

# **Business Statistics**

**MBA HA-103**

**DIRECTORATE OF DISTANCE EDUCATION**

**SWAMI VIVEKANAND**

**SUBHARTI UNIVERSITY**

Meerut (National Capital Region Delhi)



---

# CONTENTS

---

Syllabus ..... (ix)

## 1. ROLE OF STATISTICS AND MEASURES OF CENTRAL TENDENCY ..... 1-38

1.1.	Introduction .....	2
1.2.	Applications of Inferential Statistics .....	2
1.3.	Measures of Central Tendency .....	2
1.4.	Meaning of Central Tendency .....	3
1.5.	Requisites of a Good Average .....	3
1.6.	Types of Measures of Central Tendency (Averages) .....	3

### I. ARITHMETIC MEAN (A.M.)

1.7.	Definition .....	3
1.8.	Step Deviation Method .....	6
1.9.	A.M. of Combined Group .....	8
1.10.	Weighted A.M. ....	9
1.11.	Mathematical Properties of A.M. ....	10

### II. GEOMETRIC MEAN (G.M.)

1.12.	Definition .....	12
1.13.	G.M. of Combined Group .....	14
1.14.	Averaging of Percentages .....	15
1.15.	Weighted G.M. ....	16

### III. HARMONIC MEAN (H.M.)

1.16.	Definition .....	18
1.17.	H.M. of Combined Group .....	20
1.18.	Weighted H.M. ....	20

### IV. MEDIAN

1.19.	Definition .....	22
-------	------------------	----

### V. MODE

1.20.	Definition .....	29
1.21.	Mode by Inspection .....	30
1.22.	Mode by Grouping .....	30
1.23.	Empirical Mode .....	31
1.24.	Mode in Case of Classes of Unequal Widths .....	34
1.25.	Summary .....	37
1.26.	Review Exercises .....	38

## 2. MEASURES OF DISPERSION ..... 39-61

2.1.	Introduction .....	39
2.2.	Requisites of a Good Measure of Dispersion .....	40
2.3.	Methods of Measuring Dispersion .....	40

## **I. RANGE**

2.4.	Definition .....	40
------	------------------	----

## **II. QUARTILE DEVIATION (Q.D.)**

2.5.	Inadequacy of Range .....	42
2.6.	Definition .....	43

## **III. MEAN DEVIATION (M.D.)**

2.7.	Definition .....	46
2.8.	Coefficient of M.D. ....	47
2.9.	Short-cut Method for M.D. ....	49

## **IV. STANDARD DEVIATION (S.D.)**

2.10.	Definition .....	52
2.11.	Coefficient of S.D., C.V., Variance .....	52
2.12.	Short-cut Method for S.D. ....	54
2.13.	Relation between Measures of Dispersion .....	59
2.14.	Summary .....	61
2.15.	Review Exercises .....	61

## **3. SKEWNESS ..... 62-75**

3.1.	Introduction .....	62
3.2.	Meaning .....	62
3.3.	Tests of Skewness .....	63
3.4.	Methods of Measuring Skewness .....	63
3.5.	Karl Pearson's Method .....	64
3.6.	Bowley's Method .....	67
3.7.	Kelly's Method .....	71
3.8.	Method of Moments .....	73
3.9.	Summary .....	75
3.10.	Review Exercises .....	75

## **4. KURTOSIS ..... 76-81**

4.1.	Introduction .....	76
4.2.	Definitions .....	76
4.3.	Measure of Kurtosis .....	77
4.4.	Summary .....	80
4.5.	Review Exercises .....	80

## **5. ANALYSIS OF TIME SERIES ..... 82-108**

5.1.	Introduction .....	82
5.2.	Meaning .....	83
5.3.	Components of Time Series .....	83
5.4.	Secular Trend or Long-term Variations .....	83
5.5.	Seasonal Variations .....	84
5.6.	Cyclical Variations .....	84
5.7.	Irregular Variations .....	85
5.8.	Additive and Multiplicative Models of Decomposition of Time Series .....	86
5.9.	Determination of Trend .....	86
5.10.	Free Hand Graphic Method .....	87
5.11.	Semi-Average Method .....	88

5.12.	Moving Average Method .....	90
5.13.	Least Squares Method .....	95
5.14.	Linear Trend .....	95
5.15.	Non-linear Trend (Parabolic) .....	101
5.16.	Non-linear Trend (Exponential) .....	104
5.17.	Summary .....	106
5.18.	Review Exercises .....	107
<b>6.</b>	<b>INDEX NUMBERS .....</b>	<b>109-145</b>
6.1.	Introduction .....	109
6.2.	Definition and Characteristics of Index Numbers .....	109
6.3.	Uses of Constructing Index Numbers .....	109
6.4.	Types of Index Numbers .....	110
<b>I. PRICE INDEX NUMBERS</b>		
6.5.	Methods .....	110
6.6.	Simple Aggregative Method .....	110
6.7.	Simple Average of Price Relatives Method .....	112
6.8.	Laspeyre's Method .....	117
6.9.	Paasche's Method .....	117
6.10.	Dorbish and Bowley's Method .....	117
6.11.	Fisher's Method .....	118
6.12.	Marshall Edgeworth's Method .....	118
6.13.	Kelly's Method .....	118
6.14.	Weighted Average of Price Relatives Method .....	119
6.15.	Chain Base Method .....	123
<b>II. QUALITY INDEX NUMBERS</b>		
6.16.	Methods .....	126
6.17.	Index Numbers of Industrial Production .....	128
<b>III. VALUE INDEX NUMBERS</b>		
6.18.	Simple Aggregative Method .....	130
6.19.	Mean of Index Numbers .....	131
<b>IV. TESTS OF ADEQUACY OF INDEX NUMBER FORMULAE</b>		
6.20.	Meaning .....	133
6.21.	Unit Test (U.T.) .....	134
6.22.	Time Reversal Test (T.R.T.) .....	134
6.23.	Factor Reversal Test (F.R.T.) .....	136
6.24.	Circular Test (C.T.) .....	136
<b>V. CONSUMER PRICE INDEX NUMBERS (C.P.I.)</b>		
6.25.	Meaning .....	138
6.26.	Significance of C.P.I. ....	139
6.27.	Assumptions .....	139
6.28.	Procedure .....	139
6.29.	Methods .....	140
6.30.	Aggregate Expenditure Method .....	140
6.31.	Family Budget Method .....	141
6.32.	Summary .....	144
6.33.	Review Exercises .....	145



<b>7.</b>	<b>MEASURES OF CORRELATION .....</b>	<b>146-172</b>
7.1.	Introduction .....	146
7.2.	Definition .....	147
7.3.	Correlation and Causation .....	147
7.4.	Positive and Negative Correlation .....	148
7.5.	Linear and Non-linear Correlation .....	148
7.6.	Simple, Multiple and Partial Correlation .....	149
	<b>I. KARL PEARSON'S METHOD</b>	
7.7.	Definition .....	150
7.8.	Alternative Form of 'R' .....	152
7.9.	Step Deviation Method .....	158
	<b>II. SPEARMAN'S RANK CORRELATIONS METHOD</b>	
7.10.	Meaning .....	163
7.11.	Case I. Non-Repeated Ranks .....	164
7.12.	Case II. Repeated Ranks .....	167
7.13.	Summary .....	171
7.14.	Review Exercises .....	171
<b>8.</b>	<b>REGRESSION ANALYSIS .....</b>	<b>173-201</b>
8.1.	Introduction .....	173
8.2.	Meaning .....	173
8.3.	Uses of Regression Analysis .....	174
8.4.	Types of Regression .....	174
8.5.	Regression Lines .....	174
8.6.	Regression Equations .....	175
8.7.	Step Deviation Method .....	185
8.8.	Regression Lines for Grouped Data .....	191
8.9.	Properties of Regression Coefficients and Regression Lines .....	194
8.10.	Summary .....	200
8.11.	Review Exercises .....	201
<b>9.</b>	<b>PROBABILITY .....</b>	<b>202-242</b>
9.1.	Introduction .....	202
9.2.	Random Experiment .....	203
9.3.	Sample Space .....	203
9.4.	Tree Diagram .....	203
9.5.	Event .....	204
9.6.	Algebra of Events .....	204
9.7.	Equality Likely Outcomes .....	205
9.8.	Exhaustive Outcomes .....	205
9.9.	Three Approaches of Probability .....	206
9.10.	Classical Approach of Probability .....	206
9.11.	'Odds in Favour' and 'Odds Against' an Event .....	206
9.12.	Mutually Exclusive Events .....	211
9.13.	Addition Theorem (For Mutually Exclusive Events) .....	211
9.14.	Addition Theorem (General) .....	213
9.15.	Conditional Probability .....	216
9.16.	Independent Events .....	220
9.17.	Dependent Events .....	221

9.18. Independent Experiments .....	223
9.19. Multiplication Theorem .....	223
9.20. Total Probability Rule .....	231

### I. BAYE'S THEOREM

9.21. Motivation .....	235
9.22. Criticism of Classical Approach of Probability .....	238
9.23. Empirical Approach of Probability .....	239
9.24. Subjective Approach of Probability .....	240
9.25. Summary .....	240
9.26. Review Exercises .....	240

## 10. PROBABILITY DISTRIBUTIONS (Binomial, Poisson, Normal Distributions) ..... 243-284

10.1. Introduction .....	243
10.2. Empirical Distribution .....	243

### I. BINOMIAL DISTRIBUTION

10.3. Introduction .....	243
10.4. Conditions .....	244
10.5. Binomial Variable .....	244
10.6. Binomial Probability Function .....	244
10.7. Binomial Frequency Distribution .....	245

### II. PROPERTY OF BINOMIAL DISTRIBUTION

10.8. The Shape of B.D. ....	250
10.9. The Limiting Case of B.D. ....	252
10.10. Mean of B.D. ....	252
10.11. Variance and S.D. of B.D. ....	253
10.12. $\gamma_1$ and $\gamma_2$ of B.D. ....	253
10.13. Recurrence Formula for B.D. ....	254
10.14. Fitting of a Binomial Distribution .....	255

### III. POISSON DISTRIBUTION

10.15. Introduction .....	259
10.16. Conditions .....	259
10.17. Poisson Variable .....	259
10.18. Poisson Probability Function .....	259
10.19. Poisson Frequency Distribution .....	260

### IV. PROPERTY OF POISSON DISTRIBUTION

10.20. The Shape of P.D. ....	264
10.21. Special Usefulness of P.D. ....	264
10.22. Mean of P.D. ....	265
10.23. Variance and S.D. of P.D. ....	265
10.24. $\gamma_1$ and $\gamma_2$ of P.D. ....	266
10.25. Recurrence Formula for P.D. ....	266
10.26. Fitting of a Poisson Distribution .....	267

### V. NORMAL DISTRIBUTION

10.27. Introduction .....	271
10.28. Probability Function of Continuous Random Variable .....	271

10.29. Normal Distribution .....	272
10.30. Definition .....	272
10.31. Standard Normal Distribution .....	272
10.32. Area Under Normal Curve .....	273
10.33. Table of Area Under Standard Normal Curve .....	273
10.34. Properties of Normal Distribution .....	274
10.35. Fitting of a Normal Distribution .....	282
10.36. Summary .....	283
10.37. Review Exercises .....	284
<b>11. ESTIMATION THEORY AND HYPOTHESIS TESTING .....</b>	<b>285-322</b>
11.1. Introduction .....	285
11.2. Null Hypothesis and Alternative Hypothesis .....	286
11.3. Level of Significance and Confidence Limits .....	286
11.4. Type I Error and Type II Error .....	287
11.5. Power of the Test .....	288
<b>I. TEST OF SIGNIFICANCE FOR SMALL SAMPLES</b>	
11.6. Student's $t$ -Test .....	288
11.7. Assumptions for Student's $t$ -Test .....	288
11.8. Degree of Freedom .....	288
11.9. Test for Single Mean .....	288
11.10. $t$ -Test for Difference of Means .....	291
11.11. Paired $t$ -Test for Difference of Means .....	291
11.12. F-Test .....	295
11.13. Properties of F-Distribution .....	296
11.14. Procedure to F-Test .....	296
11.15. Critical Values of F-Distribution .....	297
<b>II. TEST OF SIGNIFICANCE FOR LARGE SAMPLES</b>	
11.16. Test of Significance for Proportion .....	300
11.17. Test of Significance for Single Mean .....	305
11.18. Test of Significance for Difference of Means .....	308
11.19. Chi-Square Test .....	313
11.20. Chi-Square Test to Test the Goodness of Fit .....	313
11.21. Chi-Square Test to Test the Independence of Attributes .....	314
11.22. Conditions for $\chi^2$ Test .....	315
11.23. Uses of $\chi^2$ Test .....	316
11.24. Summary .....	321
11.25. Review Exercises .....	321

# 1. ROLE OF STATISTICS AND MEASURES OF CENTRAL TENDENCY

## STRUCTURE

- 1.1. Introduction
- 1.2. Applications of Inferential Statistics
- 1.3. Measures of Central Tendency
- 1.4. Meaning of Central Tendency
- 1.5. Requisites of a Good Average
- 1.6. Types of Measures of Central Tendency (Averages)

### I. Arithmetic Mean (A.M.)

- 1.7. Definition
- 1.8. Step Deviation Method
- 1.9. A.M. of Combined Group
- 1.10. Weighted A.M.
- 1.11. Mathematical Properties of A.M.

### II. Geometric Mean (G.M.)

- 1.12. Definition
- 1.13. G.M. of Combined Group
- 1.14. Averaging of Percentages
- 1.15. Weighted G.M.

### III. Harmonic Mean (H.M.)

- 1.16. Definition
- 1.17. H.M. of Combined Group
- 1.18. Weighted H.M.

### IV. Median

- 1.19. Definition

### V. Mode

- 1.20. Definition
- 1.21. Mode by Inspection
- 1.22. Mode by Grouping
- 1.23. Empirical Mode
- 1.24. Mode in Case of Classes of Unequal Widths
- 1.25. Summary
- 1.26. Review Exercises



## NOTES

---

## 1.1. INTRODUCTION

---

In ancient times, the use of statistics was very much limited and is just confined to the collection of data regarding manpower, agricultural land and its production, taxable property of the people etc. But as the time passed, the utility of this subject increased manifold. Many researches were conducted in this field and with the result of this it started growing as a separate subject of study. Many experts in the field of mathematics and economics contributed toward the development of this subject. The word 'Statistics' which was once used in the sense of just collection of data is now considered as a full fledged subject. The knowledge of this subject is used for taking decisions in the midst of uncertainty.

---

## 1.2. APPLICATIONS OF INFERENCE STATISTICS

---

The part of the subject statistics which deals with the analysis of a given group and drawing conclusions about a larger group is called **inferential statistics**. For studying data regarding a group of individuals or objects, such as heights, weights, income, expenditure of persons in a locality or number of defective and non-defective articles produced in a factory, it is generally impracticable to collect and study data regarding the entire group. Instead of examining the entire group, we concentrate on a small part of the group called a **sample**. If this sample happen to be a true representative of the entire group, called **population**, important conclusions can be drawn from the analysis of the sample. The conditions under which the conclusions for samples can be considered valid for the corresponding populations are studied in inferential statistics. Since such conclusions cannot be absolutely certain, the language of probability is often used in stating conclusions. Theoretical distributions are also needed in inferential statistics. In the present course, we shall be studying probability and theoretical distributions. Binomial, Poisson and Normal. Inferential statistics is also known as **inductive statistics**.

---

## 1.3. MEASURES OF CENTRAL TENDENCY

---

Suppose we have the data regarding the marks obtained by all the students of a class and we are to give an impression about the performance of students, to someone. It would not be desirable rather impracticable to tell him the marks obtained by all the students of the class. Perhaps, it may not be possible for him to gather any impression about the standard of students of that class. Similarly suppose we intend to compare the wage distribution of workers in two sugar factories and to decide as to which factory is paying more to individual workers than the other. In this case also, if we proceed with comparing the wages of workers of one factory with that of the other on individual basis, we may not be able to get any "thing". Even this type of comparison may not be possible if the number of workers in two factories are different.

## 1.4. MEANING OF CENTRAL TENDENCY

In fact, such type of problems can be easily dealt with, if we could find a single value of the variable which may be considered as a representative of the entire data. This type of representative which help in describing the characteristics of the entire data is called an *average* of the data. The individual values of the variable usually cluster around it. An average is also called a *measure of central tendency*, because it tends to lie centrally with the values of the variable arranged according to magnitude. Thus, we see that an *average* or a *measure of central tendency* of a statistical data is that single value of the variable which represents the entire data.

### NOTES

## 1.5. REQUISITES OF A GOOD AVERAGE

1. It should be easy to understand.
2. It should be simple to compute.
3. It should be well-defined in the sense that it is defined algebraically and should not depend upon personal bias.
4. It should be based on all the items.
5. It should not be unduly affected by extreme items in the series.
6. It should be capable of further algebraic treatment. For example, if we are given the averages of some groups, then we should be able to find the average of all the items taken together.
7. It should have sampling stability. By this we mean that the averages of different samples, drawn from the same population, should not vary significantly. Though it cannot be claimed that all the samples would have exactly the same average, but we expect that the values of the averages, should not vary significantly.

## 1.6. TYPES OF MEASURES OF CENTRAL TENDENCY (Averages)

- |                           |                           |
|---------------------------|---------------------------|
| I. Arithmetic Mean (A.M.) | II. Geometric Mean (G.M.) |
| III. Harmonic Mean (H.M.) | IV. Median                |
| V. Mode.                  |                           |

### I. ARITHMETIC MEAN (A.M.)

## 1.7. DEFINITION

This is the most popular and widely used measure of central tendency. The popularity of this average can be judged from the fact that it is generally referred to as 'mean'. The **arithmetic mean** of a statistical data is defined as the quotient of the sum of all the values of the variable by the total number of items and is generally denoted by  $\bar{x}$ .

∴ (a) For an individual series, the A.M. is given by

$$\text{A.M.} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} \text{ or more briefly as } \frac{\Sigma x}{n}$$

## NOTES

i.e.,

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

where  $x_1, x_2, \dots, x_n$  are the values of the variable, under consideration.

(b) For a frequency distribution,

$$\text{A.M.} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} = \frac{\Sigma f x}{\Sigma f} = \frac{\Sigma f x}{N}$$

i.e.,

$$\bar{x} = \frac{\Sigma f x}{N}$$

where  $f_i$  is the frequency of  $x_i$  ( $1 \leq i \leq n$ ). For simplicity,  $\Sigma f$ , i.e., the total number of items is denoted by  $N$ .

When the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable ( $x$ ).

## WORKING RULES TO FIND A.M.

**Rule I.** In case of an individual series, first find the sum of all the items. In the second step, divide this sum by  $n$ , total number of items. This gives the value of  $\bar{x}$ .

**Rule II.** In case of a frequency distribution, find the products ( $fx$ ) of frequencies and value of items. In the second step, find the sum ( $\Sigma fx$ ) of these products. Divide this sum by the sum ( $N$ ) of all frequencies. This gives the value of  $\bar{x}$ .

**Rule III.** If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.

**Example 1.1.** Find the A.M. of the following data:

Roll No.	1	2	3	4	5	6	7	8
Marks in Maths	12	8	6	9	7	8	7	14

**Solution.** Let the variable 'marks in maths' be denoted by  $x$ .

$$\therefore \bar{x} = \frac{\text{Sum of values of } x}{\text{Number of items}} = \frac{12 + 8 + 6 + 9 + 7 + 8 + 7 + 14}{8} = \frac{71}{8} = 8.875 \text{ marks.}$$

**Example 1.2.** The A.M. of 9 items is 15. If one more item is added to this series, the A.M. becomes 16. Find the value of the 10th item.

**Solution.** Let the values of 9 items be  $x_1, x_2, \dots, x_9$ .

$$\therefore 15 = \frac{x_1 + x_2 + \dots + x_9}{9}$$

$$\therefore x_1 + x_2 + \dots + x_9 = 15 \times 9 = 135$$

Let  $x_{10}$  be the 10th item.

## NOTES

∴ A.M. of  $x_1, x_2, \dots, x_9, x_{10}$  is 16

$$\therefore 16 = \frac{x_1 + x_2 + \dots + x_9 + x_{10}}{10}$$

$$\therefore x_1 + x_2 + \dots + x_9 + x_{10} = 160$$

$$\therefore 135 + x_{10} = 160$$

$$\therefore x_{10} = 160 - 135 = 25.$$

**Example 1.3.** (a) The marks obtained by 20 students in a test were:

13, 17, 11, 5, 18, 16, 11, 14, 13, 12, 18, 11, 9, 6, 8, 17, 21, 22, 7, 6.

Find the mean marks per student.

(b) If extra 5 marks are given to each student, show that the mean marks are also increased by 5 marks.

**Solution.** (a) Mean marks =  $\frac{\text{Sum of marks obtained by 20 students}}{20}$

$$= \frac{13 + 17 + 11 + 5 + 18 + 16 + 11 + 14 + 13 + 12 + 18 + 11 + 9 + 6 + 8 + 17 + 21 + 22 + 7 + 6}{20} = \frac{255}{20} = 12.75.$$

(b) New marks are:

$$\begin{array}{llll} 13 + 5 = 18, & 17 + 5 = 22, & 11 + 5 = 16, & 5 + 5 = 10, \\ 18 + 5 = 23, & 16 + 5 = 21, & 11 + 5 = 16, & 14 + 5 = 19, \\ 13 + 5 = 18, & 12 + 5 = 17, & 18 + 5 = 23, & 11 + 5 = 16, \\ 9 + 5 = 14, & 6 + 5 = 11, & 8 + 5 = 13, & 17 + 5 = 22, \\ 21 + 5 = 26, & 22 + 5 = 27, & 7 + 5 = 12, & 6 + 5 = 11. \end{array}$$

∴ New mean marks

$$= \frac{18 + 22 + 16 + 10 + 23 + 21 + 16 + 19 + 18 + 17 + 23 + 16 + 14 + 11 + 13 + 22 + 26 + 27 + 12 + 11}{20}$$

$$= \frac{355}{20} = 17.75 = 12.75 + 5 = \text{old mean marks} + 5.$$

**Example 1.4.** Calculate the A.M. for the following data:

Marks	0—10	10—30	30—40	40—50	50—80	80—100
No. of students	5	7	15	8	3	2

**Solution.**

**Calculation of A.M.**

Marks	No. of students $f$	Mid-points of classes $x$	$fx$
0—10	5	5	25
10—30	7	20	140
30—40	15	35	525
40—50	8	45	360
50—80	3	65	195
80—100	2	90	180
	$N = 40$		$\Sigma fx = 1425$

$$\therefore \bar{x} = \frac{\Sigma fx}{N} = \frac{1425}{40} = 35.625 \text{ marks.}$$



## 1.8. STEP DEVIATION METHOD

### NOTES

When the values of the variable ( $x$ ) and their frequencies ( $f$ ) are large, the calculation of A.M. may become quite tedious. The calculation work can be reduced considerably by taking *step deviations* of the values of the variable.

Let  $A$  be any number, called **assumed mean**, then  $d = x - A$  are called the **deviations** of the values of  $x$ , from  $A$ .

If the values of  $x$  are  $x_1, x_2, \dots, x_n$ , then the values of deviations are  $d_1 = x_1 - A, d_2 = x_2 - A, \dots, d_n = x_n - A$ . We define  $u = \frac{x - A}{h}$ , where  $h$  is some suitable

common factor in the deviations of values of  $x$  from  $A$ . The definition of ' $u$ ' is meaningful, because at least  $h = 1$  is a common factor for all the values of the deviations. The

different values of  $u = \frac{x - A}{h}$  are called the **step deviations** of the corresponding values of  $x$ . In this case, the values of the step deviations are

$$u_1 = \frac{x_1 - A}{h}, u_2 = \frac{x_2 - A}{h}, \dots, u_n = \frac{x_n - A}{h}.$$

$$\therefore \text{For } 1 \leq i \leq n, \quad u_i = \frac{x_i - A}{h} \quad \text{i.e., } x_i = A + u_i h$$

$$\begin{aligned} \therefore \bar{x} &= \frac{1}{N} \sum f_i x_i = \frac{1}{N} \sum f_i (A + u_i h) = \frac{1}{N} \sum f_i A + \frac{1}{N} \sum f_i u_i h \\ &= A \cdot \frac{\sum f_i}{N} + \frac{1}{N} (\sum f_i u_i) h = A + \frac{\sum f_i u_i}{N} h \quad (\because \sum f_i = N) \end{aligned}$$

$$\therefore \bar{x} = A + \left( \frac{\sum f_i u_i}{N} \right) h.$$

In brief, the above formula is written as  $\bar{x} = A + \left( \frac{\sum fu}{N} \right) h$ .

In case of individual series, this formula takes the form  $\bar{x} = A + \left( \frac{\sum u}{n} \right) h$ .

In dealing with practical problems, it is advisable to first take deviations ( $d$ ) of the values of the variable ( $x$ ) from some suitable number ( $A$ ). Then we see, if there is any common factor, greater than one in the values of the deviations. If there is a common

factor  $h(> 1)$ , then we calculate  $u = \frac{d}{h} = \frac{x - A}{h}$  in the next column. In case, there is no

common factor other than one, then we take  $h = 1$  and  $u$  becomes  $\frac{d}{1} = d = x - A$ . In this case, the formulae reduces as given below:

$$\bar{x} = A + \frac{\sum d}{n} \quad \text{(For Individual Series)}$$

$$\bar{x} = A + \frac{\sum fd}{N} \quad \text{(For Frequency Distribution)}$$

where  $d = x - A$  and  $A$  is any constant; to be chosen suitably.

## NOTES

### WORKING RULES TO FIND A.M.

**Rule I.** In case of an individual series, choose a number  $A$ . Find deviations  $d(=x - A)$  of items from  $A$ . Find the sum ' $\Sigma d$ ' of the deviations. Divide this sum by  $n$ , the total number of items. This quotient is added to  $A$  to get the value of  $\bar{x}$ .

