Example 13.13.

Two varieties of wheat are each planted in ten localities with differences in yield as follows: 2, 4, 2, 2, 3, 6, 2, 2, 4, 3. Test the hypothesis that the population mean difference is zero, using $\alpha = 0.01$.

Solution:

1. Null hypothesis: $H_0: \mu_1 = \mu_2 \text{ or } \mu_D = \mu_1 - \mu_2 = 0$

Alternative hypothesis: $H_1: \mu_1 \neq \mu_2 \text{ or } \mu_D = \mu_1 - \mu_2 \neq 0$

2. Level of significance: $\alpha = 0.01$

3. **Test - statistic:** $t = \frac{\bar{d} - d_0}{s_d / \sqrt{n}}$
4. **Critical region:** $|t| > 3.250$ ($t < -3.250$ and $t > 3.250$)
 (From the t-table, we have $t_{\frac{\alpha}{2}(n-1)} = t_{0.005(9)} = 3.250$)
5. **Computations:** Here $n = 10$, $\Sigma d = 30$, $\Sigma d^2 = 106$, $\bar{d} = \frac{\Sigma d}{n} = \frac{30}{10} = 3$,
 $s_d^2 = \frac{1}{n-1} \left[\Sigma d^2 - \frac{(\Sigma d)^2}{n} \right] = \frac{1}{10-1} \left[106 - \frac{(30)^2}{10} \right]$
 $= 1.7778$, $s_d = 1.33$, and hence
 $t = \frac{3 - 0}{1.33 / \sqrt{10}} = \frac{3}{1.33} \sqrt{10} = 7.133$

6. **Conclusion:** Since the calculated value of $t=7.133$ falls in the critical region, so we reject our null hypothesis $H_0: \mu_1 = \mu_2$ at 1% level of significance.

13.16 TEST OF POPULATION PROPORTION p (LARGE SAMPLE)

Let us consider a binomial population with a proportion p which is unknown and we have to test a hypothesis about the unknown population parameter. A random sample of size n ($n > 30$) is selected from the population and the sample proportion \hat{p} is calculated. When sample size is large, the distribution of \hat{p} is normal with mean p and standard error $\sqrt{\frac{p q}{n}}$. The random variable Z can be calculated

from \hat{p} . Thus $Z = \frac{\hat{p} - p}{\sqrt{\frac{p q}{n}}}$.

The random variable Z is used as test statistic and the value of Z makes a base for the acceptance or rejection of the null hypothesis about the population proportion. The procedure for testing p runs as below:

1. We frame a hypothesis about the population proportion p . Let us specify a value p_0 for the population parameter p . The null hypothesis H_0 and the alternative hypothesis H_1 can take any one of the following three forms:

- (a) $H_0: p = p_0$ and $H_1: p \neq p_0$ (b) $H_0: p \leq p_0$ and $H_1: p > p_0$
 (c) $H_0: p \geq p_0$ and $H_1: p < p_0$

2. Level of significance is decided. It is denoted by α .

3. **Test-statistic:** Used in this case is $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$ where $q_0 = 1 - p_0$

The sample proportion \hat{p} can also be written as $\hat{p} = \frac{X}{n}$, where 'X' is the number of successes in the sample of size n. Putting $\hat{p} = \frac{X}{n}$ in the above formula for Z,

$$\text{we get } Z = \frac{\frac{X}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{\frac{X - np_0}{n}}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{X - np_0}{n \sqrt{\frac{p_0 q_0}{n}}} = \frac{X - np_0}{\sqrt{np_0 q_0}}$$

Thus $Z = \frac{X - np_0}{\sqrt{np_0 q_0}}$ can also be used as *test-statistic* for testing population proportion p.

4. Critical region:

The critical region depends upon the alternative hypothesis H_1 . The three forms of H_1 are:

- (a) H_1 is $p \neq p_0$. In this case the rejection region is taken in both ends of the sampling distribution. The rejection region on each side is equal to $\alpha/2$. The two critical values $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ separate

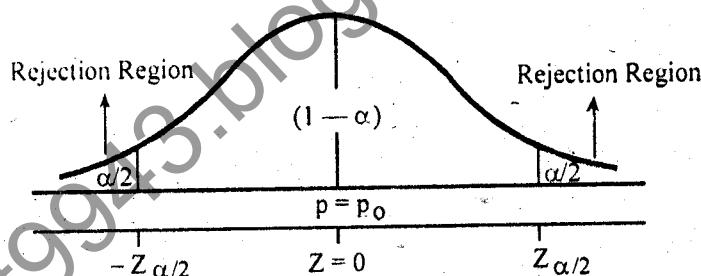


Figure 13.14

the critical region from the acceptance region as shown in Fig. 13.14. H_0 is rejected when the calculated value of Z lies in rejection region. H_0 is rejected when $Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$. The values between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ form the acceptance region. The test is called two - sided.

- (b) H_1 : $p > p_0$. In this case the rejection region is taken only in the right side of the sampling distribution. The test is called one - sided to the right. The critical value between the acceptance region and the rejection

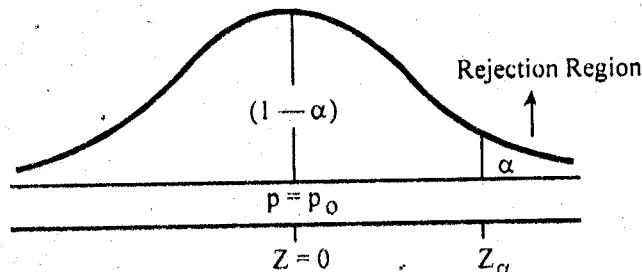


Figure 13.15

region is Z_α as shown in Fig. 13.15. The values above Z_α form the critical region and the values less than Z_α form the acceptance region where as Z_α is the critical value and should not be used for acceptance or rejection of H_0 .

- (c) When H_1 is $p < p_0$, the entire rejection region falls in the left side of the sampling distribution. The test is called one-sided to the left. The critical value $-Z_\alpha$ is a point between the critical region and the acceptance region as

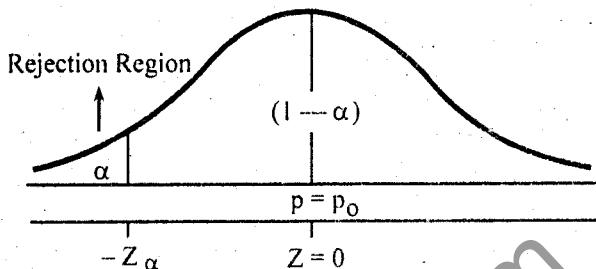


Figure 13.16

shown in Fig. 13.16. The value less than $-Z_\alpha$ form the critical region. H_0 is rejected when the Z value calculated from the sample data falls in the rejection region otherwise the null hypothesis H_0 is accepted with the usual meaning of the term 'acceptance'. The rejection region is $Z < -Z_\alpha$.

5. Computation 6. Conclusion

Example 13.14.

In a poll of 1000 voters selected at random from all the voters in a certain district, it is found that 518 voters are in favour of a particular candidate. Test the null hypothesis that the proportion of all the voters in the district who favour the candidate is equal to or less than 50 percent against the alternative that it is greater than 50 percent at $\alpha = 0.05$.

Solution:

1. Null hypothesis: $H_0: p \leq 0.50$ Alternative hypothesis: $H_1: p > 0.50$

2. Level of significance: $\alpha = 0.05$

3. Test - statistic: $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$

4. Critical region: $Z > 1.645$

(From the area table of normal distribution, we have $Z_\alpha = Z_{0.05} = 1.645$)

5. Computations: Here $n = 1000$, $X = 518$, $\hat{p} = \frac{X}{n} = \frac{518}{1000} = 0.518$

$p_0 = 0.50$, $q_0 = 1 - p_0 = 0.50$, and hence

$$Z = \frac{(0.518 - 0.50)}{\sqrt{\frac{(0.50)(0.50)}{1000}}} = \frac{0.018}{\sqrt{0.016}} = 1.125$$

6. Conclusion:

Since the calculated value of $Z = 1.125$ falls in the acceptance region, so we accept our null hypothesis $H_0: p \leq 0.50$ at 5 % level of significance.

Example 13.15.

At a certain college it is estimated that at most 25 % of the students ride bicycles to class. Does this seem to be a valid estimate, if in a random sample of 90 college students, 28 are found to ride bicycles to class? Use a 5 % level of significance.

Solution:

1. Null hypothesis: $H_0 : p \leq 0.25$ Alternative hypothesis: $H_1 : p > 0.25$
2. Level of significance: $\alpha = 0.05$

3. Test statistic: $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$

4. Critical region: $Z > 1.645$

(From the area table of normal distribution, we have $Z_\alpha = Z_{0.05} = 1.645$)

5. Computations: Here $n = 90$, $X = 28$, $\hat{p} = \frac{X}{n} = \frac{28}{90} = 0.31$,

$p_0 = 0.25$, $q_0 = 1 - p_0 = 0.75$, and hence

$$Z = \frac{0.31 - 0.25}{\sqrt{\frac{(0.25)(0.75)}{90}}} = \frac{0.06}{\sqrt{0.0456}} = 1.32$$

6. Conclusion:

Since the calculated value of $Z = 1.32$ falls in the acceptance region, so we accept our null hypothesis $H_0: p \leq 0.25$ at 5 % level of significance. On the basis of the evidence, we may conclude that at most 25 % of the students ride bicycles to class.

13.17 TEST OF DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS, $p_1 - p_2$ (LARGE SAMPLES)

Suppose there are two binomial populations with proportions p_1 and p_2 which are unknown. Two independent large random samples of sizes n_1 and n_2 are selected from the populations and sample proportion \hat{p}_1 and \hat{p}_2 are calculated. The difference $(\hat{p}_1 - \hat{p}_2)$ is a random variable and has the normal distribution with mean $p_1 - p_2$ and

standard error $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

The procedure for testing of the difference between p_1 and p_2 is given below:

1. Three forms of the hypotheses are as below:
 - (a) $H_0 : p_1 - p_2 = 0$ (or $p_1 = p_2$) and $H_1 : p_1 - p_2 \neq 0$ (or $p_1 \neq p_2$)
 - (b) $H_0 : p_1 - p_2 \leq 0$ (or $p_1 \leq p_2$) and $H_1 : p_1 - p_2 > 0$ (or $p_1 > p_2$)
 - (c) $H_0 : p_1 - p_2 \geq 0$ (or $p_1 \geq p_2$) and $H_1 : p_1 - p_2 < 0$ (or $p_1 < p_2$)
2. Level of significance is decided and is denoted by α .
3. Test statistic:

The random variable Z is used as test statistic where

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

but Z as defined above is only in theory. In actual practice when H_0 is $p_1 - p_2 = 0$ (or $p_1 = p_2$), the values of p_1 , q_1 , p_2 and q_2 are not known because these are all unknown parameters. When H_0 is $p_1 = p_2$, then we assume that the common population proportion for both populations is p_c . This proportion p_c is estimated

by \hat{p}_c by pooling the data from both samples. Thus $\hat{p}_c = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$

Thus the test - statistic used in actual practice is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_c \hat{q}_c}{n_1} + \frac{\hat{p}_c \hat{q}_c}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c \hat{q}_c \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

When H_0 is $p_1 - p_2 = \Delta$ (say), then the test statistic used is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

4. Critical region:

The critical region depends upon the alternative hypothesis H_1 . For three forms of H_1 , the rejection regions are:

- (a) When H_1 is $p_1 - p_2 = 0$ or $p_1 = p_2$, the rejection region is taken in both ends of the sampling distribution. The critical values are $-Z_{\alpha/2}$ and $Z_{\alpha/2}$. The values greater than $Z_{\alpha/2}$ and less than $-Z_{\alpha/2}$ form the rejection region. The values which lie between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ form the acceptance region. H_0 is rejected if $Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$. When H_0 is $p_1 - p_2 = 0$, then it does not make any difference whether we take $(\hat{p}_1 - \hat{p}_2)$ or $(\hat{p}_2 - \hat{p}_1)$ in the test-statistic.
- (b) When H_1 is $p_1 - p_2 > 0$ or $p_1 > p_2$, the entire rejection region is taken in the right side of the curve. It is called one - tailed test to the right. The critical value is Z_α and if Z lies in rejection region the hypothesis $(p_1 - p_2) \leq 0$ or $(p_1 \leq p_2)$ is rejected and $H_1 : p_1 > p_2$ is accepted. It is important to note that if H_1 is $p_2 > p_1$, then the difference $(\hat{p}_2 - \hat{p}_1)$ is used in the test - statistic.

Thus $Z = \frac{(\hat{p}_2 - \hat{p}_1)}{\sqrt{\hat{p}_c \hat{q}_c \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$. The rejection region is $Z > Z_\alpha$.

- (c) When H_1 is $(p_1 - p_2) < 0$ (or $p_1 < p_2$), the rejection region equal to α is taken in the extreme left side. The critical value is $-Z_\alpha$ and the hypothesis $H_0 : (p_1 - p_2) \geq 0$ is rejected and $H_1 : (p_1 - p_2) < 0$ is accepted. The critical region is $Z < -Z_\alpha$.

5. Computation

6. Conclusion

Example 13.16.

The cigarette-manufacturing firm distributes two brands of cigarettes. It is found that 56 of 200 smokers prefer brand 'A' and that 30 of 150 smokers prefer brand 'B'. Test the hypothesis at 0.05 level of significance that brand 'A' outsells brand 'B' by 10% against the alternative hypothesis that the difference is less than 10 %.

Solution:

1. Null hypothesis: $H_0: p_1 - p_2 \geq 0.10$

Alternative hypothesis: $H_1: p_1 - p_2 < 0.10$

2. Level of significance: $\alpha = 0.05$

3. Test - statistic: $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$

4. Critical region: $Z < -1.645$

(From the area table of normal distribution, we have $-Z_{\alpha} = -Z_{0.05} = -1.645$)

5. Computations: Here $n_1 = 200$, $X_1 = 56$ (No. of smokers who prefer brand A),
 $n_2 = 150$, $X_2 = 30$ (No. of smokers who prefer brand B),

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{56}{200} = 0.28, \hat{q}_1 = 1 - \hat{p}_1 = 0.72,$$

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{30}{150} = 0.2, \hat{q}_2 = 1 - \hat{p}_2 = 0.8, \text{ and hence}$$

$$Z = \frac{(0.28 - 0.2) - 0.10}{\sqrt{\frac{(0.28)(0.72)}{200} + \frac{(0.2)(0.8)}{150}}} = \frac{-0.02}{0.0455} = -0.44$$

6. Conclusion: Since the calculated value of $Z = -0.44$ falls in the acceptance region, so we accept our null hypothesis $H_0: p_1 - p_2 \geq 0.10$ at 5 % level of significance and we may conclude that the brand 'A' outsells brand 'B'.

Example 13.17.

A random sample of 150 high school students was asked whether they would turn to their fathers or their mothers for help with a home work assignment in Mathematics and another random sample of 150 high school students was asked the same question with regard to a homework assignment in English. Use the result shown in the following table at the 0.01 level of significance to test whether or not there is a difference between the true proportions of high school students who turn to their fathers rather than their mothers for help in these two subjects:

	Mathematics	English
Mother	59	85
Father	91	65

Solution:

1. Null hypothesis: $H_0: p_1 = p_2$ or $p_1 - p_2 = 0$

Alternative hypothesis: $H_1: p_1 \neq p_2$ or $p_1 - p_2 \neq 0$

2. Level of significance: $\alpha = 0.01$

3. Test - statistic: $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_c \hat{q}_c \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$

4. Critical region: $|Z| > 2.575$ ($Z < -2.575$ and $Z > 2.575$)

(From the area table of normal distribution, we have $Z_{\frac{\alpha}{2}} = Z_{0.005} = 2.575$)

5. Computations: Here $n_1 = 150$, $X_1 = 91$, $n_2 = 150$, $X_2 = 65$,

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{91}{150}, \quad \hat{p}_2 = \frac{X_2}{n_2} = \frac{65}{150}$$

$$\hat{p}_c = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{150 \left(\frac{91}{150} \right) + 150 \left(\frac{65}{150} \right)}{150 + 150} = \frac{91 + 65}{300} = \frac{156}{300} = 0.52,$$

$$\hat{q}_c = 1 - \hat{p}_c = 1 - 0.52 = 0.48, \text{ and hence}$$

$$Z = \frac{\left(\frac{91}{150} - \frac{65}{150} \right) - 0}{\sqrt{(0.52)(0.48) \left(\frac{1}{150} + \frac{1}{150} \right)}} = \frac{\left(\frac{26}{150} \right)}{\sqrt{0.003328}} = \frac{0.1733}{0.0577} = 3.003$$

6. Conclusion: Since the calculated value of $Z = 3.003$ falls in the critical region, so we reject our null hypothesis $H_0: p_1 = p_2$ at 1 % level of significance and we may conclude that there is a difference between the true proportions of high school students.

13.18 CHOICE OF PROPER TEST - STATISTIC

In a certain given situation, we have to choose the proper test-statistic. For example the population mean μ can be tested with the help of Z-test and t-test. The testing of hypotheses along with other things, mainly depends upon the sample size. The sample size plays a major role in the testing of hypothesis. The following table can be used for guidance in choosing the proper test-statistic.

	n - Large	n - Small
σ - Known	Z - test	Z - test
σ - Unknown	Z - test	t - test

SHORT DEFINITIONS**Hypothesis**

A statement about a population parameter developed for the purpose of testing.

or

Hypothesis is a statement which may or may not appear to be true after conclusion.

Hypothesis Testing

The objective of hypothesis testing is to check the validity of a statement about a population parameter.

or

A procedure based on sample evidence and probability theory to determine whether the hypothesis is a reasonable statement or not is called hypothesis testing.

Statistical Hypothesis

A statistical hypothesis is a statement about the numerical value of a population parameter.

or

A statistical hypothesis is a quantitative statement about a population.

Null Hypothesis

A null hypothesis is any hypothesis which is tested for possible rejection or acceptance under the assumption that it is true.

or

The null hypothesis is a statement about the value of a population parameter.

Alternative Hypothesis or Research Hypothesis

The alternative hypothesis is usually the hypothesis for which the researcher wants to gather supporting evidence.

or

A statement specifying that the population parameter is some value other than the one specified under the null hypothesis.

Simple Hypothesis

A hypothesis which specifies all values of parameters of a distribution is called simple hypothesis.

or

A hypothesis is said to be a simple hypothesis if the hypothesis uniquely specifies the distribution from which the sample is taken.

Composite Hypothesis

A hypothesis is said to be a composite hypothesis if it does not completely specify the probability distribution.

or

A hypothesis which does not specify all values of parameters of a distribution is called composite hypothesis.

Significance Level or Level of Significance

The probability of rejecting a true null hypothesis is called the significance level α .

or

The probability of making a type I error is called the significance level of the hypothesis test and is denoted by α (alpha).

Tests of Significance

A significance test is a statistical test laying down the procedure for deciding whether to accept or reject a statistical hypothesis.

Test Statistic

A statistic used as a basis for deciding whether the null hypothesis should be rejected is called test statistic.

or

The sample quantity on which the decision to support H_0 or H_1 is based is called the test statistic.

Rejection Region

The rejection region is the set of possible computed values of the test statistic for which the null hypothesis will be rejected.

or

The set of values for the test statistic that lead to rejection of the null hypothesis H_0 is called rejection region.

Acceptance Region

The set of values for the test statistic that lead to accept the null hypothesis is called acceptance region.

or

The portion of the area under a curve that includes those values of a statistic that lead to acceptance of the null hypothesis.

One-Tailed Test

A statistical test in which the critical region is at one end of sampling distribution is called as one-tailed test.

or

A one-tailed test of hypothesis is one in which the alternative hypothesis is directional, and includes either the symbol " $<$ " or " $>$ ".

Two-Tailed Test

A two-tailed test of hypothesis is one in which the alternative hypothesis does not specify departure from H_0 in a particular direction; such an alternative is written with the symbol " \neq ".

or

A statistical test in which the critical region is located at both ends of sampling distribution is known as two-tailed test.

Critical Value

The value which separates the rejection and acceptance regions is called the critical value of the test statistic.

or

The dividing point between the region where the null hypothesis is rejected and the region where it is accepted is said to be critical value.

Type I Error

If we reject a true null hypothesis, the error is called a type I error.

or

Type I error is the rejection of H_0 when it is true.

Type II Error

If we accept a false null hypothesis, the error is called a type II error.

or

Acceptance of H_0 when it is false is known as type II error.

Power of a Test

The power of a test is the probability of rejecting the null hypothesis when it is false.

or

The power of a test is the probability that the test will lead to a rejection of the null hypothesis H_0 when, in fact, the alternative hypothesis H_1 is true.

Power Curve

A graph of the probability of rejecting H_0 for all possible values of the population parameter not satisfying the null hypothesis is known as power curve.

MULTIPLE - CHOICE QUESTIONS

1. A statement about a population developed for the purpose of testing is called:
 - (a) hypothesis
 - (b) hypothesis testing
 - (c) level of significance
 - (d) test - statistic
2. Any hypothesis which is tested for the purpose of rejection under the assumption that it is true is called:
 - (a) null hypothesis
 - (b) alternative hypothesis
 - (c) statistical hypothesis
 - (d) composite hypothesis
3. A statement about the value of a population parameter is called:
 - (a) null hypothesis
 - (b) alternative hypothesis
 - (c) simple hypothesis
 - (d) composite hypothesis
4. Any statement whose validity is tested on the basis of a sample is called:
 - (a) null hypothesis
 - (b) alternative hypothesis
 - (c) statistical hypothesis
 - (d) simple hypothesis
5. A quantitative statement about a population is called:
 - (a) research hypothesis
 - (b) composite hypothesis
 - (c) simple hypothesis
 - (d) statistical hypothesis
6. A statement that is accepted if the sample data provide sufficient evidence that the null hypothesis is false is called:
 - (a) simple hypothesis
 - (b) composite hypothesis
 - (c) statistical hypothesis
 - (d) alternative hypothesis

7. The alternative hypothesis is also called:
- null hypothesis
 - statistical hypothesis
 - research hypothesis
 - simple hypothesis
8. A hypothesis that specifies all the values of parameter is called:
- simple hypothesis
 - composite hypothesis
 - statistical hypothesis
 - none of the above.
9. The hypothesis $\mu \leq 10$ is a:
- simple hypothesis
 - composite hypothesis
 - alternative hypothesis
 - difficult to tell.
10. If a hypothesis specifies the population distribution is called:
- simple hypothesis
 - composite hypothesis
 - alternative hypothesis
 - none of the above.
11. The probability of rejecting the null hypothesis when it is true is called:
- level of confidence
 - level of significance
 - power of the test
 - difficult to tell
12. The dividing point between the region where the null hypothesis is rejected and the region where it is not rejected is said to be:
- critical region
 - critical value
 - acceptance region
 - significant region
13. If the critical region is located equally in both sides of the sampling distribution of test - statistic, the test is called:
- one tailed
 - two tailed
 - right tailed
 - left tailed
14. The choice of one-tailed test and two-tailed test depends upon:
- null hypothesis
 - alternative hypothesis
 - none of these
 - composite hypothesis
15. A rule or formula that provides a basis for testing a null hypothesis is called:
- test-statistic
 - population statistic
 - both of these
 - none of the above
16. The test statistic is equal to:
- $$\frac{\text{Sample} - \text{Population}}{\text{Standard error}}$$
 - $$\frac{\text{Sample statistic} - \text{Parameter}}{\text{Standard error of the statistic}}$$
 - $$\frac{\text{Sample mean} - \text{Population mean}}{\text{Population standard deviation}}$$
 - $$\frac{\text{Statistic} - E(\text{Statistic})}{\text{Variance of the statistic}}$$
17. $1 - \alpha$ is also called:
- confidence coefficient
 - power of the test
 - size of the test
 - level of significance

18. If H_0 is true and we reject it is called:
- (a) type-I error
 - (b) type-II error
 - (c) standard error
 - (d) sampling error
19. The probability associated with committing type-I error is:
- (a) β
 - (b) α
 - (c) $1 - \beta$
 - (d) $1 - \alpha$
20. $1 - \alpha$ is the probability associated with:
- (a) type-I error
 - (b) type-II error
 - (c) level of confidence
 - (d) level of significance
21. Level of significance is also called:
- (a) power of the test
 - (b) size of the test
 - (c) level of confidence
 - (d) confidence coefficient
22. The probability of rejecting H_0 when it is false is called:
- (a) power of the test
 - (b) size of the test
 - (c) level of confidence
 - (d) confidence coefficient
23. In testing hypothesis $\alpha + \beta$ is always equal to:
- (a) one
 - (b) zero
 - (c) two
 - (d) difficult to tell
24. The significance level is the risk of:
- (a) rejecting H_0 when H_0 is correct
 - (b) rejecting H_0 when H_1 is correct
 - (c) rejecting H_1 when H_1 is correct
 - (d) accepting H_0 when H_0 is correct.
25. An example in a two-sided alternative hypothesis is:
- (a) $H_1 : \mu < 0$
 - (b) $H_1 : \mu > 0$
 - (c) $H_1 : \mu \geq 0$
 - (d) $H_1 : \mu \neq 0$
26. If the magnitude of calculated value of t is less than the tabulated value of t , and H_1 is two-sided, we should:
- (a) reject H_0
 - (b) accept H_1
 - (c) not reject H_0
 - (d) difficult to tell
27. Accepting a null hypothesis H_0 :
- (a) proves that H_0 is true
 - (b) proves that H_0 is false
 - (c) implies that H_0 is likely to be true
 - (d) proves that $\mu \leq 0$.
28. The chance of rejecting a true hypothesis decreases when sample size is:
- (a) decreased
 - (b) increased
 - (c) constant
 - (d) both (a) and (b)
29. The equality condition always appears in:
- (a) null hypothesis
 - (b) simple hypothesis
 - (c) alternative hypothesis
 - (d) both (a) and (b)

30. Which hypothesis is always in an inequality form?
- (a) null hypothesis
 - (b) alternative hypothesis
 - (c) simple hypothesis
 - (d) composite hypothesis
31. Which of the following is not composite hypothesis?
- (a) $\mu \geq \mu_0$
 - (b) $\mu \leq \mu_0$
 - (c) $\mu = \mu_0$
 - (d) $\mu \neq \mu_0$
32. P(Type I error) is equal to:
- (a) $1 - \alpha$
 - (b) $1 - \beta$
 - (c) α
 - (d) β
33. P(Type II error) is equal to:
- (a) α
 - (b) β
 - (c) $1 - \alpha$
 - (d) $1 - \beta$
34. The power of the test is equal to:
- (a) α
 - (b) β
 - (c) $1 - \alpha$
 - (d) $1 - \beta$
35. The degree of confidence is equal to:
- (a) α
 - (b) β
 - (c) $1 - \alpha$
 - (d) $1 - \beta$
36. $\alpha/2$ is called:
- (a) one tailed significance level
 - (b) two tailed significance level
 - (c) left tailed significance level
 - (d) right tailed significance level
37. In an unpaired samples t-test with sample sizes $n_1 = 11$ and $n_2 = 11$, the value of tabulated t should be obtained for:
- (a) 10 degrees of freedom
 - (b) 21 degrees of freedom
 - (c) 22 degrees of freedom
 - (d) 20 degrees of freedom
38. In analyzing the results of an experiment involving seven paired samples, tabulated t should be obtained for:
- (a) 13 degrees of freedom
 - (b) 6 degrees of freedom
 - (c) 12 degrees of freedom
 - (d) 14 degrees of freedom
39. The purpose of statistical inference is:
- (a) to collect sample data and use them to formulate hypotheses about a population
 - (b) to draw conclusion about populations and then collect sample data to support the conclusions
 - (c) to draw conclusions about populations from sample data
 - (d) to draw conclusions about the known value of population parameter.
40. Suppose that the null hypothesis is true and it is rejected, is known as:
- (a) a type-I error, and its probability is β
 - (b) a type-I error, and its probability is α
 - (c) a type-II error, and its probability is α
 - (d) a type-II error, and its probability is β

41. An advertising agency wants to test the hypothesis that the proportion of adults in Pakistan who read a Sunday Magazine is 25 percent. The null hypothesis is that the proportion reading the Sunday Magazine is:
- (a) different from 25 %
 - (b) equal to 25 %
 - (c) less than 25 %
 - (d) more than 25 %
42. If the mean of a particular population is μ_0 , $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$ is distributed:
- (a) as a standard normal variable, if the population is non-normal
 - (b) as a standard normal variable, if the sample is large
 - (c) as a standard normal variable, if population is normal
 - (d) as the t-distribution with $v = n - 1$ degrees of freedom
43. If μ_1 and μ_2 are means of two populations, $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ is distributed:
- (a) as a standard normal variable, if both samples are independent and less than 30
 - (b) as a standard normal variable, if both populations are normal
 - (c) as both (a) and (b) state
 - (d) as the t-distribution with $n_1 + n_2 - 2$ degrees of freedom
44. If the population proportion equals p_0 , then $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$ is distributed:
- (a) as a standard normal variable, if $n > 30$.
 - (b) as a poisson variable
 - (c) as the t-distribution with $v = n - 1$ degrees of freedom
 - (d) as a χ^2 -distribution with v degrees of freedom
45. Given $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$, $\alpha = 0.05$ and we reject H_0 ; the absolute value of the Z-statistic must have equalled or been beyond what value?
- (a) 1.96
 - (b) 1.65
 - (c) 2.58
 - (d) 2.33
46. Given $\mu_0 = 130$, $\bar{X} = 150$, $\sigma = 25$ and $n = 4$; what test statistic is appropriate?
- (a) t
 - (b) z
 - (c) χ^2
 - (d) F

ANSWERS

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

SHORT QUESTIONS

1. Given $\bar{X} = 100$, $\sigma_{\bar{X}} = 16$ and $\mu_0 = 90$. Find Z.

Ans. 0.62

2. Given $\sigma = 80$, $n = 625$, $\mu_0 = 350$ and $\bar{X} = 356$. Find Z.

Ans. 1.88

3. Given $H_0: \mu = 12$, $H_1: \mu > 12$, $n = 64$, $\bar{X} = 15$, $\sigma = 10$ and $\alpha = 0.05$. Find Z and make the statistical decision.

Ans. Z = 2.4, reject H_0

4. Given $H_0: \mu = 150$, $n = 36$, $\bar{X} = 160$, $S = 60$ and $\alpha = 0.05$. Find Z and make the statistical decision.

Ans. Z = 1, accept H_0

5. Given $\bar{X} = 120$, $\mu_0 = 100$, $s = 34.75$ and $n = 25$. Find t.

Ans. 2.88

6. Given $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$, $\alpha = 0.05$, $t = -2.08$ and $n = 26$. Make the statistical decision.

Ans. reject H_0 and assert H_1

7. Given $H_0: \mu = 10$, $H_1: \mu \neq 10$, $n = 16$, $\bar{X} = 10.5$, $s = 0.75$ and $\alpha = 0.05$. Find t and make the statistical decision.

Ans. t = 2.67, reject H_0

8. Given $\sigma_1^2 = 150$, $\sigma_2^2 = 180$, $n_1 = 30$ and $n_2 = 30$. Find $\sigma_{\bar{X}_1 - \bar{X}_2}$.

Ans. 3.32

9. Given $H_0: \mu_1 = \mu_2$, $\bar{X}_1 = 6.53$, $\bar{X}_2 = 4.44$ and $\sigma_{\bar{X}_1 - \bar{X}_2} = 0.78$. Find Z.

Ans. 2.68

10. Given $\bar{X}_1 = 26$, $\bar{X}_2 = 18$, $\sigma_{\bar{X}_1 - \bar{X}_2} = 3.41$, $H_0: \mu_1 \leq \mu_2$ and $\alpha = 0.05$. Find Z and make the statistical decision.

Ans. $Z = 2.35$, reject H_0

11. Given $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $n_1 = 100$, $\bar{X}_1 = 14$, $s_1^2 = 4$, $n_2 = 150$, $\bar{X}_2 = 11$, $s_2^2 = 9$ and $\alpha = 5\%$. Find Z and make the statistical decision.

Ans. $Z = 9.5$, reject H_0

12. Given $H_0: \mu_1 \geq \mu_2$, $H_1: \mu_1 < \mu_2$, $n_1 = 60$, $\bar{X}_1 = 75.6$, $S_1 = 25$, $n_2 = 40$, $\bar{X}_2 = 89.2$, $S_2 = 30$ and $\alpha = 0.05$. Find Z and make the statistical decision.

Ans. $Z = -2.37$, reject H_0

13. Given $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $\bar{X}_1 = 15$, $\bar{X}_2 = 23$, $s_p = 11.48$, $n_1 = 19$, $n_2 = 23$ and $\alpha = 0.05$. Find t and make the statistical decision.

Ans. $t = -2.25$, reject H_0

14. Given $s_1^2 = 1.43$, $s_2^2 = 5.21$, $n_1 = 10$ and $n_2 = 10$. Find s_p .

Ans. 1.82

15. Given $\bar{X}_1 = 84$, $\bar{X}_2 = 77$, $n_1 = 31$, $n_2 = 41$, $H_0: \mu_1 = \mu_2$ and $s_{\bar{X}_1 - \bar{X}_2} = 3.07$. Find t.

Ans. 2.28

16. Given $\sum X_1 = 671$, $\sum X_1^2 = 38275$, $n_1 = 12$, $\sum X_2 = 551$, $\sum X_2^2 = 31707$ and $n_2 = 10$. Find $s_{\bar{X}_1 - \bar{X}_2}$.

Ans. 4.4

17. Given $H_0: \mu_2 - \mu_1 = 10$, $H_1: \mu_2 - \mu_1 > 10$, $n_1 = 10$, $n_2 = 18$, $\bar{X}_1 = 10$, $\bar{X}_2 = 25$, $s_p = 31.68$ and $\alpha = 0.05$. Find t and make the statistical decision.

Ans. $t = 0.40$, accept H_0

18. Given $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $n = 10$, $\bar{d} = -0.5$, $s_d = 3.44$ and $\alpha = 0.05$. Find t and make the statistical decision.

Ans. $t = -0.46$, accept H_0

19. Given $H_0: p = 0.5$, $H_1: p \neq 0.5$, $\hat{p} = 0.54$, $n = 1340$ and $\alpha = 0.02$. Find Z and make the statistical decision.

Ans. $Z = 2.93$, reject H_0

20. Given $H_0: p \geq 0.85$, $H_1: p < 0.85$, $n = 400$, $\hat{p} = 0.81$ and $\alpha = 0.01$. Find Z and make the statistical decision.

Ans. $Z = -2.23$, accept H_0

21. Given $H_0: p_1 = p_2$, $H_1: p_1 \neq p_2$, $\hat{p}_1 = 0.30$, $\hat{p}_2 = 0.25$, $n_1 = 1200$, $n_2 = 900$ and $\alpha = 0.05$. Find Z and make the statistical decision.
Ans: $Z = 2.53$, reject H_0
22. Given $H_0: p_1 - p_2 \geq 0.10$, $H_1: p_1 - p_2 < 0.10$, $n_1 = 200$, $n_2 = 150$, $\hat{p}_1 = 0.28$, $\hat{p}_2 = 0.20$ and $\alpha = 0.05$. Find Z and make the statistical decision.
Ans. $Z = -0.44$, accept H_0
23. Describe the procedure for testing hypothesis about mean of a normal population when population standard deviation is known.
24. Explain the general procedure for testing of hypothesis regarding the population mean when population standard deviation is unknown and the sample size is large.
25. Describe the procedure for testing hypothesis about mean of a normal population when population standard deviation is unknown and the sample size is small.
26. Describe the procedure for testing equality of means of two normal populations when population standard deviations are known and sample sizes are large or small.
27. Describe the procedure for testing equality of means of two normal populations when $\sigma_1 = \sigma_2$ but unknown for small samples.
28. Describe the procedure for testing hypothesis about two means with paired observations.
29. Explain the general procedure for testing of hypothesis regarding the population proportion p for a large sample.
30. Explain the general procedure for testing of hypothesis about the difference between two population proportions for large samples.
31. Distinguish between null hypothesis and alternative hypothesis.
32. Differentiate between type I error and type II error.
33. Differentiate between one-tailed test and two-tailed test.
34. What is meant by critical region?
35. Differentiate between simple hypothesis and composite hypothesis.
36. Differentiate between acceptance region and rejection region.
37. Define null hypothesis and describe the general procedure for its testing.
38. What is meant by test-statistic?
39. Explain the terms hypothesis and tests of hypothesis.
40. Explain the terms level of significance and tests of significance.
41. What is meant by a statistical hypothesis?
42. Explain the difference between one-sided and two-sided tests. When should each be used?
43. Explain with example the difference between acceptance region and rejection region.
44. What is meant by critical value?
45. Define the terms power of a test and power curve.

EXERCISES

1. A sample of 900 plants is found to have a mean of 34 cm. Can it be reasonably regarded as a random sample from a large population with mean 32 cm. and standard deviation 23 cm. Use 5 % level of significance.

Ans. $Z = 2.61$, $H_0: \mu = 32$, $H_1: \mu \neq 32$; reject H_0

2. Suppose that the variance of the IQ'S of the high school students in a certain city is 225. A random sample of 36 students has a mean IQ of 106. If the level of significance is chosen at 0.05, should we conclude that the IQS of the high school students in this city are higher than 100?

Ans. $H_0: \mu \leq 100$, $H_1: \mu > 100$; $Z = 2.4$; reject H_0

3. Suppose that scores on an aptitude test used for determining admission to graduate study in statistics are known to be normally distributed with a mean of 500 and a population standard deviation of 100. If a random sample of 64 applicants from a college has a sample mean of 537, is there any evidence that their mean score is different from the mean expected of all applicants? Use $\alpha = 0.01$.

Ans. $H_0: \mu = 500$, $H_1: \mu \neq 500$, $Z = 2.96$; reject H_0

4. Let $X \sim N(\mu, 100)$ and \bar{X} be the mean of a random sample of 64 observations of X , given that $\bar{X} = 15$. Test $H_0: \mu = 12$ against the alternative $H_1: \mu > 12$. Use $\alpha = 0.05$.

Ans. $Z = 2.4$; reject H_0

5. A random sample of 64 drinks from a soft-drink machine has an average content of 21.9 deciliters, with a standard deviation of 1.42 deciliters. Test the hypothesis that $\mu = 22.2$ deciliters against the alternative hypothesis $\mu < 22.2$, at the 5 % level of significance.

Ans. $Z = -1.69$; reject H_0

6. A random sample of 200 trucks were driven on the average 16300 miles a year with a sample standard deviation of 3100 miles. Test the null hypothesis that the average truck mileage in the population is 17000 miles a year against the alternative hypothesis that the average is less. Use the 5 % level of significance.

Ans. $H_0: \mu = 17000$, $H_1: \mu < 17000$, $Z = -3.19$; reject H_0

7. A manufacturer of detergent claims that the mean weight of a particular box of detergent is 3.25 pounds. A random sample of 64 boxes revealed a sample average of 3.238 pounds with a standard deviation of 0.117 pounds. Using the 1 % level of significance, is there evidence that the average weight of the boxes is different from 3.25 pounds?

Ans. $H_0: \mu = 3.25$, $H_1: \mu \neq 3.25$, $Z = -0.82$; accept H_0

8. Past experience indicates that the time for high school seniors to complete a standardized test is a normal random variable with a mean of 35 minutes. If a random sample of 20 high school seniors took an average of 33.1 minutes to complete this test with a standard deviation $s = 4.3$ minutes, test the hypothesis at the 1 % level of significance that $\mu = 35$ minutes against the alternative that $\mu < 35$ minutes.

Ans. $t = -1.976$; accept H_0

9. A random sample of 10 from a population gave $\bar{X} = 20$ and sum of square of deviations from mean is 144 test $H_0: \mu = 19.5$ against $H_1: \mu > 19.5$. At $\alpha = 0.05$.

Ans. $t = 0.395$; accept H_0

10. Given the following information. What is your conclusion in testing each of the indicated null and alternative hypotheses?

	n	\bar{X}	s^2	α	H_0	H_1
(i)	25	8	64	0.05	$\mu \geq 10$	$\mu < 10$
(ii)	9	12	36	0.01	$\mu = 10$	$\mu > 10$
(iii)	16	13	64	0.05	$\mu = 10$	$\mu \neq 10$

Ans. (i) $t = -1.25$; accept H_0 (ii) $t = 1$; accept H_0 ; (iii) $t = 1.5$; accept H_0

11. Suppose you wish to estimate the difference between the daily wages for machinists and carpenters. Two independent samples of 50 people each are respectively taken, and the relevant data are shown as follows:

	Machinists	Carpenters
Sample Size	50	50
Sample mean	172.5	170.0
Population variance	98	102

Should we reject the null hypothesis that the daily wages for machinists and carpenters are the same in favor of the alternative hypothesis that they are different at $\alpha = 0.05$.

Ans. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $Z = 1.25$; accept H_0

12. A random sample of 100 workers in a large farm took an average of 14 minutes to complete a task. A random sample of 150 workers in another large farm took an average of 11 minutes to complete the task. Can it be assumed at 5 % level of significance that the average time taken by the workers in the two farms is same, if the standard deviations of all the workers of first farm and second farm are 2 minutes and 3 minutes respectively.

Ans. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $Z = 9.49$; reject H_0

13. A tire manufacturer wishes to test two types of tires. Fifty tires of type 'A' has a mean life of 24000 miles with $S^2 = 6250000$. Forty tires of type 'B' has a mean life of 26000 miles with $S^2 = 9000000$. Is there a significant difference between the two sample means? Use $\alpha = 0.05$.

Ans. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $Z = -3.38$; reject H_0

14. A carpet manufacturer is studying differences between two of its major outlet stores. The company is particularly interested in the time it takes before customers receive carpeting that has been ordered from the plant. Data concerning a sample of delivery times for the most popular type of carpet are summarized as follows:

	A	B
\bar{X}	34.3 days	43.7 days
s	2.4 days	3.1 days
n	41	31

At the 0.01 level of significance, is there evidence of a difference in the average delivery times for the two outlet stores?

Ans. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $Z = -14$, reject H_0

15. Two random samples taken independently from normal populations with an identical variance yield the following results:

	Sample I	Sample II
Size	$n_1 = 10$	$n_2 = 18$
Mean	$\bar{X}_1 = 10$	$\bar{X}_2 = 25$
Variance	$s_1^2 = 1200$	$s_2^2 = 900$

Test the hypothesis that the true difference between the population means is 10, that is, $H_0: \mu_2 - \mu_1 = 10$, against the alternative $H_1: \mu_2 - \mu_1 > 10$ at the 5 % level of significance.

Ans. $t = 0.40$; accept H_0

16. The means of two random samples of sizes 9 and 7 respectively are 196.42 and 198.82 respectively. The sums of the squares of the deviation from the mean are 26.94 and 18.73 respectively. Assume that the two samples are drawn from normal populations with identical variance. Test $H_0: \mu_1 = \mu_2$, against the alternative $H_1: \mu_1 < \mu_2$ at the 5 % level of significance.

Ans. $t = -2.631$; reject H_0

17. In an examination, a class of 18 students had a mean of 70 with $s = 6$. Another class of 21 had a mean of 77 with $s = 8$ in the same examination. Is there reason to believe that one class is significantly better than the other? Consider the students as samples from one population. Use a 5 % level of significance.

Ans. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $t = -3.05$; reject H_0

18. The weights of 4 persons before they stopped smoking and 5 weeks after they stopped smoking are as follows:

Person	1	2	3	4
Before	148	176	153	118
After	154	176	150	120

Use the t-test for paired observations to test the hypothesis at 0.05 level of significance that giving up smoking has no effect on a person's weight.