If some common factor  $h (> 1)$  is available in the values of  $d$ , then we calculate ' $u$ ' by dividing the values of  $d$  by  $h$  and find  $\bar{x}$  by using the formula :

$$\bar{x} = A + \left( \frac{\Sigma d}{n} \right) h.$$

**Rule II.** In case of a frequency distribution, choose a number  $A$ . Find deviations  $d(=x - A)$  of items from  $A$ . Find the products  $fd$  of  $f$  and  $d$ . Find the sum ' $\Sigma fd$ ' of these products. Divide this sum by  $N$ , the total number of items. This quotient is added to  $A$  to get the value of  $\bar{x}$ .

If some common factor  $h (> 1)$  is available in the values of  $d$ , then we calculate ' $u$ ' dividing  $d$  by  $h$  and find  $\bar{x}$  by using the formula :

$$\bar{x} = A + \left( \frac{\Sigma fu}{N} \right) h.$$

**Rule III.** If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.

**Example 1.5.** Find the A.M. for the following individual series:

12.36, 14.36, 16.36, 18.36, 20.36, 24.36.

**Solution.**

**Calculation of A.M.**

Variable $x$	$d = x - A$ $A = 16.36$	$u = d/h$ $h = 2$
12.36	-4	-2
14.36	-2	-1
16.36	0	0
18.36	2	1
20.36	4	2
24.36	8	4
		$\Sigma u = 4$

Now  $\bar{x} = A + \left( \frac{\Sigma u}{n} \right) h = 16.36 + \left( \frac{4}{6} \right) 2 = 16.36 + 1.33 = 17.69.$

**Example 1.6.** Calculate A.M. for the following data:

Temp. (in°C)	-40 to -30	-30 to -20	-20 to -10	-10 to 0
No. of days	10	28	30	42
Temp. (in°C)	0 - 10	10 - 20	20 - 30	
No. of days	65	180	10	

## Solution.

## Calculation of A.M.

## NOTES

Temp. (in°C)	No. of days $f$	Mid-points of classes $x$	$d = x - A$ $A = -5$	$u = d/h$ $h = 10$	$fu$
-40 to -30	10	-35	-30	-3	-30
-30 to -20	28	-25	-20	-2	-56
-20 to -10	30	-15	-10	-1	-30
-10 to 0	42	-5	0	0	0
0-10	65	5	10	1	65
10-20	180	15	20	2	360
20-30	10	25	30	3	30
	$N = 365$				$\Sigma fu = 339$

$$\text{Now } \bar{x} = A + \left( \frac{\Sigma fu}{N} \right) h = -5 + \left( \frac{339}{365} \right) 10 = -5 + 9.2877 = 4.2877^\circ\text{C}.$$

## 1.9. A.M. OF COMBINED GROUP

**Theorem.** If  $\bar{x}_1$  and  $\bar{x}_2$  are the A.M. of two groups having  $n_1$  and  $n_2$  items, then the A.M. ( $\bar{x}$ ) of the combined group is given by

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

**Proof.** Let  $x_1, x_2, \dots, x_{n_1}$  and  $y_1, y_2, \dots, y_{n_2}$  be the items in the two groups respectively:

$$\therefore \bar{x}_1 = \frac{x_1 + x_2 + \dots + x_{n_1}}{n_1}$$

$$\bar{x}_2 = \frac{y_1 + y_2 + \dots + y_{n_2}}{n_2}$$

$$\therefore x_1 + x_2 + \dots + x_{n_1} = n_1 \bar{x}_1$$

$$y_1 + y_2 + \dots + y_{n_2} = n_2 \bar{x}_2$$

$$\begin{aligned} \text{Now } \bar{x} &= \frac{\text{sum of items in both groups}}{n_1 + n_2} \\ &= \frac{x_1 + x_2 + \dots + x_{n_1} + y_1 + y_2 + \dots + y_{n_2}}{n_1 + n_2} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \end{aligned}$$

$$\therefore \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

This formula can also be extended to more than two groups.

**Example 1.7.** The mean wage of 1000 workers in a factory running two shifts of 700 and 300 workers is ₹ 500. The mean wage of 700 workers, working in the day shift, is ₹ 450. Find the mean wage of workers, working in the night shift.



## NOTES

**Solution.** No. of workers in the day shift ( $n_1$ ) = 700  
 No. of workers in the night shift ( $n_2$ ) = 300  
 Mean wage of workers in the day shift ( $\bar{x}_1$ ) = ₹ 450  
 Mean wage of all workers ( $\bar{x}$ ) = ₹ 500  
 Let mean wage of workers in the night shift =  $\bar{x}_2$

Now 
$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$\therefore 500 = \frac{700(450) + 300(\bar{x}_2)}{700 + 300} \quad \text{or} \quad 500000 = 315000 + 300\bar{x}_2$$

$$\therefore 300\bar{x}_2 = 185000$$

$$\therefore \bar{x}_2 = \frac{185000}{300} = ₹ 616.67.$$

## 1.10.WEIGHTED A.M.

If all the values of the variable are not of equal importance, or in other words, these are of varying significance, then we calculate **weighted A.M.**

$$\text{Weighted A.M.} = \bar{x}_w = \frac{\sum wx}{\sum w}$$

where  $w_1, w_2, \dots, w_n$  are the weights of the values  $x_1, x_2, \dots, x_n$  of the variable, under consideration.

**Example 1.8.** An examination was held to decide the award of a scholarship. The weights given to different subjects were different. The marks were as follows:

Subjects	Weight	Marks of A	Marks of B	Marks of C
Statistics	4	63	60	65
Accountancy	3	65	64	70
Economics	2	58	56	63
Mercantile Law	1	70	80	52

The candidate getting the highest marks is to be awarded the scholarship. Who should get it?

**Solution.**

**Calculation of weighted A.M.**

Subject	Weight $w$	Marks of A $x_1$	$wx_1$	Marks of B $x_2$	$wx_2$	Marks of C $x_3$	$wx_3$
Statistics	4	63	252	60	240	65	260
Accountancy	3	65	195	64	192	70	210
Economics	2	58	116	56	112	63	126
Mercantile Law	1	70	70	80	80	52	52
$\Sigma w = 10$			$\Sigma wx_1 = 633$		$\Sigma wx_2 = 624$		$\Sigma wx_3 = 648$



## NOTES

Weighted A.M. of

$$A = \frac{\sum wx_1}{\sum w} = \frac{633}{10} = 63.3$$

Weighted A.M. of

$$B = \frac{\sum wx_2}{\sum w} = \frac{624}{10} = 62.4$$

Weighted A.M. of

$$C = \frac{\sum wx_3}{\sum w} = \frac{648}{10} = 64.8$$

∴ The student 'C' is to get the scholarship.

### 1.11. MATHEMATICAL PROPERTIES OF A.M.

1. In a statistical data, the sum of the deviations of items from A.M. is always zero

$$\text{i.e., } \sum_{i=1}^n f_i (x_i - \bar{x}) = 0,$$

where  $f_i$  is the frequency of  $x_i$  ( $1 \leq i \leq n$ ).

2. In a statistical data, the sum of squares of the deviations of items from A.M. is always least i.e.,  $\sum_{i=1}^n f_i (x_i - \bar{x})^2$  is least, where  $f_i$  is the frequency of  $x_i$  ( $1 \leq i \leq n$ ).

#### Merits of A.M.

1. It is the simplest average to understand.
2. It is easy to compute.
3. It is well-defined.
4. It is based on all the items.
5. It is capable of further algebraic treatment.
6. It has sampling stability.
7. It is specially used in finding the average speed, when time taken at different speeds are varying, or are equal.

#### Demerits of A.M.

1. It may not be present in the given series itself. For example, the A.M. of 4, 5, 6, 6 is  $\frac{4+5+6+6}{4} = 5.25$ , which is not present in the series. So, sometimes it becomes theoretical.
2. It cannot be calculated for qualitative data.
3. It may be badly affected by the extreme item.

**EXERCISE 1.1****NOTES**

- Find the A.M. of the series 4, 6, 8, 10, 12.
- The A.M. of 25 items is found to be 78.4. If at the time of calculation, two items were wrongly taken as 96 and 43 instead of 69 and 34, find the value of the correct mean.
- Find the A.M. for the following frequency distribution:

$x$	10	11	12	13	14	15
$f$	2	6	8	6	2	6

- Find the A.M. for the following data:

Marks	18	19	20	21	22	23	24
No. of students	169	320	530	698	230	140	105

- Two hundred people were interviewed by a public opinion polling agency. The following frequency distribution gives the ages of people interviewed. Calculate A.M.

Age Groups (Years)	80—89	70—79	60—69	50—59
No. of Persons	2	2	6	20
Age Groups (Years)	40—49	30—39	20—29	10—19
No. of Persons	56	40	40	42

- Find the A.M. for the following data:

Class intervals	- 2 to 2	3—7	8—12	13—17	18—22	23—27
Frequency	3277	4096	2048	512	64	3

- From the following information, find out:
  - Which of the factor pays larger amount as daily wages.
  - What is the average daily wage of the workers of two factories taken together.

	Factory A	Factory B
No. of wage earners	250	200
Average daily wages	₹ 20	₹ 25

- The mean wage of 100 workers in a factory running two shifts of 60 and 40 workers is ₹ 38. The mean wage of 60 workers working in the day shift is ₹ 40. Find the mean wage of workers, working in the night shift.
- The average weight of 150 students in a class is 80 kg. The average weight of boys in the class is 85 kg and that of girls is 70 kg. Tell the number of boys and girls in the class separately.
- If a student gets the following marks: English 80, Hindi 70, Mathematics 85, Physics 75 and Chemistry 67, find the weighted mean marks if the weights of the subjects are 1, 2, 1, 3, 1 respectively.

## NOTES

11. The following table gives the number of students in different classes in a Government Senior Secondary School and their tuition fees. Find the average tuition fees per student.

Class	No. of students	Tuition fee (₹)
V	65	0.50
VI	80	0.75
VII	95	1.00
VIII	90	1.50
IX	70	2.00

## Answers

1. 8                                      2. 76.96                                      3. 12.6  
 4. 20.61 marks                                      5. 34.9 years                                      6. 4.9995  
 7. (i) Both factories are paying equal amount                                      (ii) ₹ 22.22                                      8. ₹ 35  
 9. Boys = 100, Girls = 50                                      10. 74.625 marks                                      11. ₹ 1.16

## II. GEOMETRIC MEAN (G.M.)

## 1.12. DEFINITION

The **geometric mean** of a statistical data is defined as the  $n$ th root of the product of all the  $n$  values of the variable.

For an individual series, the G.M. is given by

$$\text{G.M.} = (x_1 x_2 \dots x_n)^{1/n}$$

where  $x_1, x_2, \dots, x_n$  are the values of the variable, under consideration. From the definition of G.M., we see that it involves the  $n$ th root of a product, which is not possible to evaluate by using simple arithmetical tools. To solve this problem, we take the help of logarithms.

We have

$$\begin{aligned} \text{G.M.} &= (x_1 x_2 \dots x_n)^{1/n} \\ &= \text{Antilog} [\log (x_1 x_2 \dots x_n)^{1/n}] \\ &= \text{Antilog} \left[ \frac{1}{n} \log (x_1 x_2 \dots x_n) \right] \\ &= \text{Antilog} \left[ \frac{1}{n} (\log x_1 + \log x_2 + \dots + \log x_n) \right] \end{aligned}$$

$$\therefore \text{G.M.} = \text{Antilog} \left( \frac{\sum \log x}{n} \right)$$

For a frequency distribution,

$$\text{G.M.} = (x_1^{f_1} x_2^{f_2} \dots x_n^{f_n})^{1/N}$$

where  $f_i$  is the frequency of  $x_i$  ( $1 \leq i \leq n$ ).

Proceeding on the same lines, we get

$$\text{G.M.} = \text{Antilog} \left( \frac{\sum f \log x}{N} \right)$$

When the values of the variable are given in the form of classes, the mid-points are taken as the values of the variable ( $x$ ).



### WORKING RULES TO FIND G.M.

- Rule I.** In case of an individual series, first find the sum of logarithms of all the items. In the second step, divide this sum by  $n$ , the total number of items. Next, take the 'antilogarithm' of this quotient. This gives the value of the G.M.
- Rule II.** In case of a frequency distribution, find the product ( $f \log x$ ) of frequencies and logarithm of value of items. In the second step, find the sum ( $\Sigma f \log x$ ) of these products. Divide this sum by the sum ( $N$ ) of all the frequencies. Next, take the 'antilogarithm' of this quotient. This gives the value of the G.M.
- Rule III.** If the values of the variables are given in the form of classes, then their respective mid-points are taken as the values of the variable.

Role of Statistics and  
Measures of Central  
Tendency

### NOTES

**Example 1.9.** Find the G.M. for the following frequency distribution:

$x$	2	4	6	8	10	12
$f$	5	7	15	4	2	1

**Solution.**

**Calculation of G.M.**

$x$	$f$	$\log x$	$f \log x$
2	5	0.3010	1.5050
4	7	0.6021	4.2147
6	15	0.7782	11.6730
8	4	0.9031	3.6124
10	2	1.0000	2.0000
12	1	1.0792	1.0792
	$N = 34$		24.0843

Now 
$$\text{G.M.} = \text{Antilog} \left( \frac{\Sigma f \log x}{N} \right)$$

$$= \text{Antilog} \left( \frac{24.0843}{34} \right) = \text{Antilog} (0.7084) = 5.110.$$

**Example 1.10.** Find the G.M. for the data given below:

Yield of wheat (in quintals)	7.5—10.5	10.5—13.5	13.5—16.5	16.5—19.5
No. of farms	5	9	19	23
Yield of wheat (in quintals)	19.5—22.5	22.5—25.5	25.5—28.5	
No. of farms	7	4	1	



## NOTES

Class	Mid-point $x$	$f$	$\log x$	$f \log x$
7.5—10.5	9	5	0.9542	4.7710
10.5—13.5	12	9	1.0792	9.7128
13.5—16.5	15	19	1.1761	22.3459
16.5—19.5	18	23	1.2553	28.8719
19.5—22.5	21	7	1.3222	9.2554
22.5—25.5	24	4	1.3802	5.5208
25.5—28.5	27	1	1.4314	1.4314
		$N = 68$		$\Sigma f \log x$ $= 81.9092$

Now

$$G = \text{Antilog} \left( \frac{\Sigma f \log x}{N} \right) = \text{Antilog} \left( \frac{81.9092}{68} \right)$$

$$= \text{Antilog} (1.2045) = 16.02 \text{ quintals.}$$

**1.13. G.M. OF COMBINED GROUP**

**Theorem.** If  $G_1$  and  $G_2$  are the GMs of two groups having  $n_1$  and  $n_2$  items, then the G.M. ( $G$ ) of the combined group is given by

$$G = \text{Antilog} \left( \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2} \right).$$

**Proof.** Let  $x_1, x_2, \dots, x_{n_1}$  and  $y_1, y_2, \dots, y_{n_2}$  be the items in the two groups respectively.

$$\therefore G_1 = \text{Antilog} \left( \frac{\Sigma \log x}{n_1} \right)$$

$$\therefore \log G_1 = \frac{\Sigma \log x}{n_1}$$

$$\therefore n_1 \log G_1 = \Sigma \log x$$

Similarly,  $n_2 \log G_2 = \Sigma \log y$

Now  $G = \text{Antilog} \left( \frac{\text{sum of logarithms of all items}}{\text{no. of items in both groups}} \right)$

$$= \text{Antilog} \left( \frac{\Sigma \log x + \Sigma \log y}{n_1 + n_2} \right)$$

$$\therefore G = \text{Antilog} \left( \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2} \right).$$

This formula can also be extended to more than two groups.

**Example 1.11.** The G.M. of wages of 200 workers working in a factory is ₹ 700. The G.M. of wages of 300 workers, working in another factory is ₹ 1000. Find the G.M. of wages of all the workers taken together.

**Solution.** No. of workers in I factory ( $n_1$ ) = 200

No. of workers in II factory ( $n_2$ ) = 300

G.M. of wages of workers of I factory ( $G_1$ ) = ₹ 700

G.M. of wages of workers of II factory ( $G_2$ ) = ₹ 1000

Let G be the G.M. of wages of all the workers taken together.

$$\begin{aligned}\therefore G &= \text{Antilog} \left( \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2} \right) \\ &= \text{Antilog} \left( \frac{200 \log 700 + 300 \log 1000}{200 + 300} \right) \\ &= \text{Antilog} \left( \frac{200 (2.8451) + 300 (3.0000)}{500} \right) = \text{Antilog} \left( \frac{569.0200 + 900}{500} \right) \\ &= \text{Antilog} (2.9380) = \text{Rs. } 867.\end{aligned}$$

## NOTES

### 1.14. AVERAGING OF PERCENTAGES

Geometric mean is specially used to find the average rate of increase or decrease in sale, production, population, etc.

If  $V_0$  and  $V_n$  are the values of a variable at the beginning of the first and at the end of the  $n$ th period, then

$$V_n = V_0 (1 + r)^n, \text{ where } r \text{ is the average rate of growth per unit.}$$

**Example 1.12.** At what rate of interest would Rs. 100 double in 10 years.

**Solution.** Here  $V_0 = 100$  and  $V_{10} = 200$ .

Let  $r$  be the average rate of interest per rupee

$$\begin{aligned}\therefore V_{10} &= V_0 (1 + r)^{10} \\ \text{or } 200 &= 100(1 + r)^{10} \quad \text{or } (1 + r)^{10} = 2\end{aligned}$$

$$\therefore 10 \log (1 + r) = \log 2 = 0.3010$$

$$\therefore \log (1 + r) = 0.03010$$

$$\therefore 1 + r = \text{Antilog } 0.0301 = 1.074$$

$$\therefore r = 1.074 - 1 = 0.074$$

$$\therefore \text{Average percentage rate of interest} = 0.074 \times 100 = 7.4\%.$$

**Example 1.13.** The machinery of an industrial house is depreciated by 50% in the first year, 30% in the second year and by 10% in the following three years. Find out the average rate of depreciation for the entire period.

**Solution.****NOTES**

Year	Rate of depreciation	Depreciated value of the machine at the end of the year taking 100 in the beginning (x)	log x
I	50%	50	1.6990
II	30%	70	1.8451
III	10%	90	1.9542
IV	10%	90	1.9542
V	10%	90	1.9542
			$\Sigma \log x = 9.4067$

$$\therefore \text{G.M.} = \text{Antilog} \left( \frac{\Sigma \log x}{n} \right) = \text{Antilog} \left( \frac{9.4067}{5} \right)$$

$$= \text{Antilog} (1.88134) = 76.08$$

$$\therefore \text{Average rate of depreciation} = 100 - 76.08 = 23.92\%.$$

**1.15. WEIGHTED G.M.**

If all the values of the variable are not of equal importance, or in other words, these are of varying significance, then we calculate **weighted G.M.**

$$\text{Weighted G.M.} = \text{Antilog} \left( \frac{\Sigma w \log x}{\Sigma w} \right),$$

where  $w_1, w_2, \dots, w_n$  are the weights of the values  $x_1, x_2, \dots, x_n$  of the variable, under consideration.

**Example 1.14.** The G.M. of 15 observations is found to be 12. Later on, it was discovered that the item 21 was misread as 14. Calculate the correct value of G.M.

**Solution.** No. of items = 15

Incorrect G.M. = 12

Correct item = 21

Incorrect item = 14

$$\text{Now } G = \text{Antilog} \left( \frac{\Sigma \log x}{n} \right)$$

$$\therefore 12 = \text{Antilog} \left( \frac{\text{incorrect } \Sigma \log x}{15} \right)$$

$$\text{or } \log 12 = \frac{\text{incorrect } \Sigma \log x}{15}$$

$$\therefore \text{Incorrect } \Sigma \log x = 15 \log 12 = 15(1.0792) = 16.1880$$

$$\begin{aligned} \text{Now } \text{Correct } \Sigma \log x &= 16.1880 - \log 14 + \log 21 \\ &= 16.1880 - 1.1461 + 1.3222 = 16.3641. \end{aligned}$$

$$\therefore \text{Correct G.M.} = \text{Antilog} \left( \frac{16.3641}{15} \right) = \text{Antilog} (1.0909) = 12.33.$$

## Merits of G.M.

1. It is well defined.
2. It is based on all the items.
3. It is capable of further algebraic treatment.
4. It is used to find the average rate of increase or decrease in the variables like sale, production, population etc.
5. It is specially used in the construction of index numbers.
6. It is used when larger weights are to be given to smaller items and smaller weights to larger items.
7. It has sampling stability.

## NOTES

## Demerits of G.M.

1. It is not simple to understand.
2. It is not easy to compute.
3. It may become imaginary in the presence of negative items.
4. If any one item is zero, then its value would be zero, irrespective of magnitude of other items.

### EXERCISE 1.2

1. From the monthly incomes of ten families given below, calculate G.M.

S. No.	1	2	3	4	5	6	7	8	9	10
Income (in ₹)	145	367	268	73	185	619	280	115	870	315

2. Find the G.M. for the following frequency distribution:

$x$	8	10	12	14	16	18
$f$	6	10	20	8	5	1

3. Calculate G.M. for the following data:

Income (in ₹)	100—300	100—500	100—700	100—1000	100—1500
No. of employees	12	18	30	50	100

4. A firm declared bonus according to respective salary groups as given below :

Salary Group (in ₹)	60—75	75—90	90—105
Rate of Bonus	60	70	80
No. of employees	3	4	5
Salary Group (in ₹)	105—120	120—135	135—150
Rate of Bonus	90	100	110
No. of employees	5	7	6

Calculate A.M. of salaries and G.M. of the bonus payable to the employees.

## NOTES

5. The population of a country is increased from 40 crore to 70 crore in 30 years. Find out the annual average rate of growth.
6. A Principal increased the number of students in his college in the year 1983 by 15%. Then increased again in 1984 by 5% but in 1985, it decreased by 20% due to introduction of 10 + 2 system. Hence the number of students becomes the same as it was before 1983. Do you agree, if not give reasons.
7. A machine is assumed to depreciate 30% in value in the I year, 25% in the II year and 20% for the next 2 years, each percentage being calculated on the diminishing value. Find the average rate of depreciation for the four years.
8. The G.M. of 20 items was found to be 10. Later on, it was found that one item 18 was misread as 8. Find the correct value of the G.M.

## Answers

1. ₹ 252.40                      2. 11.82                      3. ₹ 794.10
4. Average salary = ₹ 111 ; Average bonus = ₹ 87.44
5. 1.9%                          6. No, G.M. is to be used, 1.14% decrease
7. 23.86%                      8. 10.41.

## III. HARMONIC MEAN (H.M.)

## 1.16. DEFINITION

The **harmonic mean** of a statistical data is defined as the quotient of the number of items by the sum of the reciprocals of all the values of the variable.

(a) For an individual series, the H.M. is given by

$$\text{H.M.} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x}}$$

where  $x_1, x_2, \dots, x_n$  are the values of the variable, under consideration.

(b) For a frequency distribution,

$$\text{H.M.} = \frac{f_1 + f_2 + \dots + f_n}{f_1 \left( \frac{1}{x_1} \right) + f_2 \left( \frac{1}{x_2} \right) + \dots + f_n \left( \frac{1}{x_n} \right)} = \frac{\sum f}{\sum f \left( \frac{1}{x} \right)} = \frac{N}{\sum \left( \frac{f}{x} \right)}$$

where  $f_i$  is the frequency of  $x_i$  ( $1 \leq i \leq n$ ).

When the values of the variable are given in the form of classes, then the mid-points of classes are taken as the values of the variable ( $x$ ).

## WORKING RULES TO FIND H.M.

- Rule I.** In case of an individual series, first find the sum of the reciprocals of all the items. In the second step, divide  $n$ , the total number of items by this sum of reciprocals. This gives the value of the H.M.
- Rule II.** In case of a frequency distribution, find the quotients ( $f/x$ ) of frequencies by the value of items. In the second step, find the sum ( $\sum(f/x)$ ) of these quotients. Divide  $N$ , the total of all frequencies by this sum of quotients. This gives the value of the H.M.
- Rule III.** If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.

**Example 1.15.** Calculate the H.M. for the following individual series:

$x$	4	7	10	12	19
-----	---	---	----	----	----

*Role of Statistics and  
Measures of Central  
Tendency*

**Solution.**

**Calculation of H.M.**

S. No.	$x$	$1/x$
1	4	0.2500
2	7	0.1429
3	10	0.1000
4	12	0.0833
5	19	0.0526
$n = 5$		$\sum \left( \frac{1}{x} \right) = 0.6288$

Now 
$$\text{H.M.} = \frac{n}{\sum \left( \frac{1}{x} \right)} = \frac{5}{0.6288} = 7.9516.$$

**Example 1.16.** Calculate the value of H.M. for the following data:

Marks	0—10	0—20	0—30	0—40	0—50	0—60	0—70
No. of students	4	8	15	23	51	60	70

**Solution.**

**Calculation of H.M.**

Class	No. of students $f$	Mid-points $x$	$\frac{f}{x}$
0—10	4	5	0.8000
10—20	4	15	0.2667
20—30	7	25	0.2800
30—40	8	35	0.2286
40—50	28	45	0.6222
50—60	9	55	0.1636
60—70	10	65	0.1538
	$N = 70$		$\sum \left( \frac{f}{x} \right) = 2.5149$

Now 
$$\text{H.M.} = \frac{N}{\sum \left( \frac{f}{x} \right)} = \frac{70}{2.5149} = 27.83 \text{ marks.}$$

**NOTES**

## 1.17. H.M. OF COMBINED GROUP

### NOTES

**Theorem.** If  $H_1$  and  $H_2$  are the H.M. of two groups having  $n_1$  and  $n_2$  items, then the H.M. of the combined group is given by

$$H = \frac{n_1 + n_2}{\frac{n_1}{H_1} + \frac{n_2}{H_2}}$$

**Proof.** Let  $x_1, x_2, \dots, x_{n_1}$  and  $y_1, y_2, \dots, y_{n_2}$  be the items in the two groups respectively.

$$\therefore H_1 = \frac{n_1}{\sum \frac{1}{x}}, \quad H_2 = \frac{n_2}{\sum \frac{1}{y}}$$

$$\therefore \sum \frac{1}{x} = \frac{n_1}{H_1}, \quad \sum \frac{1}{y} = \frac{n_2}{H_2}$$

$$\begin{aligned} \text{Now } H &= \frac{\text{no. of items in both groups}}{\text{sum of reciprocals of all the items in both groups}} \\ &= \frac{n_1 + n_2}{\sum \frac{1}{x} + \sum \frac{1}{y}} \quad \therefore H = \frac{n_1 + n_2}{\frac{n_1}{H_1} + \frac{n_2}{H_2}} \end{aligned}$$

This formula can also be extended to more than two groups.

**Example 1.17.** The H.M. of two groups containing 10 and 12 items are found to be 29 and 35. Find the H.M. of the combined group.

**Solution.** Here  $n_1 = 10, \quad n_2 = 12$   
 $H_1 = 29, \quad H_2 = 35$

Let  $H$  be the H.M. of the combined group

$$\begin{aligned} \therefore H &= \frac{n_1 + n_2}{\frac{n_1}{H_1} + \frac{n_2}{H_2}} = \frac{10 + 12}{\frac{10}{29} + \frac{12}{35}} \\ &= \frac{22}{0.3448 + 0.3429} = \frac{22}{0.6877} = 31.9907. \end{aligned}$$

## 1.18. WEIGHTED H.M.

If all the values of the variable are not of equal importance or in other words, these are of varying importance, then we calculate **weighted H.M.**

$$\text{Weighted H.M.} = \frac{\sum w}{\sum \left( \frac{w}{x} \right)}$$

where  $w_1, w_2, \dots, w_n$  are the weights of the values  $x_1, x_2, \dots, x_n$  of the variable, under consideration.

**Example 1.18.** Find the weighted H.M. of the items 4, 7, 12, 19, 25 with weights 1, 2, 1, 1, 1 respectively.

*Role of Statistics and  
Measures of Central  
Tendency*

**Solution.**

**Calculation of weighted H.M.**

$x$	$w$	$w/x$
4	1	0.2500
7	2	0.2857
12	1	0.0833
19	1	0.0526
25	1	0.0400
	$\sum w = 6$	$\sum \left(\frac{w}{x}\right) = 0.7116$

$$\text{Now weighted H.M.} = \frac{\sum w}{\sum \left(\frac{w}{x}\right)} = \frac{6}{0.7116} = 8.4317.$$

**NOTES**

### Merits of H.M.

1. It is well-defined.
2. It is based on all the items.
3. It is capable of further algebraic treatment.
4. It has sampling stability.
5. It is specially used in finding the average speed, when the distances covered at different speeds are equal or unequal.

### Demerits of H.M.

1. It is not simple to understand.
2. It is not easy to compute.
3. It gives higher weightage to smaller items, which may not be desirable in some problems.

### EXERCISE 1.3

1. Find the H.M. for the following series:  
3, 5, 6, 6, 7, 10, 12.
2. Find the H.M. for the following series:  
0.874, 0.989, 0.012, 0.008, 0.00009.
3. The following table gives the marks obtained by students in a class. Calculate the H.M.:

Marks	18	21	30	45
No. of students	6	12	9	2



## NOTES

4. Calculate the H.M. for the following:

Income (in ₹)	10	20	30	40	50
No. of persons	2	4	3	0	1

5. The following table gives the marks (out of 50) obtained by 70 students in a class. Calculate the H.M.

Marks	18	21	24	26	30	38	45
No. of students	6	12	15	19	9	7	2

6. Calculate the H.M. for the following frequency distribution:

Marks	0—10	10—20	20—30	30—40	40—50
No. of students	4	7	28	12	9

7. Following is the data regarding the marks obtained by 159 students in an examination. Find the H.M.

Marks	0—9	10—19	20—29	30—39	40—49
No. of students	19	37	61	27	15

## Answers

1. 5.9                      2. 0.0004416                      3. 23.2147 marks                      4. Rs. 19.23  
 5. 25.09 marks                      6. 20.48 marks                      7. 15.31 marks

## IV. MEDIAN

## 1.19. DEFINITION

The **median** of a statistical series is defined as the size of the middle most item (or the A.M. of two middle most items), provided the items are in order of magnitude. For example, the median for the series 4, 6, 10, 12, 18 is 10 and for the series 4, 6, 10, 12,

18, 22, the value of median would be  $\frac{10+12}{2} = 11$ . It can be observed that 50% items in the series would have value less than or equal to median and 50% items would be with value greater or equal to the value of the median.

For an individual series, the median is given by,

$$\text{Median} = \text{size of } \frac{n+1}{2} \text{th item}$$

where  $x_1, x_2, \dots, x_n$  are the values of the variable under consideration. The values  $x_1, x_2, \dots, x_n$  are supposed to have been arranged in order of magnitude. If  $\frac{n+1}{2}$

comes out to be in decimal, then we take median as the A.M. of size of  $\frac{n}{2}$ th and  $\left(\frac{n}{2} + 1\right)$ th items.

## NOTES

### WORKING RULES FOR FINDING MEDIAN FOR AN INDIVIDUAL SERIES

**Step I.** Arrange the given items in order of magnitude.

**Step II.** Find the total number 'n' of items.

**Step III.** Write: median = size of  $\frac{n+1}{2}$ th item.

**Step IV.** (i) If  $\frac{n+1}{2}$  is a whole number, then  $\frac{n+1}{2}$ th item gives the value of median.

(ii) If  $\frac{n+1}{2}$  is in fraction, then the A.M. of  $\frac{n}{2}$ th and  $\left(\frac{n}{2} + 1\right)$ th items gives the value of median.

For a frequency distribution, in which frequencies (f) of different values (x) of the variable are given, we have

Median = size of  $\frac{N+1}{2}$ th item.

**Remark.** The values of the variable are supposed to have been arranged in order of magnitude.

### WORKING RULES FOR FINDING MEDIAN FOR A FREQUENCY DISTRIBUTION

**Step I.** Arrange the values of the variable in order of magnitude and find the cumulative frequencies (c.f.).

**Step II.** Find the total 'N' of all frequencies and check that it is equal to the last c.f.

**Step III.** Write: median = size of  $\frac{N+1}{2}$ th item.

**Step IV.** (a) If  $\frac{N+1}{2}$  is a whole number, then  $\frac{N+1}{2}$ th item gives the value of median. For this, look at the cumulative frequency column and find that total which is either equal to  $\frac{N+1}{2}$  or the next higher than  $\frac{N+1}{2}$  and determine the value of the variable corresponding to this. This gives the value of median.

(b) If  $\frac{N+1}{2}$  is in fraction, then the A.M. of  $\frac{N}{2}$ th and  $\left(\frac{N}{2} + 1\right)$ th items gives the value of median.

In case, the values of the variable are given in the form of classes, we shall assume that items in the classes are uniformly distributed in the corresponding classes. We define

Median = size of  $\frac{N}{2}$ th item.

Here we shall get the class in which  $N/2$ th item is present. This is called the **median class**. To ascertain the value of median in the median class, the following formula is used.

$$\text{Median} = L + \left( \frac{N/2 - c}{f} \right) h$$

where  $L$  = lower limit of the median class

$c$  = cumulative frequency of the class preceding the median class

$f$  = simple frequency of the median class

$h$  = width of the median class.

## NOTES

**Remark.** In problems on **Averages** or in other problems in the following chapters, where we need only the mid values of class intervals in the formula, we need not convert the classes written using 'inclusive method'.

The following points must be taken care of, while calculating median:

1. The values of the variable must be in order of magnitude. In case of classes of values of the variable, the classes must be strictly in *ascending* order of magnitude.

2. If the classes are in inclusive form, then the actual limits of the median class are to be taken for finding  $L$  and  $h$ .

3. The classes may not be of equal width i.e.,  $h$  need not be the common width of all classes. It is the width of the "median class".

4. In case of open end classes, it is advisable to find average by using median.

### WORKING RULES FOR FINDING MEDIAN FOR A FREQUENCY DISTRIBUTION WITH CLASS INTERVALS

**Step I.** Arrange the classes in the ascending order of magnitude. The classes must be in 'exclusive form'. The widths of classes may not be equal. Find the cumulative frequencies (c.f.).

**Step II.** Find the total ' $N$ ' of all frequencies and check that it is equal to the last c.f.

**Step III.** Write: median = size of  $\frac{N}{2}$  th item.

**Step IV.** Look at the cumulative frequency column and find that total which is either equal to  $\frac{N}{2}$  or the next higher than  $\frac{N}{2}$  and determine the class corresponding to this. That gives the 'median class'.

**Step V.** Write: median =  $L + \left( \frac{N/2 - c}{f} \right) h$ . Put the values of  $L$ ,  $N/2$ ,  $c$ ,  $f$ ,  $h$  and calculate the value of median.

**Example 1.19.** The following are the marks obtained by a batch of 10 students in a certain class test in Statistics and Accountancy:

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks in Statistics	63	64	62	32	30	60	47	46	35	28
Marks in Accountancy	68	65	35	42	26	85	44	80	33	72

In which subject is the level of knowledge of students higher?

**Solution.** In this problem, median is the most suitable average.

The marks in Statistics arranged in ascending order are:

28, 30, 32, 35, 46, 47, 60, 62, 63, 64.

## NOTES

Here  $n = 10$ .  $\frac{n+1}{2} = \frac{10+1}{2} = 5.5$

$\therefore$  Median = size of 5.5th item

$$= \frac{\text{size of 5th item} + \text{size of 6th item}}{2}$$

$$= \frac{46 + 47}{2} = 46.5 \text{ marks.}$$

The marks in Accountancy arranged in ascending order are:

26, 33, 35, 42, 44, 65, 68, 72, 80, 85.

Here  $n = 10$ .  $\frac{n+1}{2} = \frac{10+1}{2} = 5.5$

$\therefore$  Median = size of 5.5th item

$$= \frac{\text{size of 5th item} + \text{size of 6th item}}{2}$$

$$= \frac{44 + 65}{2} = 54.5 \text{ marks.}$$

$\therefore$  Level of knowledge is higher in accountancy.

**Example 1.20.** The following table gives the weekly expenditure of 100 families. Find the median.

Weekly expenditure (in ₹)	0—10	10—20	20—30	30—40	40—50
No. of families	14	23	27	21	15

**Solution.**

**Calculation of Median**

Weekly expenditure (in ₹)	No. of families $f$	c.f.
0—10	14	14
10—20	13	37 = $c$
$L = 20—30$	27 = $f$	64
30—40	21	85
40—50	15	100 = $N$
	$N = 100$	

$$\frac{N}{2} = \frac{100}{2} = 50$$

$\therefore$  Median = size of 50th item

$\therefore$  Median class is 20—30.

$$\text{Now, median} = L + \left( \frac{\frac{N}{2} - c}{f} \right) h = 20 + \left( \frac{50 - 37}{27} \right) 10 = 20 + 4.81 = ₹ 24.81.$$

**Example 1.21.** The following table gives the ages in years of 800 persons. Find out the median age.

**NOTES**

Age (in years)	20—60	20—55	20—40	20—30
No. of persons	800	740	400	120
Age (in years)	20—50	20—45	20—25	20—35
No. of persons	670	550	50	220

**Solution.****Calculation of Median**

Age (in years)	No. of persons (f)	c.f.
20—25	50	50
25—30	120 - 50 = 70	50 + 70 = 120
30—35	220 - 120 = 100	120 + 100 = 220 = c
L = 35—40	400 - 220 = 180 = f	220 + 180 = 400
40—45	550 - 400 = 150	400 + 150 = 550
45—50	670 - 550 = 120	550 + 120 = 670
50—55	740 - 670 = 70	670 + 70 = 740
55—60	800 - 740 = 60	740 + 60 = 800
	N = 800	

$$\frac{N}{2} = \frac{800}{2} = 400$$

∴ Median = size of 400th item

∴ Median class is 35—40.

$$\begin{aligned} \therefore \text{Median} &= L + \left( \frac{\frac{N}{2} - c}{f} \right) h = 35 + \left( \frac{400 - 220}{180} \right) 5 \\ &= 35 + 5 = 40 \text{ years.} \end{aligned}$$

**Example 1.22.** Calculate the median for the following data:

Wages upto (in ₹)	15	30	45	60	75	90	105	120
No. of workers	12	30	65	107	157	202	222	230

**Solution.****Calculation of Median**

Wages (in ₹)	No. of workers f	c.f.
0—15	12	12
15—30	30 - 12 = 18	30
30—45	65 - 30 = 35	65
45—60	107 - 65 = 42	107 = c
L = 60—75	157 - 107 = 50 = f	157
75—90	202 - 157 = 45	202
90—105	222 - 202 = 20	222
105—120	230 - 222 = 8	230 = N
	N = 230	



$$\frac{N}{2} = \frac{230}{2} = 115$$

∴ Median = size of 115th item

∴ Median class is 60—75.

$$\therefore \text{Median} = L + \left( \frac{\frac{N}{2} - c}{f} \right) h = 60 + \left( \frac{115 - 107}{50} \right) 15 = 60 + 2.4 = ₹ 62.40.$$

**Example 1.23.** You are given the following incomplete frequency distribution. It is known that the total frequency is 1000 and that the median is 413.11. Estimate the missing frequencies.

Value	Frequency	Value	Frequency
300—325	5	400—425	326
325—350	17	425—450	?
350—375	80	450—475	88
375—400	?	475—500	9

**Solution.** Let the missing frequencies of the classes 375—400 and 425—450 be  $a$  and  $b$  respectively.

Value	Frequency $f$	c.f.
300—325	5	5
325—350	17	22
350—375	80	102
375—400	$a$	$102 + a = c$
$L = 400—425$	$326 = f$	$428 + a$
425—450	$b$	$428 + a + b$
450—475	88	$516 + a + b$
475—500	9	$525 + a + b = 1000$
	$N = 1000$	

Median is given to be 413.11.

∴ Median class is 400—425.

Now, 
$$\text{Median} = L + \left( \frac{N/2 - c}{f} \right) h$$

Here  $L = 400, N/2 = 500, c = 102 + a, h = 25.$

$$413.11 = 400 + \left( \frac{500 - (102 + a)}{326} \right) 25$$

$$\therefore (13.11) 326 = (500 - 102 - a) 25$$

or  $4273.86 = (398 - a) 25$

or  $398 - a = 170.9544$  or  $a = 227.0456 = 227$

Also  $525 + a + b = 1000$

$$b = 1000 - 525 - 227 = 228$$

∴ The missing frequencies are 227 and 228.

## NOTES

**Merits of Median****NOTES**

1. It is simple to understand.
2. It is easy to compute.
3. It is well-defined.
4. It is not affected by the extreme items.
5. It is best suited for open end classes.
6. It can also be located graphically.

**Demerits of Median**

1. It is not based on all the items.
2. It is not capable of further algebraic treatment.
3. It can only be calculated when the data is in-order of magnitude.

**EXERCISE 1.4**

1. Find the value of the median for the following series:

4, 6, 7, 8, 12, 10, 13, 14.

2. Find the median for the following frequency distribution:

<i>x</i>	5	10	15	20	25
<i>f</i>	2	4	6	8	10

3. Find the median for the following frequency distribution:

<i>Marks</i>	0—10	10—20	20—30	30—40	40—50	50—60
<i>No. of students</i>	15	17	19	27	19	12

4. For the following frequency distribution, find out the value of median:

<i>Marks</i>	0—7	7—14	14—21	21—28
<i>Frequency</i>	3	4	7	11
<i>Marks</i>	28—35	35—42	42—49	
<i>Frequency</i>	0	16	9	

5. Calculate median and arithmetic average for the following data:

<i>Class Interval</i>	10—20	10—30	10—40	10—50
<i>Frequency</i>	4	6	56	97
<i>Class Interval</i>	10—60	10—70	10—80	10—90
<i>Frequency</i>	124	137	146	150

6. Calculate the median for the following distribution:

Height (in inches)	60—63	63—66	66—69	69—72	72—75	75—78
No. of men	8	28	118	66	16	4

7. In a frequency distribution of 100 families given below, the median is known to be 50. Find the missing frequencies.

Expenditure (in ₹)	0—20	20—40	40—60	60—80	80—100
No. of families	14	?	27	?	15

8. Find the missing frequencies in the following distribution, if  $N = 100$  and median of the distribution is 30:

Marks	0—10	10—20	20—30	30—40	40—50	50—60
No. of students	10	?	25	30	?	10

### Answers

1. 9                      2. 20                      3. 31.2963 marks                      4. 28 marks  
 5. 44.6341, 47                      6. 68.1356 inches                      7. 22, 22                      8. 15, 10

## V. MODE

### 1.20. DEFINITION

The **mode** of a statistical series is defined as that value of the variable around which the values of the variable tend to be most heavily concentrated. It can also be defined as that value of the variable whose own frequency is dominating and at the same time, the frequencies of its neighbouring items are also dominating. Thus, we see that mode is that value of the variable around which the items of the series cluster densely. Let us consider the data regarding the sale of ready made shirts:

Size (in inches)	30	32	34	36	38	40	42
No. of shirts sold	5	22	24	38	16	8	2

Here we see that the frequency of 36 is highest and the frequencies of its neighbouring items (34, 38) are also dominating. Here the most fashionable, modal size is 36 inches. Technically, we shall say that the mode of the distribution is 36 inches.

In case of mode, we are to deal with the frequencies of values of the items, thus if we are to find the value of mode for an individual series, we will have to see the repetition of different items. i.e., we would be in a way expressing it in the form of frequency distribution. Thus, we start our discussion for evaluating mode for frequency distributions. There are two methods of finding mode of a frequency distribution.

### NOTES



## NOTES

**1.21. MODE BY INSPECTION**

Sometimes the frequencies in a frequency distribution are so distributed that we would be able to find the value of mode just, by inspection. For example, let us consider the frequency distribution:

$x$	4	5	6	7	8	9	10	11	12
$f$	1	2	1	5	12	4	2	2	1

Here we can say, at once, that mode is 8.

**1.22. MODE BY GROUPING**

Let us consider the distribution:

$x$	4	5	6	7	8	9	10	11	12
$f$	4	5	7	14	8	15	2	2	1

Here the frequency of 9 is more than the frequency of 7, whereas the frequencies of neighbouring items of 7 are more than that for 9. In such a case, we would not be able to judge the value of mode just by inspecting the data. In case there is even slight doubt as to which is the value of mode, we go for this method. In this method, two tables are drawn. These tables are called 'Grouping Table' and 'Analysis Table'. In the grouping table, six columns are drawn. The column of frequencies is taken as the column I. In the column II, the sum of two frequencies are taken at a time. In the column III, we exclude the first frequency and take the sum of two frequencies at a time. In the column IV, we take the sum of three frequencies at a time. In the column V, we exclude the first frequency and take the sum of frequencies, taking three at a time. In the last column, we exclude the first two frequencies and take the sum of three frequencies at a time. The next step is to mark the maximum sums in each of the six columns.

In the analysis table, six rows are drawn corresponding to each column in the grouping table. In this table, columns are made for those values of the variable whose frequencies accounts for giving maximum totals in the columns of the grouping table. In this table, marks are given to the values of the variable as often as their frequencies are added to make the total maximum in the columns of the grouping table. The value of the variable which get the maximum marks is declared to be the mode of the distribution.

In case, the values of the variable are given in the form of classes, we shall assume that the items in the classes are uniformly distributed in the corresponding classes. Here we shall get a 'class' either by the method of inspection or the method of grouping. This class is called the **modal class**. To ascertain the value of mode in the modal class, the following formula is used.

$$\text{Mode} = L + \left( \frac{\Delta_1}{\Delta_1 + \Delta_2} \right) h$$

where  $L$  = lower limit of modal class

$\Delta_1$  = difference of frequencies of modal class and pre-modal class

$\Delta_2$  = difference of frequencies of modal class and post-modal class

$h$  = width of the modal class.

## NOTES

The following points must be taken care of while calculating mode:

1. The values (or classes of values) of the variable must be in ascending order of magnitude.
2. If the classes are in inclusive form, then the actual limits of the modal class are to be taken for finding  $L$  and  $h$ .
3. The classes must be of equal width.

It may be noted that while analysing the analysis table, we may find two or more values (or classes of values) of the variable getting equal marks. In such a case, the grouping method fails. Such distribution is called a **multi-modal distribution**.

### 1.23. EMPIRICAL MODE

In case of a multi-modal distribution, we find the value of mode by using the relation

$$\text{Mode} = 3 \text{ Median} - 2 \text{ A.M.}$$

This mode is called **empirical mode** in the sense that this relation cannot be established algebraically. But it is generally observed that in distributions, the value of mode is approximately equal to  $3 \text{ Median} - 2 \text{ A.M.}$  That is why, this mode is called *empirical mode*.

#### WORKING RULES FOR FINDING MODE

- Step I.** If mode is not evident by the 'method of inspection', then the 'method of grouping' should be used.
- Step II.** In case, the values of variable are given in terms of classes of equal width, then **Step I**, will give the 'modal class'.
- Step III.** To find value of the mode, use the formula:

$$\text{mode} = L + \left( \frac{\Delta_1}{\Delta_1 + \Delta_2} \right) h.$$

- Step IV.** In case, the distribution is multimodal, then find the value of mode by using the formula: 'mode =  $3 \text{ median} - 2 \text{ A.M.}$ '.

**Example 1.24.** Find the mode for the following distribution:

Profit ('000 ₹)	28	29	30	31	32	33
No. of firms	4	7	10	6	2	1

**Solution.**

#### Calculation of Mode

Profit (000 ₹) $x$	No. of firms $f$
28	4
29	7
30	10
31	6
32	2
33	1

By inspection we can say that mode is ₹ 30,000. This is so because the frequency of 30,000 is very high as compared with the frequencies of other values of  $x$ . Moreover, the frequencies of the neighbouring items are also dominating.

**Example 1.25.** Find the mode for the following frequency distribution:

$x$	5	10	15	20	25	30	35	40
$y$	4	15	25	20	17	26	10	3

### NOTES

**Solution.** We find the 'mode' by using the 'method of grouping'.

**Grouping Table**

$x$	$f$	I	II	III	IV	V	VI
5	4						
10	15		19		44		
15	25			40		60	
20	20		45				
25	17			37			
30	26		43		63		62
35	10			36		53	39
40	3		13				

**Analysis Table**

Column	30	15	20	10	25
I	1				
II		1	1		
III		1		1	
IV	1		1	1	1
V		1	1	1	
VI		1	1		1
	2	4	4	3	2

Since the totals for 15 and 20 are equal, the given frequency distribution is bimodal. For this distribution, we find mode by using the formula:

$$\text{mode} = 3 \text{ median} - 2 \text{ A.M.}$$

**Calculation of  $\bar{X}$  and median**

$x$	$f$	$c.f.$	$d = x - A$ $A = 20$	$u = d/h$ $h = 5$	$fu$
5	4	4	-15	-3	-12
10	15	19	-10	-2	-30
15	25	44	-5	-1	-25
20	20	64	0	0	0
25	17	81	5	1	17
30	26	107	10	2	52
35	10	117	15	3	30
40	3	120	20	4	12
	N = 120				$\Sigma fu = 44$



## NOTES

Now,  $\bar{x} = A + \left( \frac{\sum fu}{N} \right) h$

$\therefore \bar{x} = 20 + \left( \frac{44}{120} \right) 5 = 20 + 1.833 = 21.833.$

$$\frac{N+1}{2} = \frac{20+1}{2} = 60.5$$

$\therefore$  Median = size of 60.5th item =  $\frac{20+20}{2} = 20.$

$\therefore$  Mode = 3 median - 2  $\bar{x}$  = 3(20) - 2(21.833) = **16.334.**

**Example 1.26.** Find the mode for the following frequency distribution.

Class	0—5	5—10	10—15	15—20	20—25
f	6	9	4	2	10
Class	25—30	30—35	35—40	40—45	45—50
f	8	7	5	1	3

**Solution.** We find the 'modal class' by using the 'method of grouping'.

**Grouping Table**

Class	f	I	II	III	IV	V	VI
0—5	6						
5—10	9		15		19		
10—15	4		6	13		15	
15—20	2			12			16
20—25	<b>10</b>				<b>20</b>		
25—30	8		<b>18</b>			<b>25</b>	
30—35	7			<b>15</b>			<b>20</b>
35—40	5		12		13		
40—45	1			6		9	
45—50	3		4				

**Analysis Table**

Column	20—25	25—30	30—35	15—20	35—40
I	1				
II	1	1			
III		1	1		
IV	1	1		1	
V	1	1	1		
VI		1	1		1
Total	4	5	3	1	1

Since the total is maximum for the class 25—30, the modal class is 25—30.

## NOTES

Now 
$$\text{mode} = L + \left( \frac{\Delta_1}{\Delta_1 + \Delta_2} \right) h$$

Here  $L = 25, \Delta_1 = 10 - 8 = 2, \Delta_2 = 8 - 7 = 1, h = 5.$

$\therefore \text{Mode} = 25 + \left( \frac{2}{2+1} \right) 5 = 25 + 3.333 = 28.333.$

**Example 1.27.** If the mode and mean of a moderately asymmetrical series are 16 m and 15.6 m respectively, what would be its most probable median?

**Solution.** We have mode = 16 m and mean = 15.6 m.

The formulae is mode = 3 median - 2 A.M.

$\therefore 16 = 3 \text{ median} - 2(15.6)$

$\Rightarrow 3 \text{ median} = 16 + 31.2 = 47.2$

$\therefore \text{median} = \frac{47.2}{3} = 15.73 \text{ m.}$

**Example 1.28.** What are the relationships between mathematical averages?

**Solution.** The following are the relations between mathematical averages:

(I) A.M.  $\geq$  G.M.  $\geq$  H.M.

In particular, if all the items are identical, then

$$\text{A.M.} = \text{G.M.} = \text{H.M.}$$

(II) A.M., G.M. and H.M. are in geometric progression i.e.,

$$(\text{G.M.})^2 = (\text{A.M.})(\text{H.M.})$$

(III) Mode = 3 Median - 2 A.M. (Approximately).

## 1.24. MODE IN CASE OF CLASSES OF UNEQUAL WIDTHS

When the values of the variable are given in the form of classes and the classes are not of equal width, then we would not be able to proceed directly to find the modal class either by the method of inspection or by the method of grouping. In fact, we are to compare the frequencies of different classes in order to observe the concentration of items about some item. If the classes happen to be of unequal width, then we would not be able to compare the frequencies in different classes. To make the comparison meaningful, we will first make classes of equal width by grouping two or more classes or by breaking classes, as per the need.