Ans. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $t = -0.662$; accept H_0

19. An experiment was performed with five hop plants. One half of each plant was pollinated and the other half was non-pollinated. The yield of the seed of each hop plant is tabulated as follows:

Pollinated	0.78	0.76	0.43	0.92	0.86
Non-pollinated	0.21	0.12	0.32	0.29	0.30

Determine at the 5 % level of significance whether the pollinated half of the plant gives a higher yield in seed than the non-pollinated half.

Ans. $H_0: \mu_1 \leq \mu_2$, $H_1: \mu_1 > \mu_2$, $t = 5.102$; reject H_0

20. Let X designate the defective parts produced by an automatic machine. From a randomly selected sample of 50 parts, 10 are defective. Let p be the true proportion of all the parts that are defective; test the null hypothesis $H_0: p = 0.1$ against the alternative hypothesis $H_1: p \neq 0.1$ at $\alpha = 0.01$.

Ans. $Z = 2.358$; accept H_0

21. A coin is tossed 20 times resulting in 5 heads. Is this sufficient evidence to reject the hypothesis at the 5 % level of significance that the coin is balanced in favour of the alternative that heads occur less than 50 % of the times?

Ans. $H_0: p = 0.5$, $H_1: p < 0.5$, $Z = -2.236$; reject H_0

22. A random sample of 200 workers was selected from a population and 140 workers were found to be skilled. The factory owner claimed that at least 80 % workers were skilled in his factory. Is it possible to reject the claim of the factory owner at 5 % level of significance.

Ans. $H_0: p \geq 0.80$, $H_1: p < 0.80$, $Z = -3.534$; reject H_0

23. An electric company claimed that at least 85 % of the parts which it supplied conformed to specifications. A sample of 400 parts was tested and 75 did not meet specifications. Can we accept the company's claim at 1 % level of significance?

Ans. $H_0: p \geq 0.85$, $H_1: p < 0.85$, $Z = -2.095$; accept H_0

24. An expert is interested in the proportion of males and females in a population that have a certain minor blood disorder. In a random sample of 100 males, 31 are found to be afflicted whereas only 24 of 100 females tested appear to have the disorder. Can we conclude at the 5 % level of significance that the proportion of men in the population afflicted with this blood disorder is significantly greater than the proportion of women afflicted?

Ans. $H_0: p_1 \leq p_2$, $H_1: p_1 > p_2$, $Z = 1.109$; accept H_0

25. Studies comparing the mathematical abilities of male and female students have produced conflicting conclusions. The distribution of grades earned in introductory statistics by a random sample of students at one institution is given below. Use these data to test the hypothesis that there is no difference in the population proportion of males and females that receive grade A. Let $\alpha = 0.05$.

Grade	A	B	C	D	E	Total
Males	15	18	16	8	11	68
Females	20	16	19	12	15	82

Ans. $H_0: p_1 = p_2$, $H_1: p_1 \neq p_2$, $Z = -0.331$; accept H_0

REGRESSION AND CORRELATION

14.1 INTRODUCTION

There are some statistical tools with the help of which we study a single variable. The averages, the measures of dispersion, the moments etc. are calculated for the frequency distribution of a single variable. There are certain tools with the help of which two or more than two variables or attributes are studied. What do we study when there are two or more than two variables or attributes. In Chapter association, we shall discuss mutual relationship between qualitative variables. The qualitative variables are also called attributes. The attributes are studied by a statistical tool χ^2 (read as chi-square). In the present chapter and in the next chapter we shall discuss the tools which are used for the study of two variables. Cases of more than two variables are beyond the level of this book, and therefore will not be covered in this book. There are two different techniques which are used for the study of two or more than two variables. These are *regression* and *correlation*. Both study the behaviour of the variables but they differ in their end results. Regression studies the relationship where *dependence* is necessarily involved. One variable has the dependence on a certain number of variables. Regression can be used for predicting the values of the variable which depends upon other variables. Correlation attempts to study the strength of the mutual relationship between two variables. In correlation we assume that the variables are random and dependence of any nature is not involved.

14.2 MATHEMATICAL MODEL OR EQUATION

Regression involves the study of equations. First we talk about some simple equations or models. The simplest mathematical model or equation is the equation of straight line.

Example 14.1.

Suppose a shop-keeper is selling pencils. He sells one pencil for Rs. 2. Table 14.1. gives the number of pencils sold and the sale price of the pencils.

Table 14.1.

Number of pencils sold	0	1	2	3	4	5
Sale price (Rs.)	0	2	4	6	8	10

Let us examine the two variables given in Table 14.1. For the sake of our convenience, we can give some names to the variables given in the table. Let X

denote the number of pencils sold and S (S for sale) denote the amount realised by selling X pencils. Thus,

X	0	1	2	3	4	5
S	0	2	4	6	8	10

The information written above can be presented in some other forms as well. For example we can write an equation describing the above relation between X and S. It is very simple to write the equation. The algebraic equation connecting X and S is, $S = 2X$.

It is called mathematical equation or mathematical model in which S depends upon X. Here X is called independent variable and S is called dependent variable. There is exact relation between X and S. When 2 pencils are sold, the sale price is Rs. 4. Neither less than 4 nor more than 4. The above model is called deterministic mathematical model because we can determine the value of S without any error by putting the value of X in the equation. The sale S is said to be function of X. This statement in symbolic form is written as: $S = f(X)$

It is read as 'S is function of X'. It means that S depends upon X and only X and no other element. The data in Table 14.1 can be presented in the form of a graph as shown in figure 14.1.

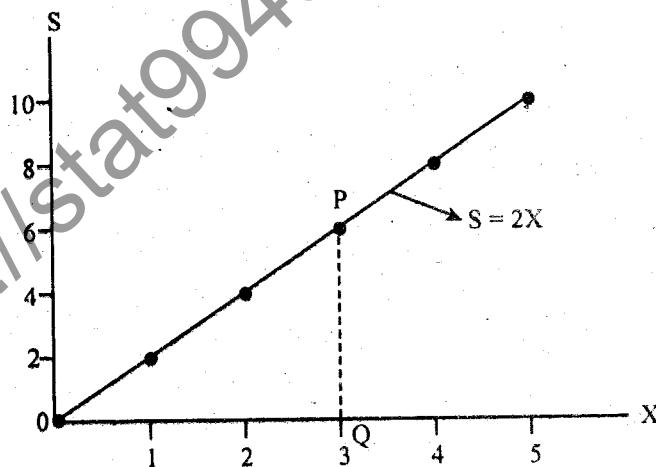


Figure 14.1

The main features of the graph in figure 14.1. are:

- (i) The graph lies in the first quadrant because all the values of X and S are positive.
- (ii) It is an exact straight line. But all graphs are not in the form of a straight line. It could be some curve also.
- (iii) All the points (pairs of X and S) lie on the straight line.
- (iv) The line passes through the origin.

- (v) Take any point P on the line and draw a perpendicular line PQ which joins P with the X-axis. Let us find the ratio $\frac{PQ}{OQ}$. Here $PQ = 6$ units and $OQ = 3$ units. Thus $\frac{PQ}{OQ} = \frac{6}{3} = 2$ units.

It is called the slope of the line and in general it is denoted by 'b'. The slope of the line is the same at all points on the line. The slope 'b' is equal to the change in Y for a unit change in X. The relation $S = 2X$ is also called *linear equation* between X and S.

Example 14.2.

Suppose a carpenter wants to make some wooden toys for the small children. He has purchased some wood and some other material for Rs. 20. The cost of making each toy is Rs. 5. Table 14.2. gives the information about the number of toys made and the cost of the toys.

Number of Toys	0	1	2	3	4	5
Cost of Toys	20	25	30	35	40	45

Let X denote the number of toys and Y denote the cost of the toys. What is the algebraic relation between X and Y. When $X = 0$, $Y = 20$. This is called fixed or starting cost and it may be denoted by 'a'. For each additional toy, the cost is Rs. 5. Thus Y and X are connected through the following equation:

$$Y = 20 + 5X$$

It is called equation of straight line. It is also mathematical model of deterministic nature. Let us make the graph of the data in Table 14.2. Figure 14.2. is the graph of the data in Table 14.2.

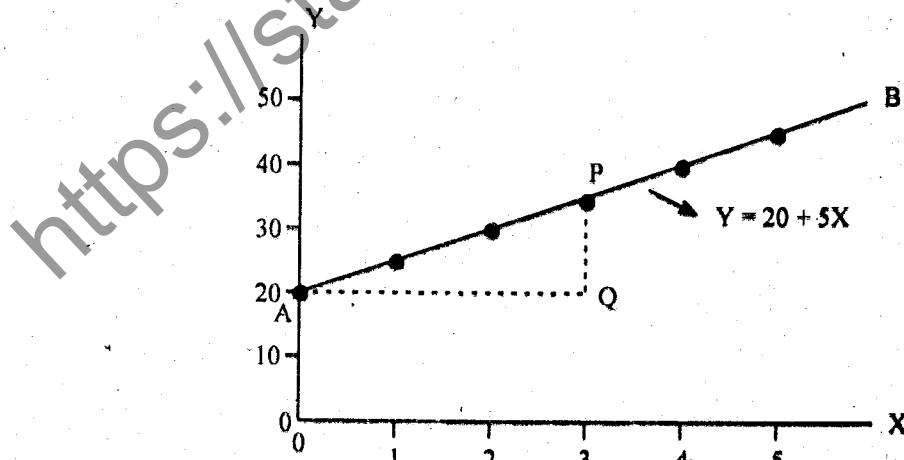


Figure 14.2

Let us note some important features of the graph obtained in figure 14.2.

- (i) The line AB does not pass through the origin. It passes through the point 'A' on Y-axis. The distance between A and the origin 'O' is called the '*intercept*' and is usually denoted by 'a'.

- (ii) Take any point P on the line and complete a triangle PQA as shown in the figure. Let us find the ratio between the perpendicular PQ and the base AQ of this triangle. The ratio is, $\frac{PQ}{AQ} = \frac{15}{3} = 5$ units.

This ratio is denoted by 'b' in the equation of straight line. Thus the equation of straight line $Y = 20 + 5X$ has the intercept $a = 20$ and slope $b = 5$. In general, when the values of *intercept* and *slope* are not known, we write the equation of straight line as $Y = a + bX$. It is also called *linear* equation between X and Y and the relation between X and Y is called *linear*. The equation $Y = a + bX$ may also be called exact linear model between X and Y or simply linear model between X and Y. The value of Y can be determined completely when X is given. The relation $Y = a + bX$ is therefore, called the deterministic linear model between X and Y. In statistics, when we shall use the term 'linear model', we shall not mean a mathematical model as described above.

Another property of the exact linear model is that the 1st differences of Y-variable are zero. The first differences of the Y-variable in Table 14.2. are calculated as below:

X	Y	First differences ΔY
0	20	$25 - 20 = 5$
1	25	$30 - 25 = 5$
2	30	$35 - 30 = 5$
3	35	$40 - 35 = 5$
4	40	$45 - 40 = 5$
5	45	

It means that when all the points of the pairs (X_i, Y_i) lie on the straight line, the first differences ΔY are exactly constant. We shall take help from this property later on. In a certain observed data, when the first differences will be constant or almost constant, we shall consider the observed data to be close to a straight line and we would like to find the equation of that line.

14.3 NON-LINEAR MODEL

Let us consider an equation $Y = 10 + 5X^2$

By putting the values of $X = 0, 1, 2, 3, 4$, in this equation, we find the values of Y as given in Table 14.3 below. The first and second differences are calculated in Table 14.3.

Table 14.3.

X	Y	First differences ΔY	Second differences $\Delta^2 Y$
0	10		
1	15	$15 - 10 = 5$	
2	30	$30 - 15 = 15$	$15 - 5 = 10$
3	55	$55 - 30 = 25$	$25 - 15 = 10$
4	90	$90 - 55 = 35$	$35 - 25 = 10$

The second differences are exactly constant. The general quadratic equation or model is written as

$$Y = a + bX + cX^2 \quad (c \neq 0)$$

It is also called second degree parabola or second degree curve. The graph of the data is shown below in figure 14.3.

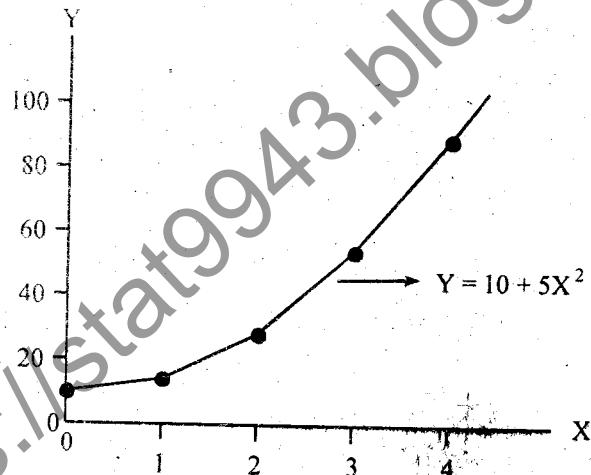


Figure 14.3

Figure 14.3 is not a straight line. It is a curve or we say that the model $Y = 10 + 5X^2$ is non-linear. The students are advised to remember that if in a certain observed data, the second differences are constant or almost constant, we find the second degree curve close to the observed data. We shall face this type of situation in time series.

14.4 STATISTICAL MODEL

Statistical model is also a mathematical model but the difference is that statistical model always contains an error term or random term in the right side of the mathematical equation. What is an error term? Let us take an example of practical life to explain this term.

Suppose there are 10 agricultural plots of the same size and the same fertility. It is assumed that the plots are similar in all aspects. The irrigation and dressing are to remain as constant as possible from plot to plot. The same seed of rice is used

in the plots. We decide to put 5 kg. of fertilizer in each plot. The yields of rice from each plot are recorded. Let X denote the amount of fertilizer and Y denote the yield of rice. For a single fixed value of X there are corresponding 10 figures of yields of plots. If 5 kg. of fertilizer is applied in very large number of plots, then the yields will form a normal distribution with some mean denoted by $\mu_{y/x}$. The mean $\mu_{y/x}$ is the mean of Y values when X is fixed at 5 kg. This mean is also denoted by $E(Y)$. Some yields are above $E(Y)$ and some are below $E(Y)$. The difference between the yield, Y and $E(Y)$ is called the error term or the random term. It is also called the residual. These residuals may be denoted by e_i . The yields of rice are a random variable with a certain probability distribution. The random errors are calculated from the random variable Y_i . A random variable calculated from another random variable is also a random variable. Thus the errors e_i are the random variable and it is a well known fact they e_i 's are normally distributed with mean zero. Thus $E(e_i) = 0$. Table 14.4. shows certain yields of rice Y_i for a given value of X and the mean $E(Y)$ and the errors e_i are calculated as below:

Table 14.4.

Amount of fertilizer, X	Yield of rice (kgs.), Y_i	Average	Error $e_i = Y_i - E(Y)$
5 kg.	40		$40 - 55 = - 15$
	40		$40 - 55 = - 15$
	50		$50 - 55 = - 5$
	50		$50 - 55 = - 5$
	50	$E(Y) = 55 \text{ kg}$	$50 - 55 = - 5$
	60		$60 - 55 = 5$
	60		$60 - 55 = 5$
	60		$60 - 55 = 5$
	70		$70 - 55 = 15$
	70		$70 - 55 = 15$
			$\sum e_i = 0$

In Table 14.4 we have taken only 10 values of Y_i . In actual practice the number of Y -values are very large corresponding to a fixed value of X . In this table we observe that there is no equation which links the X value with the given Y values.

There is in fact no relation of mathematical nature between X and individual values of Y. The individual values of Y cannot be determined by any mathematical equation. If we change the amount of fertilizer, we shall obtain another set of Y values (distribution of Y values) for the different yields of rice. Thus for each value of X, there is a normal distribution of Y values. This fact is illustrated in figure 14.4.

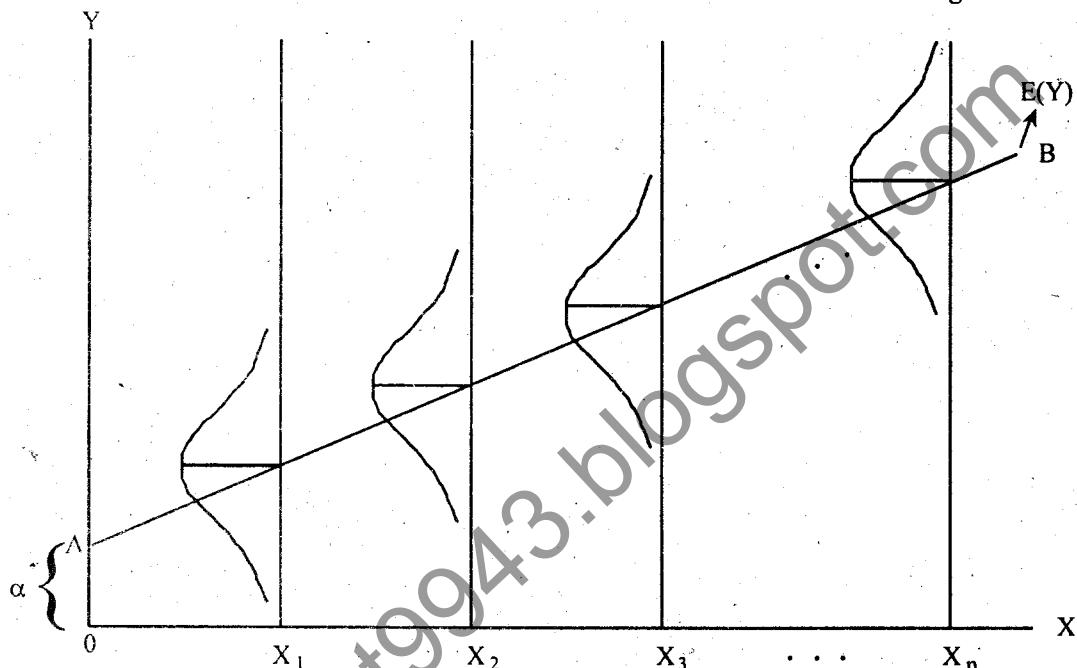


Figure 14.4

On each value of X, there is a normal distribution with mean $E(Y)$. The Y-values in the same distribution differ from their mean $E(Y)$ and the difference is called error term. If a population data on two variables X and Y is under consideration, then a linear statistical model or equation can be written as:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where α is the intercept, β is the slope of the line and ϵ_i (epsilon) is the error term and it may take positive or negative values. The line AB in figure 14.4. which passes through the $E(Y)$'s is called the regression line. The observed value Y_i can also be written as $Y_i = E(Y) + \epsilon_i$

This equation contains a random term ϵ_i on the right side. Thus the variable Y_i is random because it depends on ϵ_i .

14.4.1 INDEPENDENT AND DEPENDENT VARIABLES

The value which is decided by the experimenter is called fixed variable or independent variable. It is also called regressor or predictor. The variable which is influenced by the independent variable is called dependent variable. It is also called regressand or predictand. This variable is of random nature and cannot be determined exactly for a given value of X. It is also called random variable.

14.4.2 CAUSE AND EFFECT RELATION

In a relation, in which one variable is independent and the other is dependent, some people use the terms 'cause' and 'effect'. In the previous example of production of rice for a given dosage of fertilizer, the amount of fertilizer is the 'cause' and 'production of rice' is the 'effect'. Thus in this regression relation, we can say that there is 'cause' and 'effect' relation between the variables. Some special food may be tested on poultry birds. The amount of food is 'cause' and the weight of the birds is an 'effect'. The 'effect' variable is also called the response variable. But there may be a regression relation between two variables X and Y in which there is no cause and effect (causal) relationship between them. In some cases a change in X does cause a change in Y but it does not happen always. Sometimes the change in Y is not caused by change in X. The dependence of Y on X should not be interpreted as cause and effect relation between X and Y. In regression analysis the word dependence means that there is a distribution of Y values for a given single value of X. For a given height of 60 inches for men, there may be very large number of people with different weights. The distribution of these weights depends upon the fixed value of X. It is in this sense that the word dependence is used. Thus dependence does not mean response (effect) due to some cause. Some examples are discussed here to elaborate the idea.

- (i) The sun rises and the shining sun increases the temperature. Let temperature be denoted by X. With increase in X, the ice on the mountains melts and the average thickness of ice Y_i decreases. It is possible that the thickness of ice decreases due to increase in temperature. But this is also possible that the thickness of ice is decreasing due to weight and hardening of ice. We may be regressing the thickness Y against the temperature X only whereas another important factor is being ignored. In this type of problem, more than one regression equations are developed and then the equations are solved simultaneously to estimate the unknown parameters.
- (ii) We may think that increase in the number of workers (X) is increasing the production of fans (Y) in the factory. The increase in Y may be due to change in the administration and some changes about the leave rules and other benefits.

In a regression relation there may or may not be a causal relation between X and Y. The cause and effect relation between two variables is also called causation. It is important to note that the statistical method of regression analysis is silent about the cause and effect relation between the variables. Sometimes it is not possible to identify as to which variable is 'cause' and which one is 'effect'. In fact, the answer is to be searched not in regression analysis but in some other area of relationship between the variables.

14.5 REGRESSION

Regression is concerned with the study of relationships among variables. The aim of regression (or regression analysis) is to make models for prediction and for making other inferences. Two variables or more than two variables may be treated by regression. Historically, the word "regression" was first time used by a British

scientist, Sir Francis Galton, who analyzed the heights of sons and the average heights of their parents. Galton concluded that the sons of very tall (or short) parents were generally taller (or shorter) than the average but not as tall (or short) as their parents. His work was published in 1885 under the title "Regression Toward Mediocrity in Hereditary Stature". According to his conclusion, "regression towards mediocrity" means that the sons heights tended towards the average rather than take the extreme values. But now the word regression is used in much broader sense. It is the statistical study of the relationship among variables.

14.5.1 SIMPLE LINEAR REGRESSION

Suppose we want to study the dependence of Y variable on a single independent variable X. The variable Y depends on X and is also subject to unaccountable errors. This study is covered by simple linear regression. For a population data the simple linear regression model is written as $Y_i = \alpha + \beta X_i + \epsilon_i$ where α is the intercept, β is the slope and ϵ_i (epsilon) is the error term and on sample basis the simple linear regression model is written as $Y_i = a + b X_i + e_i$ where 'a' is the intercept in the sample and is the estimate of the population parameter α . The parameter β is estimated by the sample value 'b' and e_i is the error term in the equation:

14.5.2 PURPOSE OF REGRESSION ANALYSIS

There is no statistical problem when the parameters of the regression model are known. Statistical problem arises when some of the parameters are not known. The study of regression aims at:

- The regression models contain the unknown parameters. These parameters are estimated in regression analysis.
- The value of the dependent variable can be predicted when the value of the independent variable is fixed.
- Certain hypotheses about the parameters α and β are tested. Confidence intervals for α and β are constructed.

14.5.3 SCATTER DIAGRAM

Scatter diagram is a graphic picture of the sample data. Suppose a random sample of n pairs of observations has the values $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n)$. These points are plotted on a rectangular co-ordinate system taking independent variable on X-axis and the dependent variable on Y-axis. Whatever be the name of the independent variable, it is to be taken on X-axis. Suppose the plotted points are as shown in figure 14.5(a). Such a diagram is called *scatter diagram*. In this figure, we see that when X has a small value, Y is also small and when X takes a large value, Y also takes a large value. This is called direct or positive relationship between X and Y. The plotted points cluster around a straight line. It appears that if a straight line is drawn passing through the points, the line will be a good approximation for representing the original data. Suppose we draw a line AB to represent the scattered points. The line AB rises from left to the right and has positive slope. This line can be used to establish an approximate relation

between the random variable Y and the independent variable X. It is non-mathematical method in the sense that different persons may draw different lines. This line is called the regression line obtained by inspection or judgement.

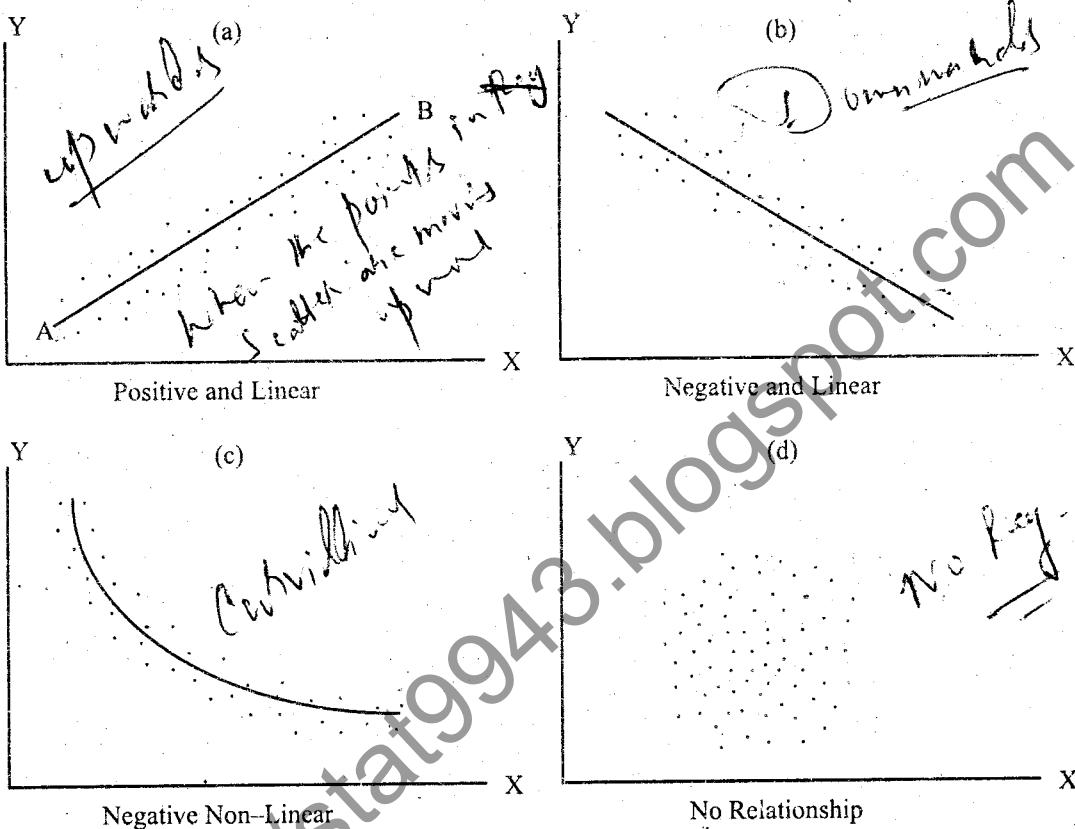


Figure 14.5 .

Making a *scatter diagram* and drawing a line or curve is the primary investigation to assess the type of relationship between the variables. The knowledge gained from the scatter diagram can be used for further analysis of the data. In most of the cases the diagrams are not as simple as in figure 14.5 (a). There are quite complicated diagrams and it is difficult to choose a proper mathematical model for representing the original data. The scatter diagram gives an indication of the appropriate model which should be used for further analysis with the help of method of least squares. Figure 14.5. (b) shows that the points in the scatter diagram are falling from the top left corner to the right. This is a relation called inverse or indirect. The points are in the neighbourhood of a certain line called the regression line.

As long as the scattered points show a closeness to a straight line of some direction, we draw a straight line to represent the sample data. But when the points do not lie around a straight line, we do not draw the regression line. Figure 14.5. (c) shows that the plotted points have a tendency to fall from left to right in the form of

a curve. This is a relation called non-linear or curvilinear. This type of relations will not be discussed in this book.

Figure 14.5 (d) shows the points which apparently do not follow any pattern. If X takes a small value, Y may take a small or large value. There seems to be no sympathy between X and Y . Such a diagram suggests that there is no relationship between the two variables. But there is one point to be remembered that the figures like 14.5 (d) have sometimes the relationship of circular nature, something which cannot be elaborated here.

14.6 FITTING A LINEAR REGRESSION LINE - THE METHOD OF LEAST SQUARES

The linear regression line is $Y = \alpha + \beta X$ which contains the parameters α and β . If we know the values of α and β , then this line is determined. But the values of α and β are usually unknown and we have to find the regression line by estimating α and β from the sample data. Suppose we have a random sample of n pairs of observations $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n)$ and we are required to find the regression line of Y on X . These observations are plotted in figure 14.6. Let us draw a line AB passing through the plotted points. The values of the dependent variable which lie on the line are denoted by \hat{Y} . Thus the estimated regression line of sample data is $\hat{Y} = a + bX$ where 'a' and 'b' represent the estimates of the true values of α and β .

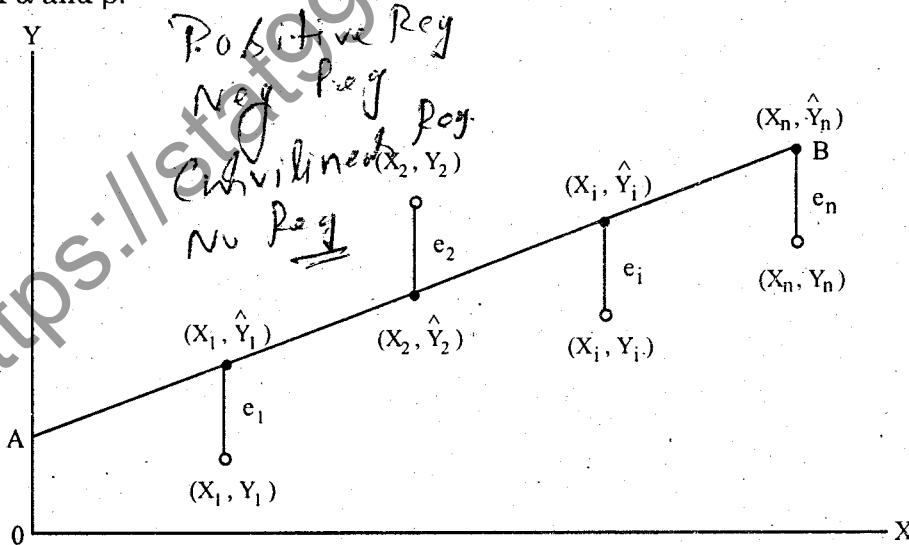


Figure 14.6

The difference between Y_i and \hat{Y}_i is called the error which is denoted by e_i . The sum of squares of errors (SSE) can be written as

$$\begin{aligned} SSE &= (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \dots + (Y_i - \hat{Y}_i)^2 + \dots + (Y_n - \hat{Y}_n)^2 \\ &= e_1^2 + e_2^2 + \dots + e_i^2 + \dots + e_n^2 = \sum e_i^2 \end{aligned}$$

We have to find that line which is *best fitting* for the sample data. This *best fitting* line is obtained by using the principle of least squares. The principle of least squares is that "the best fitting line is that one for which the sum of squares of errors is minimum". This means we have to minimize

$$SSE = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 \quad \text{but } \hat{Y}_i = a + bX_i \text{ and } SSE = \sum (Y_i - a - bX_i)^2$$

Thus SSE is a function of 'a' and 'b'. Each line has some values of 'a' and 'b'. Those values of 'a' and 'b' are required for which SSE is minimum. The values of 'a' and 'b' are calculated from the following two equations called the normal equations:

$$\Sigma Y = na + b \Sigma X \quad \text{and} \quad \Sigma XY = a \Sigma X + b \Sigma X^2$$

Solving these normal equations simultaneously, we get the values of 'a' and 'b' which minimize SSE. We can calculate the values of 'a' and 'b' by using the formulas as below:

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

If the numerator and denominator are multiplied with n, this is convenient for computational purposes.

$$\text{we get } b = \frac{n \Sigma XY - (\Sigma X)(\Sigma Y)}{n \Sigma X^2 - (\Sigma X)^2} \quad \text{and} \quad a = \frac{(\Sigma X^2)(\Sigma Y) - (\Sigma X)(\Sigma XY)}{n \Sigma X^2 - (\Sigma X)^2}$$

The slope 'b' is also called the regression coefficient of Y on X and is denoted by b_{yx} . Putting the values of 'a' and 'b' in the regression equation $\hat{Y} = a + bX$, we can find the \hat{Y} values which lie on the regression line. The \hat{Y} values are called the estimated values. The linear regression equation $\hat{Y} = a + bX$ can be used to estimate the values of the dependent variable when the value (or values) of X is known. The calculated values of 'a' and 'b' are the estimates of the unknown parameters α and β and are used for inference about α and β . The inference about α and β will not be discussed in this book.

How to Write Normal Equations

The derivation of the normal equations is beyond the level of this book. However the normal equations can be written directly as explained below:

We write the equation of straight line i.e., $Y = a + bX \dots \dots \dots (1)$

We want to write the normal equation of 'a'. The coefficient of 'a' is 1 and the above equation is multiplied with 1 and then summation Σ is applied. Thus we get

$$\Sigma Y = na + b \Sigma X \quad (a + a + \dots + a = \Sigma a = na)$$

This is called normal equation for 'a'. To find the normal equation for 'b' the equation (1) is multiplied with X which is the coefficient of b in equation (1) and then summation Σ is applied. We get

$$\Sigma XY = a \Sigma X + b \Sigma X^2$$

This is called the normal equation for 'b'.

Example 14.3.

Construct an equation of the line of regression (using normal equations) of yield of rice on water from the data given in the following table which shows the amount of water applied in inches and the yield of rice in tons per acre in an experimental farm. Estimate the most probable yield of rice of 36 inches of water.

Water (X)	10	16	22	28	34	40	46
Yield of rice (Y)	2.25	2.85	2.95	3.15	3.40	3.80	4.00

Solution:

The regression line of yield of rice (Y) on water (X) is $\hat{Y} = a + bX$

The normal equations are: $\Sigma Y = na + b \Sigma X$ and $\Sigma XY = a \Sigma X + b \Sigma X^2$

The necessary calculations are given below:

X	Y	XY	X^2
10	2.25	22.5	100
16	2.85	45.6	256
22	2.95	64.9	484
28	3.15	88.2	784
34	3.40	115.6	1156
40	3.80	152.0	1600
46	4.00	184.0	2116
$\Sigma X = 196$	$\Sigma Y = 22.4$	$\Sigma XY = 672.8$	$\Sigma X^2 = 6496$

Substituting the values in the normal equations, we have

$$22.4 = 7a + 196b \dots\dots (1) \quad 672.8 = 196a + 6496b \dots\dots (2)$$

Solving these two equations, we multiply equation (1) by 28 and subtract from equation (2), we get

$$672.8 = 196a + 6496b$$

$$627.2 = 196a + 5488b$$

$$45.6 = 1008b \text{ or } b = \frac{45.6}{1008} = 0.05$$

Substituting $b = 0.05$ in equation (1), we get

$$22.4 = 7a + 196(0.05) \text{ or } 22.4 = 7a + 9.8 \text{ or } 7a = 22.4 - 9.8$$

$$\text{or } 7a = 12.6 \text{ or } a = \frac{12.6}{7} = 1.8$$

Hence the regression line of Y on X is $\hat{Y} = 1.8 + 0.05X$

To estimate the most probable yield of rice, we put $X = 36$ in the above equation, we get

$$\hat{Y} = 1.8 + 0.05(36) = 1.8 + 1.8 = 3.6$$

Example 14.4.

Show that the sum of errors and sum of squares of errors are zero.

X	1	2	3	4	5
Y	0	1	2	3	4

Solution:

The equation of a least square line is $\hat{Y} = a + bX$

The normal equations are

$$\sum Y = na + b\sum X \quad \text{and} \quad \sum XY = a\sum X + b\sum X^2$$

The necessary calculations are given below:

X	Y	XY	X^2	$\hat{Y} = X - 1$	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
1	0	0	1	0	0	0
2	1	2	4	1	0	0
3	2	6	9	2	0	0
4	3	12	16	3	0	0
5	4	20	25	4	0	0
$\sum X = 15$	$\sum Y = 10$	$\sum XY = 40$	$\sum X^2 = 55$	$\sum \hat{Y} = 10$	$\sum (Y - \hat{Y}) = 0$	$\sum (Y - \hat{Y})^2 = 0$

Substituting the values in the normal equations, we have

$$10 = 5a + 15b \quad \dots \dots (1)$$

$$40 = 15a + 55b \quad \dots \dots (2)$$

Solving these two equations, we multiply equation (1) by 3 and subtract from equation (2), we get

$$40 = 15a + 55b$$

$$30 = 15a + 45b$$

$$10 = 10b \quad \text{or} \quad b = \frac{10}{10} = 1$$

Substituting $b = 1$ in equation (1), we get

$$10 = 5a + 15(1) \quad \text{or} \quad 5a = 10 - 15 = -5 \quad \text{or} \quad a = \frac{-5}{5} = -1$$

Hence the fitted least square line is $\hat{Y} = X - 1$

Another Form of Regression Equation

We know that the linear regression equation is

$$\hat{Y} = a + bX \quad \dots \dots (1)$$

Normal equation for 'a' is

$$\Sigma Y = na + b \Sigma X$$

Dividing both sides by n, we get

$$\frac{\Sigma Y}{n} = \frac{n a}{n} + \frac{b \Sigma Y}{n} \text{ or } \bar{Y} = a + b \bar{X} \quad \dots \dots \quad (2)$$

It means that the regression line passes through the point (\bar{X}, \bar{Y}) .

Subtracting equation (2) from equation (1), we get

$$\hat{Y} - \bar{Y} = b(X - \bar{X})$$

which is another way of writing the regression equation of Y on X. From equation (2) we can write

$$a = \bar{Y} - b \bar{X}$$

The regression coefficient b may be written as b_{yx} , which means that the coefficient belongs to an equation in which X is the independent variable and Y is the dependent variable.

Calculation of Sum of Squares of Errors

The sum of squares of errors defined by $SSE = \sum(Y - \hat{Y})^2$ can also be calculated as below:

$$SSE = \Sigma Y^2 - a \Sigma Y - b \Sigma XY$$

14.6.1 PROPERTIES OF THE REGRESSION LINE

(i) The regression line of $\hat{Y} = a + bX$ has the following properties.

- (ii) We know that $\bar{Y} = a + b \bar{X}$. This shows that the line passes through the means \bar{X} and \bar{Y} .
- (iii) The sum of errors is equal to zero. The regression equation is $\hat{Y} = a + bX$ and the sum of deviations of observed Y from estimated \hat{Y} is

$$\Sigma(Y - \hat{Y}) = \Sigma(Y - a - bX) = \Sigma Y - na - b \Sigma X = 0 \quad [\Sigma Y = na + b \Sigma X]$$

when $\Sigma(Y - \hat{Y}) = 0$; it means that $\Sigma Y = \Sigma \hat{Y}$

14.6.2 REGRESSION EQUATION OF X ON Y

There are some special cases in which X and Y can be assumed independent variable turn by turn. This is possible when both X and Y are random variables and to estimate some Y-value, X is assumed as independent and to estimate some X-value, Y is taken as independent. When X and Y can be interchanged then the regression coefficient of Y on X is denoted by b_{yx} and the intercept of Y on X is denoted by a_{yx} when Y is the independent variable, the regression equation of X on Y can be written as:

$$X = a_{xy} + b_{xy} Y$$

where b_{xy} is the regression coefficient of X on Y. The normal equations for the regression of X on Y are

$$\Sigma X = n a_{xy} + b_{xy} \Sigma Y \quad \text{and} \quad \Sigma XY = a_{xy} \Sigma Y + b_{xy} \Sigma Y^2$$

The regression coefficient b_{xy} and the intercept a_{xy} can be directly calculated by the relations:

$$b_{xy} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(Y - \bar{Y})^2} = \frac{n \Sigma XY - (\Sigma X)(\Sigma Y)}{n \Sigma Y^2 - (\Sigma Y)^2}$$

$$\text{and } a_{xy} = \frac{(\Sigma Y^2)(\Sigma X) - (\Sigma Y)(\Sigma XY)}{n \Sigma Y^2 - (\Sigma Y)^2}$$

The regression equation of X on Y can be written as $\hat{X} - \bar{X} = b_{xy}(Y - \bar{Y})$

Also $\Sigma(X - \hat{X}) = 0$, $\Sigma X = \Sigma \hat{X}$ and $\bar{X} = a_{xy} + b_{xy} \bar{Y}$ or $a_{xy} = \bar{X} - b_{xy} \bar{Y}$

Various Formulas for the Calculation of Regression Coefficients

Different forms of the formulas for regression coefficient of Y on X are:

$$\begin{aligned} b_{yx} &= \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2} = \frac{\Sigma XY - (\Sigma X)(\Sigma Y)}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}} \\ &= \frac{n \Sigma XY - (\Sigma X)(\Sigma Y)}{n \Sigma X^2 - (\Sigma X)^2} = \frac{\Sigma XY - n \bar{X} \bar{Y}}{\Sigma X^2 - n \bar{X}^2} \\ &= \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n S_x^2} = \frac{S_{xy}}{S_x^2} \quad \text{where } S_{xy} = \frac{1}{n} \Sigma(X - \bar{X})(Y - \bar{Y}) \end{aligned}$$

When the calculations are to be reduced by change of origin, then we use D_x and D_y where $D_x = X - A$ and $D_y = Y - B$, A and B are some constants b_{yx} can be calculated as below:

$$b_{yx} = \frac{\Sigma D_x D_y - \frac{(\Sigma D_x)(\Sigma D_y)}{n}}{\Sigma D_x^2 - \frac{(\Sigma D_x)^2}{n}} = \frac{n \Sigma D_x D_y - (\Sigma D_x)(\Sigma D_y)}{n \Sigma D_x^2 - (\Sigma D_x)^2}$$

When change of origin and scale is used, then $U = \frac{X - A}{h}$ and $V = \frac{Y - B}{k}$ and

$$b_{yx} = \frac{\Sigma U V - \frac{(\Sigma U)(\Sigma V)}{n}}{\Sigma U^2 - \frac{(\Sigma U)^2}{n}} = \frac{n \Sigma U V - (\Sigma U)(\Sigma V)}{n \Sigma U^2 - (\Sigma U)^2}$$

The regression equation of Y on X can be written as:

$$\hat{Y} - \bar{Y} = \frac{S_{xy}}{S_x^2} (X - \bar{X})$$

Similarly, the formulas for the regression coefficient of X on Y are

$$\begin{aligned} b_{xy} &= \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(Y - \bar{Y})^2} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}} \\ &= \frac{n \Sigma XY - (\Sigma X)(\Sigma Y)}{n \Sigma Y^2 - (\Sigma Y)^2} = \frac{\Sigma XY - n \bar{X} \bar{Y}}{\Sigma Y^2 - n \bar{Y}^2} \\ &= \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n S_y^2} = \frac{S_{xy}}{S_y^2} \text{ where } S_{xy} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n} \end{aligned}$$

When the idea of change of origin is applied, then

$$b_{xy} = \frac{\Sigma D_x D_y - \frac{(\Sigma D_x)(\Sigma D_y)}{n}}{\Sigma D_y^2 - \frac{(\Sigma D_y)^2}{n}} = \frac{n \Sigma D_x D_y - (\Sigma D_x)(\Sigma D_y)}{n \Sigma D_y^2 - (\Sigma D_y)^2}$$

When change of origin and scale is used, then

$$b_{xy} = \frac{\Sigma U V - \frac{(\Sigma U)(\Sigma V)}{n}}{\Sigma V^2 - \frac{(\Sigma V)^2}{n}} = \frac{n \Sigma U V - (\Sigma U)(\Sigma V)}{n \Sigma V^2 - (\Sigma V)^2}$$

The regression equation of X on Y can be written as:

$$\hat{X} - \bar{X} = \frac{S_{xy}}{S_y^2} (Y - \bar{Y})$$

Example 14.5.