**Example 1.29.** Calculate median and mode for the following data:

Class	2	3	4	5-7	7-10	10-15	15-20	20-25
Frequency	1	2	2	3	5	10	8	4

**Solution.** We make classes as 0—5, 5—10, 10—15, 15—20 and 20—25.

Class	Frequency $f$	c.f.
0—5	$1 + 2 + 2 = 5$	5
5—10	$3 + 5 = 8$	13
10—15	10	23
15—20	8	31
20—25	4	35 = N
	N = 35	

## NOTES

### Calculation of Median

$$\frac{N}{2} = \frac{35}{2} = 17.5$$

∴ Median = size of 17.5th item

∴ Median class is 10—15.

$$\therefore \text{Median} = L + \left( \frac{N/2 - c}{f} \right) h = 10 + \left( \frac{17.5 - 13}{10} \right) \cdot 5 = 10 + 2.25 = 12.25.$$

### Calculation of Mode

By inspection, modal class is 10—15.

$$\text{Now} \quad \text{Mode} = L + \left( \frac{\Delta_1}{\Delta_1 + \Delta_2} \right) h$$

Here,  $L = 10$ ,  $\Delta_1 = 10 - 8 = 2$ ,  $\Delta_2 = 10 - 8 = 2$ ,  $h = 5$ .

$$\therefore \text{Mode} = 10 + \left( \frac{2}{2 + 2} \right) 5 = 10 + 2.5 = 12.5.$$

### Merits of Mode

1. It is easy to compute.
2. It is not affected by the extreme items.
3. It can be located graphically.

### Demerits of Mode

1. It is not simple to understand.
2. It is not well defined. There are number of formulae to calculate mode, not necessarily giving the same answer.
3. It is not capable of further algebraic treatment.



## EXERCISE 1.5

## NOTES

1. Find the mode for the following series:

3, 5, 6, 2, 5, 4, 5, 9, 5.

2. Calculate the mode for the following frequency distribution:

<i>x</i>	4	5	6	7	8	9	10	11	12	13
<i>f</i>	2	5	8	9	12	14	14	15	11	13

3. The number of fully formed apples on 100 plants were counted with the following results:

2	plants	had	0	apples
5	"	"	1	"
7	"	"	2	"
11	"	"	3	"
18	"	"	4	"
24	"	"	5	"
12	"	"	6	"
8	"	"	7	"
6	"	"	8	"
4	"	"	9	"
3	"	"	10	"

- (i) How many apples are there?  
(ii) What is the average number of apples per plant?  
(iii) What is the modal number of apples?

4. Find the mode for the following frequency distribution:

<i>Marks</i>	0—5	5—10	10—15	15—20	20—25	25—30	30—35
<i>No. of students</i>	11	20	31	45	30	12	6

5. Calculate the modal value for the following frequency distribution:

<i>Marks</i>	<i>No. of candidates</i>	<i>Marks</i>	<i>No. of candidates</i>
0—9	6	50—59	263
10—19	29	60—69	133
20—29	87	70—79	43
30—39	181	80—89	9
40—49	247	90—99	2

6. Obtain the mean, median and mode for the following series:

<i>Marks</i>	10—25	25—40	40—55	55—70	70—85	85—100
<i>Frequency</i>	6	20	44	26	3	1

7. Find the mean, median and mode for the following distribution:

<i>Wages (in ₹)</i>	5—15	15—25	25—35	35—45	45—55	55—65
<i>No. of employees</i>	4	6	10	5	3	2

## NOTES

8. Calculate mode for the following distribution:

Class	0—4	4—6	6—8	8—12	12—18	18—20
Frequency	4	6	8	12	7	2

9. Calculate the median and mode for the following distribution:

Class	Frequency	Class	Frequency
10—20	4	10—60	124
10—30	16	10—70	137
10—40	56	10—80	146
10—50	97	10—90	150

10. Calculate median and mode from the following table:

Income	100—200	100—300	100—400	100—500	100—600
No. of persons	15	33	63	53	100

### Answers

1. 5                      2. 10                      3. (i) 486 (ii) 4.86 (iii) 5  
 4. 17.414              5. 47.5488              6. 47.95 marks, 48.18 marks, 48.57 marks  
 7. ₹ 31, ₹ 30, ₹ 29.44    8. 7                      9. 44.63, 40.67  
 10. 356.67, 354.55

## 1.25. SUMMARY

- The part of the subject statistics which deals with the analysis of a given group and drawing conclusions about a larger group is called **inferential statistics**.
- Instead of examining the entire group, we concentrate on a small part of the group called a **sample**. If this sample happen to be a true representative of the entire group, called **population**, important conclusions can be drawn from the analysis of the sample.
- This is the most popular and widely used measure of central tendency. The popularity of this average can be judged from the fact that it is generally referred to as 'mean'. The **arithmetic mean** of a statistical data is defined as the quotient of the sum of all the values of the variable by the total number of items and is generally denoted by  $\bar{x}$ .
- If  $\bar{x}_1$  and  $\bar{x}_2$  are the A.M. of two groups having  $n_1$  and  $n_2$  items, then the A.M. ( $\bar{x}$ ) of the combined group is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

- If all the values of the variable are not of equal importance, or in other words, these are of varying significance, then we calculate **weighted A.M.**

$$\text{Weighted A.M.} = \bar{x}_w = \frac{\sum wx}{\sum w}$$

where  $w_1, w_2, \dots, w_n$  are the weights of the values  $x_1, x_2, \dots, x_n$  of the variable, under consideration.

## NOTES

- The **geometric mean** of a statistical data is defined as the  $n$ th root of the product of all the  $n$  values of the variable.

For an individual series, the G.M. is given by

$$\text{G.M.} = (x_1 x_2 \dots x_n)^{1/n}$$

- If all the values of the variable are not of equal importance, or in other words, these are of varying significance, then we calculate **weighted G.M.**

$$\text{Weighted G.M.} = \text{Antilog} \left( \frac{\sum w \log x}{\sum w} \right),$$

where  $w_1, w_2, \dots, w_n$  are the weights of the values  $x_1, x_2, \dots, x_n$  of the variable, under consideration.

- The **harmonic mean** of a statistical data is defined as the quotient of the number of items by the sum of the reciprocals of all the values of the variable.
- If  $H_1$  and  $H_2$  are the H.M. of two groups having  $n_1$  and  $n_2$  items, then the H.M. of the combined group is given by

$$H = \frac{n_1 + n_2}{\frac{n_1}{H_1} + \frac{n_2}{H_2}}$$

- If all the values of the variable are not of equal importance or in other words, these are of varying importance, then we calculate **weighted H.M.**

$$\text{Weighted H.M.} = \frac{\sum w}{\sum \left( \frac{w}{x} \right)}$$

where  $w_1, w_2, \dots, w_n$  are the weights of the values  $x_1, x_2, \dots, x_n$  of the variable, under consideration.

- The **median** of a statistical series is defined as the size of the middle most item (or the A.M. of two middle most items), provided the items are in order of magnitude.
- The **mode** of a statistical series is defined as that value of the variable around which the values of the variable tend to be most heavily concentrated. It can also be defined as that value of the variable whose own frequency is dominating and at the same time, the frequencies of its neighbouring items are also dominating.
- In case of a multi-modal distribution, we find the value of mode by using the relation

$$\text{Mode} = 3 \text{ Median} - 2 \text{ A.M.}$$

This mode is called **empirical mode** in the sense that this relation cannot be established algebraically.

## 1.26. REVIEW EXERCISES

- What are the properties of median?
- What are the requisites of a good average?
- What do you mean by 'Central Tendency'? What are the desirable properties for an average to possess?
- Give different measures of central tendency with their formulae. Also state the situations where these measures are used.
- What are the desirable properties of an average? Which of the averages you know possesses most of them?



## 2. MEASURES OF DISPERSION

### STRUCTURE

- 2.1. Introduction
- 2.2. Requisites of a Good Measure of Dispersion
- 2.3. Methods of Measuring Dispersion

#### I. Range

- 2.4. Definition

#### II. Quartile Deviation (Q.D.)

- 2.5. Inadequacy of Range
- 2.6. Definition

#### III. Mean Deviation (M.D.)

- 2.7. Definition
- 2.8. Coefficient of M.D.
- 2.9. Short-cut Method for M.D.

#### IV. Standard Deviation (S.D.)

- 2.10. Definition
- 2.11. Coefficient of S.D., C.V., Variance
- 2.12. Short-cut Method for S.D.
- 2.13. Relation Between Measures of Dispersion
- 2.14. Summary
- 2.15. Review Exercises

### 2.1. INTRODUCTION

We have already seen that an average of a statistical series is a representative of the series. It tells us about the concentration of the items about an average value of the distribution. Let us consider the following series:

I	10	10	10	10	10
II	10	9	11	12	8
III	1	45	1	2	1

**NOTES**

In all the three series, there are five items in each and A.M. of each series is  $50/5 = 10$ . But there is a lot of difference in their formation. In the first series, all the items are coinciding with 10, i.e., the A.M. and there are no deviations of items from A.M. In the second series, the deviations are very small in magnitude. In the third series, we find that the deviations are very large and it is not justified to keep 10 as the average of the series. Thus, we see that the number of items and A.M. of all the series are the same, but even then there is lot of difference in their formation.

---

## **2.2. REQUISITES OF A GOOD MEASURE OF DISPERSION**

---

The requisites of a good measure of a dispersion are the same as those for a good measure of central tendency. For the sake of completeness, we list the requisites as under :

1. It should be simple to understand.
2. It should be easy to compute.
3. It should be well-defined.
4. It should be based on all the items.
5. It should not be unduly affected by the extreme items.
6. It should be capable of further algebraic treatment.
7. It should have sampling stability.

---

## **2.3. METHODS OF MEASURING DISPERSION**

---

- I. Range
- II. Quartile Deviation (Q.D.)
- III. Mean Deviation (M.D.)
- IV. Standard Deviation (S.D.)
- V. Lorenz Curve.

**I. RANGE**

---

## **2.4. DEFINITION**

---

The **range** of a statistical data is defined as the difference between the largest and the smallest values of the variable.

$$\therefore \text{Range} = L - S,$$

where  $L$  = largest value of the variable

$S$  = smallest value of the variable.

In case, the values of the variable are given in the form of classes, then  $L$  is taken as the upper limit of the largest value class and  $S$  as the lower limit of the smallest value class.



Age (in years)	16—18	18—20	20—22	22—24	24—26	26—28
No. of students	0	4	6	8	2	2

**Solution.** Here  $L = 28, S = 18$

$\therefore$  Range =  $L - S = 28 - 18 = 10$  years.

It may be noted that  $S \neq 16$ , though it is the lower limit of the smallest value class, but there is no item in this class and so this class is meaningless so far as the calculation of range is concerned.

Let us consider the market value of shares of companies A and B, during a particular week.

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
M.V. of shares of company A (in ₹)	12	11	10	13	16	20
M.V. of shares of company B (in ₹)	60	50	55	62	70	75

From the data, we see that Range (A) =  $20 - 10 = ₹ 10$  and Range (B) =  $75 - 50 = ₹ 25$ . From these results, one is likely to infer that there is more variability in the II series. But this is not so, because the M.V. of shares of A has increased by 100% in the week, whereas there is only 50% rise in the M.V. of shares of B, during that week. Thus, variability is more in the first series. Thus, we see that range may give misleading results if used for comparing two or more series for variability (scatteredness, dispersion). For comparison purpose, we use its corresponding relative measure, called 'coefficient of range'. This is defined as

$$\text{Coeff. of Range} = \frac{L - S}{L + S}.$$

$$\text{Now Coeff. of Range for A} = \frac{20 - 10}{20 + 10} = \frac{10}{30} = 0.3333.$$

$$\text{Coeff. of Range for B} = \frac{75 - 50}{75 + 50} = \frac{25}{125} = 0.2000.$$

$\therefore$  Coeff. of Range (A) > Coeff. of Range (B)

$\therefore$  Variability is more in the M.V. of shares of company A.

### Merits of Range

1. It is simple to understand.
2. It is easy to compute.
3. It is well-defined.
4. It helps in giving an idea about the variation, just by giving the lowest value and the greatest value of variable.

### NOTES

**Demerits of Range****NOTES**

1. It is not based on all the items.
2. It is highly affected by the extreme items. In fact, if extreme items are present, then range would be calculated by taking only extreme items.
3. It does not take into account the frequencies of items in the middle of the series.
4. It is not capable of further algebraic treatment.
5. It does not have sampling stability.

**EXERCISE 2.1**

1. Calculate the range for the following series:

17, 10, 12, 8, 12, 16, 19.

2. Find the value of range for the following frequency distribution:

Age (in years)	14	15	16	17	18	19	20
No. of students	1	2	2	2	6	4	0

3. Compare the following series for variability:

Days	M	T	W	T	F	S
M.V. of shares of company X (in ₹)	48	47	46	49	43	45
M.V. of shares of company Y (in ₹)	10	9	12	12	14	12

**Answers**

1. 11
2. 5 years
3.  $\left. \begin{array}{l} \text{Coeff. of Range (X)} = 0.0652 \\ \text{Coeff. of Range (Y)} = 0.2174 \end{array} \right\}$  Variability is more in the second series.

**II. QUARTILE DEVIATION (Q.D.)****2.5. INADEQUACY OF RANGE**

Consider the series

I: 4, 4, 4, 5, 5, 6, 4, 5, 5, 1000.  
 II: 4, 4, 4, 5, 5, 6, 4, 5, 5.

$$\text{For series I, Coeff. of Range} = \frac{1000 - 4}{1000 + 4} = \frac{996}{1004} = 0.992$$

$$\text{For series II, Coeff. of Range} = \frac{6 - 4}{6 + 4} = \frac{2}{10} = 0.200.$$



On comparing the values of coeff. of range for these series, one is likely to conclude that there is marked difference in variability in the series. In fact, the series II is obtained from the series I, just by ignoring the extreme item 1000. Thus, we see that extreme items can distort the value of range and even the coefficient of range. If we have a glance at the definitions of these measures, we would find that only extreme items are required in their calculation, if at all extreme items are present. Even if extreme items are present in a series, the middle 50% values of the variable would be expected to vary quite smoothly, keeping this in view, we define another measure of dispersion, called 'Quartile Deviation'.

## NOTES

## 2.6. DEFINITION

The **quartile deviation** of a statistical data is defined as

$$\frac{Q_3 - Q_1}{2} \text{ and is denoted as Q.D.}$$

This is also called *semi-inter quartile range*. We have already studied the method of calculating quartiles. The value of Q.D. is obtained by subtracting  $Q_1$  from  $Q_3$  and then dividing it by 2.

For comparing two or more series for variability, the absolute measure Q.D. would not work. For this purpose, the corresponding relative measure, called coeff. of Q.D. is calculated. This is defined as:

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

**Example 2.2.** Find Q.D. and its coefficient for the following series:

$x$  (in ₹): 4, 7, 6, 5, 9, 12, 19.

**Solution.** The values of the variable arranged in ascending order are

$x$  (in ₹): 4, 5, 6, 7, 9, 12, 19.

Here  $n = 7$ .

$$Q_1: \frac{n+1}{4} = \frac{7+1}{4} = 2 \quad \therefore Q_1 = \text{size of 2nd item} = ₹ 5$$

$$Q_3: 3 \left( \frac{n+1}{4} \right) = 3 \left( \frac{7+1}{4} \right) = 6 \quad \therefore Q_3 = \text{size of 6th item} = ₹ 12$$

$$\therefore \text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{12 - 5}{2} = ₹ 3.5.$$

$$\text{Coeff. of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{12 - 5}{12 + 5} = \frac{7}{17} = 0.4118.$$

**Example 2.3.** For the following data, calculate:

- the coefficient of range
- interquartile range, and
- percentile range

Marks	5—9	10—14	15—19	20—24
No. of students	1	3	8	5
Marks	25—29	30—34	35—39	
No. of students	4	2	2	

**Solution.** The first and the last classes in the exclusive form are 4.5—9.5 and 34.5—39.5 respectively.

$$\therefore \text{Coeff. of range} = \frac{L-S}{L+S} = \frac{39.5-4.5}{39.5+4.5} = \frac{35}{44} = 0.7955.$$

**NOTES****Calculation of  $Q_1$ ,  $Q_3$ ,  $P_{10}$ ,  $P_{90}$** 

Marks	No. of students $f$	c.f.
4.5—9.5	1	1
9.5—14.5	3	4
14.5—19.5	8	12
19.5—24.5	5	17
24.5—29.5	4	21
29.5—34.5	2	23
34.5—39.5	2	25 = N
	N = 25	

$$Q_1: \quad \frac{N}{4} = \frac{25}{4} = 6.25. \quad \therefore Q_1 = \text{size of 6.25th item}$$

$\therefore Q_1$  class is 14.5—19.5

$$\begin{aligned} \therefore Q_1 &= L + \left( \frac{N/4 - c}{f} \right) h = 14.5 + \left( \frac{6.25 - 4}{8} \right) 5 \\ &= 14.5 + 1.4063 = 15.9063 \text{ marks} \end{aligned}$$

$$Q_3: \quad 3\left(\frac{N}{4}\right) = 3\left(\frac{25}{4}\right) = 18.75 \quad \therefore Q_3 = \text{size of 18.75th item}$$

$\therefore Q_3$  class is 24.5—29.5

$$\begin{aligned} \therefore Q_3 &= L + \left( \frac{3(N/4) - c}{f} \right) h = 24.5 + \left( \frac{18.75 - 17}{4} \right) 5 \\ &= 24.5 + 2.1875 = 26.6875 \text{ marks} \end{aligned}$$

$\therefore$  Interquartile range

$$= Q_3 - Q_1 = 26.6875 - 15.9063 = \mathbf{10.7812 \text{ marks}}$$

Percentile range is defined as  $P_{90} - P_{10}$ .

$$P_{10}: \quad 10\left(\frac{N}{100}\right) = 10\left(\frac{25}{100}\right) = 2.5 \quad \therefore P_{10} = \text{size of 2.5th item}$$

$\therefore P_{10}$  class is 9.5—14.5.

$$\therefore P_{10} = L + \left( \frac{10(N/100) - c}{f} \right) h = 9.5 + \left( \frac{2.5 - 1}{3} \right) 5 = 9.5 + 2.5 = 12 \text{ marks.}$$

$$P_{90}: \quad 90\left(\frac{N}{100}\right) = 90\left(\frac{25}{100}\right) = 22.5 \quad \therefore P_{90} = \text{size of 22.5th item}$$

$\therefore P_{90}$  class is 29.5—34.5.



$$\begin{aligned}
 P_{90} &= L + \left( \frac{90(N/100) - c}{f} \right) h \\
 &= 29.5 + \left( \frac{22.5 - 21}{2} \right) 5 = 29.5 + 3.75 = 33.25 \text{ marks}
 \end{aligned}$$

$\therefore$  Percentile range =  $P_{90} - P_{10} = 33.25 - 12 = 21.25$  marks.

### Merits of Q.D.

1. It is simple to understand.
2. It is easy to calculate.
3. It is well-defined.
4. It helps in studying the middle 50% items in the series.
5. It is not affected by the extreme items.
6. It is useful in the case of open end classes.

### Demerits of Q.D.

1. It is not based on all the items.
2. It is not capable of further algebraic treatment.
3. It does not have sampling stability.

### EXERCISE 2.2

1. Find the Q.D. and its coefficient for the given data regarding the age of 7 students.  
Age (in years): 17, 19, 22, 26, 19, 28, 17.
2. Compare the following two series of figures in respect of their dispersion by quartile measures:

Height (in inches)	58	56	62	61	63	64	65	59	62	65	55
Weight (in pounds)	117	112	127	123	125	130	106	119	121	132	108

3. Calculate the coefficient of Q.D. of the marks of 39 students in statistics given below:

Marks	0—5	5—10	10—15	15—20	20—25	25—30
No. of students	4	6	8	12	7	2

4. Calculate the values of Q.D. and its coefficient for the following data:

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

### NOTES

## NOTES

5. Find Quartile deviation for the following data:

Mid-point	2	3	4	5	6	7	8	9	10	11
Frequency	2	3	5	6	8	12	16	7	5	4

## Answers

1. Q.D. = 4.5 years, Coeff. of Q.D. = 0.2093
2. Coeff. of Q.D. (Height) = 0.0492,  
Coeff. of Q.D. (Weight) = 0.0628  
Variability is more in the II series.
3. 0.3356
4. Q.D. = 5.2083, Coeff. of Q.D. = 0.2643
5. Q.D. = 1.406.

III. MEAN DEVIATION (M.D.)

---

## 2.7. DEFINITION

---

Mean deviation is also called **average deviation**. The **mean deviation** of a statistical data is defined as the arithmetic mean of the numerical values of the deviations of items from some average. Generally, A.M. and median are used in calculating mean deviation. Let ' $a$ ' stand for the average used for calculating M.D.

For an **individual series**, the M.D. is given by

$$\text{M.D.} = \frac{\sum_{i=1}^n |x_i - a|}{n} = \frac{\sum |x - a|}{n}$$

where  $x_1, x_2, \dots, x_n$  are the values of the variable, under consideration.

For a **frequency distribution**,

$$\text{M.D.} = \frac{\sum_{i=1}^n f_i |x_i - a|}{N} = \frac{\sum f |x - a|}{N}$$

where  $f_i$  is the frequency of  $x_i$  ( $1 \leq i \leq n$ ).

When the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.

Median is used in calculating M.D., because of its property that the sum of numerical values of deviations of items from median is always least. So, if median is used in the calculation of M.D., its value would come out to be least. M.D. is also calculated by using A.M. because of its simplicity and popularity. In problems, it is generally given as to which average is to be used in the calculation of M.D. If it is not given, then either of the two can be made use of.



## 2.8. COEFFICIENT OF M.D.

For comparing two or more series for variability, the corresponding relative measure, 'Coefficient of M.D.', is used. This is defined as:

$$\text{Coeff. of M.D.} = \frac{\text{M.D.}}{\text{Average}}$$

If M.D. is calculated about A.M., then M.D. is written as M.D. ( $\bar{x}$ ). Similarly, M.D. (Median) would mean that median has been used in calculating M.D.

∴ We can write

$$\text{Coeff. of M.D.}(\bar{x}) = \frac{\text{M.D.}(\bar{x})}{\bar{x}}$$

$$\text{Coeff. of M.D.}(\text{Median}) = \frac{\text{M.D.}(\text{Median})}{\text{Median}}$$

### WORKING RULES TO FIND M.D. ( $\bar{x}$ )

**Rule I.** In case of an individual series, first find  $\bar{x}$  by using the formula  $\bar{x} = \frac{\Sigma x}{n}$ .

In the second step, find the values of  $x - \bar{x}$ . In the next step, find the numerical values  $|x - \bar{x}|$  of  $x - \bar{x}$ . Find the sum  $\Sigma |x - \bar{x}|$  of these numerical values  $|x - \bar{x}|$ . Divide this sum by  $n$  to get the value of M.D. ( $\bar{x}$ ).

**Rule II.** In case of a frequency distribution, first find  $\bar{x}$  by using the formula  $\bar{x} = \frac{\Sigma fx}{N}$ . In the second step, find the values of  $x - \bar{x}$ . In the next step,

find the numerical values  $|x - \bar{x}|$  of  $x - \bar{x}$ . Find the products of the values of  $|x - \bar{x}|$  and their corresponding frequencies. Find the sum  $\Sigma f|x - \bar{x}|$  of these products. Divide this sum by  $N$  to get the value of M.D. ( $\bar{x}$ ).

**Rule III.** If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.

**Rule IV.** To find the coefficient of M.D. ( $\bar{x}$ ), divide M.D. ( $\bar{x}$ ) by  $\bar{x}$ .

**Remarks:** Similar working rules are followed to find the values of M.D. (Median) and coefficient of M.D. (Median).

**Example 2.4.** Find the M.D. from A.M. for the following data:

$x$	3	5	7	9	11	13
$f$	2	7	10	9	5	2

## NOTES

## NOTES

## Solution.

Calculation of M.D. ( $\bar{x}$ )

$x$	$f$	$fx$	$x - \bar{x}$	$ x - \bar{x} $	$f x - \bar{x} $
3	2	6	-4.8	4.8	9.6
5	7	35	-2.8	2.8	19.6
7	10	70	-0.8	0.8	8.0
9	9	81	1.2	1.2	10.8
11	5	55	3.2	3.2	16.0
13	2	26	5.2	5.2	10.4
	$N = 35$	$\Sigma fx = 273$			$\Sigma f x - \bar{x}  = 74.4$

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{273}{35} = 7.8$$

Now  $M.D.(\bar{x}) = \frac{\Sigma f|x - \bar{x}|}{N} = \frac{74.4}{35} = 2.1257.$

**Example 2.5.** Find the coeff. of M.D.(Median) for the following frequency distribution:

Marks	0—10	10—20	20—30	30—40	40—50
No. of students	5	8	15	16	6

## Solution.

## Calculation of M.D. (Median)

Marks	No. of students ( $f$ )	c.f.	Mid-points of classes $x$	$x$ -median (med. = 28)	$ x - med.  $	$f x - med.  $
0—10	5	5	5	-23	23	115
10—20	8	13	15	-13	13	104
20—30	15	28	25	-3	3	45
30—40	16	44	35	7	7	112
40—50	6	50 = N	45	17	17	102
	$N = 50$					$\Sigma f x - med.  $ $= 478$

Median = size of 50/2th item = size of 25th item.

$\therefore$  Median class is 20—30

$$\text{Median} = L + \left( \frac{N/2 - c}{f} \right) h = 20 + \left( \frac{25 - 13}{15} \right) 10 = 28$$

Now  $M.D.(\text{Median}) = \frac{\Sigma f|x - \text{median}|}{N} = \frac{478}{50} = 9.56 \text{ marks.}$

$\therefore$  Coeff. of M.D.(Median) =  $\frac{M.D.(\text{Median})}{\text{Median}} = \frac{9.56}{28} = 0.3414.$

## 2.9. SHORT-CUT METHOD FOR M.D.

We know that the calculation of M.D. involve taking of deviations of items from some average. If the value of the average under consideration is a whole number, we can easily take the deviations and proceed without any difficulty. But in case, the value of the average comes out to be in decimal like 18.6747, the calculation of M.D. would become quite tedious. In such a case, we would have to approximate the value of the average up to one or two places of decimal for otherwise we would have to bear the heavy calculation work involved. If the value of the average is in decimal, the following short-cut method is preferred.

$$\text{M.D.} = \frac{(\Sigma fx)_A - (\Sigma fx)_B - ((\Sigma f)_A - (\Sigma f)_B) a}{N}$$

where 'a' is the average about which M.D. is to be calculated. In this formula, suffixes A and B denote the sums corresponding to the values of  $x \geq a$  and  $x < a$  respectively.

This formula can also be used for an individual series, by taking 'f' equal to 1 for each x, in the series. In this case, the formula reduces to

$$\text{M.D.} = \frac{(\Sigma x)_A - (\Sigma x)_B - ((n)_A - (n)_B) a}{n}$$

where  $(n)_A$  and  $(n)_B$  are the number of items whose values are greater than or equal to a and less than a respectively.

If short-cut method is to be used to find M.D. ( $\bar{x}$ ), then it is advisable to use *direct method* to find  $\bar{x}$ , because we would be needing  $(\Sigma fx)_A$  and  $(\Sigma fx)_B$  in the calculation of M.D. ( $\bar{x}$ ).

**Example 2.6.** Calculate M.D. (Median) for the following data :

x: 4, 6, 10, 12, 18, 19.

**Solution.** Calculation of M.D. (Median)

S. No.	x		x - median	x - median
1	4	$(\Sigma x)_B$ = 20	-7	7
2	6		-5	5
3	10		-1	1
4	12	$(\Sigma x)_A$ = 49	1	1
5	18		7	7
6	19		8	8
n = 6				$\Sigma   x - \text{median}   = 29$

$$\text{Median} = \text{size of } \frac{6+1}{2} \text{ th item} = \text{size of 3.5th item} = \frac{10+12}{2} = 11.$$

**Direct Method**

$$\text{M.D. (Median)} = \frac{\Sigma | x - \text{median} |}{n} = \frac{29}{6} = 4.8333.$$

**Short-cut Method**

$$\begin{aligned} \text{M.D. (Median)} &= \frac{(\Sigma x)_A - (\Sigma x)_B - ((n)_A - (n)_B) \text{ median}}{n} \\ &= \frac{49 - 20 - (3 - 3) \cdot 11}{6} = \frac{29}{6} = 4.8333. \end{aligned}$$

## NOTES

**Example 2.7.** Calculate M.D. ( $\bar{x}$ ) and its coefficient for the following data :

## NOTES

Profit (in ₹)	No. of firms	Profit (in ₹)	No. of firms
5000—6000	10	0—1000	4
4000—5000	15	- 1000 to 0	6
3000—4000	30	- 2000 to - 1000	8
2000—3000	10	- 3000 to - 2000	10
1000—2000	5		

**Solution.**

**Calculation of M.D. ( $\bar{x}$ )**

Profit (in ₹)	No. of firms (f)	x	fx
- 3000 to - 2000	10	- 2500	- 25000
- 2000 to - 1000	8	- 1500	- 12000
- 1000 to 0	6	- 500	- 3000
0—1000	4	500	2000
1000—2000	5	1500	7500
			$(\Sigma fx)_B = - 30500$
2000—3000	10	2500	25000
3000—4000	30	3500	105000
4000—5000	15	4500	67500
5000—6000	10	5500	55000
			$(\Sigma fx)_A = 252500$
N = 98			$\Sigma fx = 222000$

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{222000}{98} = \text{Rs. } 2265.3061$$

$$\begin{aligned} \text{Now M.D.}(\bar{x}) &= \frac{(\Sigma fx)_A - (\Sigma fx)_B - [(\Sigma f)_A - (\Sigma f)_B] \bar{x}}{N} \\ &= \frac{252500 - (- 30500) - (65 - 33) 2265.3061}{98} \\ &= \frac{210510.21}{98} = ₹ 2148.0633 \end{aligned}$$

$$\text{Coeff. of M.D.}(\bar{x}) = \frac{\text{M.D.}(\bar{x})}{\bar{x}} = \frac{2148.0633}{2265.3061} = 0.9482.$$

**Merits of M.D.**

1. It is simple to understand.
2. It is easy to compute.
3. It is well-defined.
4. It is based on all the items.
5. It is not unduly affected by the extreme items.
6. It can be calculated by using any average.

## Demerits of M.D.

1. It is not capable of further algebraic treatment.
2. It does not take into account the signs of the deviations of items from the average value.

## NOTES

### EXERCISE 2.3

1. Calculate M.D. ( $\bar{x}$ ) and its coefficient for the following individual series:

21, 23, 25, 28, 30, 32, 38, 39, 46, 48.

2. Compute M.D. ( $\bar{x}$ ) for the following data:

Marks	10	15	20	25	30
No. of students	2	4	6	8	5

3. Find the mean deviation about median for the following data:

$x$	6	12	18	24	30	36	42
$f$	4	7	9	18	15	10	5

4. Find the mean deviation about the mean for the following frequency distribution:

Class	0—4	4—8	8—12	12—16	16—20
$f$	4	6	8	5	2

5. Calculate M.D. about A.M. and also about median for the following data:

Income per week (in ₹)	20—30	30—40	40—50	50—60	60—70
No. of families	120	201	150	75	25

6. Calculate coefficient of mean deviation and coefficient of median deviation for the following:

Marks	140—150	150—160	160—170
No. of students	4	6	10
Marks	170—180	180—190	190—200
No. of students	18	9	3

7. Find M.D. and coefficient of M.D. about median for the following data:

Size	5	6	7	8	9	10
Frequency	8	12	18	8	3	1

### Answers

1. M.D. ( $\bar{x}$ ) = 7.8, coeff. of M.D. ( $\bar{x}$ ) = 0.2364.
2. M.D. ( $\bar{x}$ ) = 5.12 marks.
3. 7.5
4. 3.84



5. M.D. ( $\bar{x}$ ) = ₹ 9.22, M.D. (median) = ₹ 9.07.
6. Coeff. of M.D. ( $\bar{x}$ ) = 0.062, Coeff. of M.D. (median) = 0.059.
7. M.D. (median) = 0.9, Coeff. of M.D. (median) = 0.1286.

## NOTES

## IV. STANDARD DEVIATION (S.D.)

**2.10. DEFINITION**

It is the most important measure of dispersion. It finds indispensable place in advanced statistical methods. The **standard deviation** of a statistical data is defined as the positive square root of the A.M. of the squared deviations of items from the A.M. of the series under consideration. The S.D. is often denoted by the greek letter ' $\sigma$ '.

For an **individual series**, the S.D. is given by

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

where  $x_1, x_2, \dots, x_n$  are the value of the variable, under consideration.

For a **frequency distribution**,

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum f(x - \bar{x})^2}{N}}$$

where  $f_i$  is the frequency of  $x_i$  ( $1 \leq i \leq n$ ).

When the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.

**2.11. COEFFICIENT OF S.D., C.V., VARIANCE**

For comparing two or more series for variability, the corresponding relative measure, called coefficient of S.D. is calculated. This measure is defined as:

$$\text{Coefficient of S.D.} = \frac{\text{S.D.}}{\bar{x}}$$

The product of coefficient of S.D. and 100 is called as the *coefficient of variation*.

$$\therefore \text{Coefficient of variation} = \left( \frac{\text{S.D.}}{\bar{x}} \right) 100.$$

This measure is denoted as C.V.

$$\therefore \text{C.V.} = \left( \frac{\text{S.D.}}{\bar{x}} \right) 100.$$

In practical problems, we prefer comparing C.V. instead of comparing coefficient of S.D. The coefficient of variation is also represented as percentage. The square of S.D. is called the **variance** of the distribution.

**WORKING RULES TO FIND S.D.**

**Rule I.** In case of an individual series, first find  $\bar{x}$  by using the formula  $\bar{x} = \frac{\Sigma x}{n}$ . In the second step, find the values of  $x - \bar{x}$ . In the next step, find the

squares  $(x - \bar{x})^2$  of the values of  $x - \bar{x}$ . Find the sum  $\Sigma (x - \bar{x})^2$  of the values of  $(x - \bar{x})^2$ . Divide this sum by  $n$ . Take the positive square root of this to get the value of S.D.

**Rule II.** In case of a frequency distribution, first find  $\bar{x}$  by using the formula  $\bar{x} = \frac{\Sigma fx}{N}$ . In the second step, find the values of  $x - \bar{x}$ . In the next step, find

the squares  $(x - \bar{x})^2$  of the values of  $x - \bar{x}$ . Find the products of the values of  $(x - \bar{x})^2$  and their corresponding frequencies. Find the sum  $\Sigma f(x - \bar{x})^2$  of these products. Divide this sum by  $N$ . Take the positive square root of this to get the value of S.D.

**Rule III.** If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.

**Rule IV.** (i) Coeff. of S.D. =  $\frac{S.D.}{A.M.}$

(ii) Coeff. of variation (C.V.) =  $\frac{S.D.}{A.M.} \times 100$

(iii) Variance =  $(S.D.)^2$ .

**NOTES**

**Example 2.8.** Calculate S.D. and C.V. for the following data:

$x$	5	15	25	35	45	55
$f$	12	18	27	20	17	6

**Solution.**

**Calculation of S.D. and C.V.**

$x$	$f$	$fx$	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
5	12	60	-23	529	6348
15	18	270	-13	169	3042
25	27	675	-3	9	243
35	20	700	7	49	980
45	17	765	17	289	4913
55	6	330	27	729	4374
	$N = 100$	$\Sigma fx = 2000$			$\Sigma f(x - \bar{x})^2 = 19900$

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{2000}{100} = 20.$$

$$\text{Now S.D.} = \sqrt{\frac{\sum f(x - \bar{x})^2}{N}} = \sqrt{\frac{19900}{100}} = \sqrt{199} = 14.1067.$$

$$\text{C.V.} = \left( \frac{\text{S.D.}}{\bar{x}} \right) 100 = \left( \frac{14.1067}{28} \right) 100 = 50.3811\%.$$

**NOTES**

**Example 2.9.** The mean of 5 observations is 4 and variance is 5.2. If three of the five observations are 1, 2 and 6, find the other two.

**Solution.** Given observations are 1, 2, 6. Let the other two observations be  $a$  and  $b$ .

$$\text{A.M.} = 4 \Rightarrow \frac{\sum x}{n} = 4$$

$$\Rightarrow \frac{1+2+6+a+b}{5} = 4 \Rightarrow a+b = 20-9 = 11$$

$$\therefore a+b = 11 \quad \dots(1)$$

$$\text{Also Variance} = \frac{\sum (x - \bar{x})^2}{n}$$

$$\sum (x - \bar{x})^2 = \sum (x^2 + \bar{x}^2 - 2n\bar{x}) = \sum x^2 + n\bar{x}^2 - 2\bar{x} \sum x$$

$$= \sum x^2 + n\bar{x}^2 - 2\bar{x} \left( \frac{\sum x}{n} \right) n$$

$$= \sum x^2 + n\bar{x}^2 - 2n\bar{x}^2 = \sum x^2 - n\bar{x}^2.$$

$$\therefore \text{Variance} = \frac{\sum x^2 - n\bar{x}^2}{n} = \frac{\sum x^2}{n} - \bar{x}^2.$$

$$\therefore 5.2 = \frac{1^2 + 2^2 + 6^2 + a^2 + b^2}{5} - (4)^2 \Rightarrow 5.2 = \frac{41 + a^2 + b^2}{5} - 16$$

$$\Rightarrow a^2 + b^2 + 41 = (21.2) \times 5 = 106 \Rightarrow a^2 + b^2 = 65 \quad \dots(2)$$

Solving (1) and (2), we get  $a = 4$ ,  $b = 7$ .

## 2.12. SHORT-CUT METHOD FOR S.D.

We have seen in the above examples that the calculations of S.D. involves a lot of computation work. Even if the value of A.M. is a whole number, the calculations are not so simple. In case, A.M. is in decimal, then the calculation work would become more tedious. In problems, where A.M. is expected to be in decimal, we shall use this method, which is based on deviations (or step deviations) of items in the series.

For an individual series  $x_1, x_2, \dots, x_n$ , we have

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^n u_i^2}{n} - \left( \frac{\sum_{i=1}^n u_i}{n} \right)^2} \cdot h = \sqrt{\frac{\sum u^2}{n} - \left( \frac{\sum u}{n} \right)^2} \cdot h$$

$$\text{where } u_i = \frac{x_i - A}{h}, \quad 1 \leq i \leq n.$$

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^n f_i u_i^2}{N} - \left( \frac{\sum_{i=1}^n f_i u_i}{N} \right)^2} \cdot h = \sqrt{\frac{\sum f u^2}{N} - \left( \frac{\sum f u}{N} \right)^2} \cdot h$$

## NOTES

where  $f_i$  is the frequency of  $x_i$  ( $1 \leq i \leq n$ ) and  $u_i = \frac{x_i - A}{h}$ ,  $1 \leq i \leq n$ .

$A$  and  $h$  are constants to be chosen suitably. This method is also known as *step deviation method*.

In practical problems, it is advisable to first take deviations ' $d$ ' of the values of the variable ( $x$ ) from some suitable number ' $A$ '. Then we see if there is any common factor greater than one, in the values of the deviations. If there is a common factor

$h$  ( $> 1$ ), then we calculate  $u = \frac{d}{h} = \frac{x - A}{h}$  in the next column. In case, there is no common

factor greater one, then we take  $h = 1$  and  $u$  becomes  $u = \frac{d}{1} = x - A$ .

In this case, the formula reduces as given below:

$$\text{S.D.} = \sqrt{\frac{\sum d^2}{n} - \left( \frac{\sum d}{n} \right)^2} \quad (\text{Individual Series})$$

$$\text{S.D.} = \sqrt{\frac{\sum f d^2}{N} - \left( \frac{\sum f d}{N} \right)^2} \quad (\text{Frequency Distribution})$$

where  $d = x - A$  and  $A$  is any constant, to be chosen suitably.

## WORKING RULES TO FIND S.D.

**Rule I.** In case of an individual series, choose a number  $A$ . Find deviations  $d (= x - A)$  of items from  $A$ . Find the squares ' $d^2$ ' of the values of  $d$ . Find S.D. by using the formula

$$\sqrt{\frac{\sum d^2}{n} - \left( \frac{\sum d}{n} \right)^2}$$

If some common factor  $h$  ( $> 1$ ) is available in the values of  $d$ , then we calculate ' $u$ ' by dividing the values of  $d$  by  $h$ . Find the squares ' $u^2$ ' of the

values of  $u$ . Find S.D. by using the formula:  $\sqrt{\frac{\sum u^2}{n} - \left( \frac{\sum u}{n} \right)^2} \times h$ .

**Rule II.** In case of a frequency distribution, choose a number  $A$ . Find deviations  $d (= x - A)$  of items from  $A$ . Find the products  $fd$  of  $f$  and  $d$ . Next, find the products of  $fd$  and  $d$ . Find the sums  $\sum fd$  and  $\sum fd^2$ . Find S.D. by using the formula:

$$\sqrt{\frac{\sum f d^2}{N} - \left( \frac{\sum f d}{N} \right)^2}$$

## NOTES

If some common factor  $h(> 1)$  is available in the values of  $d$ , then we calculate ' $u$ ' by dividing the values of  $d$  by  $h$ . Find the product  $fu$  of  $f$  and  $u$ . Next find the products of  $fu$  and  $u$ . Find the sums  $\Sigma fu$  and  $\Sigma fu^2$ . Find S.D. by using the formula:

$$\sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} \times h.$$

**Rule III.** If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.

**Example 2.10.** The scores of two batsmen A and B for 20 innings are tabulated below. Which of the two may be regarded as the more consistent batsman?

Score		50	51	52	53	54	55	56	57
No. of innings	A	1	0	0	4	3	6	3	3
	B	1	2	2	6	3	4	2	0

**Solution.**

**Calculation of C.V. for Batsman A**

Score $x$	No. of innings $f$	$d = x - A$ $A = 53$	$u = d$	$fu$	$fu^2$
50	1	-3	-3	-3	9
51	0	-2	-2	0	0
52	0	-1	-1	0	0
53	4	0	0	0	0
54	3	1	1	3	3
55	6	2	2	12	24
56	3	3	3	9	27
57	3	4	4	12	48
	$N = 20$			$\Sigma fu = 33$	$\Sigma fu^2 = 111$

$$\bar{x} = A + \frac{\Sigma fu}{N} = 53 + \frac{33}{20} = 54.65$$

$$S.D. = \sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} = \sqrt{\frac{111}{20} - \left(\frac{33}{20}\right)^2} = 1.6815$$

$$\therefore \text{C.V. for A} = \left(\frac{S.D.}{\bar{x}}\right) 100 = \left(\frac{1.6815}{54.65}\right) 100 = 3.0768\%.$$



### Calculation of C.V. for Batsman B

Measures of Dispersion

### NOTES

Score $x$	No. of innings $f$	$u = d = x - A$ $A = 53$	$fu$	$fu^2$
50	1	-3	-3	9
51	2	-2	-4	8
52	2	-1	-2	2
53	6	0	0	0
54	3	1	3	3
55	4	2	8	16
56	2	3	6	18
57	0	4	0	0
	$N = 20$		$\Sigma fu = 8$	$\Sigma fu^2 = 56$

$$\bar{x} = A + \frac{\Sigma fu}{N} = 53 + \frac{8}{20} = 53.4$$

$$S.D = \sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} = \sqrt{\frac{56}{20} - \left(\frac{8}{20}\right)^2} = 1.6248$$

$$\therefore \text{C.V. for B} = \left(\frac{S.D.}{\bar{x}}\right) 100 = \left(\frac{1.6248}{53.4}\right) 100 = 3.0427\%.$$

$\therefore$  C.V. for A > C.V. for B

$\therefore$  Batsman B is more consistent.

**Example 2.11.** For the following data, find out which group is more uniform:

Age group (years)	No. of persons	
	Group A	Group B
0—10	5	7
10—20	15	12
20—30	20	22
30—40	25	30
40—50	18	20
50—60	10	5
60—70	7	4

## Solution.

## Calculation of C.V. for group A

## NOTES

Age group (years)	No. of persons $f$	$x$	$d = x - A$ $A = 35$	$u = d/h$ $h = 10$	$fu$	$fu^2$
0—10	5	5	-30	-3	-15	45
10—20	15	15	-20	-2	-30	60
20—30	20	25	-10	-1	-20	20
30—40	25	35	0	0	0	0
40—50	18	45	10	1	18	18
50—60	10	55	20	2	20	40
60—70	7	65	30	3	21	63
	$N = 100$				$\Sigma fu = -6$	$\Sigma fu^2 = 246$

$$\bar{x} = A + \left( \frac{\Sigma fu}{N} \right) h = 35 + \left( \frac{-6}{100} \right) 10 = 34.4$$

$$\text{S.D.} = \sqrt{\frac{\Sigma fu^2}{N} - \left( \frac{\Sigma fu}{N} \right)^2} \cdot h = \sqrt{\frac{246}{100} - \left( \frac{-6}{100} \right)^2} \cdot 10 = 15.6729$$

$$\therefore \text{C.V. for group A} = \frac{\text{S.D.}}{\bar{x}} \times 100$$

$$= \frac{15.6729}{34.4} \times 100 = 45.5608\%$$

## Calculation of C.V. for Group B

Age group (years)	No. of person ( $f$ )	$x$	$d = x - A$ $A = 35$	$u = d/h$ $h = 10$	$fu$	$fu^2$
0—10	7	5	-30	-3	-21	63
10—20	12	15	-20	-2	-24	48
20—30	22	25	-10	-1	-22	22
30—40	30	35	0	0	0	0
40—50	20	45	10	1	20	20
50—60	5	55	20	2	10	20
60—70	4	65	30	3	12	36
	$N = 100$				$\Sigma fu = -25$	$\Sigma fu^2 = 209$

$$\bar{x} = A + \left( \frac{\Sigma fu}{N} \right) h = 35 + \left( \frac{-25}{100} \right) 10 = 32.5$$

$$\text{S.D.} = \sqrt{\frac{\Sigma fu^2}{N} - \left( \frac{\Sigma fu}{N} \right)^2} \cdot h = \sqrt{\frac{209}{100} - \left( \frac{-25}{100} \right)^2} \cdot 10 = 14.2390$$

$$\therefore \text{C.V. for group B} = \frac{\text{S.D.}}{\bar{x}} \times 100 = \frac{14.2390}{32.5} \times 100 = 43.8123\%$$

$\therefore$  C.V. for Group A > C.V. for Group B.

$\therefore$  Group B is more uniform.



**Example 2.12.** The A.M. of the runs scored by three batsmen A, B and C in the same series of 10 innings are 58, 48 and 12 respectively. The S.D. of their runs are respectively 15, 12 and 2. Who is the most consistent of the three? If one of these is to be selected, who will be selected?

**Solution.** We have

$$\bar{x} (A) = 58 \quad \sigma (A) = 15$$

$$\bar{x} (B) = 48 \quad \sigma (B) = 12$$

$$\bar{x} (C) = 12 \quad \sigma (C) = 2$$

$$\therefore \text{C.V. (A)} = \left( \frac{15}{58} \right) 100 = 25.86\%$$

$$\text{C.V. (B)} = \left( \frac{12}{48} \right) 100 = 25.00\%$$

$$\text{C.V. (C)} = \left( \frac{2}{12} \right) 100 = 16.67\%.$$

From this, we conclude that player C is most consistent, whereas the average score is highest for A. If the selection committee is to select the player on the basis of consistency of performance, then C would be selected. If on the other hand, scoring of highest runs is the basis, then A would be selected.

## NOTES

## 2.13. RELATION BETWEEN MEASURES OF DISPERSION

It has been observed that in frequency distribution, the following relations hold.

1. Q.D. is approximately equal to  $\frac{2}{3}$  S.D.

2. M.D. is approximately equal to  $\frac{4}{5}$  S.D.

### Merits of S.D.

1. It is simple to understand.
2. It is well-defined.
3. In the calculation of S.D., the signs of deviations of items are also taken into account.
4. It is based on all the items.
5. It is capable of further algebraic treatment.
6. It has sampling stability.
7. It is very useful in the study of "Tests of Significance".

### Demerits of S.D.

1. It is not easy to calculate.
2. It is unduly affected by the extreme items, because the squares of deviations of extreme items would be either extremely low or extremely high.

## EXERCISE 2.4

## NOTES

1. Calculate mean, standard deviation and its coefficient for the following data:

Wages up to (in ₹)	10	20	30	40	50	60	70	80
No. of persons	12	30	65	107	157	202	220	230

2. Find the standard deviation for the following data:

Wages (₹)	50—60	60—70	70—80	80—90	90—100	100—110
No. of workers	8	10	16	14	10	5

3. Find which of the following batsman is more consistent in scoring:

Batsman A	5	7	16	27	39	53	56	61	80	101	105
Batsman B	0	4	16	21	41	43	57	78	83	90	95

4. (a) The mean of 5 observations is 4.4 and the variance is 8.24. If three of the five observations are 1, 2 and 6, find the other two.  
 (b) Mean of 48 items is 9 and their standard deviation is 1.6. Find the sum of the squares of all items.
5. If the S.D. of a series is 7.5, find the most likely value of the mean deviation.
6. From the prices of shares of  $x$  and  $y$  given below, state which share is more stable in value:

$x$	41	44	43	48	45	46	49	50	42	40
$y$	91	93	96	92	90	97	99	94	98	95

7. In a cricket season, batsman A gets an average score of 64 runs per inning with a S.D. of 18 runs, while batsman B gets an average score of 43 runs with a S.D. of 9 runs in about an equal number of innings. Discuss the efficiency and consistency of both the batsmen.
8. The mean and S.D. of 20 items is found to be 10 and 2 respectively. At the time of checking, it was found that one item 8 was incorrect. Calculate the correct mean and S.D., if:  
 (i) the wrong item is omitted. (ii) it is replaced by 12.
9. For a group of 50 male workers, the mean and standard deviation of their weekly wages are ₹ 63 and ₹ 9 respectively. For a group of 40 female workers, these measures are respectively ₹ 54 and ₹ 6. Find the S.D. for the combined group of 90 workers.
10. Following table gives height of boys and girls studying in a college:

	Boys	Girls
Number	72	38
Average height	68 inches	61 inches
Variance	9 inches	4 inches

Find the (i) S.D. of the height of boys and girls taken together and (ii) whose heights are more variable.

## Answers

1.  $\bar{x} = ₹ 40.52$ , S.D. = ₹ 17.41, Coeff. of S.D. = 0.4296      2. ₹ 14.5079
3.  $\left. \begin{array}{l} \text{C.V. for A} = 67.0738\% \\ \text{C.V. for B} = 69.5120\% \end{array} \right\} \text{A is consistent.}$
4. (a) 4, 9      (b) 4010.9      5. M.D. = 6
6. S.D. for X = 3.2496, S.D. for Y = 2.8723  
C.V. for X = 7.2536%, C.V. for Y = 3.0395%  
Stability is more in series Y.
7. C.V. for A = 28.125%, C.V. for B = 20.9302%  
If average is the criterion, then A is efficient.  
If consistency is the criterion, then B is efficient.
8. (i) Correct  $\bar{x} = 10.1053$ , Correct S.D. = 1.997  
(ii) Correct  $\bar{x} = 10.2$ , Correct S.D. = 1.9899
9.  $\bar{x} = ₹ 59$ , S.D. = ₹ 9
10. (i) S.D. = 4.2839 inches  
(ii) C.V. for boys = 4.4118%, C.V. for girls = 3.2887%  
Heights of boys are more variable.