Compute the regression lines of X on Y and Y on X on the basis of the following informations: $\Sigma X = 50$, $\Sigma Y = 60$, $\Sigma XY = 350$, $\bar{X} = 5$, $\bar{Y} = 6$,

standard deviation of X = 2, standard deviation of Y = 3.

Solution: The necessary calculations are given below:

$$\bar{X} = \frac{\Sigma X}{n} \text{ or } n = \frac{\Sigma X}{\bar{X}} = \frac{50}{5} = 10$$

$$S_{xy} = \frac{1}{n} \left[\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} \right] = \frac{1}{10} \left[350 - \frac{(50)(60)}{10} \right] = \frac{1}{10} [50] = 5$$

The regression coefficient of X on Y is

$$b_{xy} = \frac{S_{xy}}{S_y^2} = \frac{5}{(3)^2} = \frac{5}{9}$$

The regression line of X on Y is

$$\hat{X} - \bar{X} = b_{xy} (Y - \bar{Y})$$

$$\hat{X} - 5 = \frac{5}{9}(Y - 6) = \frac{5}{9}Y - \frac{30}{9}$$

$$\hat{X} = \frac{5}{9}Y + 5 - \frac{30}{9} = \frac{5}{9}Y + \frac{15}{9}$$

$$\hat{X} = 0.56Y + 1.67$$

Example 14.6.

For 8 observations on deposits (X) and loans (Y) the following data were obtained:

$$\Sigma(X - 47) = 16, \Sigma(Y - 35) = 8, \Sigma(X - 47)^2 = 74, \Sigma(Y - 35)^2 = 68,$$

$$\Sigma(X - 47)(Y - 35) = 0, \bar{X} = 49, \bar{Y} = 36$$

- (a) Compute the regression equation of X on Y and estimate most likely value of X when Y = 36.
- (b) Compute the regression equation of Y on X and estimate most likely value of Y when X = 45.

Solution:

Here $\Sigma D_x = 16, \Sigma D_y = 8, \Sigma D_x^2 = 74, \Sigma D_y^2 = 68, \Sigma D_x D_y = 0, \bar{X} = 49$ and $\bar{Y} = 36$.

The regression coefficient of X on Y is

$$b_{xy} = \frac{\Sigma D_x D_y - \frac{(\Sigma D_x)(\Sigma D_y)}{n}}{\Sigma D_y^2 - \frac{(\Sigma D_y)^2}{n}} = \frac{0 - \frac{(16)(8)}{8}}{68 - \frac{(8)^2}{8}} = -\frac{16}{60} = -0.27$$

The regression coefficient of Y on X is

$$b_{yx} = \frac{\Sigma D_x D_y - \frac{(\Sigma D_x)(\Sigma D_y)}{n}}{\Sigma D_x^2 - \frac{(\Sigma D_x)^2}{n}} = \frac{0 - \frac{(16)(8)}{8}}{74 - \frac{(16)^2}{8}} = -\frac{16}{42} = -0.38$$

The regression coefficient of Y on X is

$$b_{yx} = \frac{S_{xy}}{S_x^2} = \frac{5}{(2)^2} = \frac{5}{4}$$

The regression line of Y on X is

$$\hat{Y} - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\hat{Y} - 6 = \frac{5}{4}(X - 5) = \frac{5}{4}X - \frac{25}{4}$$

$$\hat{Y} = \frac{5}{4}X + 6 - \frac{25}{4} = \frac{5}{4}X - \frac{1}{4}$$

$$\hat{Y} = 1.25X - 0.25$$

(a) The regression equation of X on Y is

$$\hat{X} - \bar{X} = b_{xy} (Y - \bar{Y})$$

$$\hat{X} - 49 = -0.27 (Y - 36)$$

$$\hat{X} - 49 = -0.27 Y + 9.72$$

$$\hat{X} = -0.27 Y + 9.72 + 49$$

$$= -0.27 Y + 58.72$$

Put $Y = 36$, we get

$$\hat{X} = -0.27 (36) + 58.72 = 49$$

Example 14.7.

Estimate the regression equation $\hat{Y} = a + bX$ for the following informations on lot size (X) and number of man-hours of labor (Y) for ten recent production runs performed under similar production conditions:

$$\Sigma X = 500, \Sigma Y = 1100, \Sigma XY = 61800, \Sigma X^2 = 28400, \Sigma Y^2 = 134660.$$

Estimate the expected value of Y given that $X = 50$.

Solution:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{500}{10} = 50 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{1100}{10} = 110$$

The least squares estimates a and b are calculated by using the formulas as below:

$$b = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{10(61800) - (500)(1100)}{10(28400) - (500)^2} = \frac{68000}{34000} = 2$$

$$a = \bar{Y} - b\bar{X} = 110 - 2(50) = 110 - 100 = 10$$

Hence the estimated regression equation is $\hat{Y} = 10 + 2X$.

$$\text{When } X = 50, \text{ then } \hat{Y} = 10 + 2(50) = 10 + 100 = 110$$

Example 14.8.

Fitting a straight line to a set of data yields the following regression equation:

$$\hat{Y} = 2 + 5X$$

(a) Interpret the meaning of the Y intercept 'a'.

(b) Interpret the meaning of the slope 'b'.

(c) Predict the average value of Y for $X = 3$.

(d) If the values of X range from 2 to 25, should you use this model to predict the average value of Y when X equals:

- (i) 3 ? (ii) -3 ? (iii) 0 ? (iv) 24 ? (v) 26 ?

(b) The regression equation of Y on X is

$$\hat{Y} - \bar{Y} = b_{yx} (X - \bar{X})$$

$$\hat{Y} - 36 = -0.38 (X - 49)$$

$$\hat{Y} - 36 = -0.38 X + 18.62$$

$$\hat{Y} = -0.38 X + 18.62 + 36$$

$$= -0.38 X + 54.62$$

Put $X = 45$, we get

$$\hat{Y} = -0.38(45) + 54.62 = 37.52$$

Solution:

- (a) The Y intercept $a = 2$ means that when $X = 0$, the average value of Y is 2.
 (b) The slope $b = 5$ means that for each increase of one unit of X, the value of Y is expected to increase on average by 5 units.
 (c) When $X = 3$, then $\hat{Y} = 2 + 5(3) = 17$
 (d) (i) Yes (ii) No (iii) No (iv) Yes (v) No

14.7 INTRODUCTION

Regression and correlation are two different techniques which are used for the analysis of the bi-variate data. Both these techniques sometimes use the same sample data and there are certain calculations which are common in both these techniques. None of these two can take the place of one or the other. Both have their own area of application. In a certain situation, may be that both are used and in some other situation, may be that only correlation is applicable.

14.8 CORRELATION

Correlation is a technique which measures the strength of association between two variables. Both the variables X and Y may be random or may be that one variable is independent (non-random) and the other to be correlated is dependent. When the changes in one variable appear to be linked with the changes in the other variable, the two variables are said to be correlated. When the two variables are meaningfully related and both increase or both decrease simultaneously, then the correlation is termed as positive. If increase in any one variable is associated with decrease in the other variable, the correlation is termed as negative or inverse. Suppose marks in Mathematics are denoted by X and marks in Statistics are denoted by Y. If small values of X appear with small values of Y and large values of X come with large values of Y, then correlation is said to be positive. If X stands for marks in English and Y stands for marks in Mathematics, it is possible that small values of X appear with large values of Y. It is a case of negative correlation.

14.8.1 MEASUREMENT OF CORRELATION

The degree or level of correlation is measured with the help of correlation coefficient or coefficient of correlation. For population data, the correlation coefficient is denoted by ρ . The joint variation of X and Y is measured by the covariance of X and Y. The covariance of X and Y denoted by $Cov(X, Y)$ is defined as:

$$Cov(X, Y) = E[X - E(X)][Y - E(Y)]$$

The $Cov(X, Y)$ may be positive, negative or zero. The covariance has the same units in which X and Y are measured. When $Cov(X, Y)$ is divided by σ_x and σ_y , we get the

correlation coefficient ρ . Thus $\rho = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$. ρ is free of the units of measurement.

It is a pure number and lies between -1 and +1. If $\rho = +1$, it is called perfect positive correlation. If $\rho = -1$, it is called perfect negative correlation. If there is no correlation between X and Y, then X and Y are independent and $\rho = 0$. For sample data the correlation coefficient denoted by 'r' is a measure of strength of the linear relation between X and Y variables, where 'r' is a pure number and lies between -1 and +1.

14.8.2 PERFECT POSITIVE CORRELATION

Consider the data in Table 14.5. on X and Y where X is the number of litres of oil and Y is the distance traveled by a vehicle in kilometers.

Table 14.5.

X	1	2	3	4	5
Y	20	40	60	80	100

Panel (a) of figure 14.7. illustrates a perfect positive correlation between X and Y. Here Y increases by a fixed distance of 20 kilometer when X increases by one litre. Here X and Y are assumed as random variables.

14.8.3 PERFECT NEGATIVE CORRELATION

Consider the data in Table 14.6. on X and Y where X is the number of study hours and Y is the number of sleeping hours of different students.

Table 14.6.

X	2	4	6	8	10
Y	10	9	8	7	6

Panel (b) of figure 14.7. depicts a perfect negative correlation between X and Y. Thus there is a perfect negative relationship between X and Y so that Y decreases by a fixed duration of 1 hour as X increases by 2 hours. There is a definite predictable decrease in Y whenever X increases by 2 units of time.

14.8.4 NO CORRELATION

Let us consider the data in Table 14.7. on X and Y where X is the per capita income in thousands of Rs. and Y is the crude death rate per 1000 of population in a country.

Table 14.7.

X	1	2	3	4	5
Y	15	15	15	15	15

Panel (c) of figure 14.7. shows that there is no correlation between X and Y. When X increases, the Y variable does not increase. The Y-variable is not showing any association with X-variable.

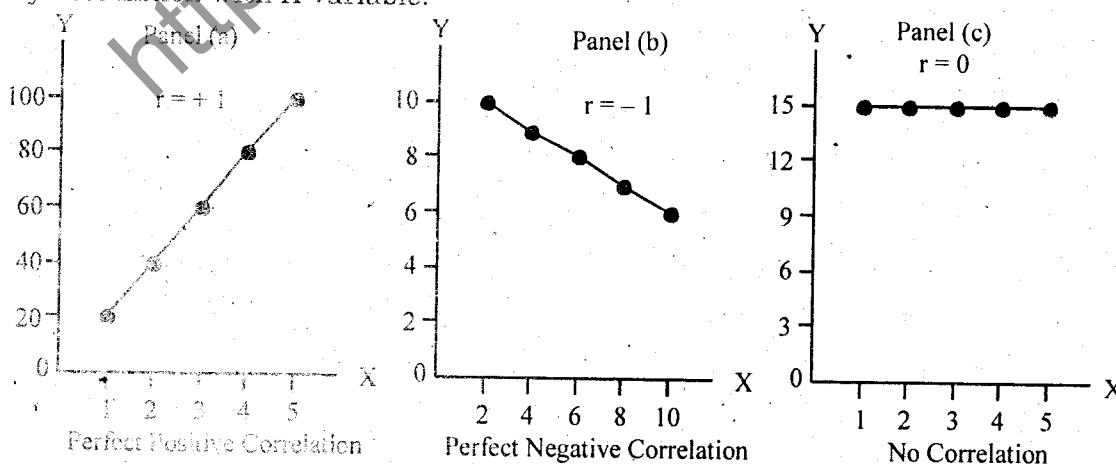


Figure 14.7

14.8.5 SCATTER DIAGRAMS

Figure 14.7. shows three cases of theoretical nature in which the plotted points lie on the exact mathematical lines. In practical life, mostly the relation between the random variables X and Y is not of the type as shown in figure 14.7. Suppose we are given a sample data in the form of pairs and the data is plotted in the form of scattered points as shown in figure 14.8.

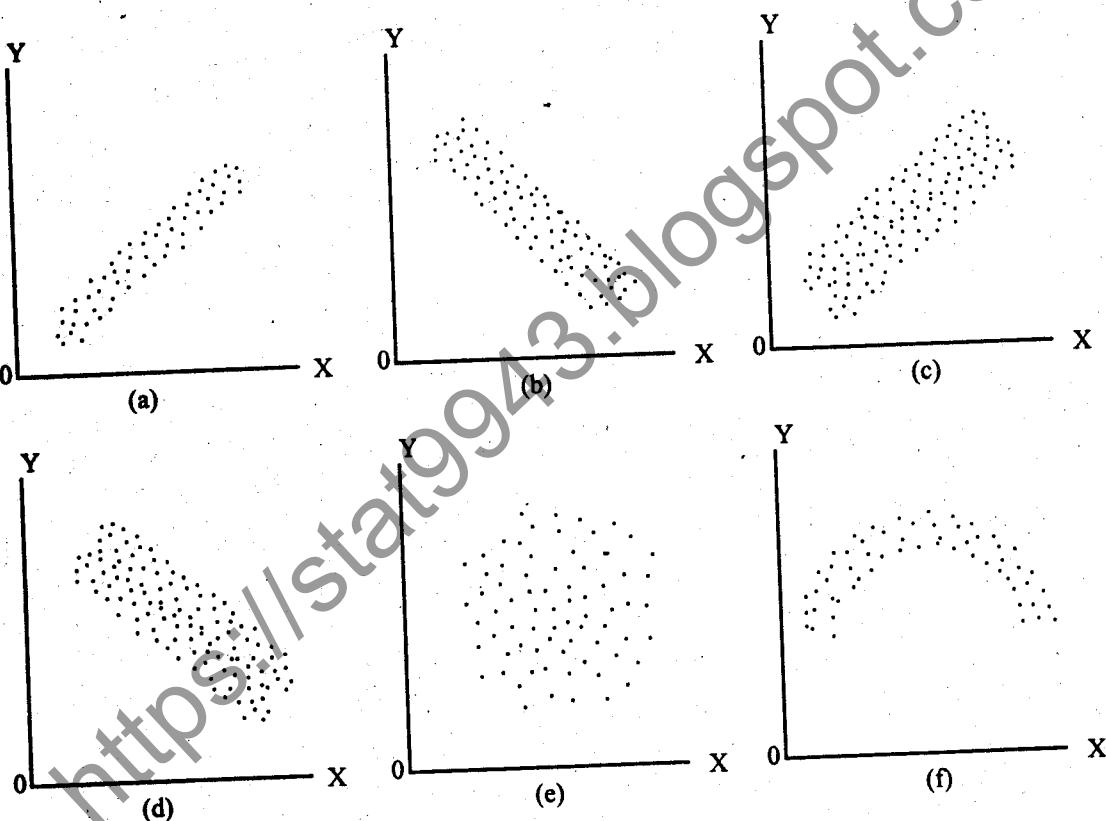


Figure 14.8

Panel (a) of figure 14.8. shows that increase in X is associated with increase in Y. The scattered points are close to a central line. The figure shows that the correlation is strong between X and Y. Panel (b) of figure 14.8. shows downward movement in Y when X increases. This shows negative correlation between X and Y. Panel (c) and (d) also indicate positive and negative correlation but the points are scattered away from some central line. Thus the correlation coefficient will have a small value. Panel (e) and (f) indicate as if there is no relation between X and Y.

14.9 CORRELATION COEFFICIENT FOR SAMPLE DATA

The correlation coefficient calculated for X and Y in a sample data is denoted by r_{xy} . It can be calculated by using any one of the formulas:

$$(i) r_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{nS_x S_y}. \text{ Where } S_x \text{ and } S_y \text{ are the standard deviations of } X \text{ and } Y$$

respectively and are given by $S_x = \sqrt{\frac{\sum(X - \bar{X})^2}{n}}$ and $S_y = \sqrt{\frac{\sum(Y - \bar{Y})^2}{n}}$

This formula is called Karl Pearson's product moment formula.

$$(ii) r_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}} \quad (iii) r_{xy} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right] \left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$

$$(iv) r_{xy} = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}} \quad (v) r_{xy} = \frac{\sum XY - n \bar{X} \bar{Y}}{\sqrt{[\sum X^2 - n \bar{X}^2][\sum Y^2 - n \bar{Y}^2]}}$$

Example 14.9.

A retail outlet for air conditioners believes that its weekly sales are dependent upon the average temperature during the week. It picks at random 6 weeks and finds that its sales are related to the average temperature in these weeks as follows:

Mean temperature (F°)	72	77	82	43	31	55
Sales (No. of air conditioners)	4	5	6	1	0	2

Calculate the correlation coefficient between the mean temperature and the retail outlet's sales.

Solution:

The necessary calculations are given below:

X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
72	4	+ 12	+1	12	144	1
77	5	+ 17	+2	34	289	4
82	6	+ 22	+3	66	484	9
43	1	- 17	-2	34	289	4
31	0	-29	-3	87	841	9
55	2	-5	-1	5	25	1
$\Sigma X = 360$	$\Sigma Y = 18$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(Y - \bar{Y}) = 0$	$\Sigma(X - \bar{X})(Y - \bar{Y}) = 238$	$\Sigma(X - \bar{X})^2 = 2072$	$\Sigma(Y - \bar{Y})^2 = 28$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{360}{6} = 60$$

$$S_x = \sqrt{\frac{\sum(X - \bar{X})^2}{n}} = \sqrt{\frac{2072}{6}} = 18.5831$$

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{n S_x S_y} = \frac{238}{6(18.5831)(2.1602)} = \frac{238}{240.8593} = 0.99$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{18}{6} = 3$$

$$S_y = \sqrt{\frac{\sum(Y - \bar{Y})^2}{n}} = \sqrt{\frac{28}{6}} = 2.1602$$

Example 14.10.

Calculate and analyze the correlation coefficient between the number of study hours and the number of sleeping hours of different students.

Number of study hours	2	4	6	8	10
Number of sleeping hours	10	9	8	7	6

Solution:

The necessary calculations are given below:

X	Y	(X - \bar{X})	(Y - \bar{Y})	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
2	10	-4	+2	-8	16	4
4	9	-2	+1	-2	4	1
6	8	0	0	0	0	0
8	7	+2	-1	-2	4	1
10	6	+4	-2	-8	16	4
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(Y - \bar{Y}) = 0$	$\Sigma(X - \bar{X})(Y - \bar{Y}) = -20$	$\Sigma(X - \bar{X})^2 = 40$	$\Sigma(Y - \bar{Y})^2 = 10$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{30}{5} = 6 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{40}{5} = 8$$

$$r_{xy} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}} = \frac{-20}{\sqrt{(40)(10)}} = \frac{-20}{20} = -1$$

There is perfect negative correlation between the number of study hours and the number of sleeping hours.

Example 14.11.

Compute and interpret the coefficient of correlation between the values of X and Y from the following table.

X	1	2	3	4	5
Y	20	40	60	80	100

Solution:

The necessary calculations are given below:

X	Y	XY	X^2	Y^2
1	20	20	1	400
2	40	80	4	1600
3	60	180	9	3600
4	80	320	16	6400
5	100	500	25	10000
$\Sigma X = 15$	$\Sigma Y = 300$	$\Sigma XY = 1100$	$\Sigma X^2 = 55$	$\Sigma Y^2 = 22000$

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} = \frac{5(1100) - 15(300)}{\sqrt{[5(55) - (15)^2][5(22000) - (300)^2]}}$$

$$= \frac{5500 - 4500}{\sqrt{(50)(20000)}} = \frac{1000}{1000} = 1$$

$r = 1$ means perfect positive correlation between X and Y.

Example 14.12.

The following are 5 pairs of values of two variables X and Y. Compute and interpret the coefficient of correlation between X and Y.

X	11	12	13	14	15
Y	15	14	13	12	16

Solution:

The necessary calculations are given below:

X	Y	XY	X^2	Y^2
11	15	165	121	225
12	14	168	144	196
13	13	169	169	169
14	12	168	196	144
15	16	240	225	256
$\Sigma X = 65$	$\Sigma Y = 70$	$\Sigma XY = 910$	$\Sigma X^2 = 855$	$\Sigma Y^2 = 990$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{65}{5} = 13 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{70}{5} = 14$$

$$r = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\sqrt{[\Sigma X^2 - n(\bar{X})^2][\Sigma Y^2 - n(\bar{Y})^2]}} = \frac{910 - 5(13)(14)}{\sqrt{[855 - 5(13)^2][990 - 5(14)^2]}}$$

$$= \frac{910 - 910}{\sqrt{(10)(10)}} = 0. \text{ So } X \text{ and } Y \text{ are uncorrelated.}$$

Example 14.13.

From the following data, compute the coefficient of correlation between X and Y:

	X series	Y series
Number of items	15	15
Arithmetic mean	25	18
Sum of square of deviations from arithmetic mean	136	138

Summation of products of deviations of X and Y series from their arithmetic means = 122.

Solution:

Here $n = 15, \bar{X} = 25, \bar{Y} = 18, \Sigma(X - \bar{X})^2 = 136, \Sigma(Y - \bar{Y})^2 = 138,$

$$\Sigma(X - \bar{X})(Y - \bar{Y}) = 122 \text{ and hence}$$

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}} = \frac{122}{\sqrt{(136)(138)}} = \frac{122}{137} = 0.89$$

Example 14.14.

In order to find the correlation coefficient between two variables X and Y from 12 pairs of observations, the following results are given:

$$\Sigma X = 30, \Sigma Y = 5, \Sigma X^2 = 670, \Sigma Y^2 = 285, \Sigma XY = 344.$$

Later it was found that one particular set of observations, namely $X = 12, Y = 6$ was wrongly taken, the correct values being $X = 21, Y = 16$. Compute the correct value of the correlation coefficient r.

Solution:

The necessary calculations are given below:

$$\text{Correct } \Sigma X = 30 - 12 + 21 = 39$$

$$\text{Correct } \Sigma Y = 5 - 6 + 16 = 15$$

$$\text{Correct } \Sigma X^2 = 670 - (12)^2 + (21)^2 = 967$$

$$\text{Correct } \Sigma Y^2 = 285 - (6)^2 + (16)^2 = 505$$

$$\text{Correct } \Sigma XY = 344 - 12(6) + 21(16) = 608$$

Thus,

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n} \right] \left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{n} \right]}} = \frac{608 - \frac{(39)(15)}{12}}{\sqrt{967 - \frac{(39)^2}{12}} \sqrt{505 - \frac{(15)^2}{12}}} \\ = \frac{608 - 48.75}{\sqrt{(840.25)(486.25)}} = \frac{559.25}{639.1960} = 0.875$$

14.9.1 CAUSATION IN CORRELATION

The use of the term 'causation' in correlation is not appropriate and should be avoided. In correlation analysis, there is no such thing as cause and effect relation between X and Y. When both are random variables, no variable is under the control of the experimenter. Both variables change simultaneously. The forces of changes are taking place in both the variables at a time. The marks of students in English and Urdu are interdependent and cannot be classified as 'cause' and 'effect'.

14.9.2 SPURIOUS CORRELATION

The numerical value of ' r ' is to be interpreted carefully. Somebody may calculate ' r ' between two variables which are not meaningfully related to each other. There is no sense in calculating correlation coefficient between the number of telephone connections over a period of time and the number of accidents on the roads. The number of telephone connections can be correlated with the per capita income of the people and the number of accidents may be correlated with variables like population, number of vehicles on the roads, the speed of vehicles etc. Any value of r between un-related variables is called *spurious* correlation or *non-sense* correlation. The observed value of r should not be used blindly. We must examine whether or not there exists any mutual relationship between the variables.

14.9.3 CHANGE OF ORIGIN

The correlation coefficient r_{xy} is not affected by change of origin. If a certain constant is added to the variable or subtracted from the variable, the correlation coefficient of the resulting variables is the same as that of X and Y. Let $D_x = X - A$, where A is a constant and $D_y = Y - B$, where B is also a constant. It can be proved that the correlation coefficient r_{xy} is equal to the correlation coefficient between D_x and D_y which may be denoted by $r_{D_x D_y}$. Thus

$$r_{xy} = r_{D_x D_y} = \frac{\sum D_x D_y - \frac{(\sum D_x)(\sum D_y)}{n}}{\sqrt{\left[\sum D_x^2 - \frac{(\sum D_x)^2}{n} \right] \left[\sum D_y^2 - \frac{(\sum D_y)^2}{n} \right]}}$$

If both numerator and denominator are multiplied with n, we get

$$r_{xy} = r_{D_x D_y} = \frac{n \sum D_x D_y - (\sum D_x)(\sum D_y)}{\sqrt{[n \sum D_x^2 - (\sum D_x)^2] [n \sum D_y^2 - (\sum D_y)^2]}}$$

Example 14.15.

The following figures show the imports and exports of a commodity, in lakhs of rupees, during the last five years:

Years	1996	1997	1998	1999	2000
Imports	82	78	75	80	95
Exports	70	74	78	75	80

Taking X = 80 and Y = 70 as origins, compute the coefficient of correlation between imports and exports.

Solution:

The necessary calculations are given below:

X	Y	$D_x = X - 80$	$D_y = Y - 70$	$D_x D_y$	D_x^2	D_y^2
82	70	+2	0	0	4	0
78	74	-2	4	-8	4	16
75	78	-5	8	-40	25	64
80	75	0	5	0	0	25
95	80	+15	10	+150	225	100
Total		10	27	102	258	205

$$r = \frac{\sum D_x D_y - \frac{(\sum D_x)(\sum D_y)}{n}}{\sqrt{\left[\sum D_x^2 - \frac{(\sum D_x)^2}{n} \right] \left[\sum D_y^2 - \frac{(\sum D_y)^2}{n} \right]}}$$

$$r = \frac{102 - \frac{(10)(27)}{5}}{\sqrt{\left[258 - \frac{(10)^2}{5} \right] \left[205 - \frac{(27)^2}{5} \right]}} = \frac{102 - 54}{\sqrt{(238)(59.2)}} = \frac{48}{118.6996} = 0.40$$

14.9.4 CHANGE OF SCALE

When X and Y are divided by some constant or they are multiplied with some constant, the operation is called change of scale. The value of r_{xy} does not change by change of scale. Let $U = \frac{X}{h}$ and $V = \frac{Y}{k}$ where h and k are some constants. The correlation coefficient between U and V is denoted by r_{UV} . It can be proved that $r_{xy} = r_{UV}$. Thus,

$$r_{xy} = r_{UV} = \frac{\sum UV - \frac{(\sum U)(\sum V)}{n}}{\sqrt{\left[\sum U^2 - \frac{(\sum U)^2}{n} \right] \left[\sum V^2 - \frac{(\sum V)^2}{n} \right]}}$$

$$= \frac{n \sum UV - (\sum U)(\sum V)}{\sqrt{[n \sum U^2 - (\sum U)^2][n \sum V^2 - (\sum V)^2]}}$$

The above formulas are also applicable when the change of scale is applied on only one variable and the other variable is not changed. When a variable is left as it is, it means the variable is multiplied with 1. May be that one variable is multiplied and the other is divided by some constant. It is also change of scale and the formulas will work if such changes are carried out.

14.9.5 CHANGE OF ORIGIN AND SCALE

Change of origin and scale may be applied simultaneously on X and Y. The value of 'r' does not change by change of origin and scale. Let $U = \frac{X-A}{h}$ and $V = \frac{Y-B}{k}$, then,

$$r_{xy} = r_{UV} = \frac{\sum UV - \frac{(\sum U)(\sum V)}{n}}{\sqrt{\left[\sum U^2 - \frac{(\sum U)^2}{n}\right] \left[\sum V^2 - \frac{(\sum V)^2}{n}\right]}} = \frac{n\sum UV - (\sum U)(\sum V)}{\sqrt{[n\sum U^2 - (\sum U)^2][n\sum V^2 - (\sum V)^2]}}$$

If the numerical values of X and Y are very large, we can reduce the calculations by applying the operators like change of origin, change of scale or change of origin and scale. The coefficient of correlation between U and V (r_{UV}) is the same thing as correlation coefficient between X and Y (r_{xy}). But $r_{xy} = r_{UV}$ when the divisors or multipliers like 'h' and 'k' have the same algebraic signs. If they have the opposite signs, then $r_{xy} = -r_{UV}$.

14.9.6 'r' IN A LINEAR REGRESSION RELATION

Suppose there is a certain linear regression relation between X and Y in which X is independent variable and Y is the dependent variable. The correlation coefficient 'r' is calculated between X and Y. This value of 'r' measures only the strength of association between the two variables. This value of r cannot be used for any type of inference about the population correlation coefficient ρ . Suppose in a linear regression problem, we regress the production of fans (Y) against the number of workers (X). We have calculated 'r' between X and Y which is 0.90. It means there is strong association between the number of workers and the production of fans.

14.9.7 'r' FOR RANDOM VARIABLES

If X and Y are both random variables and 'r' is calculated between X and Y, then the value of 'r' serves two purposes.

- (i) It measures the strength of association between the two random variables.
- (ii) It can be used for inference about ρ , the correlation coefficient in the population.

14.10 RELATION BETWEEN b_{yx} , b_{xy} AND r

For a linear regression relation, let us write the formulas used for the calculation of r, b_{yx} and b_{xy} .

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n S_x S_y}$$

$$b_{yx} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n S_x^2} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

$$b_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n S_y^2} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (Y - \bar{Y})^2}$$

The term in the numerator is the same for all the three formulas and the denominators in all the three formulas are positive. Thus the algebraic sign of r , b_{yx} and b_{xy} will be decided by the numerator. If b_{yx} is positive, then r and b_{xy} will also be positive. If any one of the three terms is negative the remaining will also have negative sign. Thus all the three will be positive, negative or zero.

We can write b_{yx} and b_{xy} in terms of r , S_x and S_y . We have,

$$b_{yx} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n S_x^2} \quad \dots \dots \quad (1)$$

The right hand side of equation (1) is multiplied and divided by S_y . Thus

$$b_{yx} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n S_x^2} \frac{S_y}{S_y} = \frac{S_y \Sigma(X - \bar{X})(Y - \bar{Y})}{S_x n S_x S_y} \quad \dots \dots \quad (2)$$

$$b_{yx} = \frac{S_y}{S_x} r_{xy}$$

In this relation b_{yx} will have the same sign as that of r_{xy} . Similarly, we can write

$$b_{xy} = \frac{S_x}{S_y} r_{xy}$$

The term $\frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n}$ is called the sample covariance and is denoted by S_{xy} .

From equation (2), we have

$$b_{yx} = \frac{S_y}{S_x} \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n} \frac{1}{S_x S_y} = \frac{S_{xy}}{S_x^2} \text{ and } b_{xy} = \frac{S_{xy}}{S_y^2}$$

The simple linear regression equation of Y on X is $\hat{Y} - \bar{Y} = b(X - \bar{X})$.

It can be written as $\hat{Y} - \bar{Y} = r \frac{S_y}{S_x} (X - \bar{X})$ or $\hat{Y} - \bar{Y} = \frac{S_{xy}}{S_x^2} (X - \bar{X})$

Similarly the regression relation of X on Y can be written as

$$\hat{X} - \bar{X} = r \frac{S_x}{S_y} (Y - \bar{Y}) \text{ or } \hat{X} - \bar{X} = \frac{S_{xy}}{S_y^2} (Y - \bar{Y})$$

14.11 PROPERTIES OF CORRELATION COEFFICIENT r

- (1) We may speak of the correlation coefficient between ' X ' and ' Y ' or between ' Y ' and ' X '. Both statements have the same meaning. Thus $r_{xy} = r_{yx}$. It is called the symmetric property of r .
- (2) r is a pure number. If X and Y are measured in kilograms, r will not be in kilograms. It is free of the units of measurement.

- (3) The value of r does not change by change of origin, change of scale or by change of origin and scale.

Thus $r_{xy} = r_{UV}$ where $U = \frac{X \pm A}{h}$ and $V = \frac{Y \pm B}{k}$ and ' h ' and ' k ' have the same algebraic signs.

If ' h ' and ' k ' have opposite signs, then r_{xy} and r_{UV} will have the same numerical values but with opposite signs. Thus $r_{xy} = -r_{UV}$.

- (4) r lies between -1 and $+1$.
 (5) In a linear regression relation, r is the geometric mean of the two regression coefficients. Thus $r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{\frac{S_y}{S_x} \cdot r \frac{S_x}{S_y}} = \sqrt{r^2} = |r|$
 r will have the same sign as that of b_{yx} and b_{xy} .
 (6) If X and Y are independent random variables, then $\text{Cov}(X, Y) = 0$. But if $\text{Cov}(X, Y) = 0$, it does not mean that X and Y are definitely independent.

Example 14.16.

Compute and interpret the coefficient of correlation. Also show that the correlation coefficient is the square root of regression coefficients on the basis of the following informations:

X = Amount of fertilizer in pounds per 100 square feet

Y = Yield of tomatoes in pounds

$$\Sigma(X - 35) = -100, \Sigma(X - 35)^2 = 3000, \Sigma(Y - 18) = 30,$$

$$\Sigma(Y - 18)^2 = 1106, \Sigma(X - 35)(Y - 18) = 900, n = 10.$$

Solution:

Here, $\Sigma D_x = -100, \Sigma D_x^2 = 3000, \Sigma D_y = 30, \Sigma D_y^2 = 1106, \Sigma D_x D_y = 900, n = 10$

The correlation coefficient is

$$r = \frac{n \Sigma D_x D_y - (\Sigma D_x)(\Sigma D_y)}{\sqrt{[n \Sigma D_x^2 - (\Sigma D_x)^2][n \Sigma D_y^2 - (\Sigma D_y)^2]}}$$

$$= \frac{10(900) - (-100)(30)}{\sqrt{[10(3000) - (-100)^2][10(1106) - (30)^2]}} = \frac{12000}{\sqrt{14254.82375}} = 0.84$$

There is a strong positive correlation between amount of fertilizer and yield of tomatoes.

The regression coefficient of Y on X is

$$b_{yx} = \frac{n \Sigma D_x D_y - (\Sigma D_x)(\Sigma D_y)}{n \Sigma D_x^2 - (\Sigma D_x)^2} = \frac{10(900) - (-100)(30)}{10(3000) - (-100)^2} = \frac{12000}{20000} = 0.6$$

The regression coefficient of X on Y is

$$b_{xy} = \frac{n \Sigma D_x D_y - (\Sigma D_x)(\Sigma D_y)}{n \Sigma D_y^2 - (\Sigma D_y)^2} = \frac{10(900) - (-100)(30)}{10(1106) - (30)^2} = \frac{12000}{10160} = 1.18$$

$$\pm \sqrt{b_{yx} \cdot b_{xy}} = \pm \sqrt{(0.6)(1.18)} = 0.84 = r$$

Hence the correlation coefficient is the square root of regression coefficients.

Example 14.17.

The equations of two regression lines obtained from ten observations are:

$$10X = 5Y - 55 \text{ and } 100Y = 200X + 1180$$

Compute and interpret the correlation coefficient r.

Solution:

The equations of regression lines can be written as:

$$X = \frac{5}{10}Y - \frac{55}{10} = 0.5Y - 5.5 \quad \text{and} \quad Y = \frac{200}{100}X + \frac{1180}{100} = 2X + 11.8$$

$$\text{Here, } b_{xy} = 0.5, b_{yx} = 2 \text{ and } r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{(0.5)(2)} = 1$$

There is perfect positive correlation between X and Y.

Example 14.18.

The following statistics have been computed:

$$\bar{X} = 14, \bar{Y} = 22, S_x = 6, S_y = 7, r = 0.72$$

Find the two regression equations.

Solution:

The regression equation of Y on X is

$$\hat{Y} - \bar{Y} = r \frac{S_y}{S_x} (X - \bar{X})$$

$$\hat{Y} - 22 = 0.72 \frac{7}{6} (X - 14)$$

$$\hat{Y} - 22 = 0.84 (X - 14)$$

$$\hat{Y} - 22 = 0.84X - 11.76$$

$$\hat{Y} = 0.84X - 11.76 + 22$$

$$\hat{Y} = 0.84X + 10.24$$

The regression equation of X on Y is

$$\hat{X} - \bar{X} = r \frac{S_x}{S_y} (Y - \bar{Y})$$

$$\hat{X} - 14 = 0.72 \frac{6}{7} (Y - 22)$$

$$\hat{X} - 14 = 0.62 (Y - 22)$$

$$\hat{X} - 14 = 0.62Y - 13.64$$

$$\hat{X} = 0.62Y - 13.64 + 14$$

$$\hat{X} = 0.62Y + 0.36$$

SHORT DEFINITIONS

Regression

When we predict the value of dependent variable with the help of one or more independent variables, it is known as regression.

or

Regression is a process by which we estimate the value of dependent variable on the basis of one or more independent variables.

Regression Analysis

The technique used to develop the equation and provide the estimates is called regression analysis.

Linear Regression

Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line is known as linear regression.

Non-Linear Regression

Non-linear regression is a procedure for fitting data to equations which are non-linear in the parameters.

Regression Line or Regression Equation

A straight line that best represents the relationship between two variables.

or

An equation that predicts the value of the dependent variable based on the value of one or more independent variables is called a regression equation.

Simple Regression Equation

A regression equation that includes one independent variable and one dependent variable.

Least Squares Method

The least squares method is the procedure used to develop the estimated regression equation.

Independent Variable

A variable that provides the basis for estimation, is called the predictor or explanatory variable or independent variable.

or

The variable that is predicting or explaining the other variable is called regressor or independent variable.

Dependent Variable

The variable that is being predicted or estimated is known as dependent variable or regressand or predictand or response variable.

Regression Coefficient or Slope of Regression Line

The average change in the dependent variable for a unit change in the independent variable is called regression coefficient. The regression coefficient may be positive or negative, depending on the relationship between the two variables.

Residual

The difference between the observed value of the dependent variable and the value predicted using the estimated regression equation is called residual.

Scatter Diagram

A graph of the pairs of observations on two variables is called a scatter diagram.

or

A graphic device used to summarize visually the relationship between two variables.

Correlation

Correlation is a measure of the degree of linear association between the two variables.

or

Correlation measures the strength of a relationship between variables.

Correlation Analysis

A group of techniques to measure the strength of the association between two variables.

Positive Correlation

When the values of two variables move in the same direction so that an increase or decrease in the value of one variable is associated with an increase or decrease in the value of the other variable, correlation is said to be positive.

Negative Correlation

When the values of two variables move in different directions, so that with an increase in the value of one variable the value of the other variable decreases, and with a decrease in the value of one variable the value of other variable increases, correlation is said to be negative.

No Correlation

When two variables have zero correlation, it is called as no correlation.

Curvilinear Correlation

When correlation between two variables represents a curve that is not a straight line, then the correlation is said to be curvilinear correlation.

Linear Correlation

Linear correlation is one where the ratio of variations in the related variables is constant.

Non-Linear Correlation

Non-linear correlation is one where the ratio of variations in the related variables is fluctuating.

Perfect Correlation

If the relationship between variables is such that with an increase or decrease in the value of one, the value of the other increase or decrease in a fixed proportion, correlation between them is said to be perfect correlation.

Perfect Positive Correlation

If both the series move in the same direction and the variations are proportionate there would be perfect correlation between them.

Perfect Negative Correlation

If the two series move in reverse directions and the variations in their values are always proportionate, it is said to be perfect negative correlation.

Correlation Coefficient

A descriptive measure of the degree of linear relationship between X and Y is called the correlation coefficient.

*A formula in which we find the strength of relationship b/w variables
or*

A measure that expresses the extent to which two variables are related.

Aims of Regression and Correlation Analysis

- Regression analysis provides estimates of the dependent variable for given values of the independent variable.
- Regression analysis provides measures of the errors that are likely to be involved in using the regression line to estimate the dependent variable.
- Regression analysis provides an estimate of the effect on the mean value of Y of a unit change in X.
- Correlation analysis provides estimates of how strong the relationship is between the two variables.