## NOTES

### 2.14. SUMMARY

- The **range** of a statistical data is defined as the difference between the largest and the smallest values of the variable.  

$$\therefore \text{Range} = L - S,$$
 where L = largest value of the variable  
 S = smallest value of the variable.
- The **quartile deviation** of a statistical data is defined as  $\frac{Q_3 - Q_1}{2}$  and is denoted as Q.D.
- Mean deviation is also called **average deviation**. The **mean deviation** of a statistical data is defined as the arithmetic mean of the numerical values of the deviations of items from some average. Generally, A.M. and median are used in calculating mean deviation. Let 'a' stand for the average used for calculating M.D.
- It is the most important measure of dispersion. It finds indispensable place in advanced statistical methods. The **standard deviation** of a statistical data is defined as the positive square root of the A.M. of the squared deviations of items from the A.M. of the series under consideration. The S.D. is often denoted by the greek letter ' $\sigma$ '.
- For comparing two or more series for variability, the corresponding relative measure, called coefficient of S.D. is calculated.

### 2.15. REVIEW EXERCISES

1. Explain the merits of quartile deviation method of measuring dispersion over the range method.
2. What is meant by dispersion? What are the requirements of a good measure of dispersion? In the light of those, comment on some of the well-known measures of dispersion.



## NOTES

### 3. SKEWNESS

#### STRUCTURE

- 3.1. Introduction
- 3.2. Meaning
- 3.3. Tests of Skewness
- 3.4. Methods of Measuring Skewness
- 3.5. Karl Pearson's Method
- 3.6. Bowley's Method
- 3.7. Kelly's Method
- 3.8. Method of Moments
- 3.9. Summary
- 3.10. Review Exercises

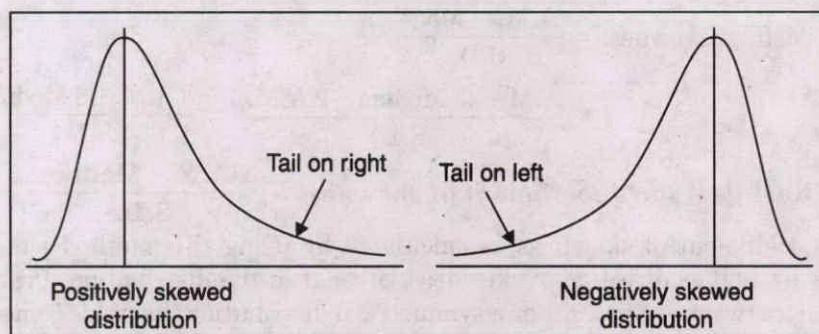
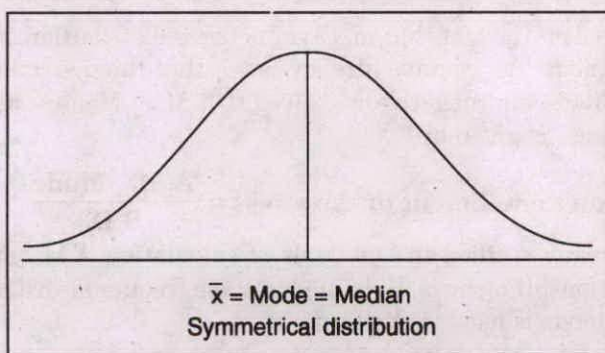
#### 3.1. INTRODUCTION

We have already seen that a single statistical measure is not capable of telling everything about a statistical distribution. A single measure cannot explore all the characteristics of a distribution. As we have already seen that an average of a distribution gives us an idea about the concentration of items about some value. Distributions with same average may differ widely in nature. We have already studied the scatter of items around some average value, in our discussion of measure of dispersion. Now we shall consider the aspect of 'symmetry' in curves of frequency distributions. The shape of the frequency curve depends upon the frequencies of different values of the variable under consideration. If the frequencies of items increases with the equally spaced increasing values of the variable and after a particular stage, the frequencies start decreasing exactly in the same way these were increased, then the frequency curve of the distribution would be *symmetrical, bell-shaped*.

#### 3.2. MEANING

In symmetrical distribution, the values of mean, mode and median, would coincide. If the curve of the distribution is not symmetrical, it may admit of tail on either side of the distribution. Such a distribution lack in symmetry. **Skewness** is the word used for lack of symmetry. A distribution which is not symmetrical is called **asymmetrical** or

**skewed**. We can define '**skewness**' of a distribution as the tendency of a distribution to depart from symmetry.



If the tail of an asymmetrical distribution is on the right side, then the distribution is called a **positively skewed distribution**. If the tail is on left side, then the distribution is defined to be **negatively skewed distribution**. Now we shall account for the situations when skewness can be expected in a distribution.

### 3.3. TESTS OF SKEWNESS

1. If A.M. = mode = median, then there is no skewness in the distribution. In other words, the curve of the frequency distribution would be symmetrical, bell-shaped.
2. If A.M. is less than (greater than), the value of mode, the tail would on left (right) side, i.e., the distribution is negatively (positively) skewed.
3. If sum of frequencies of values less than mode is equal to the sum of frequencies of values greater than mode, then there would be no skewness.
4. If quartiles are equidistant from median, then there would be no skewness.

### 3.4. METHODS OF MEASURING SKEWNESS

1. Karl Pearson's Method
2. Bowley's Method
3. Kelly's Method
4. Method of Moments

### 3.5. KARL PEARSON'S METHOD

#### NOTES

This method is based on the fact that in a symmetrical distribution, the value of A.M. is equal to that of mode. As we have already noted that the distribution is positively skewed if A.M. > Mode and negatively skewed if A.M. < Mode. The Karl Pearson's coefficient of skewness is given by

$$\text{Karl Pearson's coefficient of skewness} = \frac{\text{A.M.} - \text{Mode}}{\text{S.D.}}$$

We have already studied the methods of calculating A.M., mode and S.D. of frequency distributions. If mode is ill-defined in some frequency distribution, then the value of empirical mode is used in the formula.

$$\text{Empirical mode} = 3 \text{ Median} - 2 \text{ A.M.}$$

$$\therefore \text{Coeff. of skewness} = \frac{\text{A.M.} - \text{Mode}}{\text{S.D.}}$$

$$= \frac{\text{A.M.} - (3 \text{ Median} - 2 \text{ A.M.})}{\text{S.D.}} = \frac{3 \text{ A.M.} - 3 \text{ Median}}{\text{S.D.}}$$

$$\therefore \text{Karl Pearson's coefficient of skewness} = \frac{3 (\text{A.M.} - \text{Median})}{\text{S.D.}}$$

The coefficient of skewness as calculated by using this method would give magnitude as well as direction of skewness, present in the distribution. Practically, its value lies between -1 and 1. For a symmetrical distribution, its value comes out to be zero.

The Karl Pearson's coefficient of skewness is generally denoted by ' $SK_P$ '.

#### WORKING RULES FOR SOLVING PROBLEMS

**Rule I.** If the values of  $\bar{x}$ ,  $\sigma$  and mode are given, then find  $SK_P$  by using the formula:

$$SK_P = \frac{\bar{x} - \text{mode}}{\sigma}$$

**Rule II.** If the values of  $\bar{x}$ ,  $\sigma$  and median are given, then find  $SK_P$  by using the formula:

$$SK_P = \frac{3 (\bar{x} - \text{median})}{\sigma}$$

**Rule III.** If the values of  $\bar{x}$ ,  $\sigma$  and mode are not given, then calculate these. If mode is ill-defined, then find median.

**Rule IV.** Find  $SK_P$  by using formulae given in above rules.

**Example 3.1.** Karl Pearson's coefficient of skewness of a distribution is 0.32, its standard deviation is 6.5 and mean is 29.6. Find the mode of the distribution.

**Solution.** We have  $SK_P = 0.32$ ,  $\text{S.D.} = 6.5$ ,  $\bar{x} = 29.6$ .

$$\text{Now } SK_P = \frac{\bar{x} - \text{Mode}}{\text{S.D.}}$$

$$\therefore 0.32 = \frac{29.6 - \text{Mode}}{6.5}$$

$$\Rightarrow 29.6 - \text{Mode} = 0.32 \times 6.5 = 2.08$$

$$\Rightarrow \text{Mode} = 29.6 - 2.08 = 27.52.$$

**Example 3.2.** In a certain distribution, the following results were obtained:

A.M. = 45, Median = 48, Coefficient of Skewness = - 0.4. The person who gave you this data, failed to give the value of S.D. You are required to estimate it with the help of available data.

**Solution.** We have

$$\text{coeff. of skewness} = -0.4, \text{A.M.} = 45, \text{median} = 48.$$

$$\text{Now, coeff. of skewness} = \frac{3(\bar{x} - \text{Median})}{\text{S.D.}}$$

$$\Rightarrow -\frac{4}{10} = \frac{3(45 - 48)}{\text{S.D.}} = \frac{-9}{\text{S.D.}} \Rightarrow 4 \text{ S.D.} = 90$$

$$\Rightarrow \text{S.D.} = \frac{90}{4} = 22.5.$$

**Example 3.3.** The sum of 20 observations is 300 and sum of their squares is 5000. The median is 15. Find the Karl Pearson's coefficient of skewness and coefficient of variation.

**Solution.** Let 'x' be the variable under consideration.

We have  $n = 20$ ,  $\Sigma x = 300$ ,  $\Sigma x^2 = 5000$ , median = 15.

$$\text{Now, } \bar{x} = \frac{\Sigma x}{n} = \frac{300}{20} = 15$$

$$\text{S.D.} = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2} = \sqrt{\frac{5000}{20} - (15)^2} = \sqrt{250 - 225} = \sqrt{25} = 5.$$

Now, Karl Pearson's coeff. of skewness

$$= \frac{3(\bar{x} - \text{Median})}{\text{S.D.}} = \frac{3(15 - 15)}{5} = \frac{0}{5} = 0$$

$$\text{C.V.} = \frac{\text{S.D.}}{\bar{x}} (100) = \frac{5}{15} \times 100 = 33.33\%.$$

**Example 3.4.** Following is data regarding the position of wages in a factory before and after the settlement of an industrial dispute. Comment on the gains and losses from the point of view of the workers and management.

	Before settlement	After settlement
No. of workers	2400	2350
A.M. of wages	₹ 455	₹ 475
Median of wages	₹ 480	₹ 450
S.D. of wages	₹ 120	₹ 100

**Solution.** Let  $x$  denote the variable 'wage'.

(i) No. of workers before settlement = 2400

No. of workers after settlement = 2350.

$\therefore$  After settlement, 50 workers were thrown out of their job. This is a certain loss to the workers, who lost their job.

(ii) A.M. of wages before settlement = ₹ 455

A.M. of wages after settlement = ₹ 475.

$\therefore$  After settlement, the wages of workers have increased. This is a gain to the workers.

## NOTES

## NOTES

(iii) Median wages before settlement = ₹ 480

∴ 50% workers were getting less than or equal to ₹ 480.

Median wage after settlement = ₹ 450

After settlement, 50% workers were getting less than or equal to ₹ 450.

$$(iv) \bar{x} = \frac{\sum x}{n} \quad \therefore \sum x = n \cdot \bar{x}$$

∴ Wage bill before settlement = ₹ 2400(455) = ₹ 10,92,000

Wage bill after settlement = ₹ 2350(475) = ₹ 11,16,250

∴ Increase in wage bill = ₹ 11,16,250 – 10,92,000 = ₹ 24,250

This is a loss to the management.

$$(v) \text{C.V. before settlement} = \frac{120}{455} \times 100 = 26.374\%$$

$$\text{C.V. after settlement} = \frac{100}{475} \times 100 = 21.053\%$$

We see that C.V. has decreased after settlement.

∴ Disparity in wages has decreased after settlement.

(vi) Coeff. of skewness (before settlement)

$$= \frac{3(\bar{x} - \text{Median})}{\text{S.D.}} = \frac{3(455 - 480)}{120} = -0.625$$

∴ Tail of frequency curve is on left side.

Coeff. of skewness (after settlement)

$$= \frac{3(475 - 450)}{100} = 0.750$$

∴ Tail of frequency curve is on right side.

∴ After settlement, the management reduced the number of workers getting high wages.

## EXERCISE 3.1

- Find the coeff. of variation of a frequency distribution with the help of following information:

$$\text{A.M.} = 50$$

$$\text{Mode} = 56$$

Karl Pearson's coeff. of skewness = -0.4.

- Find Pearson's coeff. of skewness for the following frequency distribution:

Wage (in ₹)	50.00—59.99	60—69.99	70—79.99	80—89.99
No. of employees	8	10	16	14
Wage (in ₹)	90—99.99	100—109.99	110—119.99	
No. of employees	10	5	2	



3. For the following data, calculate the coefficient of skewness based on mean, median and S.D.

Skewness

Variable	100—110	110—120	120—130	130—140
Frequency	4	16	36	52
Variable	140—150	150—160	160—170	170—180
Frequency	64	40	32	11

## NOTES

4. For the following frequency distribution, calculate the value of Karl Pearson's coeff. of skewness:

Temp. (°C)	-40 to -30	-30 to -20	-20 to -10	-10 to 0
No. of days	10	28	30	42
Temp. (°C)	0—10	10—20	20—30	
No. of days	65	180	10	

5. Find the mean wage and coefficient of skewness for the following data:

35 men gets at the rate of ₹ 4.5 per man  
 40 men gets at the rate of ₹ 5.5 per man  
 48 men gets at the rate of ₹ 6.5 per man  
 100 men gets at the rate of ₹ 7.5 per man  
 125 men gets at the rate of ₹ 8.5 per man  
 87 men gets at the rate of ₹ 9.5 per man  
 43 men gets at the rate of ₹ 10.5 per man  
 22 men gets at the rate of ₹ 11.5 per man

6. Calculate Karl Pearson's coefficient of skewness for the following data:

Wage (in ₹)	70—80	80—90	90—100	100—110
No. of workers	12	18	35	42
Wage (in ₹)	110—120	120—130	130—140	140—150
No. of workers	50	45	20	8

### Answers

1. C.V. = 30%      2. 0.1454      3. - 0.0087  
 4. - 0.6617      5. Mean wage = ₹ 8.07, Coeff. of skewness = - 0.2445  
 6. - 0.3314.

## 3.6. BOWLEY'S METHOD

This method is based on the fact that in a symmetrical distribution, the quartiles are equidistant from the median. In a skewed distribution, this would not happen. The Bowley's coefficient of skewness is given by

$$\text{Bowley's coefficient of skewness} = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

For a symmetrical distribution, its values would come out to be zero. The value of Bowley's coefficient of skewness lies between - 1 and + 1. The coefficient of skewness



## NOTES

as calculated by using this, would give magnitude as well as direction of skewness present in the distribution. In problems, it is generally given as to which method is to be used. But in case, the method to be used is not specifically mentioned, then it is advisable to use Bowley's method. The calculation of Bowley's coefficient of skewness would involve the calculation of  $Q_1$ ,  $Q_3$  and median. The calculation of these measures would definitely take lesser time than for the calculation of  $\bar{x}$ , mode and S.D. It may also be noted that the values of coefficient of skewness as calculated by using different formulae may not be same. This method is also useful in case of open end classes in the distribution.

The Bowley's coefficient of skewness is generally denoted by ' $SK_B$ '.

## WORKING RULES FOR SOLVING PROBLEMS

**Rule I.** If the values of median,  $Q_1$  and  $Q_3$  are given, then find  $SK_B$  by using the formula:

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

**Rule II.** If the values of median,  $Q_1$  and  $Q_3$  are not given, then find these by using cumulative frequencies of the distribution.

**Rule III.** If the name of the method is not mentioned, then the coefficient should be calculated by Bowley's method. This method will take less time.

**Example 3.5.** For the following data, compute quartiles and the coefficient of skewness:

Income (₹)	Below 200	200—400	400—600	600—800	800—1000	above 1000
No. of persons	25	40	80	75	20	16

**Solution.**

**Calculation of  $Q_1$ ,  $Q_3$  and median**

Classes	No. of persons ( $f$ )	c.f.
Below 200	25	25
200—400	40	65
400—600	80	145
600—800	75	220
800—1000	20	240
above 1000	16	256 = N
	N = 256	

$$Q_1 : \frac{N}{4} = \frac{256}{4} = 64$$

$\therefore Q_1 = \text{size of 64th item}$

$\therefore Q_1 \text{ class is } 200-400$

$$\therefore Q_1 = L + \left( \frac{N/4 - c}{f} \right) h = 200 + \left( \frac{64 - 25}{40} \right) 200 = 200 + 195 = 395.$$

$$Q_3 : 3 \left( \frac{N}{4} \right) = 3 \left( \frac{256}{4} \right) = 192$$

$\therefore Q_3 = \text{size of 192th item}$

$\therefore Q_3 \text{ class is } 600-800.$

$$\begin{aligned}\therefore Q_3 &= L + \left( \frac{3(N/4) - c}{f} \right) h = 600 + \left( \frac{192 - 145}{75} \right) 200 \\ &= 600 + 125.33 = 725.33.\end{aligned}$$

$$\text{Median : } \frac{N}{2} = \frac{256}{2} = 128$$

$\therefore$  Median = size of 128th item

$\therefore$  Median class is 400—600.

$$\begin{aligned}\therefore \text{Median} &= L + \left( \frac{N/2 - c}{f} \right) h = 400 + \left( \frac{128 - 65}{80} \right) 200 \\ &= 400 + 157.5 = 557.5.\end{aligned}$$

$\therefore$  Bowley's coefficient of skewness

$$\begin{aligned}&= \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1} \\ &= \frac{725.33 + 395 - 2(557.5)}{725.33 - 395} = \frac{5.33}{330.33} = 0.016.\end{aligned}$$

**Example 3.6.** Calculate the Bowley's coefficient of skewness for the following frequency distribution:

Classes	1—5	6—10	11—15	16—20	21—25	26—30	31—35
Frequency	3	4	68	30	10	6	2

**Solution.**

**Calculation of  $Q_1$ ,  $Q_3$  and median**

Classes	$f$	c.f.
1—5	3	3
6—10	4	7
11—15	68	75
16—20	30	105
21—25	10	115
26—30	6	121
31—35	2	123 = N
	N = 123	

$$Q_1 : \frac{N}{4} = \frac{123}{4} = 30.75$$

$\therefore Q_1$  = size of 30.75th item

$\therefore Q_1$  class is 10.5—15.5 (actual class limits).

$$\therefore Q_1 = L + \left( \frac{N/4 - c}{f} \right) h = 10.5 + \left( \frac{30.75 - 7}{68} \right) 5 = 10.5 + 1.746 = 12.246.$$

$$Q_3 : 3 \left( \frac{N}{4} \right) = 3 \left( \frac{123}{4} \right) = 92.25$$

$\therefore Q_3$  = size of 92.25th item

$\therefore Q_3$  class is 15.5—20.5 (actual class limits)

## NOTES

$$\therefore Q_3 = L + \left( \frac{3(N/4) - c}{f} \right) h = 15.5 + \left( \frac{92.25 - 75}{30} \right) 5$$

$$= 15.5 + 2.875 = 18.375.$$

**NOTES**

**Median:**  $\frac{N}{2} = \frac{123}{2} = 61.5$

$\therefore$  Median = size of 61.5th item

$\therefore$  Median class is 10.5—15.5 (actual class limits)

$$\therefore \text{Median} = L + \left( \frac{N/2 - c}{f} \right) h = 10.5 + \left( \frac{61.5 - 7}{68} \right) 5 = 10.5 + 4.007 = 14.507.$$

Now, Bowley's coefficient of skewness

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

$$= \frac{18.375 + 12.246 - 2(14.507)}{18.375 - 12.246} = \frac{1.607}{6.129} = 0.262.$$

**EXERCISE 3.2**

1. In a frequency distribution, it is found that  $Q_1 = 14.6$  cm, median = 18.8 cm and  $Q_3 = 25.2$  cm. Find the coefficient of Q.D. and the Bowley's coefficient of skewness.
2. Calculate Bowley's coefficient of skewness for the following data:

Wage (in ₹)	85	90	95	100	105	110	115	120	125
No. of persons	15	18	25	19	15	7	28	12	11

3. Calculate the quartile coefficient of skewness for the following frequency distribution:

Weight (in kg)	No. of persons	Weight (in kg)	No. of persons
Under 100	1	150—159	65
100—109	14	160—169	31
110—119	66	170—179	12
120—129	122	180—189	5
130—139	145	190—199	2
140—149	121	200 and above	2

4. Calculate coefficient of skewness based upon quartiles for the data given below:

Marks (Less than)	10	20	30	40	50	60
No. of students	5	12	20	35	40	50

**Answers**

1. Coeff. of Q.D. = 0.2663, Coeff. of skewness = 0.2075
2. 0.5
3. 0.0233
4. 0.0397



### 3.7. KELLY'S METHOD

This method is based on the fact that in a symmetrical distribution the 10th percentile and 90th percentile are equidistant from the median. In a skewed distribution, this equality would not hold. The Kelly's coefficient of skewness is given by

$$\text{Kelly's coefficient of skewness} = \frac{P_{90} + P_{10} - 2 \text{ Median}}{P_{90} - P_{10}}$$

For a symmetrical distribution, its value would come out to be zero. This coefficient of skewness would lie between  $-1$  and  $+1$ . The coefficient of skewness as calculated by this method would give magnitude as well as direction of skewness present in the distribution.

#### WORKING RULES FOR SOLVING PROBLEMS

**Rule I.** If the values of median,  $P_{10}$  and  $P_{90}$  are given, then find Kelly's coefficient of skewness by using the formula:

$$SK = \frac{P_{90} + P_{10} - 2 \text{ Median}}{P_{90} - P_{10}}$$

**Rule II.** Kelly's coefficient of skewness is also equal to  $\frac{D_9 + D_1 - 2 \text{ Median}}{D_9 - D_1}$ .

**Rule III.** If the values of median,  $P_{10}$  and  $P_{90}$  are not given, then find these by using the cumulative frequencies of the distribution.

**Example 3.7.** In a frequency distribution,

$$P_{10} = 5, \text{ Median} = 12 \text{ and } P_{90} = 22.$$

Find Kelly's coefficient of skewness.

**Solution.** We have  $P_{10} = 5$ , median = 12 and  $P_{90} = 22$ .

Kelly's coeff. of skewness

$$= \frac{P_{90} + P_{10} - 2 \text{ Median}}{P_{90} - P_{10}} = \frac{22 + 5 - 2(12)}{22 - 5} = \frac{3}{17} = 0.1765.$$

**Example 3.8.** Calculate Kelly's coefficient of skewness for the following frequency distribution:

Daily wage (in ₹)	20—25	25—30	30—35	35—40	40—45	45—50
No. of Workers	12	16	5	4	2	1

**Solution.** Calculation of Kelly's Coefficient of Skewness

Daily wages (in ₹)	No. of workers (f)	c.f.
20—25	12	12
25—30	16	28
30—35	5	33
35—40	4	37
40—45	2	39
45—50	1	40 = N
	N = 40	

#### NOTES

## NOTES

$$P_{10}: \quad 10 \left( \frac{N}{100} \right) = 10 \left( \frac{40}{100} \right) = 4$$

$\therefore P_{10} = \text{size of 4th item}$

$\therefore P_{10} \text{ class is } 20-25$

$$\begin{aligned} \therefore P_{10} &= L + \left( \frac{10(N/100) - c}{f} \right) h \\ &= 20 + \left( \frac{4 - 0}{12} \right) 5 = 20 + 1.67 = ₹ 21.67. \end{aligned}$$

$$P_{90}: \quad 90 \left( \frac{N}{100} \right) = 90 \left( \frac{40}{100} \right) = 36$$

$\therefore P_{90} = \text{size of 36th item}$

$\therefore P_{90} \text{ class is } 35-40.$

$$\begin{aligned} \therefore P_{90} &= L + \left( \frac{90(N/100) - c}{f} \right) h = 35 + \left( \frac{36 - 33}{4} \right) 5 \\ &= 35 + 3.75 = ₹ 38.75. \end{aligned}$$

$$\text{Median:} \quad \frac{N}{2} = \frac{40}{2} = 20$$

$\therefore \text{Median} = \text{size of 20th item}$

$\therefore \text{Median class is } 25-30$

$$\begin{aligned} \therefore \text{Median} &= L + \left( \frac{N/2 - c}{f} \right) h = 25 + \left( \frac{20 - 12}{16} \right) 5 \\ &= 25 + 2.5 = ₹ 27.50 \end{aligned}$$

Now, Kelly's coefficient of skewness

$$\begin{aligned} &= \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1} \\ &= \frac{38.75 + 21.67 - 2(27.50)}{38.75 - 21.67} = \frac{5.42}{17.08} = 0.3173. \end{aligned}$$

## EXERCISE 3.3

1. In a frequency distribution,  $P_{10} = 10$ , median = 22 and  $P_{90} = 25$ . Calculate Kelly's coefficient of skewness.
2. In a frequency distribution,  $P_{10} = 17$ ,  $P_{90} = 53$  and median = 38. Find Kelly's coefficient of skewness.
3. Calculate the coefficient of skewness, using  $P_{10}$  and  $P_{90}$ , for the following data:

$x$	10	11	12	13	14	15	16	17
$f$	3	11	18	15	12	9	6	3

4. Calculate Kelly's coefficient of skewness for the following frequency distribution:

Marks less than	10	20	30	40	50	60	70
No. of students	0	5	7	10	12	18	30

## Answers

1. -0.6                      2. -0.17                      3. 0                      4. -0.51.



### 3.8. METHOD OF MOMENTS

Skewness

In this method, second and third central moments of the distribution are used. This measure of skewness is called the **Moment coefficient of skewness** and is given by

$$\text{Moment coefficient of skewness} = \frac{\mu_3}{\sqrt{\mu_2^3}}$$

For a symmetrical distribution, its value would come out to be zero. The coefficient of skewness as calculated by this method gives the magnitude as well as direction of skewness present in the distribution.

In statistics, we define  $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ .

$\therefore$  Moment coefficient of skewness can also be written as

$$= \frac{\mu_3}{\sqrt{\mu_2^3}} = \pm \sqrt{\left(\frac{\mu_3}{\sqrt{\mu_2^3}}\right)^2} = \pm \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \pm \sqrt{\beta_1}$$

The sign with  $\sqrt{\beta_1}$  is to be taken as that of  $\mu_3$ . The moment coefficient of skewness is also denoted by  $\gamma_1$ .

The moment coefficient of skewness is generally denoted by ' $SK_M$ '.

#### WORKING RULES FOR SOLVING PROBLEMS

**Rule I.** If the values of  $\mu_2$  and  $\mu_3$  are given, then find  $SK_M$  by using the formula:

$$SK_M = \frac{\mu_3}{\sqrt{\mu_2^3}}$$

**Rule II.** If raw moments  $\mu_1'$ ,  $\mu_2'$  and  $\mu_3'$  are given, then calculate:  
 $\mu_2 = \mu_2' - \mu_1'^2$  and  $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$ .

Now, find  $SK_M = \frac{\mu_3}{\sqrt{\mu_2^3}}$ .

**Rule III.** If moments are not given, then first find  $\mu_2$  and  $\mu_3$  by using the given data and then use the formula:  $SK_M = \frac{\mu_3}{\sqrt{\mu_2^3}}$ .

**Rule IV.**  $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$  and  $\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}}$ .

**Example 3.9.** The first three central moments of a distribution are 0, 15, -31. Find the moment coefficient of skewness.

**Solution.** We have  $\mu_1 = 0$ ,  $\mu_2 = 15$  and  $\mu_3 = -31$ .

Moment coefficient of skewness

$$= \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{-31}{\sqrt{(15)^3}} = -\frac{31}{\sqrt{3375}} = -\frac{31}{58.09} = -0.53.$$

NOTES

**Example 3.10.** Find the second and third central moments for the frequency distribution given below. Hence find the coefficient of skewness:

**NOTES**

Class	110.0—114.9	115.0—119.9	120.0—124.9	125.0—129.9
Frequency	5	15	20	35
Class	130.0—134.9	135.0—139.9	140.0—144.9	
Frequency	10	10	5	

**Solution.****Computation of moments**

Class	f	x	d = x - A A = 127.45	u = d/h h = 5	fu	fu <sup>2</sup>	fu <sup>3</sup>
110.0—114.9	5	112.45	-15	-3	-15	45	-135
115.0—119.9	15	117.45	-10	-2	-30	60	-120
120.0—124.9	20	122.45	-5	-1	-20	20	-20
125.0—129.9	35	127.45	0	0	0	0	0
130.0—139.9	10	132.45	5	1	10	10	10
140.0—144.9	10	137.45	10	2	20	40	80
	5	142.45	15	3	15	45	135
	N = 100				$\Sigma fu = -20$	$\Sigma fu^2 = 220$	$\Sigma fu^3 = -50$

Now

$$\mu_1' = \left( \frac{\Sigma fu}{N} \right) h = \left( \frac{-20}{100} \right) 5 = -1$$

$$\mu_2' = \left( \frac{\Sigma fu^2}{N} \right) h^2 = \left( \frac{220}{100} \right) (5)^2 = 55$$

$$\mu_3' = \left( \frac{\Sigma fu^3}{N} \right) h^3 = \left( \frac{-50}{100} \right) (5)^3 = -62.5$$

**Central moments**

$$\mu_2 = \mu_2' - \mu_1'^2 = 55 - (-1)^2 = 54$$

$$\begin{aligned} \mu_3 &= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -62.5 - 3(55)(-1) + 2(-1)^3 \\ &= -62.5 + 165 - 2 = 100.5 \end{aligned}$$

 $\therefore$  Moment coefficient of skewness

$$= \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{100.5}{\sqrt{(54)^3}} = \frac{100.5}{54 \times 7.35} = 0.253.$$

**EXERCISE 3.4**

- The first three central moments of a distribution are 0, 2.5, 0.7. Find the values of S.D. and the moment coefficient of skewness.
- In a certain distribution, the first four moments about the point 4 are -1.5, 17, -30 and 308. Calculate the moment coefficient of skewness.
- The first three moments of a frequency distribution about origin '5' are -0.55, 4.46 and -0.43. Find the moment coefficient of skewness.

4. Find the moment coefficient of skewness for the following series:

$x$	3	6	8	10	18
-----	---	---	---	----	----

5. Calculate the A.M., coefficient of variation and the moment coefficient of skewness for the following data:

$x$	0	1	2	3	4	5	6	7	8
$f$	1	8	28	56	70	56	28	8	1

### Answers

1. 1.5811, 0.1771      2. 0.7017      3. 0.7781      4. 0.7504  
 5. A.M. = 4, C.V. = 35.3553%, coefficient of skewness = 0

## NOTES

### 3.9. SUMMARY

- **Skewness** is the word used for lack of symmetry. A distribution which is not symmetrical is called **asymmetrical** or **skewed**. We can define 'skewness' of a distribution as the tendency of a distribution to depart from symmetry.
- If the tail of an asymmetrical distribution is on the right side, then the distribution is called a **positively skewed distribution**. If the tail is on left side, then the distribution is defined to be **negatively skewed distribution**.
- This method is based on the fact that in a symmetrical distribution, the value of A.M. is equal to that of mode. As we have already noted that the distribution is positively skewed if A.M. > Mode and negatively skewed if A.M. < Mode. The Karl Pearson's coefficient of skewness is given by

$$\text{Karl Pearson's coefficient of skewness} = \frac{\text{A.M.} - \text{Mode}}{\text{S.D.}}$$

- This method is based on the fact that in a symmetrical distribution, the quartiles are equidistant from the median. In a skewed distribution, this would not happen. The Bowley's coefficient of skewness is given by

$$\text{Bowley's coefficient of skewness} = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

- This method is based on the fact that in a symmetrical distribution the 10th percentile and 90th percentile are equidistant from the median. In a skewed distribution, this equality would not hold. The Kelly's coefficient of skewness is given by

$$\text{Kelly's coefficient of skewness} = \frac{P_{90} + P_{10} - 2 \text{ Median}}{P_{90} - P_{10}}$$

### 3.10. REVIEW EXERCISES

1. Define skewness. Explain the difference between positive skewness and negative skewness.
2. Explain what do you understand by "Skewness". What are the various methods of measuring skewness?
3. How does 'Skewness' differ from 'Dispersion'? Explain the different methods of studying skewness.
4. Explain the use of quartiles in studying skewness in frequency distributions.
5. Explain with formulae different measures of skewness.

## NOTES

## 4. KURTOSIS

### STRUCTURE

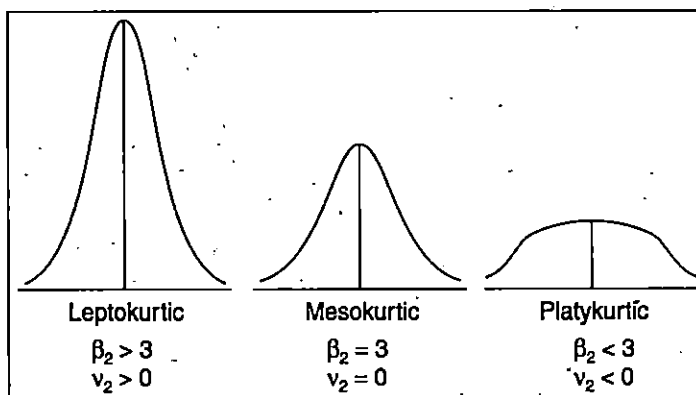
- 4.1. Introduction
- 4.2. Definitions
- 4.3. Measure of Kurtosis
- 4.4. Summary
- 4.5. Review Exercises

### 4.1. INTRODUCTION

We have already discussed some of the characteristics of statistical distributions. The measures of central tendency tells us about the concentration of the observations about an average value of the distribution whereas the measure of dispersion gives the idea of scatter of the observations about some average. The measure of skewness helps us in judging the extent of symmetry in the curves of frequency distributions. Now we shall consider the peakedness and flatness of frequency distributions. The measure of peakedness or flatness of the curve of a frequency distribution, relative to the curve of normal distribution, is called the measure of '**Kurtosis**'. Kurtosis refers to the bulginess of the curve of a frequency distribution.

### 4.2. DEFINITIONS

The curve of a frequency distribution is called '**Mesokurtic**', if it is neither flat nor sharply peaked. The curve of normal distribution is mesokurtic. The curve of a frequency



distribution is called '**Leptokurtic**', if it is more peaked than normal curve. The curve of a frequency distribution is called '**Platykurtic**', if it is more flat-topped than the normal curve.

### 4.3. MEASURE OF KURTOSIS

The measure of kurtosis is denoted by  $\beta_2$  and is defined as

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

where  $\mu_2$  and  $\mu_4$  are respectively the second and fourth moments, about mean of the distribution. If  $\beta_2 > 3$ , the distribution is Leptokurtic. If  $\beta_2 = 3$ , the distribution is Mesokurtic. If  $\beta_2 < 3$ , the distribution is Platykurtic. The kurtosis of a distribution is also measured by using Greek letter ' $\gamma_2$ ', which is defined as  $\gamma_2 = \beta_2 - 3$ .

$\therefore \gamma_2 > 0 \Rightarrow \beta_2 - 3 > 0 \Rightarrow \beta_2 > 3 \Rightarrow$  the distribution is Leptokurtic.

Similarly, if  $\gamma_2 = 0$ , then  $\beta_2 = 3$

$\therefore$  The distribution is Mesokurtic.

$\gamma_2 < 0 \Rightarrow \beta_2 < 3 \Rightarrow$  the distribution is Platykurtic.

#### WORKING RULES FOR SOLVING PROBLEMS

**Rule I.** If the values of  $\mu_2$  and  $\mu_4$  are given, then find  $\beta_2$  by using the formula:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

**Rule II.** If raw moments  $\mu_1', \mu_2', \mu_3'$  and  $\mu_4'$  are given, then calculate:

$$\mu_2 = \mu_2' - \mu_1'^2 \text{ and } \mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$\text{Now, find } \beta_2 = \frac{\mu_4}{\mu_2^2}$$

**Rule III.** If moments are not given, then first find  $\mu_2$  and  $\mu_4$  by using the given

data and then use the formula:  $\beta_2 = \frac{\mu_4}{\mu_2^2}$ .

**Rule IV.** The given distribution is leptokurtic, mesokurtic and platykurtic according as  $\beta_2 > 3$ ,  $\beta_2 = 3$  and  $\beta_2 < 3$  respectively.

**Rule V.**  $\gamma_2 = \beta_2 - 3$ . The given distribution is leptokurtic, mesokurtic and platykurtic according as  $\gamma_2 > 0$ ,  $\gamma_2 = 0$  and  $\gamma_2 < 0$  respectively.

**Example 4.1.** The first four moments about mean of a frequency distribution are 0, 100, -7 and 35000. Discuss the kurtosis of the distribution.

**Solution.** We have  $\mu_1 = 0, \mu_2 = 100, \mu_3 = -7$  and  $\mu_4 = 35000$ .

$$\text{Now } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{35000}{(100)^2} = 3.5 > 3.$$

$\therefore$  The distribution is **leptokurtic**.

**Example 4.2.** The first four moments of a distribution about the value '4' of the variable are -1.5, 17, -30 and 108. Discuss the kurtosis of the distribution.

**Solution.** We have  $\mu_1' = -1.5, \mu_2' = 17, \mu_3' = -30$  and  $\mu_4' = 108$ .

$$\therefore \mu_2 = \mu_2' - (\mu_1')^2 = 17 - (-1.5)^2 = 14.75$$



$$\begin{aligned}\mu_4 &= \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 \\ &= 108 - 4(-1.5)(-30) + 6(17)(-1.5)^2 - 3(-1.5)^4 = 142.3125.\end{aligned}$$

Now, 
$$\beta_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{142.3125}{(14.75)^2} = \frac{142.3125}{217.5625} = 0.654 < 3.$$

**NOTES**

$\therefore$  The distribution is **platykurtic**.

**Example 4.3.** Compute the coefficient of skewness and kurtosis based on moments for the following distribution:

$x$	4.5	14.5	24.5	34.5	44.5	54.5	64.5	74.5	84.5	94.5
$f$	1	5	12	22	17	9	4	3	1	1

**Solution.****Calculation of moments**

$x$	$f$	$d = x - A$ $A = 44.5$	$u = d/h$ $h = 10$	$fu$	$fu^2$	$fu^3$	$fu^4$
4.5	1	-40	-4	-4	16	-64	256
14.5	5	-30	-3	-15	45	-135	405
24.5	12	-20	-2	-24	48	-96	192
34.5	22	-10	-1	-22	22	-22	22
44.5	17	0	0	0	0	0	0
54.5	9	10	1	9	9	9	9
64.5	4	20	2	8	16	32	64
74.5	3	30	3	9	27	81	243
84.5	1	40	4	4	16	64	256
94.5	1	50	5	5	25	125	625
	$N = 75$			$\Sigma fu = -30$	$\Sigma fu^2 = 224$	$\Sigma fu^3 = -6$	$\Sigma fu^4 = 2072$

**Moments about 44.5**

$$\mu_1' = \left( \frac{\Sigma fu}{N} \right) h = \left( -\frac{30}{75} \right) 10 = -4$$

$$\mu_2' = \left( \frac{\Sigma fu^2}{N} \right) h^2 = \left( \frac{224}{75} \right) (10)^2 = 298.667$$

$$\mu_3' = \left( \frac{\Sigma fu^3}{N} \right) h^3 = \left( \frac{-6}{75} \right) (10)^3 = -80$$

$$\mu_4' = \left( \frac{\Sigma fu^4}{N} \right) h^4 = \left( \frac{2072}{75} \right) (10)^4 = 276266.667.$$

**Central moments  $\mu_2, \mu_3, \mu_4$** 

$$\mu_2 = \mu_2' - \mu_1'^2 = 298.667 - (-4)^2 = 282.667$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -80 - 3(298.667)(-4) + 2(-4)^3 = 3376.004$$

$$\begin{aligned}\mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 \\ &= 276266.667 - 4(-80)(-4) + 6(298.667)(-4)^2 - 3(-4)^4 = 302890.7.\end{aligned}$$

## Skewness

Moment coefficient of skewness,

$$\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{3376.004}{\sqrt{(282.667)^3}} = \frac{3376.004}{282.667 \sqrt{282.667}} = 0.71.$$

∴ The distribution is **positively skewed**.

## Kurtosis

$$\gamma_2 = \frac{\mu_4}{\sqrt{\mu_2^2}} - 3 = \frac{302890.7}{(282.667)^2} - 3 = 3.79 - 3 = 0.79 > 0.$$

∴ The distribution is **leptokurtic**.

**Example 4.4.** Find the measure of kurtosis for the following distribution:

Class	45—52	52—59	59—66	66—73	73—80	80—87	87—94
Frequency	4	9	12	4	3	2	1

**Solution.** In order to calculate  $\beta_2$ , the measure of kurtosis, we will have to find the values of  $\mu_2$  and  $\mu_4$ .

Class	Freq- uency	Mid- points $x$	$d = x - A$ $A = 69.5$	$u = d/h$ $h = 7$	$fu$	$fu^2$	$fu^3$	$fu^4$
45—52	4	48.5	-21	-3	-12	36	-108	324
52—59	9	55.5	-14	-2	-18	36	-72	144
59—66	12	62.5	-7	-1	-12	12	-12	12
66—73	4	69.5	0	0	0	0	0	0
73—80	3	76.5	7	1	3	3	3	3
80—87	2	83.5	14	2	4	8	16	32
87—94	1	90.5	21	3	3	9	27	81
	N = 35				$\Sigma fu =$ -32	$\Sigma fu^2 =$ 104	$\Sigma fu^3 =$ -146	$\Sigma fu^4 =$ 596

Now,  $\mu_1' = \left( \frac{\Sigma fu}{N} \right) h = \left( \frac{-32}{35} \right) 7 = -6.4$

$$\mu_2' = \left( \frac{\Sigma fu^2}{N} \right) h^2 = \left( \frac{104}{35} \right) (7)^2 = 145.6$$

$$\mu_3' = \left( \frac{\Sigma fu^3}{N} \right) h^3 = \left( \frac{-146}{35} \right) (7)^3 = -1430.8$$

$$\mu_4' = \left( \frac{\Sigma fu^4}{N} \right) h^4 = \left( \frac{596}{35} \right) (7)^4 = 40885.6$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 145.6 - (-6.4)^2 = 104.64$$

$$\begin{aligned} \mu_4 &= \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2\mu_2' - 3(\mu_1')^4 \\ &= 40885.6 - 4(-6.4)(-1430.8) + 6(-6.4)^2(145.6) - 3(-6.4)^4 \\ &= 40885.6 - 36628.48 + 35782.656 - 5033.1648 = 35006.612. \end{aligned}$$

## NOTES

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{35006.612}{(104.64)^2} = 3.1971 > 3.$$

$\therefore$  The distribution is **leptokurtic**.

## NOTES

**Example 4.5.** For a distribution, the mean is 10, variance is 16. If  $\gamma_1 = 1$ ,  $\beta_2 = 4$ , find the first four moments about the mean and about the origin.

**Solution.** We have  $\bar{x} = 10$ , variance = 16,  $\gamma_1 = 1$ ,  $\beta_2 = 4$ .

We know  $\mu_1 = 0$  (always);  $\mu_2 = \text{variance} = 16$

$$\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2}^3} \Rightarrow 1 = \frac{\mu_3}{\sqrt{(16)}^3} \Rightarrow \mu_3 = 16 \times 4 = 64$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \Rightarrow 4 = \frac{\mu_4}{(16)^2} \Rightarrow \mu_4 = 4 \times 256 = 1024.$$

**Moments about origin**

$$\gamma_1 = \bar{x} = 10$$

$$\gamma_2 = \mu_2 + \bar{x}^2 = 16 + (10)^2 = 116$$

$$\gamma_3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3 = 64 + 3(16)(10) + (10)^3 = 1544$$

$$\begin{aligned} \gamma_4 &= \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4 \\ &= 1024 + 4(64)(10) + 6(16)(10)^2 + (10)^4 = 23184. \end{aligned}$$

## 4.4. SUMMARY

- The curve of a frequency distribution is called '**Mesokurtic**', if it is neither flat nor sharply peaked. The curve of normal distribution is mesokurtic. The curve of a frequency distribution is called '**Leptokurtic**', if it is more peaked than normal curve. The curve of a frequency distribution is called '**Platykurtic**', if it is more flat-topped than the normal curve.
- The measure of kurtosis is denoted by  $\beta_2$  and is defined as

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

where  $\mu_2$  and  $\mu_4$  are respectively the second and fourth moments, about mean of the distribution. If  $\beta_2 > 3$ , the distribution is Leptokurtic. If  $\beta_2 = 3$ , the distribution is Mesokurtic. If  $\beta_2 < 3$ , the distribution is Platykurtic. The kurtosis of a distribution is also measured by using Greek letter ' $\gamma_2$ ', which is defined as  $\gamma_2 = \beta_2 - 3$ .

## 4.5. REVIEW EXERCISES

- Explain the term 'kurtosis'.
- How does kurtosis differ from skewness?
- Explain the method of studying kurtosis.
- What are Skewness and Kurtosis? Give formula for measuring them.
- Define 'Leptokurtic' distribution.
- Define Kurtosis. Give Fisher's measure of Kurtosis. Draw rough sketches for different cases.

7. The first four moments about mean of a frequency distribution are 0, 60, -50 and 8020 respectively. Discuss the kurtosis of the distribution.
8. The  $\mu_2$  and  $\mu_4$  for a distribution are found to be 2 and 12 respectively. Discuss the kurtosis of the distribution.
9. The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the kurtosis of the distribution.
10. The standard deviation of symmetric distribution is 3. What must be the value of  $\mu_4$ , so that the distribution may be mesokurtic?
11. If the first four moments about the value '5' of the variable are -4, 22, -117 and 560, find the value of  $\beta_2$  and discuss the kurtosis.
12. Compute the value of  $\beta_2$  for the following distribution. Is the distribution platykurtic?

Class	10—20	20—30	30—40	40—50	50—60	60—70	70—80
Frequency	1	20	69	108	78	22	2

13. Calculate  $\beta_1$  and  $\beta_2$  for the following distribution :

Age (in years)	25—30	30—35	35—40	40—45
Number of workers	2	8	18	27
Age (in years)	45—50	50—55	55—60	60—65
Number of workers	25	16	7	2

### Answers

7.  $\beta_2 = 2.2278$ , Platykurtic
8.  $\beta_2 = 3$ , Mesokurtic
9.  $\beta_2 = 3$ , Mesokurtic
10.  $\mu_4 = 243$
11.  $\beta_2 = 0.8889$ , Platykurtic
12.  $\beta_2 = 2.7240$ , Yes
13.  $\beta_1 = 0.033, \beta_2 = 2.7$ .

### NOTES

## NOTES

## 5. ANALYSIS OF TIME SERIES

### STRUCTURE

- 5.1. Introduction
- 5.2. Meaning
- 5.3. Components of Time Series
- 5.4. Secular Trend or Long-Term Variations
- 5.5. Seasonal Variations
- 5.6. Cyclical Variations
- 5.7. Irregular Variations
- 5.8. Additive and Multiplicative Models of Decomposition of Time Series
- 5.9. Determination of Trend
- 5.10. Free Hand Graphic Method
- 5.11. Semi-Average Method
- 5.12. Moving Average Method
- 5.13. Least Squares Method
- 5.14. Linear Trend
- 5.15. Non-linear Trend (Parabolic)
- 5.16. Non-linear Trend (Exponential)
- 5.17. Summary
- 5.18. Review Exercises

### 5.1. INTRODUCTION

We know that a **time series** is a collection of values of a variable taken at different time periods. If  $y_1, y_2, \dots, y_n$  be the values of a variable  $y$  taken at time periods  $t_1, t_2, \dots, t_n$ , then we write this time series as  $\{(t_i, y_i); i = 1, 2, \dots, n\}$ . The given time series data is arranged chronologically. If we consider the sale figures of a company for over 20 years, the data will constitute a time series. Population of a town, taken annually for 15 years, would form a time series. There are plenty of variables whose value depends on time.



## 5.2. MEANING

In a time series, the values of the concerned variable is not expected to be same for every time period. For example, if we consider the price of 1 kg tea of a particular brand, for over twenty years, we will note that the price is not the same for every year. What has caused the price to vary? In fact, there is nothing special with tea, this can happen for any variable, we consider.

There are number of economic, psychological, sociological and other forces which may cause the value of the variable to change with time. In this chapter, we shall locate, measure and interpret the changes in the values of the variable, in a time series. We shall investigate the factors, which may be held responsible for causing changes in the values of the variable with respect to time.

## NOTES

## 5.3. COMPONENTS OF TIME SERIES

We have already noted that the value of variable in a time series are very rarely constant. The graph of its time series will be a zig-zag line. The variation in the values of time series are due to psychological, sociological, economic, etc. forces. The variations in a time series are classified into four types and are called **components** of the time series. The components are as follows:

- (i) Secular trend or long-term variations
- (ii) Seasonal variations
- (iii) Cyclical variations
- (iv) Irregular variations.

## 5.4. SECULAR TREND OR LONG-TERM VARIATIONS

The general tendency of the values of the variable in a time series to grow or to decline over a long period of time is called **secular trend** of the times series. It indicates the general direction in which the graph of the time series appears to be going over a long period of time. The graph of the secular trend is either a straight line or a curve. This graph depends upon the nature of data and the method used to determine secular trend.

The secular trend of a time series depends much on factors which changes very slowly, *e.g.*, population, habits, technical development, scientific research, etc.

If the secular trend for a particular time series is upward (downward), it does not necessarily imply that the values of the variable must be strictly increasing (decreasing). For example, consider the data:

Year	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
Profit (000 ₹)	18	17	20	21	25	22	26	27	28	35

We observe that the profit figures for the years 1979 and 1983 are less than those of their corresponding previous years, but for all other years the profit figures

## NOTES

are greater than their corresponding previous years. In this time series, the general tendency of the profit figures is to grow.

If from the definition of secular trend, we drop the condition of having time series data for a long period of time, the definition will become meaningless. For example, if we consider the data:

Year	2002	2003
Price of sugar (1 kg)	₹ 14	₹ 14.50

From this time series, we cannot have the idea of the general tendency of the time series. In this connection, it is not justified to assert that the values of the variable must be taken for time periods covering 6 months or 10 years or 15 years. Rather we must see that the values of the variable are sufficient in number. Thus, in estimating trend, it is not the total time period that matters, but it is the number of time periods for which the values of the variable are known.

## 5.5. SEASONAL VARIATIONS

The **seasonal variations** in a time series counts for those variations in the series which occur annually. In a time series, seasonal variations occur quite regularly. These variations play a very important role in business activities. There are number of factors which causes such variations. We know that the demand for raincoats rises automatically during rainy season. Producers of tea and coffee feels that the demand of their products is more in winter season rather than in summer season. Similarly, there is greater demand for cold drinks during summer season. Retailers on Hill stations are also affected by the seasonal variations. Their profits are heavily increased during summer season.

Even Banks have not escaped from seasonal variations. Banks observe heavy withdrawals in the first week of every month. Agricultural yield is also seasonal and so the farmers income is unevenly divided over the year. This has direct effect on business activities.

Customs and habits also plays an important role in causing seasonal variations in time series. On the eve of festivals, we are accustomed of purchasing sweets and new clothes. Generally, people get their houses white washed before Deepawali. Sale figures of retailers dealing with fireworks immediately boost up on the eve of Deepawali and in the season of marriages.

The study of seasonal variations in a time series is also very useful. By studying the seasonal variations, the businessman can adjust his stock holding during the year. He will not feel the danger of shortfall of stock during any particular period, in the year.

## 5.6. CYCLICAL VARIATIONS

The **cyclical variations** in a time series counts for the swings of graph of time series about its trend line (curve). Cyclical variations are seldom periodic and they may or may not follow same pattern after equal interval of time.

In particular, business and economic time series are said to have cyclical variations if these variations recur after time interval of more than one year. In business and economic time series, *business cycles* are example of cyclical variations. There are four phases of a business cycle. These are:

- |                |              |
|----------------|--------------|
| (a) Depression | (b) Recovery |
| (c) Boom       | (d) Decline. |

These four phases of business cycle follows each other in this order.

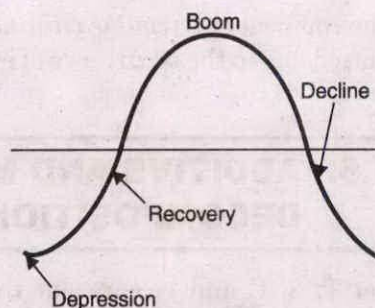
(a) **Depression.** We start with the situation of depression in business cycle. In this phase, the employment is very limited. Employees get very low wages. The purchasing power of money is high. This is the period of pessimism in business. New equilibrium is achieved in business at low level of cost, profit and prices.