MULTIPLE - CHOICE QUESTIONS

- A process by which we estimate the value of dependent variable on the basis of one or more independent variables is called:

- | | |
|-----------------|----------------|
| (a) correlation | (b) regression |
| (c) residual | (d) slope |

- The method of least squares dictates that we choose a regression line where the sum of the square of deviations of the points from the line is:

- | | |
|-------------|--------------|
| (a) maximum | (b) minimum |
| (c) zero | (d) positive |

- A relationship where the flow of the data points is best represented by a curve is called:

- | | |
|-------------------------|----------------------------|
| (a) linear relationship | (b) nonlinear relationship |
| (c) linear positive | (d) linear negative |

All data points falling along a straight line is called:

- | | |
|-------------------------|----------------------------|
| (a) linear relationship | (b) nonlinear relationship |
| (c) residual | (d) scatter diagram |

The value we would predict for the dependent variable when the independent variables are all equal to zero is called:

- | | |
|---------------|-----------------------|
| (a) slope | (b) sum of residual |
| (c) intercept | (d) difficult to tell |

The predicted rate of response of the dependent variable to changes in the independent variable is called:

- | | |
|-----------|-------------------------|
| (a) slope | (b) intercept |
| (c) error | (d) regression equation |

7. The slope of the regression line of Y on X is also called the:
- correlation coefficient of X on Y
 - correlation coefficient of Y on X
 - regression coefficient of X on Y
 - regression coefficient of Y on X
8. In simple linear regression, the number of unknown constants are:
- one
 - two
 - three
 - four
9. In simple regression equation, the number of variables involved are:
- 0
 - 1
 - 2
 - 3
10. If the value of any regression coefficient is zero, then two variables are:
- qualitative
 - correlation
 - dependent
 - independent
11. The straight line graph of the linear equation $Y = a + bX$, slope will be upward if:
- $b = 0$
 - $b < 0$
 - $b > 0$
 - $b \neq 0$
12. The straight line graph of the linear equation $Y = a + bX$, slope will be downward if:
- $b > 0$
 - $b < 0$
 - $b = 0$
 - $b \neq 0$
13. The straight line graph of the linear equation $Y = a + bX$, slope is horizontal if:
- $b = 0$
 - $b \neq 0$
 - $b = 1$
 - $a = b$
14. If regression line of $\hat{Y} = 5$, then value of regression coefficient of Y on X is:
- 0
 - 0.5
 - 1
 - 5
15. If $Y = 2 - 0.2X$, then the value of Y intercept is equal to:
- 0.2
 - 2
 - $0.2X$
 - all of the above
16. If one regression coefficient is greater than one, then other will be:
- more than one
 - equal to one
 - less than one
 - equal to minus one
17. To determine the height of a person when his weight is given is:
- correlation problem
 - association problem
 - regression problem
 - qualitative problem
18. The dependent variable is also called:
- regressor
 - regressand
 - continuous variable
 - independent

19. The dependent variable is also called:
- (a) regressand variable
 - (b) predictand variable
 - (c) explained variable
 - (d) all of these
20. The independent variable is also called:
- (a) regressor
 - (b) regressand
 - (c) predictand
 - (d) estimated
21. In the regression equation $Y = a + bX$, the Y is called:
- (a) independent variable
 - (b) dependent variable
 - (c) continuous variable
 - (d) none of the above
22. In the regression equation $X = a + bY$, the X is called:
- (a) independent variable
 - (b) dependent variable
 - (c) qualitative variable
 - (d) none of the above
23. In the regression equation $Y = a + bX$, a is called:
- (a) X-intercept
 - (b) Y-intercept
 - (c) dependent variable
 - (d) none of the above
24. The regression equation always passes through:
- (a) (X, Y)
 - (b) (a, b)
 - (c) (\bar{X}, \bar{Y})
 - (d) (\bar{X}, Y)
25. The independent variable in a regression line is:
- (a) non-random variable
 - (b) random variable
 - (c) qualitative variable
 - (d) none of the above
26. The graph showing the paired points of (X_i, Y_i) is called a:
- (a) scatter diagram
 - (b) histogram
 - (c) historigram
 - (d) pie diagram
27. The graph  represents the relationship that is:
- (a) linear
 - (b) non linear
 - (c) curvilinear
 - (d) no relation
28. The graph  represents the relationship that is :
- (a) linear positive
 - (b) linear negative
 - (c) non-linear
 - (d) curvilinear
29. When regression line passes through the origin, then:
- (a) intercept is zero
 - (b) regression coefficient is zero
 - (c) correlation is zero
 - (d) association is zero
30. When b_{xy} is positive, then b_{yx} will be:
- (a) negative
 - (b) positive
 - (c) zero
 - (d) one

31. The correlation coefficient is the _____ of two regression coefficients:
- (a) geometric mean (b) arithmetic mean
 - (c) harmonic mean (d) median
32. When two regression coefficients bears same algebraic signs, then correlation coefficient is:
- (a) positive (b) negative
 - (c) according to two signs (d) zero
33. It is possible that two regression coefficients have:
- (a) opposite signs (b) same signs
 - (c) no sign (d) difficult to tell
34. Regression coefficient is independent of:
- (a) units of measurement (b) scale and origin
 - (c) both (a) and (b) (d) none of them
35. In the regression line $Y = a + bX$:
- (a) $\sum X = \sum \hat{X}$ (b) $\sum Y = \sum \hat{Y}$
 - (c) $\sum X = \sum Y$ (d) $X = Y$
36. In the regression line $Y = a + bX$, the following is always true:
- (a) $\sum(X - \hat{X}) = 0$ (b) $\sum(Y - \hat{Y}) = 0$
 - (c) $\sum(\hat{X} - \bar{X}) = \sum(Y - \hat{Y})$ (d) $\sum(Y - \hat{Y})^2 = 0$
37. The purpose of simple linear regression analysis is to:
- (a) predict one variable from another variable
 - (b) replace points on a scatter diagram by a straight line
 - (c) measure the degree to which two variables are linearly associated
 - (d) obtain the expected value of the independent random variable for a given value of the dependent variable
38. The sum of the difference between the actual values of Y and its values obtained from the fitted regression line is always:
- (a) zero (b) positive
 - (c) negative (d) minimum
39. If all the actual and estimated values of Y are same on the regression line, the sum of squares of error will be:
- (a) zero (b) minimum
 - (c) maximum (d) unknown
40. $e_i = Y_i - \hat{Y}_i$ is called:
- (a) residual
 - (b) difference between independent and dependent variables
 - (c) difference between slope and intercept
 - (d) sum of residual

41. A measure of the strength of the linear relationship that exists between two variables is called:
- (a) slope
 - (b) intercept
 - (c) correlation coefficient
 - (d) regression equation
42. When the ratio of variations in the related variables is constant, it is called:
- (a) linear correlation
 - (b) nonlinear correlation
 - (c) positive correlation
 - (d) negative correlation
43. If both variables X and Y increase or decrease simultaneously, then the coefficient of correlation will be:
- (a) positive
 - (b) negative
 - (c) zero
 - (d) one
44. If the points on the scatter diagram indicate that as one variable increases the other variable tends to decrease the value of r will be:
- (a) perfect positive
 - (b) perfect negative
 - (c) negative
 - (d) zero
45. If the points on the scatter diagram show no tendency either to increase together or decrease together the value of r will be close to:
- (a) -1
 - (b) +1
 - (c) 0.5
 - (d) 0
46. If one item is fixed and unchangeable and the other item varies, the correlation coefficient will be:
- (a) positive
 - (b) negative
 - (c) zero
 - (d) undecided
47. In scatter diagram, if most of the points lie in the first and third quadrants, then coefficient of correlation is:
- (a) negative
 - (b) positive
 - (c) zero
 - (d) all of the above
48. If the two series move in reverse directions and the variations in their values are always proportionate, it is said to be:
- (a) negative correlation
 - (b) positive correlation
 - (c) perfect negative correlation
 - (d) perfect positive correlation
49. If both the series move in the same direction and the variations are in a fixed proportion, correlation between them is said to be:
- (a) perfect correlation
 - (b) linear correlation
 - (c) nonlinear correlation
 - (d) perfect positive correlation
50. The value of the coefficient of correlation r lies between:
- (a) 0 and 1
 - (b) -1 and 0
 - (c) -1 and +1
 - (d) -0.5 and +0.5
51. If X is measured in hours and Y is measured in minutes, then correlation coefficient r has the unit:
- (a) hours
 - (b) minutes
 - (c) both (a) and (b)
 - (d) no unit

52. If $b_{yx} = b_{xy} = 1$ and $S_x = S_y$, then r will be:
 (a) 0 (b) -1
 (c) 1 (d) difficult to calculate
53. The correlation coefficient between X and $-X$ is:
 (a) 0 (b) 0.5
 (c) 1 (d) -1
54. If $b_{yx} = b_{xy} = r_{xy}$, then:
 (a) $S_x \neq S_y$ (b) $S_x = S_y$
 (c) $S_x > S_y$ (d) $S_x < S_y$
55. If $r_{xy} = 0.4$, then $r_{(2x,2y)}$ is equal to:
 (a) 0.4 (b) 0.8
 (c) 0 (d) 1
56. r_{xx} is equal to:
 (a) 0 (b) -1
 (c) 1 (d) 0.5
57. If $r_{xy} = 0.75$, then correlation coefficient between $u = 1.5X$ and $v = 2Y$ is:
 (a) 0 (b) 0.75
 (c) -0.75 (d) 1.5
58. If $b_{yx} = -2$ and $r_{xy} = -1$, then b_{xy} is equal to:
 (a) -1 (b) -2
 (c) 0.5 (d) -0.5
59. If $b_{yx} = 1.6$ and $b_{xy} = 0.4$, then r_{xy} will be:
 (a) 0.4 (b) 0.64
 (c) 0.8 (d) -0.8
60. If $b_{yx} = -0.8$ and $b_{xy} = -0.2$, then r_{yx} is equal to:
 (a) -0.2 (b) -0.4
 (c) 0.4 (d) -0.8
61. If $\hat{Y} = 6 - X$, then r will be:
 (a) 0 (b) 1
 (c) -1 (d) both (b) and (c)
62. If $\hat{Y} = X + 10$, then r is equal to:
 (a) 1 (b) -1
 (c) 1/2 (d) difficult to tell
63. If $Y = -10X$ and $X = -0.1Y$, then r is equal to:
 (a) 0.1 (b) 1
 (c) -1 (d) 10
64. If the figure +1 signifies a perfect positive correlation and the figure -1 signifies a perfect negative correlation, then the figure 0 signifies:
 (a) a perfect correlation (b) uncorrelated variables
 (c) not significant (d) weak correlation

65. A perfect positive correlation is signified by:
- (a) 0
 - (b) -1
 - (c) +1
 - (d) -1 to +1
66. If a statistics professor tells his class: "All those who got 100 on the statistics test got 20 on the mathematics test, and all those that got 100 on the mathematics test got 20 on the statistics test", he is saying that the correlation between the statistics test and the mathematics test is:
- (a) negative
 - (b) positive
 - (c) zero
 - (d) difficult to tell
67. If $\Sigma(X - \bar{X})(Y - \bar{Y})$ is zero, the correlation is:
- (a) weak negative
 - (b) high positive
 - (c) high negative
 - (d) none of the preceding
68. If r is negative, we know that:
- (a) $\Sigma(X - \bar{X})^2$ and $\Sigma(X - \bar{X})(Y - \bar{Y})$ are negative
 - (b) $\Sigma(Y - \bar{Y})^2$ and $\Sigma(X - \bar{X})(Y - \bar{Y})$ are negative
 - (c) $\Sigma(X - \bar{X})(Y - \bar{Y})$ is negative
 - (d) either $\Sigma(X - \bar{X})^2$ or $\Sigma(Y - \bar{Y})^2$ is negative

Answers

1. (b)	2. (b)	3. (b)	4. (a)	5. (c)	6. (a)	7. (d)	8. (b)
9. (c)	10. (d)	11. (c)	12. (b)	13. (a)	14. (a)	15. (b)	16. (c)
17. (c)	18. (b)	19. (d)	20. (a)	21. (b)	22. (b)	23. (b)	24. (c)
25. (a)	26. (a)	27. (a)	28. (b)	29. (a)	30. (b)	31. (a)	32. (c)
33. (b)	34. (c)	35. (b)	36. (b)	37. (a)	38. (a)	39. (a)	40. (a)
41. (c)	42. (a)	43. (a)	44. (c)	45. (d)	46. (c)	47. (b)	48. (c)
49. (d)	50. (c)	51. (d)	52. (c)	53. (d)	54. (b)	55. (a)	56. (c)
57. (b)	58. (d)	59. (c)	60. (b)	61. (c)	62. (a)	63. (c)	64. (b)
65. (c)	66. (a)	67. (d)	68. (c)				

SHORT QUESTIONS

1. Suppose that $Y = 1$ when $X = 0$, that $Y = 2$ when $X = 1$, and that $Y = 3$ when $X = 2$. Find the least-squares estimate b .

Ans. 1.0

2. Suppose that $Y = 1$ when $X = 0$, that $Y = 2$ when $X = 1$, and that $Y = 3$ when $X = 2$. Find the least-squares estimate a .

Ans. 1

3. Given $\bar{X} = 1$, $\bar{Y} = 8$ and $b = 2$. Find the value of intercept a .

Ans. 6

4. Given $Y = 16, 18, 20$ and $X = 0, 1, 2$. Find the value of Y intercept.

Ans. 16

5. Given $Y = 6, 8, 10$ and $X = 0, 1, 2$. Find the regression coefficient of Y on X .

Ans. 2

6. If $\bar{X} = 50$, $\bar{Y} = 110$ and $a = 10$. Find the value of b .

Ans. 2

7. Find the equation for the straight line whose intercept and slope are -3 and $2/3$ respectively.

Ans. $Y = \frac{2}{3}X - 3$

8. Find the slope and Y intercept of the line whose equation is $3X - 5Y = 20$.

Ans. $\frac{3}{5}$ and -4

9. Find the equation of a regression line whose X and Y intercepts are 3 and -5 .

Ans. $5X - 3Y = 15$

10. If the regression lines of Y on X and X on Y are respectively given by $2X - 3Y = 0$ and $4Y - 5X = 8$. Find the values of two regression coefficients of Y on X and X on Y .

Ans. $\frac{2}{3}$ and $\frac{4}{5}$

11. If $Y = 2+3X$, and if the expected value of X is 10 . Find the expected value of Y .

Ans. 32

12. If $Y = 30 - 2X$, and if the variance of X is 8 . Find the variance of Y .

Ans. 32

13. Given the equation of the straight line $\hat{Y} = a + bX$, and the values of $a = 45$, $b = -10$ and $X = 3$. Find the value of \hat{Y} .

Ans. 15

14. Given $a_{yx} = \bar{Y} - b_{yx} \bar{X}$ and $\bar{Y} = 1.87$, $b_{yx} = 0.25$ and $\bar{X} = 12.45$. Find a_{yx} .

Ans. -1.24

15. Given $\hat{Y} = 109.73 + 1.58 X$ and $X = 100$. Find \hat{Y} .

Ans. 267.73

16. Given $\hat{X} = 0.6 - 0.5 Y$ and $Y = 0.8$. Find \hat{X} .

Ans. -0.2

17. In the equation $\hat{Y} = \bar{Y} + r \frac{S_y}{S_x} (X - \bar{X})$, if $\bar{Y} = 18.0$, $r = -0.95$, $S_x = 5$, $S_y = 3.2$, $\bar{X} = 30$, and $X = 20$. Find \hat{Y} .

Ans. 11.92

18. In the equation $\hat{X} = \bar{X} + r \frac{S_x}{S_y} (Y - \bar{Y})$, if $Y = 192.50$, $\bar{X} = 15.65$, $r = 0.55$, $S_x = 4.3$, $S_y = 25$, and $\bar{Y} = 170$. Find \hat{X} .

Ans. 17.8

19. Suppose that $Y = 1$ when $X = 0$, that $Y = 2$ when $X = 1$, and that $Y = 3$ when $X = 2$. In this case, find the sample correlation coefficient r .

Ans. 1

20. If $Y = 10, 8, 6$ and $X = 2, 1, 0$. Find the coefficient of correlation.

Ans. 1

21. Given $Y = 10, 8, 6$ and $X = 0, 1, 2$. Find the sample correlation coefficient.

Ans. -1

22. The equations of two regression lines obtained from ten observations are: $10X = 5Y - 55$ and $100Y = 200X + 1180$. Find the correlation coefficient between X and Y .

Ans. 1

23. For 8 observations on deposits (X) and loans (Y), two regression equations are established which are $\hat{Y} = 54.62 + 0.38X$ and $\hat{X} = 58.72 - 0.27Y$. Find the correlation coefficient between deposits and loans.

Ans. -0.32

24. A set of data yields the following regression equation: $\hat{Y} = 16 - 0.5X$. Find the average value of Y for $X = 6$.

Ans. 13

25. If $b_{yx} = 1.6$ and $b_{xy} = 0.4$, find the value of r_{xy} .

Ans. 0.8

26. If $b_{yx} = -1.6$ and $b_{xy} = -0.4$, find the value of r_{xy} .

Ans. - 0.8

27. For two variables X and Y the regression equation of X on Y is $\hat{X} = 5Y - 7$ and the regression equation of Y on X is $\hat{Y} = 0.1X + 1.7$. Find the coefficient of correlation between X and Y.

Ans. 0.71

28. Given $S_{xy} = 16$ and $S_x S_y = 81$. Find r.

Ans. 0.20

29. Given $b_{yx} = 0.82$ and $r_{xy} = 0.97$. Find b_{xy} .

Ans. 1.15

30. Given $b_{xy} = -1.4$ and $r_{xy} = -0.87$. Find b_{yx} .

Ans. -0.54

31. Given $r_{xy} = 0.8$, $S_x = 4$, $S_{xy} = 20$. Find the standard deviation of X.

Ans. 6.25

32. Given $r_{xy} = -0.75$, $S_y = 5$, $\Sigma(X - \bar{X})(Y - \bar{Y}) = -15n$. Find S_x .

Ans. 4

33. Given $r = 0.605$, $\Sigma(X - \bar{X})(Y - \bar{Y}) = 24$, $S_x = 2.12$ and $S_y = 2.34$. Find the number of items.

Ans. 8

34. Given $\Sigma(X - \bar{X})(Y - \bar{Y}) = 0$, $\Sigma(X - \bar{X})^2 = 10$, $\Sigma(Y - \bar{Y})^2 = 10$ and $n = 5$. Find the coefficient of correlation.

Ans. 0

35. Interpret the meaning when:

- (i) $r = +1$
- (ii) $r = -1$
- (iii) $r = 0$
- (iv) $r = -0.98$
- (v) $r = 0.2$
- (vi) $r = 2$

Ans.(i) perfect positive correlation **(ii)** perfect negative correlation

- (iii) no correlation
- (iv) high degree of negative correlation

- (v) week positive correlation

- (vi) not possible because r lies between -1 and +1

36. Explain the difference between $r = -0.80$ and $r = 0.80$.

Ans. $r = -0.80$ indicates that the two variables have a strong negative relationship, whereas $r = +0.80$ indicates that two variables have a strong positive relationship. The two coefficients indicate equally strong relationship.

37. What is meant by regression?
38. Explain the terms regressand and regressor.
39. Differentiate between linear regression and curvilinear regression.
40. Explain the difference between fixed variable and random variable.
41. Explain the terms regression and linear regression.
42. Write a short note on scatter diagram.
43. Define simple linear regression.
44. Define correlation.
45. Distinguish between positive and negative correlation.
46. Differentiate between perfect positive and perfect negative correlation.
47. Write down the properties of the correlation coefficient.
48. Sketch the scatter diagrams for the following terms:
 - (a) perfect positive linear correlation (b) perfect negative linear correlation
 - (c) strong positive linear correlation (d) strong negative linear correlation
49. Sketch the scatter diagrams for the following terms:
 - (a) no linear correlation (b) weak positive linear correlation
 - (c) weak negative linear correlation (d) positive correlation
50. Differentiate between linear and non-linear correlation.
51. Define the terms no correlation and curvilinear correlation.
52. Differentiate between regression and correlation.
53. Explain the meanings of correlation and perfect correlation.
54. What is meant by correlation coefficient and write its properties?
55. Write down the properties of the least squares regression line.
56. Discuss the method of least squares for fitting the regression lines of Y on X and X on Y.
57. Write down the aims of regression and correlation analysis.
58. Define the terms regression analysis and correlation analysis.
59. What is meant by residual?

EXERCISES

1. Compute the regression equation of Y on X from the following data using normal equations.

X	25	30	40	50	65
Y	6	5	4	8	7

Ans: $\hat{Y} = 3.774 + 0.053 X$

2. Determine the regression equation to the following data taking X as the independent variable. Also find the difference between the actual values of Y and the values obtained from the fitted line and show that $\sum(Y - \hat{Y}) = 0$.

X	5	10	15	20	25
Y	25	20	15	10	5

Ans: $\hat{Y} = 30 - X$.

3. The following sample observations were randomly selected:

X	4	5	3	6	12
Y	4	6	5	7	8

Determine the value of \hat{Y} when X is 7.

Ans: $\hat{Y} = 3.72 + 0.38X; 6.38$

4. Show that the sum of errors equals zero and the sum of squares of the errors is equal to 1.1

X	1	2	3	4	5
Y	1	1	2	2	4

Ans: $\sum(Y - \hat{Y}) = 0$ $\sum(Y - \hat{Y})^2 = 1.1$

5. On the basis of figures recorded below for supply and price for five years, construct a regression equation of price on supply. Compute from the equation established the most likely price when supply is 92 units.

Year	1996	1997	1998	1999	2000
Supply	87	90	98	95	90
Price	132	125	115	123	140

Ans: $\hat{Y} = 281.56 - 1.68 X; 127$

7. A market research firm wishes to develop a model to predict purchases of tennis balls by city, based on the number of tennis courts in a city. A simple random sample of 50 cities developed the following data:

X = number of courts in a city

Y = thousands of tennis balls sold in the city

$$\bar{X} = 235, \bar{Y} = 375, \Sigma XY = 4435650, \Sigma X^2 = 2780850$$

What is the equation of the estimated regression line that you would use to predict Y from X ?

Ans: $\hat{Y} = 1.5 X + 22.5$

7. A university administrator studied the relationship between the cost of operating an academic department and the total student-hours of teaching and supervision undertaken in the department, for eleven departments for the most recent academic year. The results are summarized below in a convenient form:

X = number of student-hours (in thousands), Y = cost (Rs. thousands),

$$n = 11, n\Sigma X^2 - (\Sigma X)^2 = 16810000, \Sigma X = 7040, n\Sigma Y^2 - (\Sigma Y)^2 = 8910000, \Sigma Y = 4235, n\Sigma XY - (\Sigma X)(\Sigma Y) = 9245500. \text{ Compute the two regression equations.}$$

Ans. $\hat{X} = 1.04 Y + 239.6; \hat{Y} = 0.55 X + 33$

8. For 9 observations on supply (X) and price (Y) the following data was obtained:

$$\Sigma(X - 90) = -25, \Sigma(X - 90)^2 = 301, \Sigma(Y - 127) = 12, \Sigma(Y - 127)^2 = 1006,$$

$\Sigma(X - 90)(Y - 127) = -469$. Obtain the line of regression of X on Y and estimate the supply when the price is Rs. 125.

Ans. $\hat{X} = -0.44 Y + 143.6852; 88.69$

9. Given the following information, estimate:

(i) the value of X when $Y = 30$ (ii) the value of Y when $X = 55$

The mean value of $X = 54$. The mean value of $Y = 28$. The regression coefficient of X on $Y = -0.2$. The regression coefficient of Y on $X = -1.5$.

Ans. (i) $\hat{X} = -0.2 Y + 59.6; 53.6$ (ii) $\hat{Y} = -1.5 X + 109; 26.5$

10. Computing from a data set of (X, Y) values, the following summary statistics were recorded. $n = 18, \bar{X} = 1.2, \bar{Y} = 5.1, S_x^2 = 14.10, S_y^2 = 2.01, S_{xy} = 2.31$. Construct the regression lines of Y on X and X on Y .

Ans. $\hat{Y} = 0.164 X + 4.903; \hat{X} = 1.149 Y - 4.66$

11. Fitting a straight line to a set of data yields the following regression equation is obtained:

$$\hat{Y} = 16 - 0.5 X$$

- (i) Interpret the meaning of the Y-intercept 'a'.
- (ii) Interpret the meaning of the slope 'b'.
- (iii) Predict the average value of Y for X = 6.

Ans. (i) The Y intercept $a = 16$ means that when $X = 0$, the average value of Y is 16.
 (ii) The slope $b = -0.5$ means that for each increase of one unit of X, the value of Y is expected to decrease on average by 0.5 unit. (iii) 13

12. The following statistics have been computed:

	X series	Y series
Number of items	14	14
Arithmetic mean	83	2.52
Sum of square of deviations from arithmetic mean	1544	5.6198

Summation of products of deviations of X and Y series from their arithmetic means = 90.91.

- (i) Compute the regression line of Y on X and estimate the value of Y when X = 88.
- (ii) Compute the regression line of X and Y and estimate the value of X when Y = 3.25.

Ans. (i) $\hat{Y} = 0.06X - 2.46$; 2.82 (ii) $\hat{X} = 16.18 Y + 42.23$; 94.815

13. Compute the regression coefficients in each of the following cases:

- (i) $n=24$, $\Sigma X = 5402$, $\Sigma Y = 4378$, $\Sigma X^2 = 1388656$, $\Sigma Y^2 = 911032$, $\Sigma XY = 1118516$
- (ii) $n = 10$, $\Sigma(X - \bar{X})^2 = 170$, $\Sigma(Y - \bar{Y})^2 = 140$, $\Sigma(X - \bar{X})(Y - \bar{Y}) = 92$
- (iii) $\Sigma D_x = 12$, $\Sigma D_y = -5$, $\Sigma D_x D_y = 390$, $\Sigma D_x^2 = 2830$, $\Sigma D_y^2 = 91$, $n = 10$
- (iv) $\Sigma(X - \bar{X})(Y - \bar{Y}) = 148$, $S_x = 7.933$, $S_y = 16.627$, $n = 15$

Ans. (i) $b_{yx} = 0.77$, $b_{xy} = 1.18$ (ii) $b_{yx} = 0.54$, $b_{xy} = 0.66$

(iii) $b_{yx} = 0.14$, $b_{xy} = 4.47$ (iv) $b_{yx} = 0.16$, $b_{xy} = 0.04$

14. The following data were obtained in a study of the relationship between the weight and chest size of infants at birth:

Weight (kg)	3.7	3.2	2.7	5.0	4.4
Chest size (cm)	28.7	27.2	29.5	36.4	32.2

Compute and interpret the sample correlation coefficient.

Ans: 0.854. There is a strong positive relationship between the weight and chest size.

15. A personnel officer is studying performances of job applicants on two tests given when the applicant contacts the firm. The first test measures mental ability; the second measures potential for success in the job. The test-score results of a sample of six applicants are shown below:

Applicant	A	B	C	D	E	F
Mental ability (X)	37	40	36	49	36	40
Potential (Y)	63	42	41	39	38	49

Calculate the sample correlation coefficient.

Ans: - 0.27

16. The production manager of a factory would like to develop a model to predict performance time for a manual assembly task based upon the amount of time spent in training. A sample of 5 recent employees was selected; the training time in hours and the performance time in minutes are presented below:

Observation	1	2	3	4	5
Training time (hours)	27	24	12	22	13
Performance time (minutes)	19	16	12	17	10

Compute the coefficient of correlation between training time and performance time.

Ans: 0.95

17. Compute and interpret the coefficient of correlation between the values of X and Y from the following table.

X	21	22	23	24	25
Y	25	24	23	22	26

Ans. $r_{xy} = 0$. So X and Y are uncorrelated.

18. Calculate the correlation coefficient between X and Y and the correlation coefficient between X and Z.

X	2	4	6	8	10
Y	10	15	20	25	30
Z	40	36	32	28	24

Ans. $r_{xy} = 1$, $r_{xz} = -1$

19. From the following table, compute the coefficient of correlation by Karl Pearson's method:

X	4	6	?	2	8
Y	8	9	5	11	7

Arithmetic means of X and Y series are 6 and 8 respectively.

Ans. Missing observation = 10, $r = -0.92$

20. Given the following information:

Number of pairs of observations of X and Y series	= 15
Arithmetic mean of X series	= 25
Standard deviation of X series	= 3.01
Arithmetic mean of Y series	= 18
Standard deviation of Y series	= 3.03
Sum of products of deviations from means of X and Y series	= 122
Compute coefficient of correlation between X and Y.	

Ans. 0.89

21. Compute the coefficient of correlation between X and Y from the following data:

Sum of deviations of X series	= 5
Sum of deviations of Y series	= 4
Sum of square of deviations of X series	= 40
Sum of square of deviations of Y series	= 50
Sum of the product of deviations of series X and Y	= 32
Number of pairs of observations of X and Y series	= 10

Ans. 0.704

22. (i) Coefficient of correlation between two variates X and Y is 0.8. The variance of X is 16 and $S_{xy} = 20$. Find the standard deviation of Y variate.
- (ii) If the coefficient of correlation between X and Y is - 0.75, the standard deviation of Y series is 5 and $\Sigma(X - \bar{X})(Y - \bar{Y}) = - 15n$. What will be the standard deviation of X series.
- (iii) From the following information, calculate the number of items for which $r = 0.5$, $\Sigma(X - \bar{X})(Y - \bar{Y}) = 120$, standard deviation of Y series = 8 and $\Sigma(X - \bar{X})^2 = 90$.

Ans. (i) 6.25 (ii) 4 (iii) 10

23. In order to find the correlation coefficient between two variables X and Y, the following results were obtained:

$$n = 10, \Sigma X = 500, \Sigma Y = 1100, \Sigma X^2 = 28400, \Sigma Y^2 = 134660, \Sigma XY = 61800.$$

It was however later discovered at the time of checking that two particular sets of observations, namely X = 57 and 35, Y = 121 and 114 was wrongly taken, the correct values being X = 67 and 15, Y = 112 and 124. Compute the correct value of the correlation coefficient.

Ans. 0.794

24. In order to find the coefficient of correlation between two variables X and Y from 7 pairs of observations, the following results are given:

$$\Sigma X = 220, \Sigma Y = 47.56, \Sigma XY = 1584.98, \Sigma X^2 = 7888, \Sigma Y^2 = 341.1628.$$

Later it was found that one particular set of observations, namely $X = 34$, $Y = 4.25$ was wrongly taken, the correct values being $X = 24$, $Y = 6.25$. Compute and interpret the correct value of the coefficient of correlation.

Ans. 0.9724. There is high degree of positive correlation between X and Y.

25. Compute and interpret the correlation coefficient on the basis of the following informations:

X = maximum temperature (C°) during each day and Y = sales figures per day of lemonade.

$$\Sigma (X - 17) = -2, \Sigma (X - 17)^2 = 218, \Sigma (Y - 15) = -48, \Sigma (Y - 15)^2 = 1366, \Sigma (X - 17)(Y - 15) = 464, n = 30.$$

Ans. 0.8695. There is strong positive correlation between temperature and sale.

26. The following summary statistics were recorded:

$$n = 20, \bar{X} = 25, \bar{Y} = 35, \Sigma(X - \bar{X})^2 = 80, \Sigma(Y - \bar{Y})^2 = 170, \Sigma(X - \bar{X})(Y - \bar{Y}) = -100$$

Show that the coefficient of correlation is the geometric mean of regression coefficients.

Ans: $r_{xy} = -0.86$, $b_{yx} = -1.25$, $b_{xy} = -0.59$

27. The equations of two regression lines of Y on X and X on Y respectively are given from paired observations of two variables X and Y:

$$8X - 10Y + 66 = 0 \text{ and } 40X - 18Y - 214 = 0$$

- (i) Compute and interpret the coefficient of correlation.
(ii) Compute the values of \bar{X} and \bar{Y} .

Ans: (i) $r = 0.6$ (ii) $\bar{X} = 13$ and $\bar{Y} = 17$

28. Computed from a data set of (X, Y) values, the following summary statistics were recorded:

$$n = 5, \Sigma X = 15, \Sigma Y = 25, \Sigma(X - \bar{X})^2 = 10, \Sigma(Y - \bar{Y})^2 = 26, \Sigma(X - \bar{X})(Y - \bar{Y}) = 13.$$

Compute the two regression equations. Also compute the value of correlation coefficient

Ans: $\hat{Y} = 1.3X + 1.1$ $\hat{X} = 0.5Y + 0.5$ $r = 0.81$

29. The following statistics have been computed:

$$\bar{X} = 14.19, \bar{Y} = 21.65, S_x = 5.71, S_y = 6.73, r = 0.651$$

Construct the two regression equations.

Ans: $\hat{Y} = 10.766 + 0.767X$ $\hat{X} = 2.239 + 0.552Y$

80. The following information are given:

$$\bar{X} = 20, \bar{Y} = 40, S_x = 2, S_y = 4, r = 0.70$$

Predict the most probable value of Y when X = 25 and most probable value of X when Y = 30.

Ans: $\hat{Y} = 1.4X + 12; 47 \quad \hat{X} = 0.85Y + 6; 16.5$

81. If the mean weight of 200 fathers is 140 pounds with standard deviation of 6 pounds and the mean weight of their youngest sons is 142 pounds with standard deviation of 8 pounds. The coefficient of correlation between them is 0.9. Estimate the two regression equations.

Ans: $\hat{Y} = 1.2X - 26 \quad \hat{X} = 0.675Y + 44.15$

82. The mean monthly index numbers of prices of animal feeding stuff during certain years is Rs. 120 with standard deviation 10 and the mean monthly index numbers of prices of home grown oats is 100 with standard deviation 8, the coefficient of correlation between them being 0.8. Construct the regression equations of Y on X and X on Y.

Ans: $\hat{Y} = 0.64X + 23.2 \quad \hat{X} = Y + 20$

ASSOCIATION

15.1 VARIABLE AND ATTRIBUTE

There are 4 persons and their heights in inches are 55, 56, 72 and 74. Here height is a characteristic and the figures 55, 56, 72 and 74 are the values of a variable. These figures are the result of measurements. You know that the measurements generate the continuous variable. Thus the variable on heights is a continuous variable. Suppose we select 4 bulbs from a certain lot and inspect them. The lot contains good as well as defective bulbs. The sample may contain 0, 1, 2, 3, 4 defective bulbs. The values 0, 1, 2, 3 and 4 are the values of a discrete variable.

Out of 4 persons whose heights are given above, 2 are tall with heights 72 and 74 inches and 2 are short with heights 55 and 56 inches. When we use the words, tall and short, any variable is not under consideration. We do not make any measurements. We only see who is tall and who is short. Here level of height *tall* or *short* is not a variable, it is called an *attribute*. Out of 4 bulbs 2 are good and 2 are defective. Here also any variable is not under consideration. We only count the defective bulbs and good bulbs. We examine whether the quality of being defective is present in a bulb or not. The status of the bulb is an attribute with two outcomes good and defective. Thus attribute is a quality and the data is collected to see how many objects possess the quality of being defective and how many elements do not possess this quality. Other famous examples of the attributes are level of education, level of smoking, level of social work, level of income, religion and colour etc. The data on the attribute is the result of recording the presence and absence of a certain quality (attribute) in the individuals. The data on the variables are called the quantitative data whereas the data on the attributes are called qualitative data or count data. As the data on the variables is collected for the purpose of analysis of data and for inference about the population parameters, similarly the data on the attribute or attributes is collected for the purpose of analysis of data and for testing of hypotheses about the attributes. We shall discuss the hypothesis testing about attributes in the subsequent topic in this Chapter.

15.1.1 NOTATION FOR ATTRIBUTES

For a single variable we use the symbol X and if there are two variables, we use the symbols X and Y for them. When there is a single attribute like height, the word, 'tall' may be denoted by A and 'short' may be denoted by α . If the tall and the short persons are divided into intelligent and 'non-intelligent' persons, then 'intelligent'

may be denoted by B and β may be used for the opposite attribute 'non-intelligent'. It may be noted that the word *attribute* is used for the main group like intelligence and the sub-groups 'intelligent' and 'non-intelligent' are also called attributes.

15.1.2 ONE ATTRIBUTE

Suppose that there are 100 individuals in a certain sample, the sample size is denoted by n . These 100 individuals are divided into two mutually exclusive groups on the basis of the attribute of height. Out of 100, 60 are tall and 40 are short. If 'tall' are denoted by A and short are denoted by α , we can write:

$$\begin{array}{ccc} A & \alpha \\ 60 & 40 & n = 100 \end{array}$$

There are two groups and we say that there are two classes A and α and the class frequency under A is 60. It is written as $(A) = 60$, similarly the number of individuals under α is written as $(\alpha) = 40$. Thus the attributes written within the brackets show their class frequencies. In this example the sample is divided into two groups i.e; two classes 'tall' and 'short'. Dividing the data into two groups is called *dichotomy* which means *cutting into two*. In this example a single attribute '*height*' divides the data in two groups. As only one attribute is involved, the data is called *one-way classification*. We can make a small table as below:

One-Way Classification

$$\begin{array}{ccc} A & \alpha \\ 60 = (A) & 40 = (\alpha) & n = 100 \end{array}$$

Clearly $(A) + (\alpha) = n$

The symbols (A) and (α) are used to denote the frequency of individuals who possess A and who do not possess A (α means not ' A '). It may be noted that the symbol ' A ' is not necessarily fixed for 'tall'. In some other discussion 'short' may be denoted by A .

15.1.3 TWO ATTRIBUTES

The tall and short persons may further be divided into intelligent and non-intelligent persons. Intelligence may be denoted by B and β may be used for non-intelligence. The following table shows different attributes and their combinations. When two attributes are involved, the division of the sample as below is called two-way classification.

Table 15.1.
Two-Way Classification

	A	α	Total
B	(AB)	(αB)	(B)
β	$(A\beta)$	$(\alpha\beta)$	(β)
Total	(A)	(α)	n

The column totals are denoted by (A) and (α) and the row totals are denoted by (B) and (β). The above table contains 2 rows and 2 columns and is therefore called 2×2 contingency table or 2×2 cross-tabulation briefly written as 2×2 cross-table.

There may be more than two attributes. The symbols A, B, C are used for the attributes and α , β , γ are used for the absence of the attributes A, B, C. Thus α means not A and β means not B and γ means not C.

Suppose that out of 60 tall persons, 30 are intelligent and out of 40 short persons, 20 are intelligent. We can write these frequencies in the following 2×2 contingency Table 15.2.

**Table 15.2.
 2×2 Contingency Table**

	A	α	Total
B	(AB) = 30	(α B) = 20	(B) = 50
β	(A β) = 30	(α β) = 20	(β) = 50
Total	(A) = 60	(α) = 40	n = 100

From table 15.2. we can write some relations immediately.

- (i) (A) + (α) = n
- (ii) (B) + (β) = n
- (iii) (A) = (AB) + (A β)
- (iv) (α) = (α B) + (α β)
- (v) (B) = (AB) + (α B)
- (vi) (β) = (A β) + (α β)

15.1.4 POSITIVE AND NEGATIVE CLASSES

The classes A, B, AB are called positive classes because they contain all positive attributes, the classes α , β , $\alpha\beta$ are called negative classes because they have negative attributes. The classes αB and $A\beta$ contain both positive and negative attributes, they are called mixed or contrary classes.

If we have three attributes A, B, C with their opponents or complements as α , β and γ , then we can write the different class frequencies as below in Table 15.3.

Table 15.3.

	A		α		
	C	γ	C	γ	Total
B	(ABC)	(AB γ)	(α BC)	(α B γ)	(B)
β	(A β C)	(A β γ)	(α β C)	(α β γ)	(β)
	(AC)	(A γ)	(α C)	(α γ)	
Total	(A)		(α)		n

In this table the positive classes are A, B, C, AB, AC, BC and ABC whereas the negative classes are α , β , γ , $\alpha\beta$, $\alpha\gamma$, $\beta\gamma$, $\alpha\beta\gamma$. All other classes are mixed.

15.1.5 ORDER OF CLASSES

The order of the class depends upon the number of attributes going into that class. If a certain class can give us information about only one attribute, it is called class of order one.

The classes A, α , B and β are classes of the order one and the frequencies (A), (α), (B) and (β) are the frequencies of order one. The classes AB, $\alpha\beta$, αB and $\alpha\beta$ are the classes of order two and the frequencies (AB), ($\alpha\beta$), (αB) and ($\alpha\beta$) are the frequencies of order two. The classes ABC, $\alpha B\gamma$, αBC , $\alpha B\gamma$, $A\beta C$, $A\beta\gamma$, $\alpha\beta C$ and $\alpha\beta\gamma$ are the classes of order three and (ABC), ($\alpha B\gamma$) ... ($\alpha\beta\gamma$) are the frequencies of order three. The sample size n does not contain any attribute and is therefore called frequency of order zero.

15.1.6 ULTIMATE CLASS FREQUENCIES

In a certain given situation, the *ultimate class frequencies* are the frequencies with the highest order. For two attributes, the ultimate class frequencies are (AB), ($\alpha\beta$), (αB) and ($\alpha\beta$).

For three attributes, the ultimate class frequencies are of order 3 which are (ABC), ($\alpha B\gamma$), (αBC), ($\alpha B\gamma$), ($A\beta C$), ($A\beta\gamma$), ($\alpha\beta C$) and ($\alpha\beta\gamma$).

15.1.7 LOWER ORDER FREQUENCIES IN TERMS OF HIGHER ORDER FREQUENCIES

Let us discuss the relation of the lower order frequencies in terms of higher order frequencies. Let us consider the different cases.

(i) Single Attribute

$$n = (A) + (\alpha)$$

(ii) Two Attributes

Let us consider 2×2 contingency Table 15.2. for the frequencies of the two attributes. Clearly $n = (A) + (\alpha)$ $n = (B) + (\beta)$

$$(A) = (AB) + (A\beta) \quad (\alpha) = (\alpha B) + (\alpha\beta)$$

$$(B) = (AB) + (\alpha B) \quad (\beta) = (A\beta) + (\alpha\beta)$$

(iii) Three Attributes

Let us take help from Table 15.3. to write lower order frequencies into higher order frequencies. Clearly

$$n = (A) + (\alpha) \quad n = (B) + (\beta)$$

$$(A) = (AC) + (A\gamma). \text{ But } (AC) = (ABC) + (A\beta C) \text{ and } (A\gamma) = (AB\gamma) + (A\beta\gamma)$$

$$\text{Thus } (A) = (ABC) + (A\beta C) + (AB\gamma) + (A\beta\gamma)$$

$$\text{Similarly } (\alpha) = (\alpha C) + (\alpha\gamma) = (\alpha BC) + (\alpha\beta C) + (\alpha B\gamma) + (\alpha\beta\gamma)$$

$$n = (A) + (\alpha) = (ABC) + (A\beta C) + (AB\gamma) + (A\beta\gamma) + (\alpha BC) + (\alpha\beta C) + (\alpha B\gamma) + (\alpha\beta\gamma)$$

$$(B) = (AB) + (\alpha B) = (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma)$$

$$(\beta) = (A\beta) + (\alpha\beta) = (A\beta C) + (A\beta\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$$

$$\text{and } n = (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma) + (A\beta C) + (A\beta\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$$

With the help of the Tables 15.2. and 15.3. we can easily write any lower order frequency in terms of higher orders.

15.1.8 HIGHER ORDER FREQUENCIES INTO LOWER ORDER FREQUENCIES

Sometimes we have to express the frequency of a higher order into frequencies of lower order. For this purpose we use the following operators. The frequency (A) is written as $n \cdot A$ as if $n \cdot A$ means A 's out of n . Similarly the frequency (α) is written as $n \cdot \alpha$ and (AB) is written as $n \cdot AB$ and (ABC) is written as $n \cdot ABC$.

$$\text{We know } (A) + (\alpha) = n$$

$$n \cdot A + n \cdot \alpha = n \quad \dots \dots \text{ equation (1)}$$

using the operators, we shall assume that algebraic operations are applicable on these operators. Dividing equation (1) by n , we get

$$A + \alpha = 1 \text{ or } A = 1 - \alpha \text{ and } \alpha = 1 - A$$

Similarly we can establish with the help of operators that

$$B = 1 - \beta \text{ and } \beta = 1 - B \quad C = 1 - \gamma \text{ and } \gamma = 1 - C$$

Example 15.1.

Express (AB) in terms of lower order frequencies with the help of operators.

Solution:

$$\text{We write } (AB) = n \cdot AB$$

$$\text{Putting } A = 1 - \alpha \text{ and } B = 1 - \beta$$

$$(AB) = n(1 - \alpha)(1 - \beta) = n [1 - \beta - \alpha + \alpha\beta] = n - n\beta - n\alpha + n\alpha\beta$$

Writing the original symbols for $n\beta$, $n\alpha$ and $n\alpha\beta$, we have

$$(AB) = n - (\beta) - (\alpha) + (\alpha\beta) \quad \text{or} \quad (AB) = n - (\alpha) - (\beta) + (\alpha\beta)$$

It is to be noted that the left hand side contains positive attributes and all attributes on the right side are negative except one frequency which is n . Any attribute on the left side does not appear on the right side in this type of relation.

Example 15.2.

Express $(\alpha\beta\gamma)$ in terms of lower order frequencies.

Solution:

$$(\alpha\beta\gamma) \text{ can be written as } n \cdot \alpha\beta\gamma. \text{ Thus } (\alpha\beta\gamma) = n \cdot \alpha\beta\gamma$$

Using the relations $\alpha = 1 - A$, $\beta = 1 - B$, $\gamma = 1 - C$ we get

$$(\alpha\beta\gamma) = n(1 - A)(1 - B)(1 - C)$$

$$= n - n \cdot A - n \cdot B - n \cdot C + n \cdot AB + n \cdot AC + n \cdot BC - n \cdot ABC$$

$$(\alpha\beta\gamma) = n - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC)$$

The attributes on the left side are negative and all attributes on the right side are positive except one frequency of order zero that is n .

Example 15.3.

Given the following frequencies: $n = 100$, $(AB) = 30$, $(A) = 40$, $(B) = 70$. Calculate all the remaining frequencies.

Solution:

We know $(A) + (\alpha) = n$. Thus $40 + (\alpha) = 100$ or $(\alpha) = 60$

We know $(B) + (\beta) = n$, hence $70 + (\beta) = 100$ or $(\beta) = 30$

Also $(B) = (AB) + (\alpha B)$, hence $70 = 30 + (\alpha B)$ or $(\alpha B) = 40$

Also $(AB) + (A\beta) = (A)$, hence $30 + (A\beta) = 40$ or $(A\beta) = 10$

Also $(\beta) = (A\beta) + (\alpha\beta)$, hence $30 = 10 + (\alpha\beta)$ or $(\alpha\beta) = 20$

These frequencies can be calculated very easily if the given frequencies are substituted in the 2×2 contingency table. The unknown frequencies can be calculated by simple addition or subtraction. Thus

	A	α	Total
B	$(AB) = 30$	$(\alpha B) = 40$	$(B) = 70$
β	$(A\beta) = 10$	$(\alpha\beta) = 20$	$(\beta) = 30$
Total	$(A) = 40$	$(\alpha) = 60$	$n = 100$

The unknown frequencies within the rectangles have been calculated by simple subtraction to complete the table.

Example 15.4.

There are three attributes and their ultimate class frequencies are:

$$(ABC) = 10 \quad (AB\gamma) = 30 \quad (\alpha BC) = 15 \quad (\alpha B\gamma) = 60$$

$$(A\beta C) = 20 \quad (A\beta\gamma) = 15 \quad (\alpha\beta C) = 40 \quad (\alpha\beta\gamma) = 70$$

Calculate all the negative class frequencies of order one and order two.

Solution:

$$(\alpha) = (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma) = 15 + 60 + 40 + 70 = 185$$

$$(\beta) = (A\beta C) + (A\beta\gamma) + (\alpha\beta C) + (\alpha\beta\gamma) = 20 + 15 + 40 + 70 = 145$$

$$(\gamma) = (AB\gamma) + (A\beta\gamma) + (\alpha B\gamma) + (\alpha\beta\gamma) = 30 + 15 + 60 + 70 = 175$$

$$(\alpha\beta) = (\alpha\beta C) + (\alpha\beta\gamma) = 40 + 70 = 110$$

$$(\alpha\gamma) = (\alpha B\gamma) + (\alpha\beta\gamma) = 60 + 70 = 130$$

$$(\beta\gamma) = (A\beta\gamma) + (\alpha\beta\gamma) = 15 + 70 = 85$$

These unknown frequencies can be calculated with the help of the following table.

	A		α		Total
	C	γ	C	γ	
B	$(ABC) = 10$	$(AB\gamma) = 30$	$(\alpha BC) = 15$	$(\alpha B\gamma) = 60$	$(B) = 115$
β	$(A\beta C) = 20$	$(A\beta\gamma) = 15$	$(\alpha\beta C) = 40$	$(\alpha\beta\gamma) = 70$	$(\beta) = 145$
	$(AC) = 30$	$(A\gamma) = 45$	$(\alpha C) = 55$	$(\alpha\gamma) = 130$	
Total		$(A) = 75$		$(\alpha) = 185$	$n = 260$

Clearly $(\alpha) = 185$ $(\beta) = 145$
 $(\gamma) = 30 + 15 + 60 + 70 = 175$ $(\alpha\beta) = 40 + 70 = 110$
 $(\alpha\gamma) = 60 + 70 = 130$ $(\beta\gamma) = 15 + 70 = 85$

15.2 CONSISTENCY

If the class frequencies are observed in a certain sample data and all class frequencies are recorded correctly then there will be no error in them and they will be called consistent. But sometimes the class frequencies are not recorded correctly and their column total and row total do not agree with the grand total. If there is some error in any class frequency, then we say that the frequencies are inconsistent. If one class frequency is wrong, it will affect some other frequencies as well. A simple test of consistency is that all frequencies should be positive. If any frequency is negative, it means that there is inconsistency in the sample data. If the data is consistent, all the ultimate class frequencies will be positive.

Example 15.5.

Given the frequencies: $n = 115$, $(B) = 45$, $(A) = 50$ and $(AB) = 50$.

Check for consistency of the data.

Solution:

The data is called consistent if all the ultimate class frequencies are positive. Let us calculate some frequencies of order two.

We know $(A) = (AB) + (A\beta)$

Here $(A) = 50$ and $(AB) = 50$

Thus $50 = 50 + (A\beta)$ or $(A\beta) = 0$

It does not indicate inconsistency because some frequency can be zero.

We know $(B) = (AB) + (\alpha B)$

$45 = 50 + (\alpha B)$ or $(\alpha B) = -5$

The data is inconsistent. It means the given frequencies are wrong. If we make a table of (2×2) , we get

	A	α	Total
B	$(AB) = 50$	$(\alpha B) = -5$	$(B) = 45$
β	$(A\beta) = 0$	$(\alpha\beta) = 70$	$(\beta) = 70$
Total	$(A) = 50$	$(\alpha) = 65$	$n = 115$

One frequency (αB) is negative in the table. Thus the sample data is inconsistent.

Example 15.6.

In a certain big college, 600 students of intermediate level were interviewed. They were asked to give their opinion about liking or disliking in the subjects of Mathematics, Statistics and Physics. The sample data sent by the enumerator was:

300 liked Mathematics.

350 liked Statistics.

340 liked Physics.

130 liked Mathematics and Statistics.

160 liked Mathematics and Physics. 180 liked Physics and Statistics.

100 liked all the three subjects. Examine the data for consistency.

Solution:

All the given frequencies can be written in the form of attributes. Let A, B, C denote liking Mathematics, Statistics and Physics respectively and α , β , γ are their opponents for disliking of the subjects. We are given

$$n = 600 \quad (A) = 300 \quad (B) = 350 \quad (C) = 340$$

$$(AB) = 130 \quad (AC) = 160 \quad (BC) = 180 \quad (ABC) = 100$$

All the given frequencies are positive, we can therefore calculate a negative class frequency of order three which is $(\alpha\beta\gamma)$.

$$\begin{aligned} \text{Now } (\alpha\beta\gamma) &= n \cdot \alpha\beta\gamma \\ &= n(1 - A)(1 - B)(1 - C) \\ &= n - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC) \\ &= 600 - 300 - 350 - 340 + 130 + 160 + 180 - 100 = -20 \end{aligned}$$

A negative frequency indicates that the sample data sent by the enumerator is incorrect (inconsistent).

15.3 INDEPENDENCE OF ATTRIBUTES

Let us consider certain examples before we discuss the *independence* in a formal manner.

Example 15.7.

Consider the following sample data on the liking of males and females for fish.

Gender

	Males	Females	Total
Like Fish	80	80	160
Do not like Fish	20	20	40
Total	100	100	200

Discussion: There are 100 males out of which 80 like fish and out of 100 females 80 like fish. Males and females have the same liking for fish. We say that there is independence between gender and liking or disliking of fish. Another way of saying the same thing is that there is no relation between the gender and liking for fish.

Example 15.8.

Consider the following sample data on smoking by adult males and adult females:

	Males	Females	Total
Smokers	20	1	21
Non-smokers	80	99	179
Total	100	100	200

There are 100 males out of which 20 are smokers and out of 100 females there is only one smoker. It means that the smokers in males are 20 times more than the smokers among females. Males have a strong relation or association with smoking. Thus males and smoking are strongly associated. We say that there is positive association between males and smoking. There are 99 females who are non-smokers as compared to 80 male non-smokers. Thus females are inclined towards non-smoking. The association between females and non-smoking is also of positive type. There is only 1 female smoker as compared to 20 male smokers. Thus there is negative association between females and smoking and there is also negative association between males and non-smoking. Thus in a certain contingency table, when there is positive association between two attributes, then in the same table there exists the negative association between some other pairs of attributes. If there is positive association between A and B, then α and β are also positively associated. In this case there is negative association between A and β , and between α and B.

The data in the Example 15.8. may be written as

	Males	Females	Total
	A	α	
Non-smokers, B	80	99	179
	(AB)	(α B)	
Smokers, β	20	1	21
	(A β)	(α β)	
Total	100	100	200

In this table 80 is less than 99 and 20 is greater than 1 (or 1 is less than 20). There is negative association between A and B and between α and β . There is positive association between A and β and between α and B. If the attributes in the one diagonal have positive association, then the attributes in the other diagonal have negative association.

15.3.1 DEFINITION OF INDEPENDENCE

We know that in probability, the two events A and B are called independent if the joint probability of $A \cap B$ is equal to the product of the marginal probabilities of A and B. Thus for independence of A and B

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

The same logic applies for defining independence of attributes. The two attributes are called independent if the probability of (AB) is equal to the product of the probability A and the probability of B. Consider a 2×2 contingency table as below:

	A	α	Total
B	(AB)	(αB)	(B)
β	$(A\beta)$	$(\alpha\beta)$	(β)
Total	(A)	(α)	n

If one individual is selected out of this table, then

$$P(AB) = \frac{(AB)}{n} \quad P(A) = \frac{(A)}{n} \quad P(B) = \frac{(B)}{n}$$

For independence $P(AB) = P(A) \cdot P(B)$

$$\frac{(AB)}{n} = \frac{(A)}{n} \cdot \frac{(B)}{n} \quad \text{or} \quad (AB) = \frac{(A)(B)}{n}$$

This is called rule or criterion of independence of two attributes A and B. The class frequency (AB) is called observed frequency and $\frac{(A)(B)}{n}$ is called expected frequency when A and B are independent. For independence of A and B, the rule is $(AB) = \frac{(A)(B)}{n}$. But this rule is applicable only on the attributes A and B. Similarly for independence of other attributes, we have the rules:

$$(\alpha B) = \frac{(\alpha)(B)}{n} \quad (A\beta) = \frac{(A)(\beta)}{n} \quad \text{and} \quad (\alpha\beta) = \frac{(\alpha)(\beta)}{n}$$

When $(AB) > \frac{(A)(B)}{n}$, then there is positive association between A and B.

Positive association between A and B means that proportion of A's in B's is greater than the proportion of A's in β 's.

When $(AB) < \frac{(A)(B)}{n}$, there is negative association between A and B.

Negative association between A and B means that proportion of A's in B's is less than the proportion of A's in β 's.

It is important to note that if A and B are associated in a positive manner, then α and β are also associated in the positive manner and other pairs $A\beta$ and αB will have negative association.

15.3.2 ANOTHER DEFINITION OF INDEPENDENCE

The two attributes A and B are called independent if the proportion of A's in B's is the same as in non B's (β 's).

$$\text{Proportion of A's in B's} = \frac{(AB)}{(B)} \quad \text{Proportion of A's in } \beta\text{'s} = \frac{(A\beta)}{(\beta)}$$

For independence these two proportions are equal.

$$\text{Thus } \frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \quad \left[\text{If } \frac{a}{b} = \frac{c}{d}, \text{ then } \frac{a}{b} = \frac{c}{d} = \frac{a+c}{b+d} \right]$$

$$\text{Therefore } \frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(AB) + (A\beta)}{(B) + (\beta)} = \frac{(A)}{n}$$

$$\text{Thus } \frac{(AB)}{(B)} = \frac{(A)}{n} \text{ or } (AB) = \frac{(A)(B)}{n}$$

This is called a simple rule of independence between A and B.

If A and B are independent then all the other pairs in the table are also independent. But if there is positive association between two pairs AB and $\alpha\beta$, then the other two pairs $A\beta$ and αB will have negative association as explained earlier.

Let us consider the data of Example 15.7.

	A	α	Total
B	$(AB) = 80$	$(\alpha B) = 80$	$(B) = 160$
β	$(A\beta) = 20$	$(\alpha\beta) = 20$	$(\beta) = 40$
Total	$(A) = 100$	$(\alpha) = 100$	$n = 200$

$$\text{Here } (AB) = 80 \quad \text{and} \quad \frac{(A)(B)}{n} = \frac{100 \times 160}{200} = 80$$

$$\text{Thus } (AB) = \frac{(A)(B)}{n}. \text{ Hence A and B are independent.}$$

It also implies independence between A and β , α and B and α and β . Let us check another pair.

$$(\alpha\beta) = 20 \quad \text{and} \quad \frac{(\alpha)(\beta)}{n} = \frac{100 \times 40}{200} = 20$$

$$(\alpha\beta) = \frac{(\alpha)(\beta)}{n}, \text{ there is independence between } \alpha \text{ and } \beta.$$

The students may check the other classes. The independence in this table means that men and women have the same liking for fish.

Example 15.9.

Men and women go to a certain store for buying the articles. They make the payment in cash or purchase on credit (loan). Investigate if there is any relation between mode of payment and the sex of the customer. Given the data below:

Payment

Sex	Cash	Credit
Males	80	40
Females	20	60

Solution:

Let us write the table along with the symbols

	A	α	Total
B	$(AB) = 80$	$(\alpha B) = 40$	$(B) = 120$
β	$(A\beta) = 20$	$(\alpha\beta) = 60$	$(\beta) = 80$
Total	$(A) = 100$	$(\alpha) = 100$	$n = 200$

$$(AB) = 80 \text{ and } \frac{(A)(B)}{n} = \frac{100 \times 120}{200} = 60, (AB) > \frac{(A)(B)}{n}$$

There is positive association between A and B. It means that males make the cash payments with greater frequency than the females. If we check the pair $(A\beta)$, we will find negative association.

$$(A\beta) = 20 \text{ and } \frac{(A)(\beta)}{n} = \frac{100 \times 80}{200} = 40, (A\beta) < \frac{(A)(\beta)}{n}$$

Thus there is negative association between A and β . Females are less inclined to make the cash payments. It is also clear from the given data. Out of 120 males, 80 make the payment on cash. $80 / 120 \times 100 = 66.7\%$ males make cash payment. 20 out of 80 means that $20 / 80 \times 100 = 25\%$ females make the cash payment. Thus males and cash payment go together with high frequency and are called positively related or associated.

Example 15.10.

We wish to determine if there is any difference in the popularity of football between college educated males and non college educated males. A sample of 100 college educated males showed that 55 were football fans. A sample of 200 non college educated males revealed that 125 were football fans. Is there any evidence of a difference in football popularity between college educated and non college educated males.

Solution:

We put the data in the following table.

	College educated males	Non college educated males	Total
	A	α	
Football fans, B	$55 = (AB)$	125	$180 = (B)$
Not football fans, β	45	75	$120 = (\beta)$
Total	100 (A)	200 (α)	$300 = n$

$$\text{Here } (AB) = 55, \frac{(A)(B)}{n} = \frac{100 \times 180}{300} = 60. (AB) < \frac{(A)(B)}{n}$$

Thus there is negative association between A and B. College-educated males show less of interest for football as compared to non college educated males. There is positive association between α and B. More of non college educated males are football fans as compared to college-educated males. Thus football is more popular among non college educated males. But here we are comparing only one observed frequency with the corresponding expected frequency. In Example 15.12, we shall compare all the observed frequencies with the corresponding expected frequencies. In Example 15.12, our inference will be different and we shall decide that whether there is independence between the attributes or not.

15.4 COEFFICIENT OF ASSOCIATION

When it is desired to calculate the level of association, we can calculate coefficient of association denoted by Q, where

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

This is called Yule's coefficient of association. It lies between -1 and $+1$. It is explained in the same manner as the coefficient of correlation r_{xy} between the two random variable X and Y.

If $Q = -1$ it is perfect negative association between the attributes on the top left corner in the 2×2 cross table.

If $Q = 0$ it means independence

If $Q = 1$ it means perfect positive association between attributes.

Let us calculate Q from the data given in Example 15.10.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{55 \times 75 - 45 \times 125}{55 \times 75 + 45 \times 125} = \frac{4125 - 5625}{4125 + 5625} = \frac{-1500}{9750} = -0.15$$

This indicates negative association between A and B. It is the same result as obtained earlier in Example 15.10.

15.5 χ^2 -DISTRIBUTION

Chi-square written as χ^2 is a statistic which has a positively skewed distribution as shown below. The value of χ^2 varies from 0 to ∞ . χ^2 cannot take any negative value. The shape of the distribution depends upon the degrees of freedom which is calculated from the given sample. χ^2 -distribution can be used for various purposes. One of the applications of χ^2 is to test the independence between the attributes.

15.5.1 TEST OF INDEPENDENCE

With the help of χ^2 -distribution, we can test whether the attributes are independent or there is association between them. The procedure runs as below:

1. The null hypothesis H_0 is framed

We assume that there is independence between the attributes.

The alternative hypothesis H_1 is that there is association between the attributes

2. Level of significance α is decided.

3. Test-statistic used is $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

where f_o stands for observed frequency and f_e stands for expected frequency calculated under the assumption that attributes are independent.

4. Computations:

The χ^2 -statistic can be used to check the independence in a table of attributes containing any number of columns and rows. Let us first explain the application of χ^2 on a 2×2 contingency table. The observed frequencies are given below in the form of a table. It is only for our convenience that we write the class frequencies in the form of a table having columns and rows.

2×2 Contingency Table

		A	α	Total
		(AB)	($\alpha\beta$)	(B)
B	β	(A β)	($\alpha\beta$)	(β)
Total		(A)	(α)	n

Our null hypothesis is that the attributes are independent. If A and B are independent then the observed frequency (AB) is equal to $\frac{(A)(B)}{n}$. By using this approach, we calculate the expected frequencies for all the four entries (AB), ($\alpha\beta$), β , and ($\alpha\beta$). The expected frequencies are calculated under the assumption that null hypothesis is true.

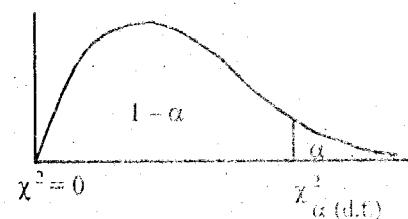


Figure 15.1

Expected Frequencies Calculated

	A	α	Total
B	$\frac{(A)(B)}{n}$	$\frac{(\alpha)(B)}{n}$	(B)
β	$\frac{(A)(\beta)}{n}$	$\frac{(\alpha)(\beta)}{n}$	(β)
Total	(A)	(α)	n

It may be noted that the column and row totals in the table of observed frequencies are the same as in the table of expected frequencies. The observed frequencies are denoted by f_o and the expected frequencies are denoted by f_e . For the calculation of χ^2 we write the expected frequencies corresponding to their observed frequencies. The necessary calculations are done as shown in the following columns:

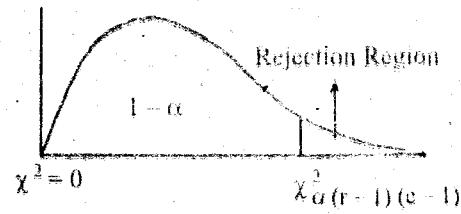
Observed frequencies f_o	Expected frequencies f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
(AB)	$\frac{(A)(B)}{n}$			
(A β)	$\frac{(A)(\beta)}{n}$			
(α B)	$\frac{(\alpha)(B)}{n}$			
(α β)	$\frac{(\alpha)(\beta)}{n}$			
n	n			$\chi^2 = \sum \left(\frac{(f_o - f_e)^2}{f_e} \right)$

5. Critical region:

The critical region in this test always lies in the right side of the distribution. It depends upon the level of significance α and the degrees of freedom. In tests of independence, the degree of freedom is calculated as below:

$$\text{degrees of freedom (d.f.)} = (r - 1)(c - 1)$$

where r is the number of rows and c is the number of columns in the contingency table. The critical value of χ^2 is seen from the table of χ^2 . For level of significance α , and degrees of freedom $(r - 1)(c - 1)$, the table value is



denoted by $\chi^2_{\alpha(r-1)(c-1)}$. In a χ^2 -table, under the column heading α and against d.f. = $(r - 1)(c - 1)$ given in the left column, we read the value of $\chi^2_{\alpha(d.f.)}$. When $\alpha = 0.05$, d.f. = 1, then $\chi^2_{0.05(1)} = 3.841$.

6. Conclusion:

The hypothesis of independence is rejected if the calculated value of χ^2 lies in the rejection region. The rejection of hypothesis means that the attributes are associated.

Example 15.11.

Calculate χ^2 by using the data given in Example 15.7. to test the independence between the gender and liking for fish. Use $\alpha = 0.05$.

Solution:

The data of Example 15.7. is reproduced here

	Males	Females	Total
Like Fish	80	80	160
Do not like Fish	20	20	40
Total	100	100	200

We write the hypotheses as below:

1. H_0 : There is independence between gender and liking for fish.
2. H_1 : There is association between gender and liking for fish.
3. Level of significance α is given, $\alpha = 0.05$.
4. Test statistic used is χ^2 where $\chi^2 = \sum \left(\frac{(f_o - f_e)^2}{f_e} \right)$
5. Computation:

The given table of observed frequencies is written as

	A	a	Total
B	(AB) = 80	(aB) = 80	(B) = 160
β	(A β) = 20	(a β) = 20	(β) = 40
Total	(A) = 100	(a) = 100	n = 200

The corresponding expected frequencies are calculated as below:

	A	a	Total
B	$\frac{(A)(B)}{n} = \frac{100 \times 160}{200} = 80$	$\frac{(a)(B)}{n} = \frac{100 \times 160}{200} = 80$	(B) = 160
β	$\frac{(A)(\beta)}{n} = \frac{100 \times 40}{200} = 20$	$\frac{(a)(\beta)}{n} = \frac{100 \times 40}{200} = 20$	(β) = 40
Total	(A) = 100	(a) = 100	n = 200

The necessary columns are as below:

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
80	80	0	0	0
20	20	0	0	0
80	80	0	0	0
20	20	0	0	0
200	200	0	0	0

$$\text{Here } d.f. = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$$

5. **Critical region:** $\chi^2 > \chi^2_{0.05(1)} = 3.841$
6. **Conclusion:** The calculated value of $\chi^2 = 0$ which falls in the acceptance region. Thus hypothesis H_0 of independence is accepted. When $\chi^2 = 0$, it means perfect independence between the attributes. Males and females have exactly equal liking for eating fish.

Example 15.12.

Let us consider the data of Example 15.10. and calculate χ^2 to examine the independence between college education and liking for football.

Solution:

The data of Example 15.10. is written as below:

	College educated males	Non college educated males	Total
A		α	
Football fans, B	$(AB) = 55$	$(\alpha B) = 125$	$(B) = 180$
Not football fans, β	$(A\beta) = 45$	$(\alpha\beta) = 75$	$(\beta) = 120$
Total	$(A) = 100$	$(\alpha) = 200$	$n = 300$

1. We frame the hypotheses as:

H_0 : There is independence between type of education and interest for football.

H_1 : There is association between type of education and their liking for football.

2. Level of significance, α is decided. Let $\alpha = 0.05$

3. Test – statistic used is χ^2 where $\chi^2 = \sum \left(\frac{(f_o - f_e)^2}{f_e} \right)$

4. Computations:

The expected frequencies under the assumption that H_0 is true are calculated as below:

$$(AB) = \frac{(A)(B)}{n} = \frac{100 \times 180}{300} = 60 \quad (A\beta) = \frac{(A)(\beta)}{n} = \frac{100 \times 120}{300} = 40$$

$$(\alpha B) = \frac{(\alpha)(B)}{n} = \frac{200 \times 180}{300} = 120 \quad (\alpha\beta) = \frac{(\alpha)(\beta)}{n} = \frac{200 \times 120}{300} = 80$$

Calculation of χ^2

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
55	60	-5	25	0.4167
45	40	5	25	0.6250
125	120	5	25	0.2083
75	80	-5	25	0.3125
300	300	0		$\chi^2 = 1.5625$

5. Critical region: $\chi^2 > \chi^2_{0.05(1)} = 3.841$

6. Conclusion: The calculated value of $\chi^2 = 1.5625$ is less than the critical value. Thus the hypothesis of independence is accepted. It means that college-educated males and non college educated males have the same liking for football. This result is different from the result given in Example 15.10. In Example 15.12, only one observed frequency of $(AB) = 55$ was compared with its corresponding expected frequency $\frac{(A)(B)}{n} = \frac{100 \times 180}{300} = 60$. The difference between 55 and 60 is not very large. They are very close. In χ^2 all the expected frequencies are compared with their observed frequencies. The χ^2 -test is a very powerful test for test of independence. We shall admit the result or conclusion based on the χ^2 -test. Thus H_0 is accepted.

15.5.2 DIRECT FORMULA FOR CALCULATING χ^2 IN 2×2 CONTINGENCY TABLE

In a 2×2 contingency table the value of χ^2 can be calculated without calculating the expected frequencies. Suppose a 2×2 contingency table has four cell frequencies as distributed below:

		1st Attribute		Total
2nd Attribute		a	b	a + b
		c	d	c + d
Total		a + c	b + d	a + b + c + d

The value of χ^2 can be calculated directly by using the formula:

$$\chi^2 = \frac{(a + b + c + d)(ad - bc)^2}{(a + b)(b + d)(c + d)(a + c)}$$

The proof of this formula is beyond the level of this book.

Let us calculate χ^2 by using the above formula from the data given in Example 15.12. From the data given in example 15.12, we have

$$a = 55, \quad b = 125, \quad c = 45 \quad \text{and} \quad d = 75$$

$$\begin{aligned} \text{Thus } \chi^2 &= \frac{(55 + 125 + 45 + 75)(55 \times 75 - 125 \times 45)^2}{(55 + 125)(125 + 75)(45 + 75)(55 + 45)} \\ &= \frac{(300)(2250000)}{(180)(200)(120)(100)} = \frac{675}{432} = 1.5625 \end{aligned}$$

This answer is the same as calculated in Example 15.12.

15.6 CONTINGENCY TABLE OF HIGHER ORDER

Sometimes a certain characteristic or attribute has more than two categories. For example when we are taking about heights of persons, the population or sample can be divided into four classes or categories like, very tall, tall, medium and short. In general if the attribute is A, then its different levels are denoted by A_1, A_2, \dots, A_r if it has r categories. The same population or sample may also be divided according to another characteristic say B with its levels B_1, B_2, \dots, B_c with c categories. The sample data on two attributes can be written in the form of two-way classification as below:

**Table 15.4.
Two-way Classification**

Attribute A						Row Totals	
	B_1	B_2	...	B_j	...	B_c	
A_1	$(A_1 B_1)$	$(A_1 B_2)$...	$(A_1 B_j)$...	$(A_1 B_c)$	(A_1)
A_2	$(A_2 B_1)$	$(A_2 B_2)$...	$(A_2 B_j)$...	$(A_2 B_c)$	(A_2)
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
A_i	$(A_i B_1)$	$(A_i B_2)$...	$(A_i B_j)$...	$(A_i B_c)$	(A_i)
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
A_r	$(A_r B_1)$	$(A_r B_2)$...	$(A_r B_j)$...	$(A_r B_c)$	(A_r)
Column Totals	(B_1)	(B_2)	...	(B_j)	...	(B_c)	n

Table 15.4. contains r rows and c columns, it is therefore called $r \times c$ contingency table. Each frequency in the table is called cell frequency. It is the extension of 2×2 contingency table and χ^2 -statistic is used to test the independence between the attributes given in the rows and columns.

The procedure is the same as explained earlier. For each observed frequency in the sample data, the corresponding expected frequency is calculated. It is calculated on the assumption that there is independence between the two characteristics. For each observed frequency $(A_i B_j)$ the expected frequency is $\frac{(A_i)(B_j)}{n}$ where (A_i) is the total of the row A_i and (B_j) is the total of the column B_j . For expected frequency E, a more general formula may be written as

$$E = \frac{R \times C}{n} \text{ where } R \text{ is the row total and } C \text{ is the column total.}$$

χ^2 is calculated by the formula

$$\Sigma \left(\frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}} \right) = \Sigma \left(\frac{(f_o - f_e)^2}{f_e} \right)$$

15.7 LIMITATIONS OF χ^2

The χ^2 -test of independence gives very good results or conclusions when all the cell frequencies are very large. For small cell frequencies the test is not very reliable. χ^2 -test should not be used if any expected frequency is less than 5. If any expected frequency is less than 5, then something is to be done about it. One column containing the small frequency/frequencies is added to the adjacent column before calculating χ^2 . Similarly if some row has expected frequencies less than 5, the entire row is added to the adjacent row by adding the corresponding cell frequencies. If we have the choice to reduce the number of rows or columns, we should choose that column or row which we think is least important in the given data and this column or row should be added to the adjacent column or row.

Example 15.13.

In a public opinion survey, 2000 persons were interviewed to give their opinion. The individuals interviewed are classified according to their attitude on a certain social scheme and according to sex. The data is given in the table below:

	Favour	Oppose	Undecided	Total
Men	600	320	280	1200
Women	450	280	70	800
Total	1050	600	350	2000

Calculate χ^2 to examine whether men and women differ in their opinion about the social scheme.

Solution:

1. The null hypothesis H_0 is that there is independence between the sex and their attitude towards the social scheme.
2. The alternative hypothesis H_1 is that there is association between the two characteristics.
3. Level of significance: Let $\alpha = 0.05$

3. Test-statistic: $\chi^2 = \sum \left(\frac{(f_o - f_e)^2}{f_e} \right)$
4. Computations: Let A_1 and A_2 denote the rows and B_1 , B_2 and B_3 denote the columns. The given table can be written as:

	B_1	B_2	B_3	Total
A_1	600	320	280	$(A_1) = 1200$
A_2	450	280	70	$(A_2) = 800$
Total	$(B_1) = 1050$	$(B_2) = 600$	$(B_3) = 350$	$n = 2000$

Expected frequencies f_e are calculated as below:

	B_1	B_2	B_3	Total
A_1	$\frac{1050 \times 1200}{2000}$ = 630	$\frac{600 \times 1200}{2000}$ = 360	$\frac{350 \times 1200}{2000}$ = 210	$(A_1) = 1200$
A_2	$\frac{1050 \times 800}{2000}$ = 420	$\frac{600 \times 800}{2000}$ = 240	$\frac{350 \times 800}{2000}$ = 140	$(A_2) = 800$
Total	$(B_1) = 1050$	$(B_2) = 600$	$(B_3) = 350$	$2000 = n$

It is important to note that the column and row totals are equal in the original table and table of expected frequencies.

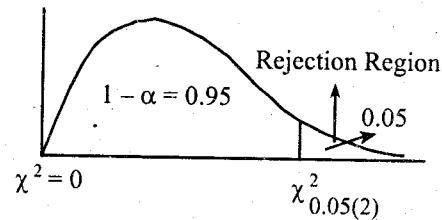
χ^2 -Calculated

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
600	630	-30	900	1.43
450	420	30	900	2.14
320	360	-40	1600	4.44
280	240	40	1600	6.67
280	210	70	4900	23.33
70	140	-70	4900	35.00
2000	2000	0		$\chi^2 = 73.01$

5. Region of rejection:

$$d.f. = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$$

$$\chi^2 > \chi^2_{0.05(2)} = 5.991$$



6. Conclusion: The calculated value of χ^2 is 73.01 and the critical value of χ^2 is 5.991. The χ^2 calculated from the sample data falls in the rejection region. Thus hypothesis of independence is rejected. It means that men and women have different opinions about the social scheme. Sex is associated with the attitude towards the social scheme.

Example 15.14.

Given the following table. Calculate χ^2 to examine whether there is evidence of relationship between the intelligence level of fathers and sons. Use $\alpha = 0.05$

Sons	Fathers			Total
	Very Intelligent	Average	Non-Intelligent	
Very Intelligent	10	35	5	50
Average	150	140	15	305
Non-Intelligent	40	95	20	155
Total	200	270	40	510

Solution:

1. The null hypothesis to be tested is that there is no relationship between the intelligence of fathers and sons.

The alternative hypothesis is that there is relationship (association) between the intelligence level of fathers and sons.

2. Level of significance: $\alpha = 0.05$

3. Test-statistic: $\chi^2 = \sum \left(\frac{(f_o - f_e)^2}{f_e} \right)$

4. Computations:

Let A_1, A_2, A_3 be used for the rows and B_1, B_2, B_3 be used for columns headings.

Table of expected frequencies calculated

	B_1	B_2	B_3	Total
A_1	$\frac{200 \times 50}{510}$ = 19.6	$\frac{270 \times 50}{510}$ = 26.5	$\frac{40 \times 50}{510}$ = 3.9	$(A_1) = 50$
A_2	$\frac{200 \times 305}{510}$ = 119.6	$\frac{270 \times 305}{510}$ = 161.5	$\frac{40 \times 305}{510}$ = 23.9	$(A_2) = 305$
A_3	$\frac{200 \times 155}{510}$ = 60.8	$\frac{270 \times 155}{510}$ = 82.0	$\frac{40 \times 155}{510}$ = 12.2	$(A_3) = 155$
Total	$(B_1) = 200$	$(B_2) = 270$	$(B_3) = 40$	$n = 510$

One expected frequency under the column B_3 and against row A_1 is 3.9 which is less than 5. This frequency cannot be used in the calculation of χ^2 . Now we have two options (i) column B_3 is added to column B_2 (ii) Row A_1 is added to row A_2 . But the total of column B_3 is 40 which is minimum of all the column and row totals. It means column B_3 is less important as compared to row A_1 . Thus column B_3 is added to column B_2 . This is equivalent to combining a small sample data with a large sample data. Thus the tables of observed frequencies and the expected frequencies would become:

Observed Frequencies

	B_1	$B_2 + B_3$	Total
A_1	10	$35 + 5 = 40$	50
A_2	150	$140 + 15 = 155$	305
A_3	40	$95 + 20 = 115$	155
Total	200	310	510

Expected Frequencies

	B_1	$B_2 + B_3$	Total
A_1	19.6	$26.5 + 3.9 = 30.4$	50
A_2	119.6	$161.5 + 23.9 = 185.4$	305
A_3	60.8	$82.0 + 12.2 = 94.2$	155
Total	200	310.0	510

χ^2 -Calculated

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
10	19.6	-9.6	92.16	4.70
150	119.6	30.4	924.16	7.73
40	60.8	-20.8	432.64	7.12
40	30.4	9.6	92.16	3.03
155	185.4	-30.4	924.16	4.98
115	94.2	20.8	432.64	4.59
510	510	0		$\chi^2 = 32.15$

5. Region of rejection:

$$d.f. = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$$

Critical region is $\chi^2 > \chi^2_{0.05(2)} = 5.991$

6. Conclusion: Since the calculated value of $\chi^2 = 32.15$ which is greater than the critical value of 5.991, the null hypothesis H_0 is rejected and H_1 is accepted. It means that there is relationship (association) between the intelligence levels of fathers and sons. Intelligent fathers have usually intelligent sons. This is what the sample data indicates through the χ^2 as test of independence.

15.8 RANK CORRELATION:

We are often confronted with situations where the basic data are not available in numerical magnitudes but where the rankings can be developed and used to examine the relationship between data sets. To calculate the Spearman's rank correlation coefficient, we first rank the X's among themselves, giving rank 1 to the largest or smallest value, rank 2 to the second largest or second smallest, and so on; then we rank the Y's similarly among themselves. The Spearman's rank correlation coefficient, r_s , is given by the following formula

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

where d = difference between the ranks for the paired observations,

n = number of paired observations

When there are tied observations, the mean rank is given to each observation in the set of ties. For example, if the fourth and fifth largest values of a variable are the same, we assign each the rank $(4+5)/2 = 4.5$, and if the sixth, seventh and eighth largest values of a variable are the same, we assign each the rank $= (6+7+8) / 3 = 7$. The possible range of values for Spearman's rank correlation coefficient r_s is -1 to $+1$. If $r_s = +1$, there is perfect positive rank correlation and if $r_s = -1$, there is perfect negative rank correlation. If X and Y are independent of each other, there is no relationship and thus the rank correlation coefficient $r_s = 0$.

Example 15.15.

The following were the "performance under stress" rankings of 10 honor students before and after mid-semester:

Student	A	B	C	D	E	F	G	H	I	J
Rank before	1	2	3	4	5	6	7	8	9	10
Rank after	6	5	8	9	3	4	10	1	7	2

Compute the Spearman's rank correlation coefficient for this data set.

Solution:

Rank before (X)	Rank after (Y)	$d = X - Y$	d^2
1	6	-5	25
2	5	-3	9
3	8	-5	25
4	9	-5	25
5	3	+2	4
6	4	+2	4
7	10	-3	9
8	1	+7	49
9	7	+2	4
10	2	+8	64
			$\sum d^2 = 218$

$$\text{Spearman's rank correlation coefficient, } r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6(218)}{10(100 - 1)} = 1 - 1.32 = -0.32$$

Example 15.16.

A Statistics instructor wants to know whether there is a correlation between students' midterm averages and their final examination scores. The instructor takes a random sample of nine students from previous Statistics courses and obtains the following data:

Midterm average X	72	96	80	77	67	92	90	74	60
Final examination score Y	49	97	80	78	71	86	95	48	52

- (i) Determine the rank correlation coefficient, r_s , of the data.
- (ii) Interpret the value of r_s obtained in part (i).

Solution:

Midterm average (X)	Final examination score (Y)	Rank of X	Rank of Y	$d = X - Y$	d^2
72	49	3	2	1	1
96	97	9	9	0	0
80	80	6	6	0	0
77	78	5	5	0	0
67	71	2	4	-2	4
92	86	8	7	1	1
90	95	7	8	-1	1
74	48	4	1	3	9
60	52	1	3	-2	4
					$\sum d^2 = 20$

$$(i) r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6(20)}{9(81 - 1)} = 1 - 0.17 = 0.83$$

- (ii) The rank correlation coefficient, $r_s = 0.83$ suggests that there is a strong positive correlation between midterm average and final-examination score in Statistics courses.

Example 15.17.

The number of hours of study for an examination and the grades received by a random sample of 10 students are:

Number of hours studied, X	8	5	11	13	10	5	18	15	2	8
Grade in examination, Y	56	44	79	72	70	54	94	85	33	65

Compute and interpret the Spearman's rank correlation coefficient.

Solution:

Number of hours studied (X)	Grade in examination (Y)	Rank of X	Rank of Y	$d = X - Y$	d^2
8	56	4.5	4	0.5	0.25
5	44	2.5	2	0.5	0.25
11	79	7	8	-1.0	1.00
13	72	8	7	1.0	1.00
10	70	6	6	0.0	0.00
5	54	2.5	3	-0.5	0.25
18	94	10	10	0.0	0.00
15	85	9	9	0.0	0.00
2	33	1	1	0.0	0.00
8	65	4.5	5	-0.5	0.25
					$\sum d^2 = 8$

$$\text{Spearman's rank correlation coefficient, } r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6(3)}{10(100 - 1)} = 1 - \frac{18}{990} = 1 - 0.02 = 0.98$$

$r_s = 0.98$, indicating strong positive correlation between the number of hours of study and the grade in examination.

SHORT DEFINITIONS

Attributes

If a characteristic which is being measured is of qualitative nature, is called an attribute. Attributes are denoted by A, B, C, α , β , γ , etc. The attributes A, B, C are the positive attributes and α , β , γ are negative attributes.

or

The characteristic being studied is nonnumeric, is known as an attribute.

Variable

A variable is a phenomenon that may vary from one individual or object to another.

or

A characteristic that can have different values is called a variable.

Independence

Two attributes A and B are said to be independent if $(AB) = \frac{(A)(B)}{N}$

Expt.

Association

If two attributes are not independent, they are said to be associated.

Obs +

Positive Association

Two attributes A and B are said to be positively associated if $(AB) > \frac{(A)(B)}{N}$

Negative Association

Two attributes A and B are said to be negatively associated if $(AB) < \frac{(A)(B)}{N}$

Contingency Table

A table showing the cross-tabulation or joint distribution of two variables is known as contingency table.

or

A table used to classify sample observations according to two or more identifiable characteristics is called contingency table.

Rank Correlation

The rank correlation describes the relationship between the two sets of rankings that is, between the rankings of the one variable and the rankings of the other variable.

Spearman's Rank Correlation Coefficient

A method of measuring and testing the degree of association between the two variables measured at the ordinal level is called Spearman's rank correlation coefficient.

or

A rank correlation provides a measure of the degree of linearity between the ranking variables.

Properties of Spearman's Rank Correlation Coefficient

- The value of r_s is always between -1 and +1, i.e. $-1 \leq r_s \leq +1$.
- r_s is positive when the ranks of the pairs of sample observations tend to increase together.
- $r_s = 0$ when the ranks are not correlated.
- r_s is negative when the ranks of one variable tend to decrease as the other variable's ranks increase.
- A value of +1 or -1 indicates perfect association between X and Y, the plus sign occurring for identical rankings and the minus sign occurring for reverse rankings.

MULTIPLE - CHOICE QUESTIONS

- In order to carry out a χ^2 -test on data in a contingency table, the observed values in the table should be:
 - close to the expected values
 - all greater than or equal to 5
 - frequencies
 - quantitative
- The χ^2 -test should not be used if any expected frequency is:
 - less than 10
 - less than 5
 - equal to 5
 - more than 5
- If $(AB) = \frac{(A)(B)}{n}$, the two attributes A and B are:
 - independent
 - dependent
 - correlated
 - quantitative
- To calculate the level of association, we can calculate coefficient of association, the coefficient of association always lies between:
 - 1 and +1
 - 0 and 1
 - 1 and 0
 - 0 and 5
- If two attributes A and B are independent, then the coefficient of association is:
 - 1
 - +1
 - 0
 - 0.5
- If $(AB) < \frac{(A)(B)}{n}$, the association between two attributes A and B is:
 - negative
 - positive
 - zero
 - symmetrical
- Two attributes A and B are said to be positive, if:
 - $(AB) = \frac{(A)(B)}{n}$
 - $(AB) \neq \frac{(A)(B)}{n}$
 - $(AB) > \frac{(A)(B)}{n}$
 - $(AB) < \frac{(A)(B)}{n}$

8. If two attributes A and B have perfect positive association, the value of coefficient of association is equal to:
- (a) +1
 - (b) -1
 - (c) 0
 - (d) $(r-1)(c-1)$
9. The degrees of freedom for χ^2 are $(r-1)(c-1)$ for a contingency table with r -rows and c -columns. So for a 2×2 contingency table there are:
- (a) one degrees of freedom
 - (b) two degrees of freedom
 - (c) three degrees of freedom
 - (d) four degrees of freedom
10. For an $r \times c$ contingency table the number of degrees of freedom equals:
- (a) $r c$
 - (b) $r + c$
 - (c) $(r-1)+(c-1)$
 - (d) $(r-1)(c-1)$
11. For a 3×3 contingency table, the number of cells in the table are:
- (a) 3
 - (b) 6
 - (c) 9
 - (d) 4
12. The null hypothesis of independence between the variables is tested using the χ^2 -statistic where calculated $\chi^2 = \sum(O-E)^2/E$, if the degrees of freedom, $(r-1)(c-1)$, are greater than:
- (a) 1
 - (b) 2
 - (c) 3
 - (d) 4
13. The shape of the chi-square distribution depends upon:
- (a) parameters
 - (b) degrees of freedom
 - (c) number of cells
 - (d) standard deviation
14. The total area under the curve of a chi-square distribution is:
- (a) 1
 - (b) 0.5
 - (c) 0 to ∞
 - (d) $-\infty$ to $+\infty$
15. Chi-square curve ranges from:
- (a) $-\infty$ to $+\infty$
 - (b) 0 to ∞
 - (c) $-\infty$ to 0
 - (d) 0 to 1
16. The value of chi-square statistic is always:
- (a) negative
 - (b) zero
 - (c) non-negative
 - (d) one
17. In testing independence in a 2×3 contingency table, the number of degrees of freedom in χ^2 -distribution is:
- (a) 1
 - (b) 2
 - (c) 3
 - (d) 5

18. Given $\chi^2 = 5.8$, df = 1, $\chi^2_{0.05(1)} = 3.841$, $\chi^2_{0.01(1)} = 6.635$, we make the following statistical decision:
- We reject H_0 at $\alpha = 0.05$ but not at $\alpha = 0.01$.
 - We reject H_0 at $\alpha = 0.01$.
 - We fail to reject H_0 at $\alpha = 0.05$.
 - We reject H_0 at $\alpha = 0.01$ but not at $\alpha = 0.05$.
19. If $\chi^2 = 13.95$, df = 4, $\chi^2_{0.05(4)} = 9.488$, $\chi^2_{0.01(4)} = 13.277$, we make the following statistical decision:
- We accept H_0 at $\alpha = 0.01$ and $\alpha = 0.05$
 - We reject H_0 at $\alpha = 0.05$ but not at $\alpha = 0.01$
 - We reject H_0 at $\alpha = 0.01$ but not at $\alpha = 0.05$
 - We reject H_0 at $\alpha = 0.01$ and $\alpha = 0.05$
20. In converting the scores 18, 24, 12, 14, 22, 29 to ranks (assigning rank 1 to the highest score), the score of 12 has a corresponding rank of:
- | | |
|-------|-------|
| (a) 1 | (b) 2 |
| (c) 6 | (d) 7 |
21. In converting the scores 8, 20, 14, 7, 11, 14, 3 to ranks (assigning rank 1 to the lowest score), the score of 14 has a corresponding rank of:
- | | |
|---------|---------|
| (a) 5 | (b) 6 |
| (c) 5.5 | (d) 4.5 |
22. If a person ranks lowest on beauty and highest on intelligence and another person ranks highest on beauty and lowest on intelligence, the Spearman's coefficient of rank correlation is probably:
- | | |
|----------------------|----------------------|
| (a) zero | (b) weak positive |
| (c) perfect positive | (d) perfect negative |
23. If $\frac{6\sum d^2}{n(n^2 - 1)}$ is zero, the value of r_s is:
- | | |
|---------|--------|
| (a) 0.5 | (b) -1 |
| (c) -1 | (d) 0 |

Answers

1. (c)	2. (b)	3. (a)	4. (a)	5. (c)	6. (a)	7. (c)	8. (a)
9. (a)	10. (d)	11. (c)	12. (a)	13. (b)	14. (a)	15. (b)	16. (c)
17. (b)	18. (a)	19. (d)	20. (c)	21. (c)	22. (d)	23. (b)	

SHORT QUESTIONS

1. Given $n = 100$, $(A) = 40$. Find (α) .

Ans. 60

2. Given $(AB) = 30$, $(A) = 40$. Find $(A\beta)$.

Ans. 10

3. Given $(\alpha BC) = 15$, $(\alpha B\gamma) = 60$, $(\alpha\beta C) = 40$ and $(\alpha\beta\gamma) = 70$. Find (α) .

Ans. 185

4. Given $a = 55$, $b = 125$, $c = 45$ and $d = 75$. Find χ^2 .

Ans. 1.5625

5. Given $(A) = 20$, $(B) = 13/50$, $(AB) = 1/50$ and $n = 250$. Whether attributes A and B are negatively associated?

Ans. Yes, $(AB) < \frac{(A)(B)}{n}$

6. Given $(A) = 304$, $(B) = 1024$, $(AB) = 256$ and $n = 1216$. Show that attributes A and B are independent.

Ans. $(AB) = \frac{(A)(B)}{n} = 256$

7. Given $(A) = 34000$, $(B) = 6000$, $(AB) = 5300$ and $n = 70000$. Whether attributes A and B are positively associated?

Ans. Yes, $(AB) > \frac{(A)(B)}{n}$

8. Given $(AB) = 150$, $(\alpha B) = 106$, $(A\beta) = 272$, $(\alpha\beta) = 1132$, and $n = 1660$. Find the coefficient of association.

Ans. $Q = 0.71$

9. Given $\chi^2 = 20.178$, $df = 4$ and $\alpha = 0.01$. Find the table value of χ^2 and make the statistical decision.

Ans. $\chi^2_{0.01(4)} = 13.277$, reject H_0

10. Given $f_0 = 30, 75, 45, 30, 75, 45$, $f_e = 52.5, 52.5, 37.5, 37.5, 60.0, 60.0$, $df = 2$ and $\alpha = 0.05$. Find χ^2 and make the statistical decision.

Ans. $\chi^2 = 29.786$, reject H_0

11. If the respective values of $(f_0 - f_e)^2/f_e$ are 0.129, 2.414, 3.214 and 8.929. Find χ^2 .

Ans. 14.69

12. Given $f_0 = 25$, $f_e = 15$ and $f_0 = 35$, $f_e = 45$. Find $\sum (|f_0 - f_e| - 0.5)^2/f_e$.

Ans. 8.03

13. If the respective values of $f_0 = 21, 38, 32, 29, 36, 25, 41, 23$ and $f_e = 31.31, 27.69, 32.37, 28.63, 32.37, 28.63, 33.96, 30.04$, then find χ^2 .

Ans. 11.22

14. Given the pairs of ranks $(4, 2), (1, 3), (2, 1), (5, 6), (6, 5), (3, 4)$. Find Σd^2 .

Ans. 12

15. Given $\Sigma d^2 = 440$ and $n = 11$. Find the value of r_s .

Ans. -1

16. Given $\Sigma d^2 = 99$ and $n = 10$. Find the coefficient of rank correlation.

Ans. 0.4

17. Differentiate between attribute and variable.

18. Explain the consistency of the data.

19. When two attributes are said to be positively associated?

20. When two attributes are said to be negatively associated?

21. When two attributes are said to be associated?

22. Explain what is meant by independence of attributes?

23. What do you understand by association?

24. Explain the terms independence and association as applied to attributes.

25. Differentiate between positive association and negative association.

26. Define a contingency table.

27. Explain the positive and negative association.

28. What is meant by attribute?

29. Interpret the meaning of coefficient of association Q when:
(a) $Q = -1$ (b) $Q = +1$ (c) $Q = 0$
30. Explain the general procedure for test of independence between the attributes.
31. Write down the direct formula for calculating χ^2 in a 2×2 contingency table.
32. Define χ^2 -distribution.
33. Explain the coefficient of association.
34. Explain the terms positive and negative attributes.
35. Explain the term rank correlation.
36. Interpret the meaning when: (a) $r_s = +1$ (b) $r_s = -1$ (c) $r_s = 0$
37. What is meant by Spearman's rank correlation coefficient?
38. Write down the properties of Spearman's rank correlation coefficient.

EXERCISES

1. Compute all the remaining possible class frequencies from the following data:
 $(\alpha) = 50$, $(B) = 70$, $(A\beta) = 20$ and $n = 100$.

Ans. $(A) = 50$, $(\beta) = 30$, $(AB) = 30$, $(\alpha\beta) = 10$, $(\alpha B) = 40$

2. Given that: $(AB) = 150$, $(A\beta) = 250$, $(\alpha B) = 260$, $(\alpha\beta) = 2340$. Find the other frequencies and the value of n.

Ans. $(A) = 400$, $(\alpha) = 2600$, $(B) = 410$, $(\beta) = 2590$, $n = 3000$

3. Given that: $(A) = 304$, $(AB) = 256$, $(\alpha\beta) = 144$, $(\alpha B) = 768$, $(A\beta) = 48$. Show that attributes A and B are independent.

Ans. $(B) = 1024$, $n = 1216$, $\frac{(A)(B)}{n} = 256$. A and B are independent.

4. Whether attributes A and B are negatively associated, positively associated or independent.

$$(i) \quad (A) = \frac{340}{17}, \quad (B) = \frac{13}{50}, \quad (AB) = \frac{1}{50}, \quad n = 250$$

$$(ii) \quad n = 154, \quad (\beta) = 88, \quad (AB) + (A\beta) = 35, \quad (A\beta) = 20$$

$$(iii) \quad (AB) = 5300, \quad (\alpha B) = 700, \quad (\alpha) = 36000, \quad (A) = 34000.$$

Ans. (i) A and B are negatively associated (ii) A and B are independent

(iii) A and B are positively associated

5. Test the independence by a simplest approach between gender and intelligence.

Gender

Level of Intelligence	Males	Females	Total
Intelligent	150	75	225
Non - intelligent	50	25	75
Total	200	100	300

Ans. $(AB) = 150$, $\frac{(A)(B)}{n} = 150$. Independence between males and intelligence

6. Test the independence by a simple approach between intelligence of fathers and sons.

Fathers

Sons	Intelligent	Not intelligent	Total
Intelligent	300	200	500
Not intelligent	100	400	500
Total	400	600	1000

Ans. Positive association between intelligent fathers and intelligent sons.

7. Find coefficient of association from the following data:

		Height of fathers	
Height of sons		Tall	Short
Tall		500	100
Short		100	400

Ans. $Q = 0.905$, positive association between height of fathers and height of sons.

8. Test the independence by a simple approach between attack of disease and vaccination.

	Vaccinated	Not vaccinated	Total
Attacked	50	500	550
Not attacked	350	500	850
Total	400	1000	1400

Ans. Negative association between vaccination and attack of disease.

9. From the following table, test the hypothesis that the flower colour is independent of flatness of leaf. Use $\alpha = 0.05$.

	Flat leaves	Lean leaves	Total
White flowers	99	36	135
Red flowers	20	5	25
Total	119	41	160

Ans. $\chi^2 = 0.494$, accept H_0 .

10. In a locality, 300 persons were randomly selected and asked about their educational attainment. The results are given as follows:

Sex	Education		
	Middle	Secondary school	College
Male	30	45	75
Female	75	30	45

Can we say that education depends on sex? Use $\alpha = 0.05$.

Ans. $\chi^2 = 29.786$, reject H_0 .

11. Find chi-square (χ^2) for the following table to examine the association between the subjects and their result. Use $\alpha = 0.05$.

Subjects	Result	
	Passed	Failed
Mathematics	60	70
Statistics	210	190
English	360	410
History	160	220
Education	60	60

Ans. $\chi^2 = 8.959$, accept H_0 :

12. The following data show initial training program performance and a job rating by a supervisor 12 months later for a sample of 400 employees of a telephone company:

Job Rating	Training Program Performance		
	Below average	Average	Above average
Below average	63	49	9
Average	60	79	28
Above average	29	60	23

Is there any relationship between performance in the training program and job rating? Use the 1% level of significance.

Ans. $\chi^2 = 20.178$, reject H_0 :

13. Find the value of chi-square (χ^2) from the following data and test the hypothesis that there is no relation between the level of intelligence and the social status. Use $\alpha = 0.05$.

Social status	Level of intelligence		
	Brilliant	Intelligent	Dull
Upper middle	20	20	60
Middle	32	70	38
Lower middle	23	35	22

Ans. $\chi^2 = 35.163$, reject H_0 :

14. The data given below are the rankings of a simple random sample of companies with regard to total sales and return on equity:

Company	A	B	C	D	E	F	G	H	I	J	K	L
Sales Rank	8	3	10	11	5	9	7	1	4	2	12	6
Return on Equity Rank	6	1	10	12	2	11	8	5	7	3	9	4

Compute the rank correlation coefficient for this set of data.

Ans. $r_s = 0.78$

15. A group of ten workers of a factory is ranked according to their efficiency by two different judges as follows:

Name of worker	A	B	C	D	E	F	G	H	I	J
Judgement of Judge I	5	8	2	9	10	7	1	4	3	6
Judgement of Judge II	7	9	4	10	8	5	1	3	2	6

- (i) Compute the Spearman's rank correlation coefficient.
(ii) Interpret the value of your result.

Ans. (i) $r_s = 0.88$ (ii) $r_s = 0.88$ means the opinion of the two judges with regard to the efficiency of the workers shows great similarity.

16. The following table shows how 10 students, arranged in alphabetical order, were ranked according to their achievements in both laboratory and lecture portions of a Statistics course. Compute and interpret the Spearman's rank correlation coefficient:

Laboratory	8	3	9	2	7	10	4	6	1	5
Lecture	9	5	10	1	8	7	3	4	2	6

Ans. $r_s = 0.85$, indicating that there is a marked relationship between achievements in laboratory and lecture portions.

17. The following table shows the first two marks, denoted by X and Y respectively, of 8 students on two quizzes in Statistics.

Marks on first Quiz (X)	50	65	75	100	125	140	170	195
Marks on second Quiz (Y)	45	60	80	95	120	150	145	190

Compute and interpret the Spearman's rank correlation coefficient.

Ans. $r_s = 0.98$. There is a high degree of positive correlation between the marks of two quizzes.

18. A chemical reaction is timed at several different temperatures; the results are:

Temperature (F°)	100	150	200	250	300	350
Time needed (seconds)	843	211	164	69	22	17

- (i) Determine the Spearman's rank correlation coefficient between the variables.
- (ii) Comment on the value of your result.

Ans. (i) $r_s = -1$

- (ii) $r_s = -1$, there is perfect negative rank correlation between the variables.

19. Consider the situation where a panel of 11 financial experts is requested to examine financial data from two corporations, corporation A and corporation B. Calculate and interpret the rank correlation coefficient between financial strength scores for corporations A and B.

Panelist	1	2	3	4	5	6	7	8	9	10	11
Corporation A	4	5	3	3	3	4	5	5	4	2	3
Corporation B	3	3	4	2	2	3	2	4	2	4	2

Ans. $r_s = 0.1$, indicating weak positive correlation between the scores given corporation A and those given corporation B.

20. In a study between the amount of rainfall and the quantity of air pollution removed, the following data were collected:

Daily Rainfall, X (0.01 centimeter)	4.3	4.5	5.9	5.6	6.1	5.2	3.8	2.1	7.5
Particulate Removed, Y (micrograms per cubic meter)	126	121	116	118	114	118	132	141	108

Calculate the rank correlation coefficient for the daily rainfall and amount of particulate removed.

Ans. $r_s = -0.99$

21. The ranks of the same 4 students in two subjects A and B were as follows:

(4, 3), (2, 4), (1, 2), (3, 1).

Two numbers within brackets denote the ranks of the students in A and B respectively. Calculate and interpret the Spearman's rank correlation coefficient.

Ans. $r_s = 0$. The ranks are uncorrelated.

Chapter 16

TIME SERIES

16.1 INTRODUCTION

A time series is a set of observations recorded according to some period of time. The observations are usually recorded at equal intervals of time. The enrolment of students in a certain college for a number of years is a yearly time series. The imports and exports of a country on monthly basis make an important monthly time series. The population of a country may be counted after every 10 years. The population figures at intervals of 10 years is called a decennial time series. The interval of the time series depends upon the nature of the observations. At what interval the observations would become practically meaningful and important? The population figures are usually not recorded on daily or weekly basis. For very small interval of time, the population data does not make a good piece of information. For prices of wheat, meat and other commodities, we do not use intervals of 5 or 10 years. The prices of fruits may be studied on daily or weekly basis and the prices of hard commodities may be studied on monthly or yearly basis.

16.2 PURPOSE OF TIME SERIES

The time series are constructed and maintained for some practical purposes. Some of their purposes are:

- (i) The previous pattern of the time series enables us to determine or estimate the future values of the time series. This is an important purpose of the time series.
- (ii) We can apply a control on the time series. If the time series is rapidly increasing (like prices of meat), we may like to control the increase in the time series. The control can be applied if we have a detailed information about the time series.

16.2.1 GRAPH OF THE TIME SERIES

A graph of the time series plays an important role in the study of the time series. To make a graph, suppose the time periods are denoted by $t_1, t_2, t_3, \dots, t_k$ and the corresponding figures of the Y-variable are denoted by $Y_1, Y_2, Y_3, \dots, Y_k$. The time $t_1, t_2, t_3, \dots, t_k$ are taken on X-axis and the Y-values $Y_1, Y_2, Y_3, \dots, Y_k$ are plotted against their respective time. The plotted points are joined together by straight lines to get a graph called the historigram. A time series is also called an historical series where historical is from history. A graph of time series about production of firewood in Punjab is given in Figure 16.1.

Example 16.1.

Table 16.1. Production of Firewood in Punjab

Years	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993
	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Values million Rs.	18.6	22.6	38.1	40.9	41.4	40.1	46.6	60.7	57.2	53.4

The graph of the above data is shown in Fig. 16.1.

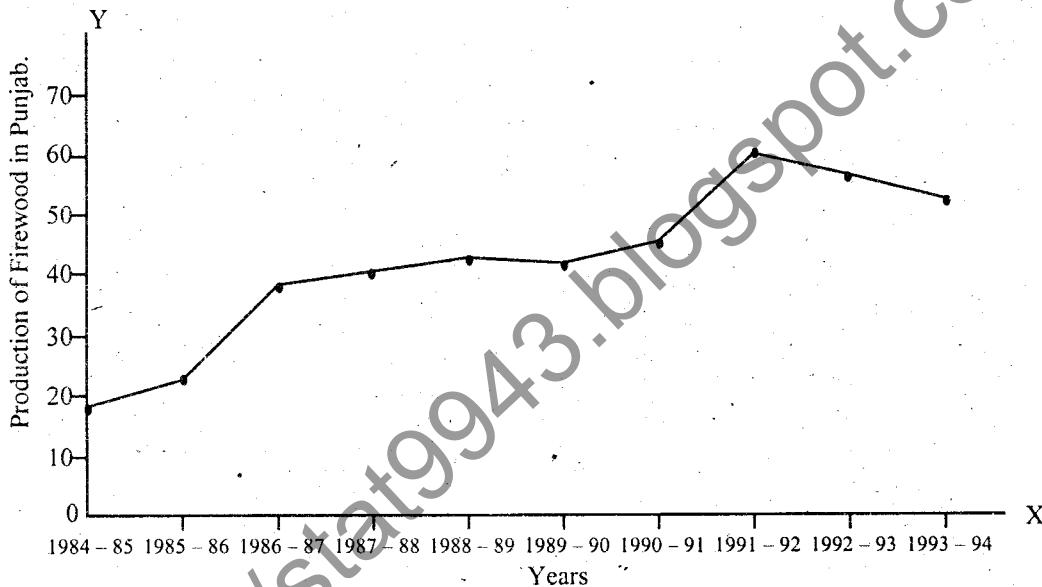


Figure 16.1

(Source: Page No. 108, Statistical Pocket Book of the Punjab, 1995)

16.3 COMPONENTS OF A TIME SERIES

A time series usually changes with the passage of time. There are many reasons which bring changes in the time series. These changes are called components or variations in a time series. The words movements or fluctuations are also used for these changes. These movements are:

- | | |
|---------------------------|----------------------------|
| (i) Secular trend | (ii) Seasonal variations |
| (iii) Cyclical variations | (iv) Irregular variations. |

16.3.1 SECULAR TREND

Secular trend is the regular component of the time series. The time series moves regularly in some direction over a long period of time. This regular movement of the series in some direction, upward or downward is called the trend or secular trend of the time series. These movements may be slow or fast but they are systematic in nature. The changes follow some rule. These movements are free of sudden jumps

and swings. The Y values fluctuate round an average and these fluctuations are called 'noise'. The trend is called linear if the change, increase or decrease, for a certain period is almost the same throughout the time series.

Figure 16.2. shows that there is upward growth of the time series. This upward movement is called the trend. As the graph of the observed time series is close to a straight line, we say that trend is linear. Thus the smooth line AB in Figure 16.2. shows the linear trend. Trend is not always linear. Figure 16.3. shows that the overall movement of the time series is close to a curve. Here the trend is non-linear or curvilinear. The curve AB shows the trend in the form of a curve.

Example 16.2.

Table 16.2. In-door patients treated in hospitals in the Punjab.

Years	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993
No. of patients (in thousands)	632	676	749	852	946	1000	1135	1160	1207	1101

The graph of the data in Table 16.2 is shown in Figure. 16.2..

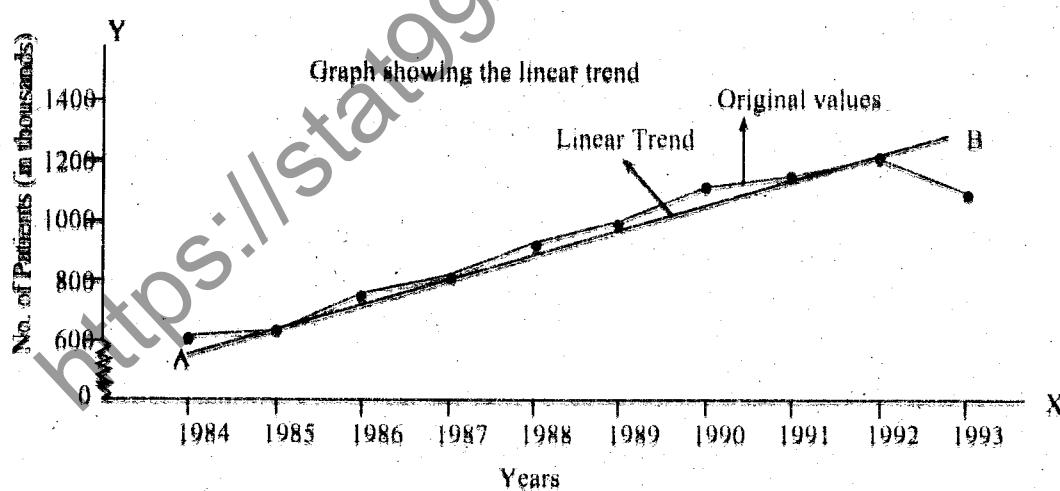


Figure 16.2

Example 16.3.

Table 16.3. Sales of a Utility Store

Years	1986	1987	1988	1989	1990	1991	1992	1993	1994
Sales (million Rs.)	5.5	7.2	13.1	28.5	37.5	55.2	77.5	105.0	135.5

The graph of the data in Table 16.3 is shown in Figure 16.3.

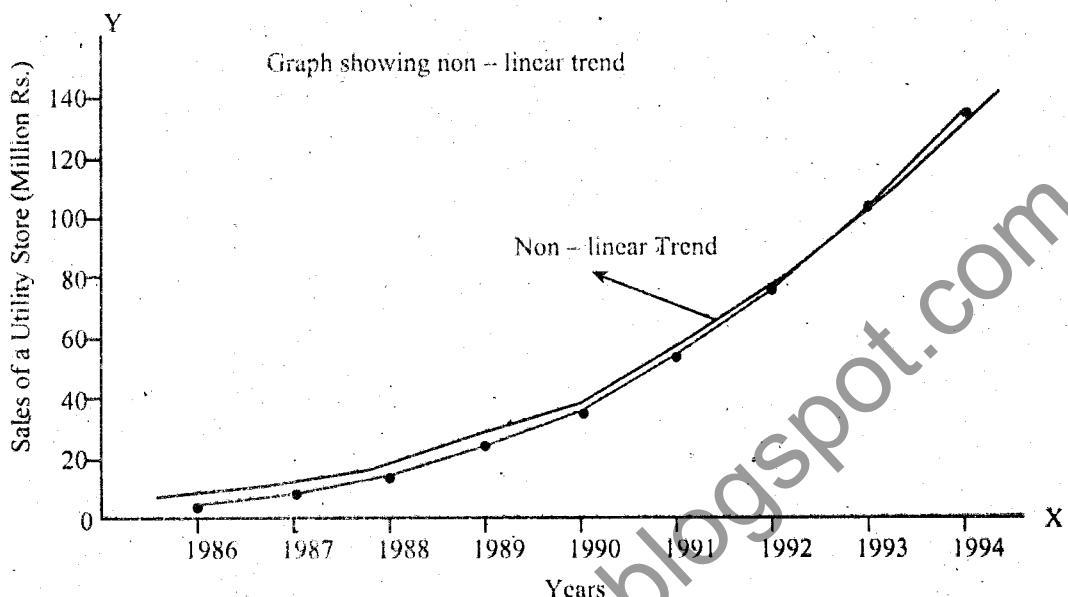


Figure 16.3

16.3.2 SEASONAL VARIATIONS

The element of increase or decrease in a time series which is purely due to changes in seasons is called seasonal variations. The changing seasons have their influence on the time series and the same influence is experienced every year in the corresponding seasons. These changes are regular in nature. A year may be divided into two, three, four or more than four seasons. The word season is a relative term and the division of the year into different seasons depends upon the variable under consideration. When we talk about the prices of warm clothing and winter clothing, a year is divided into two seasons i.e. winter and summer. The prices of winter clothing are high in winter and low in summer. The prices of cold drinks and ice-creams are low in winter and high in summer. The effect of the season prevails for the duration of the season and the time series may be influenced by the next coming season. If there is no seasonal effect, the time series remains stable in all seasons of the year. The effect of seasons is measured by seasonal indices. If the seasonal index for winter is 140 %, it means that the level of the time variable shows an increase of 40 % in winter as compared to the overall average for all the years. The seasonal variation are not necessarily linked with a season. The changes which take place regularly every year but have no concern with any season are also called seasonal variation. The increase in the prices of shoes on Eids and Xmas are examples of seasonal variations. The changes which are regularly repeated during one year are called seasonal variations.

Fig. 16.4. shows the seasonal variations in the quarterly prices of wheat (Maxi-Pak) during 1991 and 1992. The prices are low in quarters I and II in 1991 and 1992. Prices have increased in quarters III and IV and the increase has taken place in both the years.

Example 16.4.

Table 16.4. Quarterly Prices (Rs. per Quintal) of Wheat for 1991 and 1992

Years	Quarters			
	I	II	III	IV
1991	307	300	323	350
1992	371	364	390	420

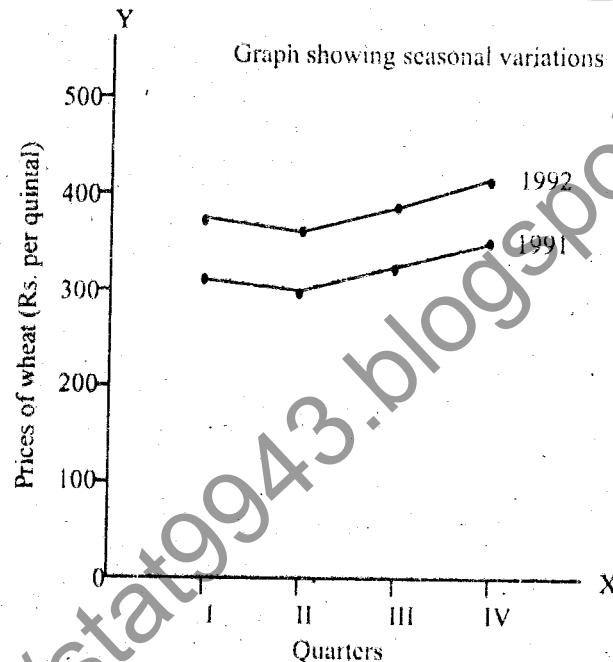


Figure 16.4

16.3.3 CYCLICAL VARIATIONS

These movements are in the form of waves. The waves are formed about the average variation of a time series. In graph, these movements are like business cycles which pass through the stages of boom, recession, depression, recovery and back to prosperity. Fig. 16.5. shows the stages of cycles.

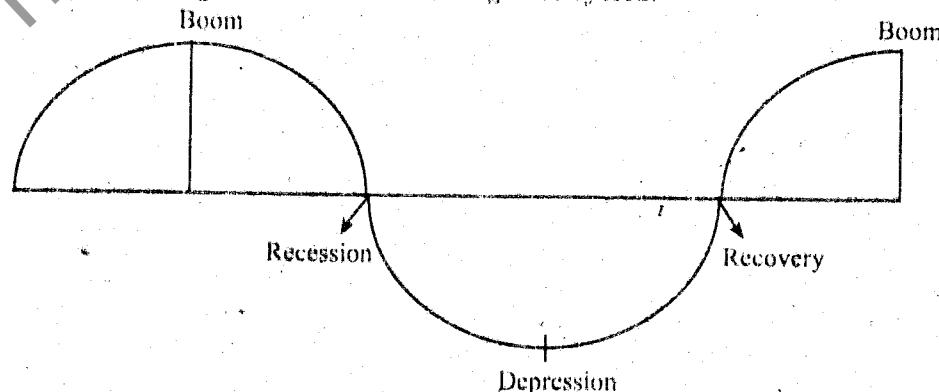


Figure 16.5

The distance between the two booms is called a cycle and the time period involved between two booms is called span of the cycle. The span of the cycle is usually very long. The time series are influenced by the economic conditions in a country. Thus the business cycles, if any, in the economic activity are responsible for the cycles in the related time series. The effect of these changes is quite feeble and it is therefore difficult to measure these variations. The numerical magnitude of these variations is insignificant and they are mixed with the trend. It is difficult to separate them from the trend.

16.3.4 IRREGULAR VARIATIONS

Sometimes there are sudden movements in the time series. These movements are due to sudden causes like floods, strikes, epidemics, wars etc. The time series is disturbed by some unpredictable forces. The time series comes to its original position when the effect of irregular or random causes is over. These variations cannot be controlled. They are also called erratic or accidental variations.

Example 16.5.

Explain the following movements with respect to variations in time series.

- (i) Increase in literacy rate in a country.
- (ii) Increase in the prices of school uniform in the start of the academic year.
- (iii) The number of T.V's damaged due to fall in voltage of electric supply.
- (iv) Non-availability of cement in the market due to depression in the country.
- (v) Decrease in demand for ice in winter season.
- (vi) Increase in the prices of ghee in the month of Ramazan.

Answers:

- (i) Increase in literacy rate is a regular and gradual process. A high level of literacy cannot be achieved in a short period of time. It requires a long time to make the buildings for schools and to train the teaching staff. This type of increase is called secular trend in a time series.
- (ii) Increase in the prices of school uniforms takes place every year in the start of the academic session. It is a regular feature. Though this change is not related to any season, yet it is called seasonal because it occurs every year.
- (iii) Fall in the voltage is an irregular feature. It is sudden and is usually unpredictable. It is of the type of irregular variation in the time series.
- (iv) Non-availability of cement due to depression is comparable with movements of cyclical nature in the time series.
- (v) People do not require ice during winter season. It is clearly something which is related to the winter season.
- (vi) Increase in the prices of ghee in the month of Ramazan is a regular feature in our country. It happens every year in the month of Ramazan. As this change is regular, it is of the seasonal nature. In some country, the increase may not take place regularly in the month of Ramazan. If any increase takes place in the month of Ramazan, it will not be called seasonal. It will be seasonal if it is repeated every year.

16.4 ANALYSIS OF TIME SERIES

A time series is the combined effect of many forces. Some of these forces are very powerful and some are weak. The forces acting on a time series are called the variations or fluctuations in a time series. They may also be termed as movements or components of a time series. The time variable Y_t is made up of other variables which are T, S, C and I where T stands for trend, S for seasonal, C is for cyclical and I is for irregular. Thus Y_t can be written as: $Y_t = TSCI$ which is called the multiplicative model of the time series. The additive model of the time series is written as: $Y = T + S + C + I$

Anyone of these two models is used for the detailed study of the changes in the time series. Which model is better? It depends upon the assumptions about the given time series. Usually the model $Y = TSCI$ is used in the study of the time series. Sometimes we are interested in the trend of the time series, sometimes our interest is about the seasonal forces. If we are interested in the trend of the time series, we have to remove the effects of other forces. Similarly the measurement of seasonal, cyclical and irregular variations is possible. A study regarding any component of the time series is called the *analysis of time series*. In this book we shall discuss the measurement of trend only. The measurement of seasonal, cyclical and irregular variations is beyond the scope of this book. A general statistical model for the time series can be written as:

$$Y_t = f(t) + u_t \quad t = 1, 2, 3, \dots, k$$

It is an old concept in which the observed time series is assumed to consist of two parts which are systematic part $f(t)$ and random part u_t . The systematic part $f(t)$ of the time series is also called the 'signal'. The random sequence u_t is also sometimes called the 'noise'.

According to this model, the time series consists of $f(t)$, a slowly moving function of time which may be considered combination of trend and cyclical variations. The random term u_t consists of all forces other than trend and cyclical. The measurement of u_t is not possible because it is combination of various actions on the time series.

16.5 MEASUREMENT OF SECULAR TREND

For the measurement of trend, we have to eliminate the short-term fluctuations from the time series. If we are interested in short-term fluctuations, the long-term trend is to be removed from the time series. The removal of the trend from the time series is possible only if trend has been measured.

Secular trend is a smooth line or a curve and is a continuous function of the time. It can be measured by the following methods:

- (i) The method of free - hand curve. (ii) The method of semi - averages.
- (iii) The method of moving averages. (iv) The method of least squares.

16.5.1 THE METHOD OF FREE-HAND CURVE

We make a graph of the observed time series taking time on X-axis and the variable on the Y-axis. The plotted points are joined together by straight lines to get a graph called the histogram. We carefully examine the shape of the graph. The graph of the original data may be in the neighbourhood of a straight line or it may be close to some curve. We draw a straight line or a freehand curve passing through the plotted points such that the growth of the time series is indicated by the trend line or the trend curve. The line or curve drawn smoothes out short term fluctuations and the remaining portion of the time series represents the trend. We can read the trend values from the trend line or trend curve for all the time periods of the time series. The trend thus drawn can be extended to estimate the values of the time variable for some periods beyond the given time series. It is called *forecasting*. The future values can be estimated very easily.

MERITS

This method is very simple. It is applicable for linear and non-linear trends. It gives us a quick idea about the rise and fall of the time series. For very long time series, the graph of the original data enables us to decide about the application of more mathematical models for the measurement of trend. A monthly data of 5 years has 60 values. A graph of these values may suggest that the trend is linear for the first two years (24 values) and for the next 3 years, it is non-linear. We accordingly apply the linear approach on the first 24 values and the curvilinear technique on the next 36 values.

DEMERITS

It is not mathematical in nature. Different persons may draw a different trend. The method does not appeal to a common man because it seems as if it is something rough and crude.

Example 16.6.

Measure the trend by method of free-hand curve from the data given in Table 16.5.

Table 16.5. Production of wheat in the Punjab.

Years	1981-82	1982-83	1983-84	1984-85	1985-86
Production million metric tons	8.6	8.9	7.6	8.3	10.4
Years	1986-87	1987-88	1988-89	1989-90	1990-91
Production million metric tons	9.2	9.2	10.5	10.5	10.5

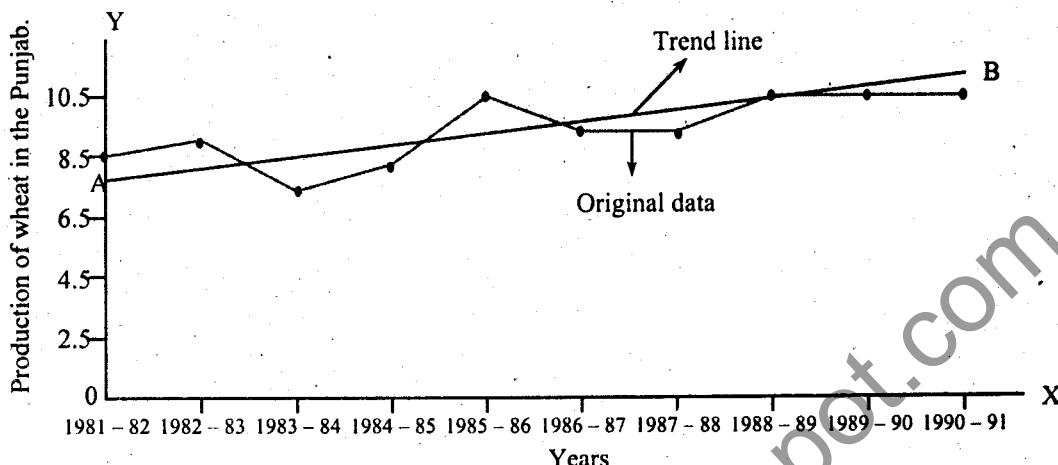


Figure 16.6

We observe that the graph of the original data does not show any closeness to any type of curve. It looks like increasing very slowly in straight (linear) manner. Thus we draw a line AB as an approximation to the original graph. The line AB represents the trend line and from this line we read the trend values for the given years. The trend values are: 8, 8.3, 8.6, 8.9, 9.2, 9.5, 9.8, 10.1, 10.4, 10.7.

16.5.2 THE METHOD OF SEMI-AVERAGES

This method consists of dividing the time series into two equal or almost equal parts and finding the arithmetic mean for each part.

If the time series contains odd number of periods, the middle value is omitted so that both the parts of the series contain equal number of values. The average for each part is calculated and is written against the middle period of the respective part. If a part contains odd number of values the average is written against the middle period of that part. Suppose there are 5 years in a part, the average is written against the 3rd year. When one half of the series contains even number of years, say 4 years, the average is written against the centre of the 2nd and 3rd year. The calculated averages are plotted on the graph paper and are joined together by a straight line. This straight line measures the trend of the data. This method of finding trend is called semi-averages method. We can read the trend values for all periods from the graph paper. The trend line can be extended on both sides to give the trend for the entire time series. The line can also be used for future prediction but only for the near future.

MERITS

This method is very simple and does not require much of calculations.

DEMERITS

The method is used only when the trend is linear or almost linear. For non-linear trend this method is not applicable. It is based on the calculation of average and the average is affected by extreme values. Thus if there is some very large value or very small value in the time series, that extreme value should either be omitted or this method should not be applied. We can also write the equation of the trend line.

Example 16.7.

Measure the trend by the method of semi-averages by using the data of Table 16.1. Also write the equation of the trend line with origin at 1984 – 85.

Years	Values (million Rs.)	Semi-totals	Semi-averages	Trend values (\hat{Y})
1984 – 85	18.6			$28.664 - 3.656 = 25.008$
1985 – 86	22.6			$32.32 - 3.656 = 28.664$
1986 – 87	38.1	161.6	32.32	32.32
1987 – 88	40.9			$32.32 + 3.656 = 35.976$
1988 – 89	41.4			$35.976 + 3.656 = 39.632$
1989 – 90	40.1			$39.632 + 3.656 = 43.288$
1990 – 91	46.6			$43.288 + 3.656 = 46.944$
1991 – 92	60.7	253.0	50.60	50.60
1992 – 93	52.2			$50.60 + 3.656 = 54.256$
1993 – 94	53.4			$54.256 + 3.656 = 57.912$

$$\text{Trend for } 1991-92 = 50.60 \quad \text{Trend for } 1986-87 = 32.32$$

$$\text{Increase in trend in 5 years} = 18.28 \quad \text{Increase in trend in 1 year} = 3.656$$

Trend for one year is 3.656. It is called slope of the trend line and is denoted by 'b'. Thus $b = 3.656$. The trend for 1987 – 88 is calculated by adding 3.656 to 32.32 and similar calculations are done for the subsequent years. Trend for 1985 – 86 is less than the trend for 1986 – 87. Thus trend for 1985 – 86 is $= 32.32 - 3.656 = 28.664$. Trend for the year 1984 – 85 = 25.008. This is called the intercept because 1984 – 85 is the origin. Intercept is the value of Y when X = 0. Intercept is denoted by 'a'. The equation of trend line is $\hat{Y} = a + bX = 25.008 + 3.656 X (1984 - 85 = 0)$ where \hat{Y} shows the trend values. This equation can be used to calculate the trend values of the time series. It can also be used for forecasting the future values of the variable.

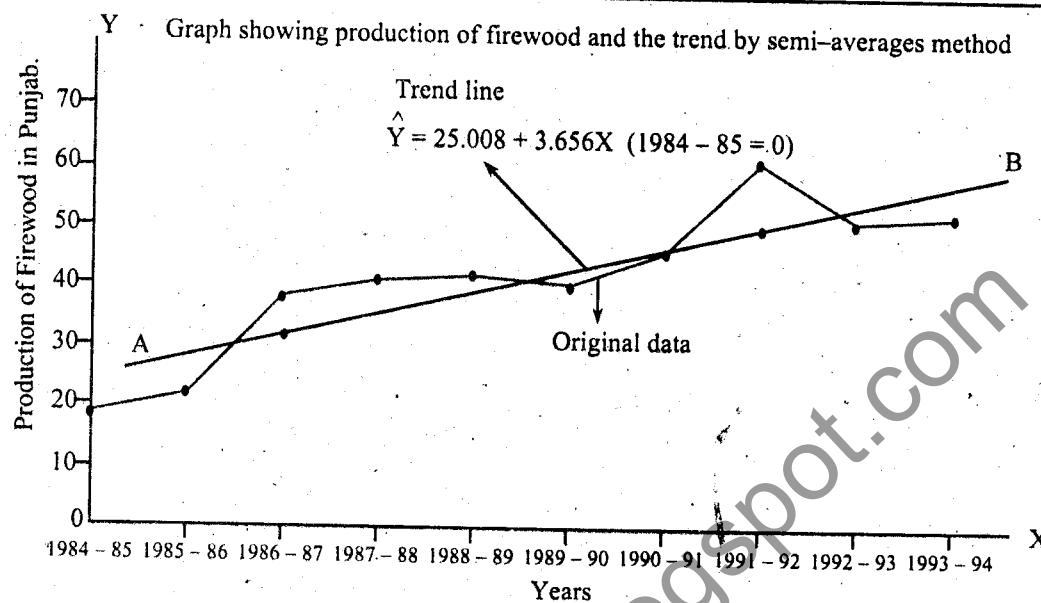


Figure 16.7

16.5.3 THE METHOD OF MOVING AVERAGES

Suppose that there are n time periods denoted by $t_1, t_2, t_3, \dots, t_n$ and the corresponding values of Y variable are $Y_1, Y_2, Y_3, \dots, Y_n$. First of all we have to decide the period of the moving averages. For short time series, we use period of 3 or 4 values. For long time series, the period may be 7, 10 or more. For quarterly time series, we always calculate averages taking 4-quarters at a time. In monthly time series, 12-monthly moving averages are calculated. Suppose the given time series is in years and we have decided to calculate 3-years moving average. The moving averages denoted by a_1, a_2, \dots, a_{n-2} are calculated as below:

Years (t)	Variable (Y)	3-Year moving totals	3-Year moving averages
t_1	Y_1	—	—
t_2	Y_2	$Y_1 + Y_2 + Y_3$	$\frac{Y_1 + Y_2 + Y_3}{3} = a_1$
t_3	Y_3	$Y_2 + Y_3 + Y_4$	$\frac{Y_2 + Y_3 + Y_4}{3} = a_2$
t_4	Y_4	•	•
•	•	•	•
•	•	•	•
t_{n-2}	Y_{n-2}	•	•
t_{n-1}	Y_{n-1}	$Y_{n-2} + Y_{n-1} + Y_n$	$\frac{Y_{n-2} + Y_{n-1} + Y_n}{3} = a_{n-2}$
t_n	Y_n	—	—

The average of the first 3 values is $\frac{Y_1 + Y_2 + Y_3}{3}$ and is denoted by a_1 . It is written against the middle year t_2 . We leave the first value Y_1 and calculate the average for the next three values. This average is $a_2 = \frac{Y_2 + Y_3 + Y_4}{3}$ and is written against the middle year t_3 . The process is carried out to calculate the remaining moving averages. 4-years moving averages are calculated as under:

Years (t)	Variable (Y)	4 – Years moving averages	4 – Years moving averages centred
t_1	Y_1		
t_2	Y_2	$\frac{Y_1 + Y_2 + Y_3 + Y_4}{4} = a_1$	
t_3	Y_3		$\frac{a_1 + a_2}{2} = A_1$
t_4	Y_4	$\frac{Y_2 + Y_3 + Y_4 + Y_5}{4} = a_2$	$\frac{a_2 + a_3}{2} = A_2$
t_5	Y_5	$\frac{Y_3 + Y_4 + Y_5 + Y_6}{4} = a_3$	
.	.	.	.
.	.	.	.
.	.	.	.

The first average is a_1 which is calculated as $a_1 = \frac{Y_1 + Y_2 + Y_3 + Y_4}{4}$. It is written against the middle of t_2 and t_3 . The second average is a_2 which is calculated as $a_2 = \frac{Y_2 + Y_3 + Y_4 + Y_5}{4}$. It is written against the middle of t_3 and t_4 . The two averages a_1 and a_2 are further averaged to get an average $A_1 = \frac{a_1 + a_2}{2}$, which refers to the centre of t_3 and is written against t_3 . This is called centering of the 4-years moving averages. The process is continued till the end of the series to get 4-years moving averages centred. The moving averages of some proper period smooth out the short term fluctuations and the trend is measured by the moving averages.

MERITS

Moving averages can be used for measuring the trend of any time series. The method is applicable for linear as well as non-linear trends.

DEMERITS

The trend obtained by moving averages is, in general, neither a straight line nor some standard curve. For this reason the trend cannot be extended for forecasting the future values. Trend values are not available for some periods in the start and some values at the end of the time series. The method is not applicable for short time series.

Example 16.8

Compute 5-year, 7-year and 9-year moving averages for the following data.

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Value	2	4	6	8	10	12	14	16	18	20	22

Solution:

The necessary calculations are given below:

Year	Value	5-year moving		7-year moving		9-year moving	
		Total	Average	Total	Average	Total	Average
1990	2	-	-	-	-	-	-
1991	4	-	-	-	-	-	-
1992	6	30	6	-	-	-	-
1993	8	40	8	56	8	-	-
1994	10	50	10	70	10	90	10
1995	12	60	12	84	12	108	12
1996	14	70	14	98	14	126	14
1997	16	80	16	112	16	-	-
1998	18	90	18	-	-	-	-
1999	20	-	-	-	-	-	-
2000	22	-	-	-	-	-	-

Example 16.9

Compute 4-year moving average centred for the following time series:

Year	1995	1996	1997	1998	1999	2000	2001	2002
Production (million pounds)	80	90	92	83	87	96	100	110

Solution:

The necessary calculations are given below:

Col. 1 Year	Col. 2 Production (million pounds)	Col. 3 4-year moving total	Col. 4 4-year moving average	Col. 5 2-values moving total of Col. 4	Col. 6 4-year moving average centred
1995	80			—	—
1996	90	345	86.25	—	—
1997	92	352	88.00	174.25	87.125
1998	83	358	89.50	177.50	88.750
1999	87	366	91.50	181.00	90.500
2000	96	393	98.25	189.75	94.875
2001	100			—	—
2002	110			—	—

Alternative Calculations:

Col. 1 Year	Col. 2 Production (million pounds)	Col. 3 4-year moving total	Col. 4 2-values moving total of Col. 3	Col. 5 4-year moving average centred $Col. 4 \div 8$
1995	80		—	—
1996	90	345	—	—
1997	92	352	697	87.125
1998	83	358	710	88.750
1999	87	366	724	90.500
2000	96	393	759	94.875
2001	100		—	—
2002	110		—	—

Example 16.10

The following data give the annual sales in Rs. (00000) of a utility store.

Year	1995	1996	1997	1998	1999	2000
Sales	50	60	75	73	80	85

Show by direct numerical calculation that the 2-year centred moving averages are equivalent to 3-year weighted moving averages with weights 1, 2, 1 respectively.

Solution:

The necessary calculations are given below:

Year	Sales	2-year moving total	2-values moving total of Col. 3	2-year moving averages centred Col. 4 ÷ 4	3-year weighted moving total (weights 1, 2, 1)	3-year weighted moving averages (weights 1, 2, 1)
1995	50	—	—	—	—	—
1996	60	110	245	61.25	245	61.25
1997	75	135	283	70.75	283	70.75
1998	73	148	301	75.25	301	75.25
1999	80	153	318	79.50	318	79.50
2000	85	165	—	—	—	—

Hence 2-year centred moving averages are equivalent to 3-year weighted moving averages with weights 1, 2, 1 respectively.

16.5.4 METHOD OF LEAST SQUARES

This method has already been explained in detail in the chapter of regression analysis of this book. We examine the given time series to decide whether the trend is linear or not. If the difference between two adjacent values is almost same in most of the cases, then the trend is linear and we fit a straight line to the time series for measurement of trend.

16.5.5 FITTING A STRAIGHT LINE

The equation of straight line is $Y = a + bX$ where 'a' and 'b' are unknowns to be determined. The variable X denotes time and Y denotes the dependent variable. Our job is to calculate the values of 'a' and 'b' from the given data. For this purpose we write the normal equations of 'a' and 'b' which are

$$\Sigma Y = na + b\Sigma X \quad \text{and} \quad \Sigma XY = a\Sigma X + b\Sigma X^2$$

Putting the values of various summations and solving the equations simultaneously, we get the values of 'a' and 'b'. Putting these values in the equation $Y = a + bX$, we get the equation of best fitted straight line. The fitted equation is written as $\hat{Y} = a + bX$ where \hat{Y} denotes the trend values which make the best fitted line.

16.5.6 CODING OF THE TIME PERIODS

The time is the X-variable and if the given time series consists of years, the given years can be denoted by X, but it is never advisable. The years are assigned some smaller units to simplify the calculations. This process is called coding. The initial period may be taken as 0, the value of the next period will be 1 and so on. Thus 0, 1, 2, 3, ... will be used as codes for the time periods which may be years, quarters, months, weeks or any other period of time. It is further illustrated in the following columns.

Years	X	Years	X	Years	X	or	Years	X	or
					X	X			X
1988	0	1988 - 89	0	1988	0	0	1951	0	0
1989	1	1989 - 90	1	1990	1	2	1961	1	10
1990	2	1990 - 91	2	1992	2	4	1971	2	20
1991	3	1991 - 92	3	1994	3	6	1981	3	30

Years	X	Years (un-equal interval)	X	or	Years and Quarters	X	Months	X
			X	X				
1988	0	1951	0	0	1988 I	0	January	0
1989	1	1961	10	1	II	1	February	1
1990	2	1972	21	2.1	III	2	March	2
*		1981	30	3.0	IV	3	April	3
1992	4				1989 II	4		
						5		

* a missing value.

16.5.7 CHANGE OF ORIGIN IN CODING

The calculations can be further simplified if we use the coding such that $\Sigma X = 0$. When $\Sigma X = 0$, the normal equations are reduced to $\Sigma Y = na$ and $\Sigma XY = b\Sigma X^2$.

From these equations we can very easily find the values of 'a' and 'b'. Taking $\Sigma X = 0$, becomes very important if we have to fit 2nd degree curve or curves involving higher powers of X in the normal equations. To make $\Sigma X = 0$, for odd number of periods we take 0 against the middle period and the succeeding periods are taken as 1, 2, 3 and so on. The periods preceding the middle period are denoted by -1, -2, -3 and so on. If we have seven years, their codes will be -3, -2, -1, 0, 1, 2, 3 with $\Sigma X = 0$. For even number of periods, we assume 0 in the centre of the two middle most periods and after 0 (we do not write 0) the codes will be 0.5, 1.5, 2.5 and so on. Before 0, the codes will be -0.5, -1.5, -2.5 and so on. The idea of change of origin is further explained in the following columns.

Odd No. of Years	X	Even No. of Years	X	or	Years with Quarters	X	Months	X
			X	X				
1989	-2	1989	-2.5	-5	1989 I	-7	January	-5
1990	-1	1990	-1.5	-3	II	-5	February	-3
1991	0	1991	-0.5	-1	III	-3	March	-1
1992	1	1992	0.5	1	IV	-1	April	1
1993	2	1993	1.5	3	1990 I	1	May	3
		1994	2.5	5	II	3	June	5
					III	5		
					IV	7		
$\Sigma X = 0$		$\Sigma X = 0$		$\Sigma X = 0$		$\Sigma X = 0$		$\Sigma X = 0$

Example 16.11

Fit a straight line with the help of least squares method to the following data taking the origin at the middle of the time period and unit of measurement for X being one year.

Year	1998	1999	2000	2001	2002
Sales (million Rs.)	28	32	40	44	56

Solution:

The equation of the straight line is $Y = a + bX$

The normal equations are: $\Sigma Y = na + b\Sigma X$ and $\Sigma XY = a\Sigma X + b\Sigma X^2$

The necessary calculations are given below:

Year	X	Y	XY	X^2
1998	-2	28	-56	4
1999	-1	32	-32	1
2000	0	40	0	0
2001	+1	44	+44	1
2002	+2	56	+112	4
Total	0	200	68	10

Since $\Sigma X = 0$, the normal equations become

$$\Sigma Y = na \text{ or } a = \frac{\Sigma Y}{n} \text{ and } \Sigma XY = b\Sigma X^2 \text{ or } b = \frac{\Sigma XY}{\Sigma X^2}$$

Substituting the values, we get

$$a = \frac{200}{5} = 40 \text{ and } b = \frac{68}{10} = 6.8$$

Hence the fitted straight line is $\hat{Y} = 40 + 6.8X$

Example 16.12

Fit a linear trend equation by the method of least squares for the following time series, taking the origin at the middle of 1997 and 1998, the unit of X being half year.

Year	1995	1996	1997	1998	1999	2000
Profits (0000 Rs.)	4	6	7	5	8	12

Also compute the sum of the residuals and the sum of squares of residuals.

Solution:

The equation of the linear trend is $\hat{Y} = a + bX$

The normal equations are: $\Sigma Y = na + b\Sigma X$ and $\Sigma XY = a\Sigma X + b\Sigma X^2$

The necessary calculations are given below:

Year	X	Y	XY	X^2	$\hat{Y} = 7 + 0.63X$	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
1995	-5	4	-20	25	3.85	0.15	0.0225
1996	-3	6	-18	9	5.11	0.89	0.7921
1997	-1	7	-7	1	6.37	0.63	0.3969
1998	+1	5	+5	1	7.63	-2.63	6.9169
1999	+3	8	+24	9	8.89	-0.89	0.7921
2000	+5	12	+60	25	10.15	1.85	3.4225
Total	0	42	44	70	42	0	12.343

Since $\Sigma X = 0$, the normal equations will be

$$\Sigma Y = na \quad \text{or} \quad a = \frac{\Sigma Y}{n} \quad \text{and} \quad \Sigma XY = b\Sigma X^2 \quad \text{or} \quad b = \frac{\Sigma XY}{\Sigma X^2}$$

Substituting the values, we get

$$a = \frac{42}{6} = 7 \quad \text{and} \quad b = \frac{44}{70} = 0.63$$

Hence the fitted linear trend is $\hat{Y} = 7 + 0.63X$.

The trend values are computed from the above equation by substituting $X = -5, -3, -1, 1, 3, 5$, which are shown in the above table.

Hence sum of residuals $= \sum(Y - \hat{Y}) = 0$ and sum of squares of residuals $= \sum(Y - \hat{Y})^2 = 12.343$.

Example 16.13

If the straight line fitted to the data for the years 1995 to 2000 (both inclusive) with origin at the middle of 1997 and 1998 is $\hat{Y} = 195 + 8.5X$, the unit of being $1/2$ year. Determine the trend values for the years 1995 to 2000. Also determine the straight line by shifting the origin to 1995.

Solution:

Year	X	$\hat{Y} = 195 + 8.5 X$	u
1995	-5	$195 + 8.5(-5) = 152.5$	0
1996	-3	$195 + 8.5(-3) = 169.5$	1
1997	-1	$195 + 8.5(-1) = 186.5$	2
1998	+1	$195 + 8.5(+1) = 203.5$	3
1999	+3	$195 + 8.5(+3) = 220.5$	4
2000	+5	$195 + 8.5(+5) = 237.5$	5

When we have to shift the origin at 1995, on the new scale the years are coded as 0, 1, 2, 3, 4 and 5. Let u denote these values. There is a relation between X and u, clearly $X = 2u - 5$. In the equation $\hat{Y} = 195 + 8.5 X$; X is replaced by $2u - 5$ and we get the equation in which the origin is at 1995. Thus $\hat{Y} = 195 + 8.5(2u - 5) = 195 + 17u - 42.5 = 152.5 + 17u$.

16.6 FITTING OF SECOND DEGREE PARABOLA

We fit a straight line to the given time series when the Y-values move in a linear manner i.e; the increase or decrease per given time period is almost constant. If the increase or decrease do not show this pattern, we do not fit the straight line. In that case we fit a curve to the data. A simple curve to be discussed here is the second degree parabola:

$$Y = a + bX + cX^2$$

It is also called second degree curve. For fitting of this curve, three unknowns 'a', 'b', 'c' are to be estimated from the given data. This is done by means of the following three normal equations:

$$\Sigma Y = na + b\Sigma X + c\Sigma X^2$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

$$\Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

The solution of these equations requires lot of computational work. To reduce the computations, we use the codes for the time series such that $\Sigma X = \Sigma X^3 = 0$. When ΣX and ΣX^3 are zero, the normal equations are reduced to

$$\Sigma \hat{Y} = na + b\Sigma X^2 \quad \dots \dots \quad (1)$$

$$\Sigma XY = b\Sigma X^2 \quad \dots \dots \quad (2)$$

$$\Sigma X^2 Y = a\Sigma X^2 + c\Sigma X^4 \quad \dots \dots \quad (3)$$

Equation (2) directly gives the value of b. Equations (1) and (3) are solved simultaneously to get the values of 'a' and 'c'. Having calculated the values of 'a', 'b' and 'c', we put them in the equation and get the equation of the fitted curve written as $\hat{Y} = a + bX + cX^2$. The origin must be mentioned with the fitted equation.

Example 16.14

Fit a second degree parabola to the following results for the years 1998 to 2002; both inclusive:

$$\Sigma X = 0, \Sigma Y = 250, \Sigma XY = 600, \Sigma X^2 = 150, \Sigma X^3 = 0, \Sigma X^2 Y = 8200, \Sigma X^4 = 5500.$$

Solution:

The equation of the second degree parabola is $Y = a + bX + cX^2$

The normal equations are:

$$\Sigma Y = na + b\Sigma X + c\Sigma X^2$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

$$\Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

Since $\Sigma X = \Sigma X^3 = 0$, the normal equations become

$$\Sigma Y = na + c\Sigma X^2$$

$$\Sigma XY = b\Sigma X^2$$

$$\Sigma X^2 Y = a\Sigma X^2 + c\Sigma X^4$$

Substituting the values, we get

$$250 = 5a + 150c \quad \dots \dots (1)$$

$$600 = 150b \quad \dots \dots (2)$$

$$8200 = 150a + 5500c \quad \dots \dots (3)$$

$$\text{From equation (2), we get } b = \frac{600}{150} = 4$$

Solving equations (1) and (3), we multiply equation (1) by 30 and subtract from equation (3), we have

$$8200 = 150a + 5500c$$

$$7500 = 150a + 4500c$$

$$700 = 1000c \text{ or } c = \frac{700}{1000} = 0.7$$

Substituting $c = 0.7$ in equation (1), we get

$$250 = 5a + 150(0.7) \text{ or } 5a = 250 - 105 = 145 \text{ or } a = \frac{145}{5} = 29$$

Hence the fitted second degree parabola is $\hat{Y} = 29 + 4X + 0.7X^2$

SHORT DEFINITIONS**Time Series**

A time series is the measurement of a variable at regular intervals of time.

or

An arrangement of statistical data with respect to their time of occurrence is called time series.

Histogram

A histogram is a vertical bar chart in which the rectangular bars are constructed at the boundaries of each class.

Historigram

A graph of time series or historical series is called historigram.

Signal

Signal is the sequence following a regular pattern of variations.

Noise

Noise is the sequence following an irregular pattern of variations.

Secular Trend

Secular trend in a time series is a long term, smooth, underlying pattern of change from time to time in the series.

or

A long-term increase or decrease in a time series in which the rate of change is relatively constant is called secular trend.

Seasonal Variation

Seasonal variation is the repetitive pattern of variation occurring within a year in a time series.

or

A pattern that is repeated throughout a time series and has a recurrence period of at most one year is called seasonal variation.

Cyclical Variation

The cyclical variation of a time series is the wavelike or oscillating pattern about the trend that is attributable to business and economic conditions at the time. It is also known as a business cycle.

or

A pattern within the time series that repeats itself throughout the time series and has a recurrence period of more than one year.

Irregular Variation

Irregular variation in a time series is composed of changes that cannot be described as secular trend, cyclical variation, or seasonal variation.

or

Changes in the time series data that are unpredictable and cannot be associated with the secular trend, seasonal variation, or cyclical variation are called irregular variation.

Business Cycle

A business cycle has four stages:

- (i) *Prosperity or Boom* When production of a thing is maximum, this stage is called as prosperity stage.
- (ii) *Recession* When production of a thing is decreasing, this stage is called as recession.
- (iii) *Depression* When the production of a thing is minimum, this stage is called as depression.
- (iv) *Recovery* When the production of a thing is increasing towards prosperity, this stage is called recovery stage.

Multiplicative Time Series Model

A model whereby the separate components of the time series are multiplied together to identify the actual time series value. When the four components of time series are assumed to be present in a multiplicative form, the model is called a multiplicative time series model that is $Y = TSCI$.

Additive Time Series Model

When the four components of a time series are assumed to be present in an additive form that is $Y = T + S + C + I$, the model is called an additive time series model.

Analysis of Time Series

Analysis of the time series is study of various components that are present in a time series and to analyze them. By analysis of time series, we mean to concentrate in finding the effect of one component after eliminating the effects of other three components from the time series data.

Graphic or Freehand Method

The given data are plotted on a graph paper and trend line is fitted to the data just by inspection.

Semi-Average Method

The data for which the trend values are to be computed by dividing the values into two equal parts. If there are odd number of years the middle year is left and the two equal parts are formed. The average is computed for each part and is written against the midpoint of the each part. These two averages are shown on the graph along with the original values and a straight line is drawn through these two points which describes the trend.

Moving Average Method

A method of forecasting or smoothing a time series by averaging each successive group of data points is called moving average method.

or

The successive averages of 'n' consecutive values in a time series is known as moving average method.

Method of Least Squares

The method of least squares is a widely used method of fitting a curve to data and is the most popular method of computing the secular trend of time series. The method locates the trend value for the middle of the time period at the average for the data to which the trend line is being fitted. The objective of least squares method is to minimize the sum of squares of residuals.

Residual

Residual is the difference between the actual value of the time series and the forecast value. It is also called the forecast error.

Principle of Least Squares

According to the principle of least squares "the best or most plausible value of any observed quantity is that for which sum of squares of residuals is least".

Forecasting

The process of predicting the magnitude that a variable will assume in future is called forecasting.

GIVEN THE FOLLOWING STATEMENTS. LINK THEM WITH THE MOVEMENTS OF THE TIME SERIES

1. Decrease in death rates in Pakistan.

Ans. Secular trend

2. Increase in prices of shoes for children before Eid.

Ans. Seasonal

3. Non-availability of transport due to heavy rains.

Ans. Irregular

4. Shortage of sugar due to strikes in sugar Mills.

Ans. Irregular

5. Depression in business.

Ans. Cyclical

6. Decrease in prices of cold drinks in winter.

Ans. Seasonal

7. Demand for umbrellas during rainy season.

Ans. Seasonal

8. Increase in literacy rate in Pakistan.

Ans. Secular trend

9. Development in agricultural sector in Pakistan.

Ans. Secular trend

10. Increase in population in Pakistan.

Ans. Secular trend

11. Break in supply of fruits and vegetables due to heavy rains.

Ans. Irregular

12. Attendance of children in the school due to rain.

Ans. Seasonal

13. Number of marriages during the months of April and October.

Ans. Seasonal

14. Number of vehicles on the roads in a city at school time.

Ans. Seasonal

15. Number of vehicles on the roads in a city at 2 P.M. in summer season.

Ans. Seasonal

16. Damage to the crops due to floods.

Ans. Irregular

17. Increase in stationery items in the start of the new academic session.

Ans. Seasonal

18. Fashion in the dress.

Ans. Secular trend

19. Number of vehicles on the roads.

Ans. Secular trend

20. Increase in the general level of prices.

Ans. Secular trend

MULTIPLE - CHOICE QUESTIONS

1. The graph of time series is called:

- | | |
|----------------|-------------------|
| (a) histogram | (b) straight line |
| (c) historgram | (d) ogive |

2. An orderly set of data arranged in accordance with their time of occurrence is called:

- | | |
|-----------------------|---------------------|
| (a) arithmetic series | (b) harmonic series |
| (c) geometric series | (d) time series |

3. The secular trend is measured by the method of semi-averages when:

- | | |
|---|---------------------|
| (a) time series based on yearly values | (b) trend is linear |
| (c) time series consists of even number of values | (d) none of them |

4. Increase in the number of patients in the hospital due to heat stroke is:

- | | |
|------------------------|-------------------------|
| (a) secular trend | (b) irregular variation |
| (c) seasonal variation | (d) cyclical variation |

5. The systematic components of time series which follow regular pattern of variations are called:

- | | |
|--------------------|--------------------------|
| (a) signal | (b) noise |
| (c) additive model | (d) multiplicative model |

6. The unsystematic sequence which follows irregular pattern of variations is called:

- | | |
|------------|----------------|
| (a) noise | (b) signal |
| (c) linear | (d) non-linear |

7. In time series seasonal variations can occur within a period of:

- | | |
|----------------|-----------------|
| (a) four years | (b) three years |
| (c) one year | (d) nine years |

8. Wheat crops badly damaged on account of rains is:

- | | |
|-----------------------|-----------------------|
| (a) cyclical movement | (b) random movement |
| (c) secular trend | (d) seasonal movement |

9. In a straight line equation $Y = a + bX$; a is the:

- | | |
|-----------------|------------------|
| (a) X-intercept | (b) slope |
| (c) Y-intercept | (d) none of them |

10. In a straight line equation $Y = a + bX$; b is the:
- Y-intercept
 - slope
 - X-intercept
 - trend
11. A second degree parabola is fitted to the time series when the variations are:
- linear
 - non-linear
 - upward
 - downward
12. If a straight line is fitted to the time series, then:
- $\Sigma Y = \Sigma \hat{Y}$
 - $\Sigma Y < \Sigma \hat{Y}$
 - $\Sigma Y > \Sigma \hat{Y}$
 - $\Sigma(Y - \hat{Y})^2 = 0$
13. Moving average method is used for measurement of trend when:
- trend is linear
 - trend is non linear
 - trend is curvilinear
 - none of them
14. When the trend is of exponential type, the moving averages are to be computed by using:
- arithmetic mean
 - geometric mean
 - harmonic mean
 - weighted mean
15. The long term trend of a time series graph appears to be:
- straight-line
 - upward
 - downward
 - parabolic curve or third degree curve
16. Indicate which of the following is an example of seasonal variations:
- Death rate decreased due to advance in science
 - The sale of air condition increases during summer
 - Recovery in business
 - Sudden causes by wars
17. The most commonly used mathematical method for measuring the trend is:
- moving average method
 - semi average method
 - method of least squares
 - none of them
18. A trend is the better fitted trend for which the sum of squares of residuals is:
- maximum
 - minimum
 - positive
 - negative
19. Decomposition of time series is called:
- histogram
 - analysis of time series
 - histogram
 - detrending
20. The fire in a factory is an example of:
- secular trend
 - seasonal movements
 - cyclical variations
 - irregular variations
21. Increased demand of admission in the subject of computer in Pakistan is:
- secular trend
 - cyclical trend
 - seasonal trend
 - irregular trend

22. Damages due to floods, droughts, strikes fires and political disturbances etc.
- trend
 - seasonal
 - cyclical
 - irregular
23. The general pattern of increase or decrease in economics or social phenomena is shown by:
- seasonal trend
 - cyclical trend
 - secular trend
 - irregular trend
24. In moving average method, we can not find the trend values of some:
- middle periods
 - end periods
 - starting periods
 - between extreme periods
25. The best fitting trend is one in which the sum of squares of residuals is:
- negative
 - least
 - zero
 - maximum
26. In fitting of a straight line, the value of slope remains unchanged by changing:
- scale
 - origin
 - both origin and scale
 - none of them
27. Depression in business is:
- secular trend
 - cyclical
 - seasonal
 - irregular
28. In fitting of straight line $\sum(Y - \hat{Y})^2 = 0$ when:
- all the observed Y values lie on the line
 - all the Y values are greater than the corresponding \hat{Y} values
 - all the Y values are positive
 - none of them
29. Semi-averages method is used for measurement of trend when:
- trend is linear
 - observed data contains yearly values
 - the given time series contains odd number of values
 - none of them
30. Moving-averages:
- give the trend in a straight line
 - measure the seasonal variations
 - smooth-out the time series
 - none of them
31. The rise and fall of a time series over periods longer than one year is called:
- secular trend
 - seasonal variation
 - cyclical variation
 - irregular variation
32. A time series has:
- two components
 - three components
 - four components
 - five components
33. The multiplicative time series model is:
- $Y = T + S + C + I$
 - $Y = TSCI$
 - $Y = a + bX$
 - $Y = a + bX + cX^2$

34. The additive model of the time series is:
- $(Y = T + C + I + S)$
 - $Y = TSCI$
 - $Y = a + bX$
 - $Y = a + bX + cX^2$
35. The difference between the actual value of the time series and the forecasted value is called:
- residual
 - sum of residual
 - sum of squares of residual
 - all of the above
36. A pattern that is repeated throughout a time series and has a recurrence period of at most one year is called:
- cyclical variation
 - irregular variation
 - seasonal variation
 - long term variation
37. A business cycle has:
- one stage
 - two stages
 - three stages
 - four stages
38. When the production of a thing is maximum, this stage is called:
- boom
 - recovery
 - recession
 - depression
39. When the production of a thing is minimum, this stage is called:
- prosperity
 - recession
 - recovery
 - depression
40. When the production of a thing is increasing towards prosperity, this stage is called as:
- recession
 - recovery
 - boom
 - depression
41. When the production of a thing is decreasing, this stage is called as:
- recession
 - recovery
 - prosperity
 - depression
42. The straight line is fitted to the time series when the movements in the time series are:
- nonlinear
 - linear
 - irregular
 - upward
43. If an annual time series consisting of even number of years is coded, then each coded interval is equal to:
- half year
 - one year
 - both (a) and (b)
 - two years
44. The second degree parabola is fitted to the time series when the variations are:
- linear
 - nonlinear
 - random
 - downward

45. For odd number of years, formula to code the values of X by taking origin at centre is:
- $X = \text{year} - \text{average of years}$
 - $X = \text{year} - \text{first year}$
 - $X = \text{year} - \text{last year}$
 - $X = \text{year} - \frac{1}{2} \text{average of years}$
46. For even number of years when origin is in the centre and the unit of X being one year, then X can be coded as:
- $X = \frac{\text{year} - \text{average of years}}{2}$
 - $X = \text{year} - \text{average of years}$
 - $X = \text{year} - 0.5 \text{average of years}$
 - $X = \text{average of years} - \text{year}$
47. For even number of years when origin is in the centre and the unit of X being half year, then X can be coded as:
- $X = \text{year} - \text{average of years}$
 - $X = 2(\text{year} - \text{average of years})$
 - $X = \frac{\text{year} - \text{average of years}}{2}$
 - $X = \text{year} - \frac{1}{2} \text{average of years}$
48. In semi averages method, if the number of values is odd then we drop:
- first value
 - last value
 - middle value
 - middle two values
49. The trend values in freehand curve method are obtained by:
- equation of straight line
 - graph
 - second degree parabola
 - all of the above
50. $\sum X = \sum X^3 = 0$, if origin is:
- at the end of time period
 - any where
 - at the middle of time period
 - at the beginning of time period

Answers

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

SHORT QUESTIONS

1. Given $Y = 127, 101, 130, 132, 126, 142, 138$ and $\hat{Y} = 116, 120, 124, 128, 132, 136, 140$. Find $e = (Y - \hat{Y})$.

Ans. 11, -19, 6, 4, -6, 6, -2

2. Given $\hat{Y} = 128 + 4X$ and $X = -3, -2, -1, 0, 1, 2, 3$. Find $\Sigma \hat{Y}$.

Ans. 896

3. Given $\Sigma X = 0$, $\Sigma Y = 27.1$, $\Sigma XY = 29.5$, $\Sigma X^2 = 330$. Determine the value of b.

Ans. b = 0.09

4. Given $\Sigma X = 0$, $\Sigma Y = 41172$ and $n = 10$. Find the value of X intercept a.

Ans. a = 4117.2

5. Given $(Y - \hat{Y}) = 0.5, -0.5, 1, -1, 0.5, -0.5$. Find sum of squares of residuals.

Ans. $\Sigma(Y - \hat{Y})^2 = 3$

6. Given $Y = 6, 8, 10, 12, 14$, $X = 0, 1, 2, 3, 4$, and $\hat{Y} = 6 + 2X$. Compute the sum of residuals.

Ans. $\Sigma(Y - \hat{Y}) = 0$

7. If $Y = 16, 18, 20, 22, 24$, $X = -2, -1, 0, 1, 2$, and $\hat{Y} = 20 + 2X$. Compute the sum of squares of residuals.

Ans. $\Sigma(Y - \hat{Y})^2 = 0$

8. Given $\Sigma X = 0$, $\Sigma Y = 245$, $\Sigma X^2 = 28$, $\Sigma XY = 66$ and $n = 7$. Fit a linear trend.

Ans. $\hat{Y} = 35 + 2.4X$

9. Suppose that a corporation finds that a linear trend for its sales is $\hat{Y} = 5 + 0.1X$, where Y is the firm's monthly sales (in millions rupees) and X is measured in months from January 1982. Based on this trend alone, what are the forecasted sales for the firm in February 1990?

Ans. If January 1982 is X = 0, then February 1990 is X = 97. Thus, the forecasted value of sales is $\hat{Y} = 5 + 0.1(97) = 5 + 9.7 = 14.7$ million rupees.

10. Suppose the following data represent the 7-year moving total over the 11-year period 1990 to 2000.
56, 70, 84, 98, 112. Compute the 7-year moving averages.

Ans. 8, 10, 12, 14, 16

11. Given the following data:

Year	1995	1996	1997	1998	1999	2000
Value	207	210	216	213	220	218

Applying the method of semi-averages, the trend values are 209, 211, 213, 215, 217 and 219. Write the equation of the trend line taking 1987 as origin.

Ans. $\hat{Y} = 209 + 2X$

12. Suppose the least-squares trend line to an annual time series containing 10 observations from 1991 to 2000 on real total revenues (in millions rupees) is $\hat{Y} = 3 + 1.2X$ (1991 as origin). Interpret the Y intercept 'a' and slope 'b' in this linear trend model.

Ans. The Y intercept a = 3 is the fitted trend value reflecting the real total revenues during the origin 1991. The slope b = 1.2 indicates that the real total revenues are increasing at a rate of 1.2 million rupees per year.

13. Suppose that the least-squares trend line to an annual time series containing 15 observations from 1986 to 2000 on real net sales is $\hat{Y} = 2.4 + 0.5 X$ (1986 as base year). What is the fitted trend value for this time series on real net sales for the 5th year?

Ans. $\hat{Y} = 2.4 + 0.5 (4) = 2.4 + 2 = 4.4$

14. Suppose that the least squares trend line to an annual time series containing 7 observations from 1987 to 1993 on real total profits (in thousands rupees) is $\hat{Y} = 600 + 75X$ (1990 as origin). What is the fitted trend value for this time series on real total profits for the most recent recorded year?

Ans. $\hat{Y} = 600 + 75 (3) = 825$

15. Suppose that the least squares trend line to an annual time series containing 7 observations from 1980 to 1986 on the world production of gold (in million ounces) is $\hat{Y} = 128 + 4X$ (1983 as origin). What is the trend forecast for this time series on the world production of gold 3 years after the last recorded value on total production of gold?

Ans. $\hat{Y} = 128 + 4(6) = 152$

16. Distinguish between short term and long term variations.
 17. Write down the various methods of measuring secular trend in a time series.
 18. What are the different components of a time series?
 19. Define a time series.
 20. What is meant by analysis of time series?
 21. Distinguish between regular and irregular fluctuations in time series.
 22. Define cyclical movements.
 23. Distinguish between additive model and multiplicative model of a time series.
 24. Explain the term secular trend.
 25. What is meant by seasonal variations?
 26. What is meant by business cycle?
 27. What do you understand by parabolic trend? How would you fit a parabola of the second degree to a time series to obtain trend values?
 28. Differentiate between histogram and historigram.
 29. Define the method of semi-averages.
 30. Explain the method of least squares in a time series.
 31. Differentiate between time series and analysis of time series.
 32. Differentiate between signal and noise.
 33. Discuss the meaning and purpose of moving averages.
 34. Describe the different components of time series.
 35. Distinguish between secular trend, seasonal variations and cyclical variations.
 36. Define irregular variations.

EXERCISES

1. Plot the following data on a graph paper and draw the trend passing through the given data by free-hand drawing and read the trend values.

Years	1980	1981	1982	1983	1984	1985	1986	1987	1988
Sales (Lakhs of Rs.)	20	22	24	26	30	34	40	46	49

Ans. 15, 19, 23, 27, 31, 35, 39, 43, 47

2. Compute the trend values by method of semi-averages for the following data. Also write the equation of the trend line.

Years	1985	1986	1987	1988	1989	1990	1991
Values	874	1024	1168	1405	1664	1958	2258

Ans. 787.5, 1022.0, 1256.5, 1491.0, 1725.5, 1960.0, 2194.5,

$$\hat{Y} = 787.5 + 234.5 X \quad (1985 = 0)$$

3. Plot the following data on a graph paper and compute trend values by the semi-averages method.

Years	1985	1986	1987	1988	1989	1990	1991	1992	1993
Values	305	315	300	290	280	285	285	279	295

Ans. 307.45, 304.15, 300.85, 297.55, 294.25, 290.95, 287.65, 284.35, 281.05

4. Applying the method of semi-averages for the following data, determine the trend values and also write the equation of the trend line.

Years	1987	1988	1989	1990	1991	1992
Price (Rs.)	207	210	216	213	220	218

Ans. 209, 211, 213, 215, 217, 219; $\hat{Y} = 209 + 2X$ (1987 as origin)

5. Obtain the trend values for the following data by using the method of semi-averages. Construct a graph illustrating the results obtained.

Years	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Consumption of cotton (thousands of bales)	706	854	886	815	827	761	805	797	746	727

Ans. 837.76, 827.68, 817.60, 807.52, 797.44, 787.36, 777.28, 767.20, 767.12, 747.04

6. Compute 3-year and 5-year moving averages from the following data:

Year	1982	1983	1984	1985	1986	1987	1988	1989
Factory Sales (in millions)	6.2	7.8	8.3	9.3	8.6	7.8	8.1	7.9

Ans: 7.43, 8.47, 8.73, 8.57, 8.17, 7.93 and 8.04, 8.36, 8.42, 8.34

7. The data given below represent the annual number of employees (in thousands) in an oil supply company for the years 1982 to 1991.

Year	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Number of employees (in thousands)	1.73	1.77	1.90	1.82	1.65	1.73	1.88	2.00	2.08	1.88

Compute 7-year moving averages to the data and plot the actual and trend values on the same graph paper.

Ans: 1.78, 1.82, 1.87, 1.86

8. Compute 5 yearly moving averages of students in a college as shown by the following figures:

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000
Number of students	1332	1397	1457	1592	1662	1805	1910	2027	2050

Ans. 1488.0, 1582.6, 1685.2, 1799.2, 1890.8

9. Compute 4-year centred moving average for the following time series:

Year	1993	1994	1995	1996	1997	1998	1999	2000
Production (in million kilograms)	331	344	349	332	364	395	400	410

Ans: 343.125, 353.625, 366.375, 382.500

10. Compute four-quarterly moving averages of the data for 2000 to 2001. Compare the moving averages with the original data by plotting both on the same graph paper.

Year	Quarter			
	I	II	III	IV
2000	20	26	33	45
2001	24	30	36	48

Ans: 31.5, 32.5, 33.4, 34.1

11. Measure the secular trend by calculating 2-yearly moving average centred for the following time series:

Year	1994	1995	1996	1997	1998	1999	2000
Value	170	250	325	425	520	600	720

Ans: 248.75, 331.25, 423.75, 516.25, 610.00

12. Compute 7-day moving averages for the following record of attendances:

Week	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
I	24	50	30	48	54	55	62
II	28	52	41	42	50	41	42

Plot the actual and trend values on the same graph paper.

Ans. 46.14, 46.71, 47.00, 48.57, 47.71, 47.14, 45.14, 42.29

13. Fit a straight line $Y = a + bX$ to the following data taking the origin at the middle of the time period and unit of measurement for X being one year.

Year	1996	1997	1998	1999	2000
Profits (000 Rs.)	50	70	80	120	160

Ans: $\hat{Y} = 96 + 27X$.

14. For the time series given below, relating to the world production of gold, compute the trend values for each year by fitting a straight line.

Year	1981	1982	1983	1984	1985	1986
Production (in million pounds)	6.3	8.1	8.2	7.9	8.9	8.6

Ans: $\hat{Y} = 8 + 0.2X$; 7.0, 7.4, 7.8, 8.2, 8.6, 9.0

15. A manufacturer of computers for the industrial market builds each unit to specifications after a firm order is received. The number of units scheduled for delivery over a 6-year period are listed below:

(i) Fit a linear trend equation by the method of least squares.

(ii) Determine the estimated value for 1990.

Year	1982	1983	1984	1985	1986	1987
Units scheduled for shipment in hundreds	42	45	49	40	43	45

Ans: (i) $\hat{Y} = 44 + 0X$ (ii) 44

16. Using the method of least squares derive a linear trend to the following results for the years 1985 to 94 (both inclusive):

$\Sigma X = 0$, $\Sigma Y = 322$, $\Sigma XY = 1550$, $\Sigma X^2 = 330$. Find out the trend values as well.

Ans. $\hat{Y} = 32.2 + 4.7X$; -10.1, -0.7, 8.7, 18.1, 27.5, 36.9, 46.3, 55.7, 65.1, 74.5.

17. Fit a linear trend to the following information for the years 1986 to 92 (both inclusive):

$\Sigma X = 0$, $\Sigma Y = 245$, $\Sigma X^2 = 28$, $\Sigma XY = 66$. Also compute the trend values.

Ans. $\hat{Y} = 35 + 2.4X$; 27.8, 30.2, 32.6, 35.0, 37.4, 39.8, 42.2.

18. If the straight line fitted to the data for the years 1996 to 2002 (both inclusive) with origin at 1999 and unit of measurement for X being one year is $\hat{Y} = 130 + 14X$. Find the trend values corresponding to the years 1996 to 2002. What would be the equation of the straight line if the origin is shifted to 1996.

Ans: 88, 102, 116, 130, 144, 158, 172; $\hat{Y} = 88 + 14u$

19. If the straight line fitted to the data for the years 1993 to 2000; both inclusive, with origin at the middle of 1996 and 1997 is $\hat{Y} = 34 + 2.6 X$, the unit of measurement for X being 1/2 year. Determine the trend values for the years 1993 to 2000. Also determine the straight line by shifting the origin to 1994.

Ans: 15.8, 21.0, 26.2, 31.4, 36.6, 41.8, 47.0, 52.2; $\hat{Y} = 21 + 5.2 u$

20. If the linear trend in the data for the years 1998 to 2003; both inclusive, with origin at the middle of 2000 and 2001 and unit of measurement for X being one year is $\hat{Y} = 40.5 + 5.8 X$. Determine the trend line by shifting the origin to the year 1998 and hence determine the trend values.

Ans: $\hat{Y} = 26 + 5.8 u$; 26.0, 31.8, 37.6, 43.4, 49.2, 55.0

21. Fit a second degree curve to the following results for the years 1993 to 2000 (both inclusive):

$$\Sigma X = 0, \Sigma Y = 220, \Sigma XY = 588, \Sigma X^2 = 168, \Sigma X^2Y = 5292, \Sigma X^3 = 0, \Sigma X^4 = 6216.$$

Ans: $\hat{Y} = 22.25 + 3.5X + 0.25 X^2$

22. Fit a second degree parabola to the following results for the years 1985 to 95 (both inclusive)

$$\Sigma X = 0, \Sigma X^2 = 110, \Sigma X^3 = 0, \Sigma X^4 = 1958, \Sigma Y = 410, \Sigma XY = 601, \Sigma X^2Y = 4587.$$

Also compute the trend values.

Ans. $\hat{Y} = 31.57 + 5.46 X + 0.57 X^2$; 18.52, 18.85, 20.32, 22.93, 26.68, 31.57, 37.60, 44.77, 53.08, 62.53, 73.12.

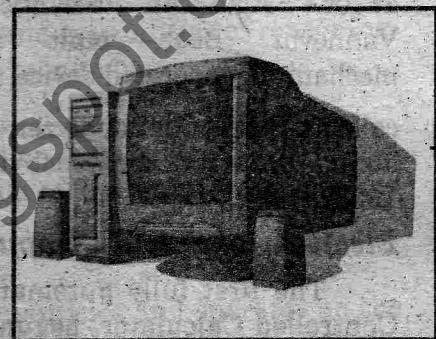
23. The fitted second degree parabola for the years 1996 to 2001; both inclusive, with origin at the middle of 1998 and 1999 is $\hat{Y} = 80 + 10X - 2X^2$, the unit of X being half year. What would be the equation of the second degree parabola by shifting the origin to the year 1996.

Ans: $\hat{Y} = -20 + 60u - 8u^2$ (1996 = 0)

ORIENTATION OF COMPUTERS

17.1 INTRODUCTION TO COMPUTERS

A *computer* is an electronic machine that accepts information, stores it until the information is needed, processes the information according to the instructions provided by the user, and finally returns the results to the user. The computer can store and manipulate large quantities of data at very high speed, but a computer cannot think. A computer makes decisions based on simple comparisons such



as one number being larger than another. Although the computer can help to solve a tremendous variety of problems, it is simply a machine. It cannot solve problems on its own.

17.1.1 COMPUTER CAPABILITIES AND ITS USES

Computer is a machine capable of performing tasks brilliantly. Following are some of the computer characteristics.

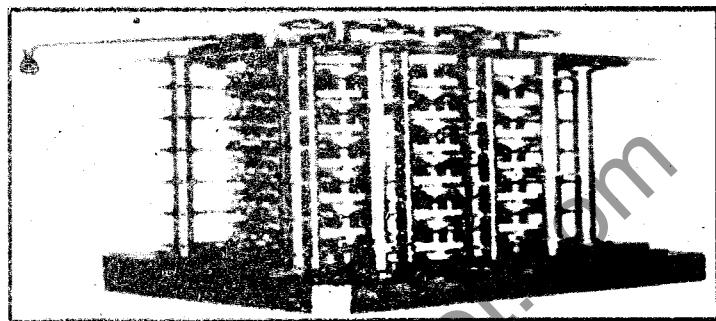
- (i) **Speed:** Computer is a very fast device. A powerful computer can perform 3 to 4 million arithmetic operations in a second.
- (ii) **Accuracy:** Computer always produces results 100% accurate, depending upon the input data and the instructions to carry out the given task.
- (iii) **Diligence:** Unlike other machines, computer can work for long hours without making any complaint. The longer time duration, never affect its working.
- (iv) **Versatility:** Computer is a versatile machine and can perform variety of tasks, like, Mathematical/ Statistical Data Manipulation, Word Processing, Spread Sheets, Data Bases, Graphics, Communication etc.

People from different fields are using computers in their professions due to its versatility, speed, accuracy and consistency and many other features. In today's world computers are being used in business, banking, stock exchange, education, medical and many other fields.

17.2 COMPUTER HISTORY

Early Development

Although the development of digital computers is rooted in the abacus and early mechanical calculating devices, Charles Babbage is credited with the design of the first modern computer, the "analytical engine," during the 1830s. American Scientist Vannevar Bush built a mechanically operated device,



Analytical Engine

called a differential analyzer, in 1930; it was the first general-purpose analog computer. John Atanasoff constructed the first semi electronic digital computing device in 1939.

Electromechanical Computing Machines

The first fully automatic calculator was the Mark I, or Automatic Sequence Controlled Calculator, begun in 1939 at Harvard by Howard Aiken; while the first all-purpose electronic digital computer, ENIAC (Electronic Numerical Integrator and Calculator), which used thousands of vacuum tubes, was completed in 1946 at the University of Pennsylvania. UNIVAC (UNIVersal Automatic Computer) became (1951) the first computer to handle both numeric and alphabetic data with equal facility; this was the first commercially available computer.



Computer Generations

From 1942 on wards the development of electronic computers are divided into five generations, depending upon the technologies used.

First-generation computers (1942-1959) utilized vacuum tubes and magnetic cores. These computers were very large in size, slow in speed, non reliable, and difficult to maintain. **Second-generation machines** (1959-1965) were come into existence with the advent of transistor technology. These computers were smaller, used less power, and could perform a million operations per second. They, in turn, were replaced by the **third-generation** (1965-1970) integrated-circuit machines. These were even smaller and were far more reliable machines. **Fourth generation computers** (1970-1980) were characterized by the development of the microprocessor and the evolution of increasingly smaller but powerful computers, such as the personal computer, which ushered in a period of rapid growth in the computer industry. **Fifth Generation** (1980-onward) is a generation of computers, to which scientists want to equip with thinking power and capabilities of reasoning, learning, drawing inferences and making decisions like human beings.

17.3 TYPES OF COMPUTER

Computers are divided into three types. The division is based upon the design and working of the computer which differs on the type of the input data and the form of its output.

17.3.1 ANALOG COMPUTER

An analog computer represents data as physical quantities and operates on the data by manipulating the quantities. It is designed to process data in which the variable quantities vary continuously; it translates the relationships between the variables of a problem into analogous relationships between electrical quantities, such as current and voltage, and solves the original problem by solving the equivalent problem, or analog, that is set up in its electrical circuits. Because of this feature, analog computers are especially useful in the simulation and evaluation of dynamic situations, such as the flight of a space capsule or the changing weather patterns over a certain area. Analog computers are commonly found in such forms as speedometers and watt-hour meters. These are also known as ***Special Purpose Computers***.

17.3.2 DIGITAL COMPUTER

A digital computer is designed to process data in numerical form; its circuits perform directly the mathematical operations of addition, subtraction, multiplication, and division. The numbers operated by a digital computer are expressed in the binary system.

The results of digital computers are more accurate, precise and repeatable than the results of analog computers. Modern digital computers are capable to store large amount of data and information and also compute data at very high speed. These computers are used in almost every field of life and also known as ***General Purpose Computers***.

17.3.3 HYBRID COMPUTER

A Hybrid computer is the combination of the best features of Analog and Digital computer, e.g. it possesses the faster speed like analog computer and

accuracy like digital computer. It can measure both in terms of physical as well as digital quantities. These are mainly used in specialized applications where both kind of information needed to be processed. These computers are used in the fields, like, air defense system, different laboratory equipment for medicine.

17.4 CLASSIFICATIONS OF COMPUTERS

The computers are classified into the following four main categories, depending upon the processing power, speed, size of main memory and other capabilities possessed by a computer.

- | | |
|-------------------------|----------------------|
| (i) Mainframe Computers | (ii) Minicomputers |
| (iii) Microcomputers | (iv) Super computers |

17.4.1 MAINFRAME COMPUTERS

Mainframe computers are very large, often filling an entire room. They can store enormous of information, can perform many tasks at the same time, can communicate with many users at the same time, and are very expensive. The price of a mainframe computer frequently runs into the millions of dollars. Mainframe computers usually have many terminals connected to them. These terminals look like small computers but they are only devices used to send and receive information from the actual computer using wires. Terminals can be located in the same room with the mainframe computer, but they can also be in different rooms, buildings, or cities. Large businesses, government agencies, and universities usually use this type of computer. Example of mainframe computers are Amdahl 580, Burroughs B 7800 Control Data CYBER 176 and IBM 4341.

17.4.2 MINICOMPUTERS

Minicomputers are much smaller than mainframe computers and they are also less expensive. These computers use integrated circuits. The cost of these computers can vary from a few thousand dollars to several hundred thousand dollars. They possess most of the features found on mainframe computers, but on a more limited scale. They can still have many terminals, but not as many as the mainframes. They can store a tremendous amount of information, but again usually not as much as the mainframe. Medium and small businesses typically use these computers. The examples of Minicomputers are PRIME 9755, VAX 8650, IBM System 36, etc.

17.4.3 MICROCOMPUTERS

Microcomputers are also known as *Personal Computers*. These computers are usually divided into desktop models and laptop models. They are terribly limited in what they can do when compared to the larger models discussed above because they can only be used by one person at a time, they are much slower than the larger computers, and they cannot store nearly as much information, but they are excellent when used in small businesses, homes, and school classrooms. These computers are inexpensive and easy to use. They have become an essential part of modern life. Examples of microcomputers are IBM PC, AT, PS\2 and Apple Macintosh, TRS - 80.

17.4.4 SUPER COMPUTERS

Super Computers are the most expensive computers. These process billions of instructions per second. Most people do not have a direct need for the speed and power of a supercomputer. The supercomputers are mainly used for tasks that require mammoth data manipulation, such as worldwide weather forecasting and weapons research. But now supercomputers are also moving toward the mainstream, for activities as varied as stock analysis, automobile design, special effects for movies, and even sophisticated artworks. Examples of super computers are Cray-1, Cray -2 and CYBER 205.

17.5 COMPUTER COMPONENTS

The computer is divided into two main components, namely, computer hardware and computer software.

17.6 COMPUTER HARDWARE

The electronic and mechanical components of a computer are known as computer hardware. These components can be physically handled. The function of these components is typically divided into three main categories: input, output, and storage. Keyboard, mouse, monitor, system unit are some of the common examples of computer hardware. The computer hardware is further sub divided into four main categories

- (i) Input Unit
- (ii) Central Processing Unit
- (iii) Secondary Storage
- (iv) Output Unit

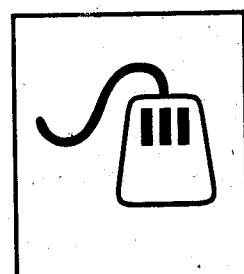
17.6.1 INPUT UNIT

Input Unit provides communication between the user and the computer. The input unit performs the following three functions.

- (a) At first, it accepts data from user.
- (b) The accepted data is then converted into coded data i.e. data in computer readable form.
- (c) The coded data is then moved to the system for processing.

Examples of input unit

- (1) **Mouse** is a pointing device designed to be gripped by one hand. It has a detection device (usually a ball) on the bottom that enables the user to control the motion of an on-screen pointer, or cursor, by moving the mouse on a flat surface. As the device moves across the surface, the cursor moves across the screen. To select items or choose commands on the screen, the user presses a button on the mouse.



Following are the important terms which are associated with the actions of mouse.

- (i) **Pointer:** Pointer is a symbol that moves on monitor's screen as the mouse is rolled over a flat surface.
- (ii) **Point:** To point means to set the pointer on a particular spot.

- (iii) **Click:** To press and to release left mouse button is known as click.
- (iv) **Double Click:** To press and to release left mouse button twice, with a quick action is known as double click.
- (v) **Right Click:** To press and release the right mouse button is known as right click.
- (vi) **Drag & Drop:** Drag and drop means to place the mouse pointer on an object, hold down the left mouse button and then release the button by taking the mouse pointer to another place on the screen.
- (2) **Keyboard** is a typewriter-like device that allows the user to type in text and commands to the computer. The keyboard is divided into four different groups.
 - (i) **Alphanumeric Keypad** consists of alphabet keys, number keys, punctuation keys, special character keys and space bar.
 - (ii) **Numeric Keypad** serves the dual purpose. One is to enter numeric data and the other is to move around the screen.
 - (iii) **Function Keypad** consists of 12 keys labeled F1, F2, F3.....F12. These keys perform specific functions.
 - (iv) **Screen Navigation and Editing Keys**, the group consists of arrow keys along with some special keys. These keys help to move around the screen and perform particular tasks. Page up, Page Down, Delete and Insert keys are some of the examples of screen navigation and editing keys.
- (3) **Light Pen** has a light sensitive tip that is used to draw directly on a computer's video screen or to select information on the screen by pressing a clip in the light pen or by pressing the light pen against the surface of the screen. The pen contains light sensors that identify which portion of the screen it is passed over.
- (4) **Joystick** is a pointing device composed of a lever that moves in multiple directions to navigate a cursor or other graphical object on a computer screen.

17.6.2 CENTRAL PROCESSING UNIT

Central Processing Unit is the most important unit of the computer system. It is the electronic brain of the computer. In addition to processing data, it controls the function of all the other components.

Main Components and Structure of the Central Processing Unit

The main components of the CPU are as under:

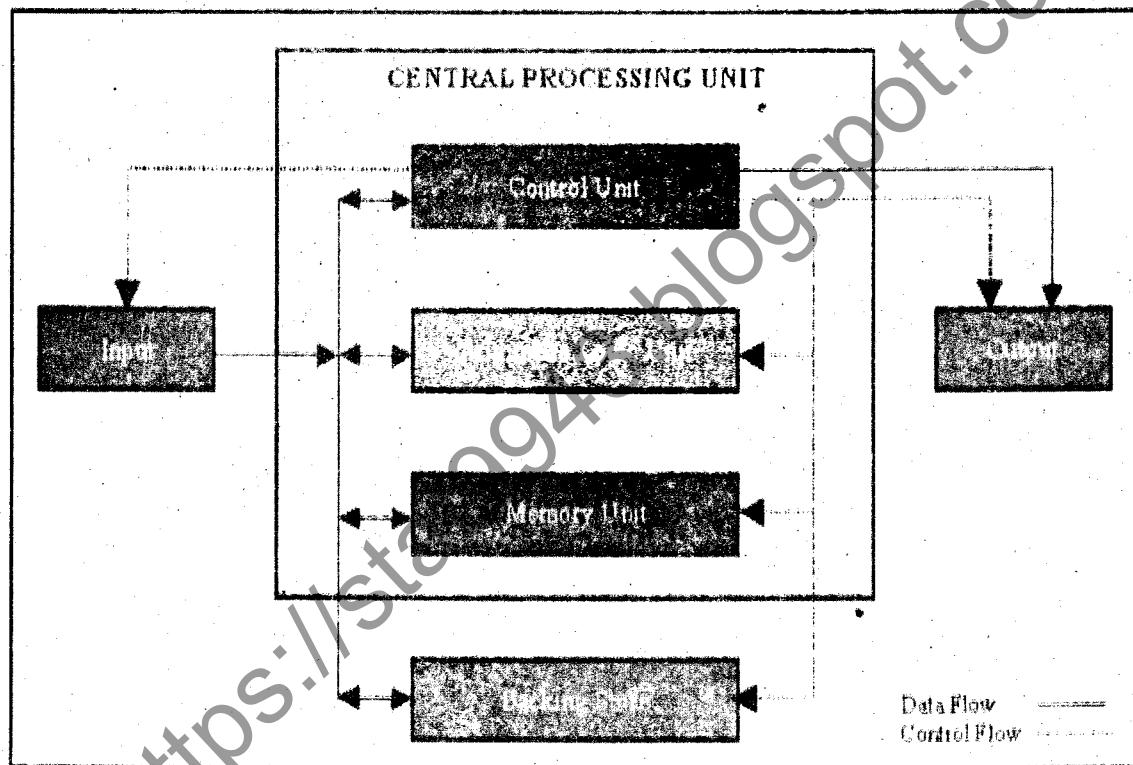
- (i) Control Unit
- (ii) Arithmetic Logic Unit
- (iii) Main Memory (It is closely associated with the CPU, in fact it is separated from it)

Working of the Central Processing Unit

- (i) It carries out instructions and tells the rest of the computer system what to do. This is done by the control unit of the CPU which sends command signals to the other components of the system.

- (ii) Performs arithmetic calculations and data manipulation, e.g. comparisons, sorting, combining, etc. The computer's calculator is a part of the CPU known as the arithmetic logic unit.
- (iii) Holds data and instructions which are in current use. These are kept in the main store or memory.

To understand how the whole system works, consider the diagram shown below. This diagram shows the basic components of a generalized CPU. An actual CPU may have these components or other with different names that provide the same functions.



Control Unit

The control unit controls and directs the operations of the entire computer system. The main features of control unit are discussed as under:

- (i) The control unit directs the entire computer system to carry out stored program instructions.
- (ii) The control unit communicates with both the arithmetic logic unit and main memory.
- (iii) The control unit uses the instruction contained in the Instruction Register to decide which circuits need to be activated.
- (iv) The control unit co-ordinates the activities of the other two units as well as all peripheral and auxiliary storage devices linked to the computer.
- (v) The control unit instructs the arithmetic logic unit which arithmetic operation or logical operation is to be performed.

Arithmetic Logic Unit

The arithmetic logic unit contains the circuits that perform arithmetic and logical operations. The main features of arithmetic logic unit are discussed as under:

- (i) The arithmetic logic unit executes arithmetic and logical operations.
- (ii) Arithmetic operations include addition, subtraction, multiplication and division.
- (iii) Logical operations compare numbers, letters and special characters.
- (iv) Comparison operations test for three conditions:
 - equal-to (=) condition in which two values are the same
 - less-than (<) condition in which one value is smaller than the other
 - greater-than (>) condition in which one value is larger than the other

The arithmetic logic unit also performs logic functions such as AND, OR and NOT.

Main Memory

The Main Memory is the part of the computer that holds data and instructions for processing. Although it is closely associated with the CPU, in actual it is separated from it. Memory associated with the CPU is also called primary storage, primary memory, main storage, internal storage and main memory.

When we load software from a floppy disk, hard disk or CD-ROM, it is stored in the main memory.

There are two types of computer memory inside the computer, RAM and ROM.

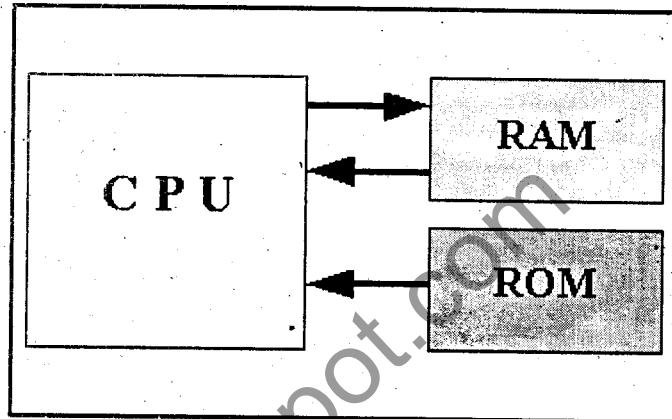
RAM

RAM stands for *Random Access Memory*. This is really the main store and is the place where the programs and software we load gets stored. When the central processing unit runs a program, it fetches the program instructions from the RAM and carries them out. If the central processing unit needs to store the results of calculations it can store them in RAM. The more RAM in your computer, the larger the programs you can run. When we switch a computer off, whatever is stored in the RAM gets erased. The following is a photo of a common RAM chip.



ROM

ROM stands for *Read Only Memory*. The CPU can only fetch or read instructions from read only memory (or ROM). ROM comes with instructions permanently stored inside and these instructions cannot be over-written by the computer's CPU. ROM memory is used for storing special sets of instructions which the computer needs.



when it starts up. When we switch the computer off, the contents of the ROM does not become erased but remains stored permanently. Therefore it is non-volatile. The following is a diagram showing the relationship between the central processing unit and the main memory (RAM and ROM).

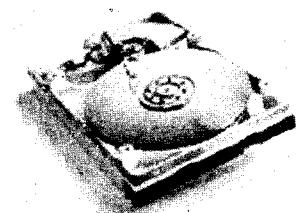
17.6.3 SECONDARY STORAGE

The secondary storage is a device which provides the opportunity to store and retrieve data and information according to the requirement of the user. The data and information stored on this device can be erased too, if required. The Secondary Storage is also called as auxiliary storage, external storage, backup storage or long term storage. The main features of the secondary storage are as follows:

- (i) It provides long term storage to data and programs.
- (ii) It provides additional memory to the computer to save data.
- (iii) It provide backup to main memory.
- (iv) It provides permanent storage so that electricity failure or switching off the computer does not affect data.

Some of the popular storage devices are discussed below:

Hard Disk: The hard disk (HD) is the main secondary storage device used to permanently store information and consists of one or more magnetic disks contained in a box. This is also called fixed disk and used for more storage i.e. it can store a huge amount of data and it has faster access speed.



Floppy Disk: A floppy disk is a removable storage device that reads and writes information magnetically onto floppy diskettes. There are two types of floppy diskettes. One is $5 \frac{1}{4}$ inch disk capable of storing 1.2 MB data and other is $3 \frac{1}{2}$ inch capable of storing 1.44 MB data. The $3 \frac{1}{2}$ inch diskettes are most commonly used diskettes and were introduced in 1987. These have a plastic exterior shell in order to protect the thin, flexible disk inside.

A hard disk is mounted inside the system unit and only removed for repairs or upgrades. The floppy disk provides removable storage, giving user the ability to take

their files with them. The drawback to the floppy diskette is that it only holds 1.44 MB of information, although very few PC's are without one. This has plenty of space for most text documents (Word & Excel files), but if file contains pictures, a floppy's capacity may be insufficient.

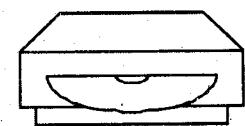
Memory Units:

The computer is a wonderful machine and can perform variety of task; its all due to that computer has a capability to store data/information into its memory. The capacity of computer memory and storage devices is expressed as bits or bytes. So, we can say that bits\bytes are measuring units for computer memory and storage.

Computers represent data digitally. They use binary digits, which are numbers using a base 2 number system rather than a decimal (or base 10) number system. A binary digit, commonly called a bit, has a value of either 0 (zero) or 1 (one). Eight bits are grouped together to represent a character--a letter, number, or special character. This group is called a byte. The terms character and byte mean the same thing.

One byte equals only one character, the storage devices must be capable of storing thousands, millions, or even billions of bytes. To describe these large capacities, the terms kilobyte(K), megabyte (M), and gigabyte (G)are used. A kilobyte equals approximately one thousand bytes, a megabyte equals approximately one million bytes, and a gigabyte equals approximately one billion bytes. (The actual number of bytes in a megabyte is slightly higher because computer storage amounts are actually measured in base 2 numbers.)

Compact Disk: A compact disk commonly known as CD is a secondary storage device that reads information stored on a compact disk. It is 4.75 inches in diameter and is capable of storing 650 MB. Floppy and hard disks are magnetic media; the compact disk is an optic media. Magnetism can simply fade away in time; however, the life span of optic media is counted in tens of years, which makes CD-ROM a very useful tool. CD-ROM drives can be housed inside the computer case (internal), or connected to the computer by a cable (exterior). CD ROM's are useful for installing programs and for running applications that install some of the files to the hard drive and execute the program by transferring the data from CD-ROM to memory, while the program is running.

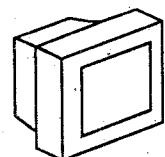


17.6.4 OUTPUT UNIT

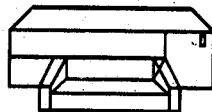
The devices which are used to get information or result from computer are called output devices. These are a communication link between a computer and user. The output can be of two types

- (i) Soft Copy
- (ii) Hard Copy

Soft Copy: The results appear on monitor screen are known as soft copy. This form of output is known as temporary output as it cannot be retained if the computer is turned off. Monitors and PC projectors are two common examples of soft copy output devices.



Hard Copy: The results printed on paper are known as hard copy. This form of output is known as permanent output as it retained even if the computer is turned off. Printer and plotters are the examples of hardcopy output devices.



17.7 COMPUTER SOFTWARE

Computer software is the set of detailed, step-by-step instructions (also called a program) which give a computer the capacity to perform a specific task. The computer software is mainly divided into the following three types

- (i) Programming Languages (ii) System Software (iii) Application Software

17.7.1 PROGRAMMING LANGUAGES

Software programs must be written in programming languages. Programmers (People qualified in programming language) write programs. Before 1952, the only available programming language was machine language, now called a low level language. A machine language is recognized by a given brand or design of computer processor. Machine language consists of nothing but the 0s and 1s with which the computer works. Machine language is difficult to learn, and early programs were few and short.

In 1952, a new low-level programming language called assembly language was introduced. In assembly language, programmers use short letter codes, such as ADD, that stands for addition.

In the 1960s, high-level programming languages emerged. With a high-level language, the programmer uses simple English words and familiar mathematical expressions. Examples of high-level languages are C, C++, PASCAL, and FORTRAN

17.7.2 SYSTEM SOFTWARE

System software are the programs that manage computer resources at a low level. System Software are categorized into the following two types:

- (i) Operating system
- (ii) Language processor

Operating System

Operating system is the most important program that runs on a computer. Every general-purpose computer must have an operating system to run other programs. Operating systems perform basic tasks, such as recognizing input from the keyboard, sending output to the display screen, keeping track of files and directories on the disk, and controlling peripheral devices such as disk drives and printers.

For large systems, the operating system has even greater responsibilities and powers. It makes sure that different programs and users running at the same time do not interfere with each other. The operating system is also responsible for security, ensuring that unauthorized users do not access the system.

Operating systems provide a software platform on top of which other programs, called application *programs*, can run. The

application programs must be written to run on top of a particular operating system. For PCs, the most popular operating systems are DOS, OS/2, and Windows.

Disk Operating System (Dos)

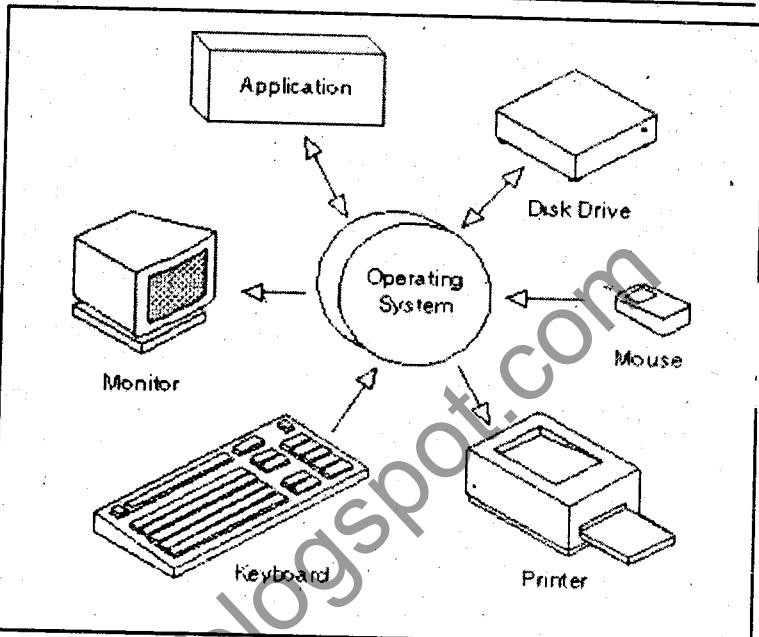
DOS stands for "Disk Operating System". It is a generic name for the basic IBM PC operating system. Several variants of DOS are available, including Microsoft's version of DOS (MS-DOS), IBM's version (PC-DOS), and several others. There's even a free version of DOS called Open DOS.

There are actually several levels to DOS. At the lowest level is the **BIOS** (Basic Input/Output System) which is responsible for managing devices like the keyboards and disk drives at the simplest possible level. The second layer provides a set of higher level services implemented using the low-level BIOS services. The third layer is the **command interpreter** (or shell). The shell's job is to display a **command prompt** on the screen to let user to type a command, then to read and interpret the typed command. The user can carry out any assignment e.g. writing/composing documents, graphics and games etc with in the environment of DOS.

Language Processor

Language processor also known as translator is a software that performs language translation. It has three types:

- (i) **Assembler** is a program that converts assembly language into machine language.
- (ii) A **compiler** is a computer program that translates a high-level programming language into machine language. The compiler usually converts the high-level language into assembly language first, and then translates the assembly language into machine language. The program fed into the compiler is called the source program; the generated machine language program is called the object program.



(iii) **Interpreter** is a program which executes source code by reading it one line at a time and performing each instruction immediately. An interpreter is different from a compiler, which does not execute the source code, but translates it into object code (machine language) which is stored in a file and executed later. Some programming languages must be interpreted; some can be both interpreted and compiled.

17.7.3 APPLICATION SOFTWARE

Application software is a complete, self-contained program that performs a specific function directly for the user. These are also known as packages. Some of the examples for Application Software are word processors, spread sheets, data bases, graphics and game packages etc.

17.8 BASIC IDEA OF WRITING AND RUNNING A COMPUTER PROGRAM

A program is an ordered set of computer instructions which enable a computer to perform a specific task. Programming is a process in which a program is written. This process is made up of following steps

- (i) Program design (ii) Program writing
- (iii) Testing and debugging
- (iv) Documentation, implementation and maintenance

17.8.1 PROGRAM DESIGN

Program design has two phases

- (a) **Function Definition:** Clearly define the data requirements (input & output) of the program and the purpose of the program.
- (b) **First Level of Refinement:** The data to be used is defined in detail and the main operations of the program are also defined.

17.8.2 PROGRAM WRITING

- (a) Describe each operation in detail. This can be done with the help of detailed program flowcharts or structure diagrams.
- (b) The next step is to code the program. Code means to write down the program in suitable programming language.
- (c) Then, feed the program into the computer. (This may be done as the program is coded)

17.8.3 TESTING AND DEBUGGING

- (a) In this phase the first step is to eliminate the compile time errors of the program.
- (b) Then the run time errors are eliminated.
- (c) The program is run with test data.
- (d) The test data provides a check on every possible situation.
- (e) The accurate results of running the program with the test data have been solved manually in advance.
- (f) The programmer uses various debugging aids to correct the program until the expected output is produced when the test data is used.

17.8.4 DOCUMENTATION, IMPLEMENTATION AND MAINTENANCE:

- (a) The documents are produced for programmers and users before the program is made available.
- (b) The new system of which program forms part then has to be implemented and maintained.

17.9 NUMBER SYSTEM

The organization and design of a computer is dependent upon the number systems, as, data is manipulated and stored in the computer in the coded numeral formats. To understand these coded numeral formats, it is necessary to have knowledge of different number systems.

17.9.1 DECIMAL NUMBER SYSTEM

Decimal number system is most widely used number system of the world. It consists of ten digits, ranging from 0 -9. The base of decimal system is considered to be 10. In this system, the successive positions to the left of the decimal point represent unit, tens, hundreds etc.

17.9.2 BINARY NUMBER SYSTEM

Binary number system suits to electronic machines. It is composed of two digits, 0 and 1. The base of this system is two.

17.9.3 OCTAL NUMBER SYSTEM

The octal number system is composed of eight digits, ranging from 0 to 7. The base of this system is eight.

17.9.4 HEXADECIMAL NUMBER SYSTEM

The hexadecimal number system is composed of sixteen digits. The first 10 digits of this system are 0 to 9 and the remaining six digits are denoted by A, B, C, D, E, F representing the decimal values 10, 11, 12, 13, 14, 15 respectively.

17.10 BINARY NUMBER SYSTEM AS A FOUNDATION OF COMPUTER PROGRAMMING

A computer is capable to carry our mathematical computations and it can also store data. The computers' tasks, like, mathematical computations, storing data are all about manipulating, numbers.

In our routine life we use the number system of ten digits 0,1,2,3,4,5,6,7,8,9. It's a base 10 number system and known as decimal number system. Computers don't use the ten digits of the decimal system for counting and arithmetic. Their CPU and memory are made up of millions of tiny switches that can be either ON or OFF. Two digits, 0 and 1, can be used to stand for the two states of ON and OFF. So, the computers could work with a number system based on two digits. This type of system is called a binary number system.

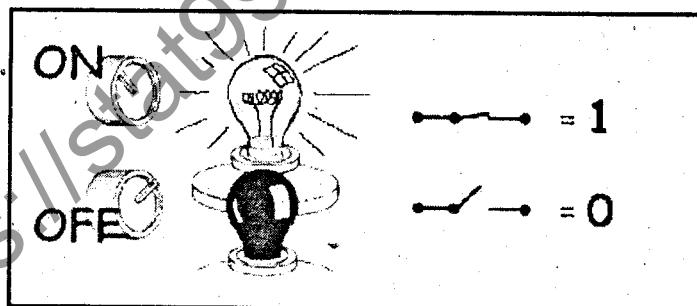
The decimal system is based on place, or location. That is, the place of each digit tells the value of that digit. For example, the number 17 has a 7 in the one's place and a 1 in the ten's place. In other words, 1 ten plus 7 ones equals 17. The number 138 has a 1

in the hundred's place, a 3 in the ten's place, and an 8 in the one's place. Written in numerals this is $(1 \times 100) + (3 \times 10) + (8 \times 1) = 138$.

The binary system works in exactly the same way, except that its place value is based on the number two. In the binary system, we have the one's place, the two's place, the four's place, the eight's place, the sixteen's place, and so on. Each place in the number represents two times (2^X 's) the place to its right.

Here's a comparison of decimal and binary numbers:

Decimal	Binary	Decimal	Binary
0	0	6	110
2	10	7	111
3	11	8	1000
4	100	9	1001
5	101	10	1010



Since the computer is really made up of tiny switches that can be either OFF or ON, binary numbers can be seen as a series of light switches. A 1 represents a switch that is ON, and a 0 means a switch that is OFF.

Numbers can become rather long in the binary system. For example, to show the number 10, we need four light switches, or four places. However, the real switches inside a computer are tiny and they are able to turn on and off very rapidly. The binary number system suits a computer extremely well and it's the main foundation of computer programming.

MULTIPLE - CHOICE QUESTIONS

1. Interpreter is a type of:
(a) language processor (b) application software
(c) storage device (d) computer hardware
2. One byte equals:
(a) 8 bits (b) 4 bits
(c) 6 bits (d) 12 bits
3. General purpose computers are also known as:
(a) hybrid computers (b) digital computers
(c) analog computers (d) super computers
4. PRIME 9755 is one of the examples of:
(a) minicomputers (b) super computers
(c) microcomputers (d) mainframe computers
5. The electronic and mechanical components of a computer are known as:
(a) computer software (b) computer hardware
(c) none of the above (d) both (a) and (b)
6. Drag and drop is a term associated with:
(a) mouse (b) keyboard
(c) printer (d) scanner
7. All the arithmetic and logical data manipulation is done by the:
(a) hard disk (b) arithmetic logic unit
(c) control unit (d) main memory
8. When we switch the computer off, the data vanishes which lies on:
(a) hard disk (b) compact disk
(c) RAM (d) floppy disk
9. The 3 ½ inches diskette can store data of the size:
(a) 1.44 MB (b) 1.2 MB
(c) 2.1 MB (d) 1.54 MB
10. The diameter of a compact disk is:
(a) 4.75 inches (b) 4.85 inches
(c) 4.65 inches (d) 4.55 inches
11. The time period of third generation of computers is:
(a) 1965-70 (b) 1980-onwards
(c) 1959-65 (d) 1942-65

12. The currently used data and instructions are held in:
- (a) main memory
 - (b) control unit
 - (c) arithmetic logic unit
 - (d) hard disk
13. The secondary storage is also known as:
- (a) long term storage
 - (b) backup storage
 - (c) none of the above
 - (d) both (a) and (b)
14. C++ is an example of:
- (a) high level language
 - (b) low level language
 - (c) assembly language
 - (d) none of the above
15. The analytical engine was invented in:
- (a) 1730s
 - (b) 1830s
 - (c) 1930s
 - (d) 1935s
16. A hybrid computer is the combination of the best features of:
- (a) analog and digital computers
 - (b) microcomputers and minicomputers
 - (c) mainframe and super computers
 - (d) digital and super computers
17. Super computers can process billion of instructions:
- (a) per second
 - (b) per micro second
 - (c) per minute
 - (d) per hour
18. Function keypad consists of:
- (a) 12 keys
 - (b) 6 keys
 - (c) 8 keys
 - (d) 14 keys
19. Joystick is an example of:
- (a) input devices
 - (b) output devices
 - (c) processing devices
 - (d) storage devices
20. Which of the following is not an arithmetic operation:
- (a) addition
 - (b) greater than
 - (c) subtraction
 - (d) multiplication
21. A binary digit is commonly called:
- (a) bit
 - (b) byte
 - (c) kilobyte
 - (d) gigabyte
22. Before 1952, the only available programming language was:
- (a) machine Language
 - (b) assembly Language
 - (c) C Language
 - (d) none of the above
23. The first commercially available computer was:
- (a) UNIVAC
 - (b) ENIAC
 - (c) Mark I
 - (d) analytical engine

24. First generation computers utilized:
- (a) vacuum tubes
 - (b) transistors
 - (c) integrated circuit
 - (d) none of the above
25. Program design is made up of two phases, including:
- (a) function design and first level of refinement
 - (b) function design and program writing
 - (c) testing and debugging
 - (d) documentation and implementation

Answers

1. (a)	2. (a)	3. (b)	4. (a)	5. (b)	6. (a)	7. (b)	8. (c)
9. (a)	10. (a)	11. (a)	12. (a)	13. (d)	14. (a)	15. (b)	16. (a)
17. (a)	18. (a)	19. (a)	20. (b)	21. (a)	22. (a)	23. (a)	24. (a)
25. (a)							

SHORT QUESTIONS

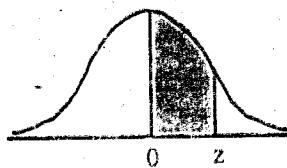
1. Differentiate between hardware and software.
2. Differentiate between control unit and arithmetic logic unit.
3. Differentiate between ram and rom.
4. Differentiate between main memory and secondary storage.
5. Differentiate between hard copy and soft copy.
6. Differentiate between system software and application software.
7. Differentiate between assembler and compiler.
8. Differentiate between decimal number system and binary number system.
9. Differentiate between floppy disk and compact disk.
10. Differentiate between low level languages and high level languages.
11. Differentiate between analog computers and digital computers.
12. Write down the different types of computers.
13. Write a short note on computer history.
14. Define the central processing unit.
15. Write a note on disk operating system.
16. What do you know about classification of computers?
17. What are the main components of central processing unit?
18. Write down different types of language processors.
19. What is meant by programming?

STATISTICAL TABLES

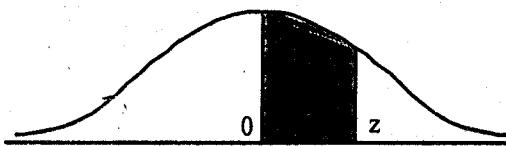
RANDOM NUMBERS

10 09 73 25 33	76 52 01 35 86	34 67 35 48 76	80 95 90 91 17	39 29 27 49 45
37 54 20 48 05	64 89 47 42 96	24 80 52 40 37	20 63 61 04 02	00 82 29 16 65
08 42 26 89 53	19 64 50 93 03	23 20 90 25 60	15 95 33 47 64	35 08 03 36 06
99 01 90 25 29	09 37 67 07 15	38 31 13 11 65	88 67 67 43 97	04 43 62 76 59
12 80 79 99 70	80 15 73 61 47	64 03 23 66 53	98 95 11 68 77	12 17 17 68 33
66 06 57 47 17	34 97 27 68 50	36 69 73 61 70	65 81 33 98 85	11 19 92 91 70
31 06 01 08 05	45 57 18 24 06	35 30 34 26 14	86 79 90 74 39	23 40 30 97 32
85 26 97 76 02	02 05 16 56 92	68 66 57 48 18	73 05 38 52 47	18 62 88 85 79
63 57 33 21 35	05 32 54 70 48	90 55 35 75 48	28 46 82 87 09	83 49 12 56 24
73 79 64 57 53	03 52 96 47 78	35 80 83 42 82	60 93 52 03 44	35 27 38 84 35
98 52 01 77 67	14 90 56 86 07	22 10 94 05 58	60 97 09 34 33	50 50 07 39 98
11 80 50 54 31	39 80 82 77 32	50 72 56 82 48	29 40 52 42 01	52 77 56 78 51
83 45 29 96 34	06 28 89 80 83	13 74 67 00 78	18 47 53 06 10	68 71 17 78 17
88 68 54 02 00	86 50 74 84 01	36 76 66 79 51	90 36 47 64 93	29 60 91 10 62
99 59 46 73 48	87 51 76 49 69	91 82 60 89 28	93 78 56 13 68	23 47 83 41 13
65 48 11 76 74	17 46 85 09 50	58 04 77 69 74	73 03 95 71 86	40 21 81 65 44
80 12 43 56 35	17 72 70 80 15	45 31 82 23 74	21 11 57 82 53	14 38 55 37 63
74 35 09 98 17	77 40 27 72 14	43 28 60 02 10	45 52 16 42 37	96 28 60 26 55
69 91 62 68 03	66 25 22 91 48	36 93 68 72 03	76 62 11 39 90	94 40 05 64 18
09 89 32 05 05	14 22 56 85 14	46 42 75 67 88	96 29 77 88 22	54 38 21 45 98
91 49 91 45 23	68 47 92 76 86	46 16 28 35 54	94 75 08 99 23	37 08 92 00 48
80 33 69 45 98	26 94 03 68 58	70 29 73 41 35	53 14 03 33 40	42 05 08 23 41
44 10 48 19 49	85 15 74 79 54	32 97 92 65 75	57 60 04 08 81	22 22 20 64 13
12 55 07 37 42	11 10 00 20 40	12 86 07 46 97	96 64 48 94 39	28 70 72 58 15
63 60 64 93 29	16 50 53 44 84	40 21 95 25 63	43 65 17 70 82	07 20 73 17 90
61 19 69 04 46	26 45 74 77 74	51 92 43 37 29	65 39 45 95 93	42 58 26 05 27
15 47 44 52 66	95 27 07 99 53	59 36 78 38 48	82 39 61 01 18	33 21 15 94 66
94 55 72 85 73	67 89 75 43 87	54 62 24 44 31	91 19 04 25 92	92 92 74 59 73
42 48 11 62 13	97 34 40 87 21	16 86 84 87 67	03 07 11 20 59	25 70 14 66 70
23 52 37 83 17	73 20 8 98 37	68 93 59 14 16	26 25 22 96 63	05 52 28 25 62
04 49 35 24 94	75 24 63 38 24	45 86 25 10 25	61 96 27 93 35	65 33 71 24 72
00 54 99 76 54	64 05 18 81 59	96 11 96 38 96	54 69 28 23 91	23 28 72 95 29
35 96 31 53 07	26 89 80 93 54	33 35 13 54 62	77 97 45 00 24	90 10 33 93 33
59 80 80 83 91	45 42 72 68 42	83 60 94 97 00	13 02 12 48 92	78 56 52 01 06
46 05 88 52 36	01 39 00 22 86	77 28 14 40 77	93 91 08 36 47	70 61 74 29 41
32 17 90 05 97	87 37 92 52 41	05 56 70 70 07	86 74 31 71 57	85 39 41 18 38
69 23 46 14 06	20 11 74 52 04	15 95 66 00 00	18 74 39 24 23	97 11 89 63 38
19 56 54 14 30	01 75 87 53 79	40 41 92 15 85	66 67 43 68 06	84 99 28 52 07
45 15 51 49 38	19 47 60 72 46	43 66 79 45 43	59 04 79 00 33	20 82 66 95 41
94 86 43 19 94	36 16 81 08 51	34 88 88 15 53	01 54 03 54 56	05 01 45 11 76

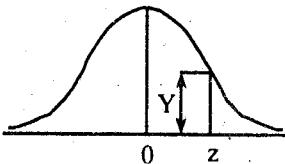
**AREAS
UNDER THE
STANDARD
NORMAL CURVE
from 0 to z
4-places of decimals)**



AREA UNDER THE STANDARD NORMAL CURVE



**Ordinates (Y)
of the
Standard
Normal Curve
at z**



z	0	1	2	3	4	5	6	7	8	9
0.0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	0.3970	0.3965	0.3961	0.3956	0.3951	0.3945	0.3939	0.3932	0.3925	0.3918
0.2	0.3910	0.3902	0.3894	0.3885	0.3876	0.3867	0.3857	0.3847	0.3836	0.3825
0.3	0.3814	0.3802	0.3790	0.3778	0.3765	0.3752	0.3739	0.3725	0.3712	0.3697
0.4	0.3683	0.3668	0.3653	0.3637	0.3621	0.3605	0.3589	0.3572	0.3555	0.3538
0.5	0.3521	0.3503	0.3485	0.3467	0.3448	0.3429	0.3410	0.3391	0.3372	0.3352
0.6	0.3332	0.3312	0.3292	0.3271	0.3251	0.3230	0.3209	0.3187	0.3166	0.3144
0.7	0.3123	0.3101	0.3079	0.3056	0.3034	0.3011	0.2989	0.2966	0.2943	0.2920
0.8	0.2897	0.2874	0.2850	0.2827	0.2803	0.2780	0.2756	0.2732	0.2709	0.2685
0.9	0.2661	0.2637	0.2613	0.2589	0.2565	0.2541	0.2516	0.2492	0.2468	0.2444
1.0	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
1.1	0.2179	0.2155	0.2131	0.2107	0.2083	0.2059	0.2036	0.2012	0.1989	0.1965
1.2	0.1942	0.1919	0.1895	0.1872	0.1849	0.1826	0.1804	0.1781	0.1758	0.1736
1.3	0.1714	0.1691	0.1669	0.1647	0.1626	0.1604	0.1582	0.1561	0.1539	0.1518
1.4	0.1497	0.1476	0.1456	0.1435	0.1415	0.1394	0.1374	0.1354	0.1334	0.1315
1.5	0.1295	0.1276	0.1257	0.1238	0.1219	0.1200	0.1182	0.1163	0.1145	0.1127
1.6	0.1109	0.1092	0.1074	0.1057	0.1040	0.1023	0.1006	0.0989	0.0973	0.0957
1.7	0.0940	0.0925	0.0909	0.0893	0.0878	0.0863	0.0848	0.0833	0.0818	0.0804
1.8	0.0790	0.0775	0.0761	0.0748	0.0734	0.0721	0.0707	0.0694	0.0681	0.0669
1.9	0.0656	0.0644	0.0632	0.0620	0.0608	0.0596	0.0584	0.0573	0.0562	0.0551
2.0	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0.0440	0.0431	0.0422	0.0413	0.0404	0.0396	0.0387	0.0379	0.0371	0.0363
2.2	0.0355	0.0347	0.0339	0.0332	0.0325	0.0317	0.0310	0.0303	0.0297	0.0290
2.3	0.0283	0.0277	0.0270	0.0264	0.0258	0.0252	0.0246	0.0241	0.0235	0.0229
2.4	0.0224	0.0219	0.0213	0.0208	0.0203	0.0198	0.0194	0.0189	0.0184	0.0180
2.5	0.0175	0.0171	0.0167	0.0163	0.0158	0.0154	0.0151	0.0147	0.0143	0.0139
2.6	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.7	0.0104	0.0101	0.0099	0.0096	0.0093	0.0091	0.0088	0.0086	0.0084	0.0081
2.8	0.0079	0.0077	0.0075	0.0073	0.0071	0.0069	0.0067	0.0065	0.0063	0.0061
2.9	0.0060	0.0058	0.0056	0.0055	0.0053	0.0051	0.0050	0.0048	0.0047	0.0046
3.0	0.0044	0.0043	0.0042	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0025
3.2	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020	0.0019	0.0018	0.0018
3.3	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013
3.4	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009
3.5	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006
3.6	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004
3.7	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003
3.8	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002
3.9	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001

OUR PUBLICATIONS ON STATISTICS

Basic Statistics Part-I & Part-II

(For Intermediate Classes)

Solution Basic Statistics Part-I & Part-II

(For Intermediate Classes)

Solved Practical Problems Part-I & Part-II

(For Intermediate Classes)

A Hand Book of Objective Statistics

(For Intermediate Classes)

A Hand Book of Statistics

(For Intermediate Classes)

Ideal Practical Note Book Part-I & Part-II

(For Intermediate Classes)

Ideal Practical Note Book

(For B.A. / B.Sc. & M.A. / M.Sc.)

Basic Statistics Part-I & Part-II

Federal Board (For Intermediate Classes)

Solution Basic Statistics Part-I & Part-II

Federal Board (For Intermediate Classes)

Solved Practical Problems Part-I & Part-II

Federal Board (For Intermediate Classes)

A Hand Book of Objective Statistics

Federal Board (For Intermediate Classes)

Probability Distributions For M.Sc

Order Statistics For M.Sc

**MAJEEED
BOOK DEPOT**

URDU BAZAR, LAHORE. Ph. 37311484-37355187

AMIN PUR BAZAR, FAISALABAD. Ph. 647841

6th ROAD, RAWALPINDI. Ph. 4423948