(b) **Recovery.** The new equilibrium in the depression phase of a cycle; last for few years. This phase is not going to continue for ever. In the phase of depression, even efficient workers are available at very low wages. In the depression period, prices are low and the costs also too low. These factors replaces pessimism by optimism. Businessman, with good financial support is optimistic in such circumstances. He invests money in repairing plants. New plants are purchased. This also boost the business of allied industries. People get employment and spend money on consumers good. So, the situation changes altogether. This is called the phase of recovery in business cycle.

(c) **Boom.** There is also limit to recovery. Investment is revived in recovery phase. Investment in one industry affects investment in other industries. People get employment. Extension in demand is felt. Prices go high. Profits are made very easily. All these leads to over development of business. This phase of business cycle is described as *boom*.

(d) **Decline.** In the phase of boom, the business is over developed. This is because of heavy profits. Wages are increased and on the contrary their efficiency decreases. Money is demanded everywhere. This results in the increase in rate of interest. In other words, the demand for production factors increases very much and this results in increase in their prices. This results in the increases in the cost of production. Profits are decreased. Banks insists for repayment of loans under these circumstances. Businessmen give concession in prices so that cash may be secured. Consumers start expecting more reduction in prices. Condition become more worse. Products accumulates with businessmen and repayment of loan does not take place. Many business houses fails. All these leads to depression phase and the business cycle continues itself.

The length of a business cycle is in general between 3 to 10 years. Moreover, the lengths of business cycles are not equal.



## NOTES

### 5.7. IRREGULAR VARIATIONS

The **irregular variations** in a time series counts for those variations which cannot be predicted before hand. This component is different from the other three components in the sense that irregular variations in a time series are very irregular. Nothing can be predicted about the occurrence of irregular variations. It is very true that floods, famines, wars, earthquakes, strikes, etc. do affect the economic and business activities.

The component *irregular variations* refers to the variations in time series which are caused due to the occurrence of events like flood, famine, war, earthquake, strike, etc.

## NOTES

## 5.8. ADDITIVE AND MULTIPLICATIVE MODELS OF DECOMPOSITION OF TIME SERIES

Let T, S, C and I represent the trend component, seasonal component, cyclical component and irregular component of a time series, respectively. Let the variable of the time series be denoted by Y. There are mainly two models of decomposition of time series.

(i) **Additive model.** In this model, we have

$$Y = T + S + C + I.$$

In this case, the components T, S, C and I represent absolute values. Here S, C and I may admit of negative values. In this model, we assume that all the four components are independent of each other.

(ii) **Multiplicative model.** In this model, we have

$$Y = T \times S \times C \times I.$$

In this case, the component T is in absolute value whereas the components S, C and I represent relative indices with base value unity. In this model, the four components are not necessarily independent of each other.

## 5.9. DETERMINATION OF TREND

Before we go in the detail of methods of measuring secular trend, we must be clear about the purpose of measuring trend. We know that the secular trend is the tendency of time series to grow or to decline over a long period of time. By studying the trend line (or curve) of the profits of a company for a number of years, it can be well-decided as to whether the company is progressing or not. Similarly, by studying the trend of *consumer price index numbers*, we can have an idea about the rate of growth (or decline) in the prices of commodities.

We can also make use of trend characteristics in comparing the behaviour of two different industries in India. It can equally be used for comparing the growth of industries in India with those functioning in some other country.

The secular trend is also used for forecasting. This is achieved by projecting the trend line (curve) for the required future value.

The secular trend is also measured in order to eliminate itself from the given time series. After this, only three components are left and these are studied separately. The following are the methods of measuring the secular trend of a time series:

- (i) Free Hand Graphic Method
- (ii) Semi-Average Method
- (iii) Moving Average Method
- (iv) Least Squares Method.



## 5.10. FREE HAND GRAPHIC METHOD

This is a graphic method. Let  $\{(t_i, y_i): i = 1, 2, \dots, n\}$  be the given time series. On the graph paper, time is measured horizontally, whereas the values of the variable  $y$  are measured vertically. Points  $(t_1, y_1), (t_2, y_2), \dots, (t_n, y_n)$  are plotted on the graph paper. These plotted points are joined by straight lines to get the graph of actual time series data.

In this method, trend line (or curve) is fitted by inspection. This is a subjective method. The trend line (or curve) is drawn through the graph of actual data so that the following are satisfied as far as possible:

(i) The algebraic sum of the deviations of actual values from the trend values is zero.

(ii) The sum of the squares of the deviations of actual values from the trend values is least.

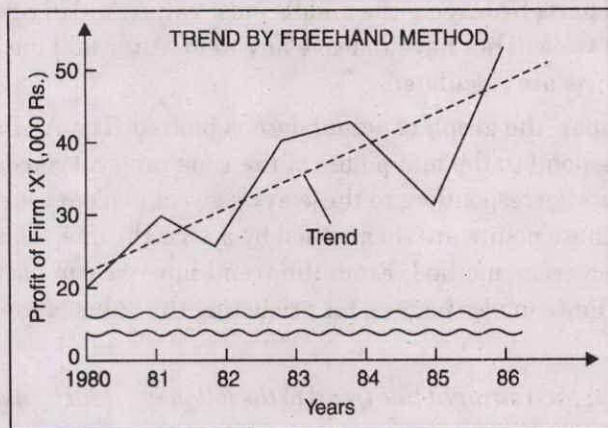
(iii) The area above the trend is equal to area below it.

(iv) The trend line (or curve) is smooth.

**Example 5.1.** Fit a straight line trend to the following data, by using free hand graphic method:

Year	1980	1981	1982	1983	1984	1985	1986
Profit of Firm X ('000 ₹)	20	30	25	40	42	30	50

**Solution.**



### Merits of Free Hand Graphic Method

1. This is the simplest of all the methods of measuring trend.
2. This is a non-mathematical method and it can be used by any one who does not have mathematical background.
3. This method proves very useful for one who is well acquainted with the economic history of the concern, under consideration.
4. For rough estimates, this method is best suited.

### NOTES



## Demerits of Free Hand Graphic Method

### NOTES

1. This method is not rigidly defined.
2. This method is not suited when accurate results are desired.
3. This is a subjective method and can be affected by the personal bias of the person, drawing it.

### EXERCISE 5.1

1. Fit a straight line trend to the following data by using free hand graphic method:

Year	1992	1993	1994	1995	1996	1997
Profit (in ₹)	27000	28000	30000	35000	42000	40000

2. Fit a straight line trend to the following data by using free hand graphic method:

Year	1992	1993	1994	1995	1996	1997	1998
X	10	8	7	15	16	25	30

## 5.11. SEMI-AVERAGE METHOD

This is a method of fitting trend line to the given time series. In this method, we divide the given values of the variable (y) into two parts. If the number of items is odd, then we make two equal parts by leaving the middle most value. And in case, the number of items is even, then we will not have to leave any item. After making two equal parts, the A.M. of both parts are calculated.

On graph paper, the graph of actual data is plotted. The A.M. of two parts are considered to correspond to the mid-points of the time interval considered in making the parts. The points corresponding to these averages of two parts are also plotted on the graph paper. These points are then joined by a straight line. This line represents the trend by semi-average method. From the trend line, we can easily get the trend values. This trend line can also be used for predicting the value of the variable for any future period.

**Example 5.2.** Fit a straight line trend to the following data by using semi-average method :

Year	1981	1982	1983	1984	1985	1986
Cost of Living Index No.	100	110	120	118	130	159

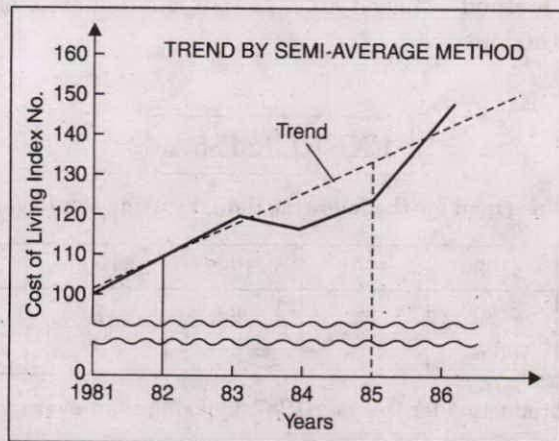
**Solution.**

### Trend Line by Semi-Average Method

Analysis of Time Series

Year	Cost of Living Index	Year	Cost of Living Index
1981	100	1984	118
1982	110	1985	130
1983	120	1986	159

**NOTES**



**Example 5.3.** Fit a straight line trend by using the following data:

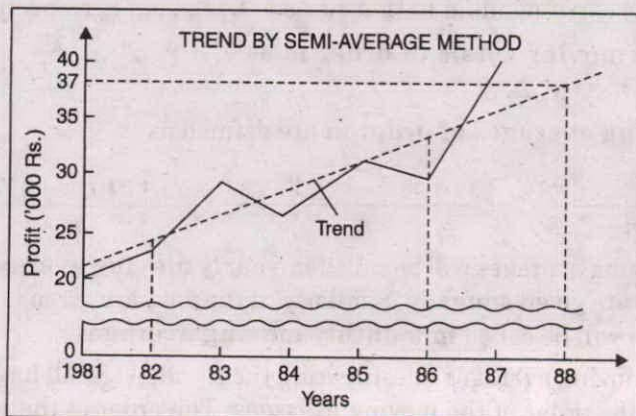
Year	1981	1982	1983	1984	1985	1986	1987
Profit ('000 ₹)	20	22	27	26	30	29	40

*Semi-average Method is to be used. Also estimate the profit for the year 1988.*

**Solution.**

### Trend Line by Semi-Average Method

Year	Profit ('000 ₹)	Year	Profit ('000 ₹)
1981	20	1985	30
1982	22	1986	29
1983	27	1987	40
1984	26		



The estimated profit for the year 1988 is ₹ 37000.

## Merits of Semi-average Method

1. This method is rigidly defined.
2. This method is simple to understand.

## NOTES

## Demerits of Semi-average Method

1. This method assumes a straight line trend, which is not always true.
2. Since this method is based on A.M., all the demerits of A.M. becomes the demerits of this method also.

### EXERCISE 5.2

1. Fit a straight line trend for the following data, by using semi-average method:

Year	1990	1991	1992	1993	1994	1995
Profit ('000 ₹)	80	82	85	70	89	95

2. Estimate the production for the year 1987, by using semi-average method:

Year	1980	1981	1982	1983	1984	1985	1986
Production	50	40	45	55	75	70	72

3. Apply the method of semi-averages for determining trend of the following data and estimate the value for 1990:

Year (March-ending)	1983	1984	1985	1986	1987	1988
Sale (in '000 units)	20	24	22	30	28	32

If the actual figure of sale for 1990 is 35000 units, how do you account for the difference between the figure you obtain and the actual figure given to you.

## 5.12. MOVING AVERAGE METHOD

Let  $\{(t_i, y_i): i = 1, 2, \dots, n\}$  be the given time series. Here  $y_1, y_2, \dots, y_n$  are the values of the variable ( $y$ ) corresponding to time periods  $t_1, t_2, \dots, t_n$  respectively.

We define **moving totals of order  $m$**  as  $y_1 + y_2 + \dots + y_m, y_2 + y_3 + \dots + y_{m+1}, y_3 + y_4 + \dots + y_{m+2}, \dots$

The **moving averages of order  $m$**  are defined as

$$\frac{y_1 + y_2 + \dots + y_m}{m}, \frac{y_2 + y_3 + \dots + y_{m+1}}{m}, \frac{y_3 + y_4 + \dots + y_{m+2}}{m}, \dots$$

These moving averages will be called  **$m$  yearly moving averages** if the values,  $y_1, y_2, \dots, y_n$  of  $y$  are given annually. Similarly, if the data are given monthly, then the moving averages will be called  **$m$  monthly moving averages**.

In using moving averages in estimating the trend, we shall have to decide as to what should be the order of the moving averages. The order of the moving averages

should be equal to the length of the cycles in the time series. In case, the order of the moving averages is given in the problem itself, then we shall use that order for computing the moving averages. The order of the moving averages may either be odd or even.

Let the order of moving averages be 3. The moving averages will be

$$\frac{y_1 + y_2 + y_3}{3}, \frac{y_2 + y_3 + y_4}{3}, \frac{y_3 + y_4 + y_5}{3}, \dots, \frac{y_{n-2} + y_{n-1} + y_n}{3}.$$

These moving averages will be considered to correspond to 2nd, 3rd, 4th, .....  $(n - 1)$ th years respectively.

Similarly, the 5 yearly moving averages will be

$$\frac{y_1 + y_3 + y_3 + y_4 + y_5}{5}, \frac{y_2 + \dots + y_6}{5}, \dots, \frac{y_{n-4} + \dots + y_n}{5}.$$

These 5 yearly moving averages will be considered to correspond to 3rd, 4th, .....  $(n - 2)$ th years respectively. These moving averages are called the trend values.

Calculation of trend values, by using moving averages of *even* order is slightly complicated. Suppose we are to find trend values by using 4 yearly moving averages. The 4 yearly moving averages are:

$$\frac{y_1 + y_2 + y_3 + y_4}{4}, \frac{y_2 + y_3 + y_4 + y_5}{4}, \dots, \frac{y_{n-3} + y_{n-2} + y_{n-1} + y_n}{4}.$$

These moving averages will not correspond to time periods, under consideration. The first moving average will correspond to the mid of  $t_2$  and  $t_3$ . Similarly, others.

In order that these moving averages may correspond to original periods, we will have to resort to a process, called *centering of moving averages*. There are two methods of finding centered moving averages. Suppose we are to find 4 yearly centered moving averages for the times series:

$$\{(t_i, y_i)\}: i = 1, 2, \dots, n\}.$$

### Method I

In this method, we first calculate 4 yearly moving totals from the given data. Of these 4 year moving totals, 2 yearly moving totals are computed. These 2 yearly moving totals are then divided by 8 to get 4 yearly *centered moving averages*. These centered moving averages will correspond to 3rd, 4th, .....  $(n - 2)$ th years, in the table.

### Method II

In this method, we first calculate 4 yearly moving averages. The first 4 yearly moving average will correspond to the mid of 2nd and 3rd years. Similarly, others. We now calculate 2 yearly moving averages of these 4 yearly moving averages. These averages will be 4 yearly *centered moving averages*. These averages will correspond to 3rd, 4th, .....  $(n - 2)$ th years, in the table.

It may be carefully noted that the centered moving averages as calculated by using these methods will be exactly same.

## NOTES

In the moving average method of finding trend, the moving averages will be the trend values. These trend values may be plotted on the graph. The graph of the trend values will not be a straight line, in general.

## NOTES

**Example 5.4.** Compute 5 yearly, 7 yearly and 9 yearly moving averages for the following time series:

Year	Value of the Variable	Year	Value of the Variable
1955	8	1965	9
1956	10	1966	11
1957	11	1967	13
1958	10	1968	9
1959	10	1969	10
1960	9	1970	8
1961	9	1971	11
1962	11	1972	9
1963	7	1973	12
1964	9	1974	11

### Solution.

#### Trend by Moving Average Method

Year	Value of the Variable	5 Yearly m.l.	5 Yearly m.a.	7 Yearly m.l.	7 Yearly m.a.	9 Yearly m.l.	9 Yearly m.a.
1955	8	—	—	—	—	—	—
1956	10	—	—	—	—	—	—
1957	11	49	9.8	—	—	—	—
1958	10	50	10	67	9.57	—	—
1959	10	49	9.8	70	10	85	9.44
1960	9	49	9.8	67	9.57	86	9.55
1961	9	46	9.2	65	9.29	85	9.44
1962	11	45	9	64	9.14	85	9.44
1963	7	45	9	65	9.29	88	9.78
1964	9	47	9.4	69	9.86	87	9.67
1965	9	49	9.8	69	9.86	88	9.78
1966	11	51	10.2	68	9.71	87	9.67
1967	13	52	10.4	69	9.86	87	9.67
1968	9	51	10.2	71	10.14	89	9.89
1969	10	51	10.2	71	10.14	92	10.22
1970	8	47	9.4	72	10.29	94	10.44
1971	11	50	10	70	10	—	—
1972	9	51	10.2	—	—	—	—
1973	12	—	—	—	—	—	—
1974	11	—	—	—	—	—	—

**Example 5.5.** Following figures relate to output of cloth in a factory (output in lakhs of metres):

Year	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976
Output	72	68	64	60	68	72	72	76	72	68

Calculate 4 yearly moving averages.



**Solution.****Trend by Moving Average Method***Analysis of Time Series*

<i>Year</i>	<i>Output</i>	<i>4 yearly moving total</i>	<i>2 yearly moving total of column 3</i>	<i>4 yearly centered moving average</i>
1967	72		—	—
1968	68		—	—
1969	64	264	524	<b>65.5</b>
1970	60	260	524	<b>65.5</b>
1971	68	264	536	<b>67</b>
1972	72	272	560	<b>70</b>
1973	72	288	580	<b>72.5</b>
1974	76	292	580	<b>72.5</b>
1975	72	288	—	—
1976	68		—	—

**NOTES****Merits of Moving Average Method**

1. This method is rigidly defined, so it cannot be affected by the personal prejudice of the person computing it.

2. If the order of the moving averages is exactly equal to the length of the cycle in the time series, the cyclical variations are eliminated.

3. If some more values of the variable are added at the end of the time series, the entire calculations are not changed.

4. This method is best suited for the time series whose trend is not linear. For such series, the general movement of the variable will be best shown by moving averages.

**Demerits of Moving Average Method**

1. Moving averages are strongly affected by the presence of extreme items, in the series.

2. It is difficult to decide the order of the moving averages, because the cycles in time series are seldom regular in duration.

3. In this method, we lose trend values at each end of the series. For example, if the order of the moving averages is five, we lose trend values for two years at each end of the series.

4. Forecasting is not possible in this method, because we cannot objectively project the graph of the trend values, for a future period.

## EXERCISE 5.3

## NOTES

1. Find trend values for the following data, by using 3 yearly moving averages:

<i>Year</i>	<i>Production (Lakh tonnes)</i>	<i>Year</i>	<i>Production (Lakh tonnes)</i>
1973	17.2	1981	25.3
1974	17.3	1982	24.9
1975	17.7	1983	23.2
1976	18.9	1984	24.3
1977	19.2	1985	25.2
1978	19.3	1986	26.3
1979	18.1	1987	27.3
1980	20.2		

2. Calculate a 7 yearly moving average for the following data on the number of commercial and industrial failures in a country during 1929–44:

<i>Year</i>	<i>No. of failures</i>	<i>Year</i>	<i>No. of failures</i>
1929	23	1937	9
1930	26	1938	13
1931	28	1939	11
1932	32	1940	14
1933	20	1941	12
1934	12	1942	9
1935	12	1943	3
1936	10	1944	1

3. Work out the centered 4 yearly moving averages for the following data:

<i>Year</i>	<i>Tonnage of cargo cleared</i>	<i>Year</i>	<i>Tonnage of cargo cleared</i>
1957	1102	1963	1452
1958	1250	1964	1549
1959	1180	1965	1586
1960	1440	1966	1476
1961	1212	1967	1625
1962	1317	1968	1586

4. Obtain the trend of bank clearances by the method of moving averages (assume a five yearly cycle):

<i>Year</i>	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962
<i>Bank Clearance (in crores of rupees)</i>	53	79	76	66	69	94	105	87	79	104	97	92

5. Find the trend values for the following data, by using 4 yearly moving averages:

<i>Year</i>	1980	1981	1982	1983	1984	1985	1986	1987
<i>Sale (in lakhs of rupees)</i>	20	22	25	24	26	30	35	40

6. Calculate trend from the following data by using four yearly moving averages:

Year	Production	Year	Production
1	52.7	8	87.2
2	79.4	9	79.3
3	76.3	10	103.6
4	66.0	11	97.3
5	68.6	12	92.4
6	93.8	13	100.7
7	104.7		

## NOTES

### Answers

- 17.4, 17.967, 18.6, 19.133, 18.867, 19.2, 21.2, 23.467, 24.467, 24.133, 24.233, 25.267, 26.267
- 21.857, 20, 17.571, 15.429, 12.429, 11.571, 11.571, 11.143, 10.143, 9
- 1256.75, 1278.875, 1321.25, 1368.875, 1429.25, 1495.875, 1537.375, 1563.625
- 68.6, 76.8, 82, 84.2, 86.8, 93.8, 94.4, 91.8
- 23.5, 25.25, 27.5, 30.75
- 70.59, 74.38, 79.73, 85.93, 89.91, 92.48, 92.78, 92.50, 95.82

## 5.13. LEAST SQUARES METHOD

This is a mathematical method. Let  $\{(t_i, y_i): i = 1, 2, \dots, n\}$  be the given time series. By using this method, we can find linear trend as well as non-linear trend of the corresponding data.

In this method, trend values ( $y_e$ ) of the variable ( $y$ ) are computed so as to satisfy the following two conditions:

(i) The sum of the deviations of values of  $y$  ( $= y_1, y_2, \dots, y_n$ ) from their corresponding trend values, is zero, i.e.,  $\Sigma(y - y_e) = 0$ .

(ii) The sum of the squares of the deviations of the values of  $y$  from their corresponding trend values is least i.e.,  $\Sigma(y - y_e)^2$  is least.

On the graph paper, we shall measure the actual values and the estimated values (trend values) of the variable  $y$ , along the vertical axis. Let  $x$  denote the deviations of the time periods ( $t_1, t_2, \dots, t_n$ ) from some fixed time period. The fixed time period is called the *origin*.

## 5.14. LINEAR TREND

From the knowledge of coordinate geometry, we know that the equation of the required trend line can be expressed as

$$y_e = a + bx,$$

where  $a$  and  $b$  are constants. We have already mentioned that our trend line will satisfy the conditions:

(i)  $\Sigma(y - y_e) = 0$  and

(ii)  $\Sigma(y - y_e)^2$  is least.

In order to meet these requirements, we will have to use those values of  $a$  and  $b$  in the trend line equation which satisfies the following *normal equations*:

$$\Sigma y = an + b\Sigma x \quad \dots(1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \dots(2)$$

## NOTES

In the equation  $y_e = a + bx$ , of the trend,  $a$  represents the trend value of the variable when  $x = 0$  and  $b$  represents the *slope* of the trend line. If  $b$  is positive, the trend will be upward and if  $b$  is negative, the trend of the time series will be downward.

It is very important to mention the origin and the  $x$  unit with the trend line equation. If either of the two is not given with the equation of the trend, we will not be able to get the trend values of the variable, under consideration.

**Example 5.6.** Calculate trend values by the method of least squares and estimate sales for 1983:

Year	1975	1976	1977	1978	1979	1980	1981
Sale (₹)	800	900	920	930	940	980	930

### Solution. Trend Line by Least Squares Method

S. No.	Year	Sales $y$	$x = \text{year} - 1976$	$x^2$	$xy$	$y_e = a + bx$
1	1975	800	-1	1	-800	$873.572 + 20.357(-1)$ <b>= 853.215</b>
2	1976	900	0	0	0	$873.572 + 20.357(0)$ <b>= 873.572</b>
3	1977	920	1	1	920	$873.572 + 20.357(1)$ <b>= 893.929</b>
4	1978	930	2	4	1860	$873.572 + 20.357(2)$ <b>= 914.286</b>
5	1979	940	3	9	2820	$873.572 + 20.357(3)$ <b>= 934.643</b>
6	1980	980	4	16	3920	$873.572 + 20.357(4)$ <b>= 955.000</b>
$n = 7$	1981	930	5	25	4650	$873.572 + 20.357(5)$ <b>= 975.357</b>
Total		6400	14	56	13370	

Let the equation of the trend line by  $y_e = a + bx$ .

The normal equations are:

$$\Sigma y = an + b\Sigma x \quad \dots(1)$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \dots(2)$$

$$(1) \Rightarrow 6400 = 7a + 14b \quad \dots(3)$$

$$(2) \Rightarrow 13370 = 14a + 56b \quad \dots(4)$$

$$(3) \times 2 \Rightarrow 12800 = 14a + 28b \quad \dots(5)$$

$$(4) - (5) \Rightarrow 570 = 28b \Rightarrow b = 570/28 = 20.357.$$

$$\therefore (3) \Rightarrow 6400 = 7a + 14(570/28) \Rightarrow a = 6115/7 = 873.572.$$

$\therefore$  The equation of the trend line is  $y_e = 873.572 + 20.357x$ , with origin 1976 and  $x$  unit = 1 year.

For 1983,  $x = 1983 - 1976 = 7$ .

$$\therefore y_e(1983) = 873.572 + (20.357)7 = ₹ 1016.071.$$

In the above two examples, we have seen that no particular rule is applied in choosing the origin. It is generally observed that the time periods in the time series are of uniform duration. If this is so, we prefer to take the origin in such a way so as to make  $\Sigma x = 0$ .

If the known values of the variable are *odd* in number, then we take the middle most time period as the origin. This choice would make  $\Sigma x = 0$ .

If the known values of the variable are *even* in number, then we take the A.M. of the two middle most time periods as the origin. Here also, this choice of origin would make  $\Sigma x = 0$ .

If for a time series, the origin is chosen so that  $\Sigma x = 0$ , then the normal equations reduces to

$$\Sigma y = an + b.0 \text{ and } \Sigma xy = a.0 + b\Sigma x^2.$$

$$\therefore a = \frac{\Sigma y}{n} \text{ and } b = \frac{\Sigma xy}{\Sigma x^2}.$$

In practical problems, we prefer to choose origin in such a way that  $\Sigma x = 0$ . This will facilitate the computation of constants  $a$  and  $b$ .

**Example 5.7.** Below are given figures of production (in '000 tonnes) of a sugar factory:

Year	1981	1982	1983	1984	1985	1986	1987
Production	80	90	92	83	94	99	92

Find the slope of a straight line trend to these figures by the method of least squares. (Plot the trend values on the graph).

**Solution.** Here the number of periods is equal to seven. Therefore, we shall take 1984 (the middle most period) as the origin.

#### Linear Trend by Least Square Method

S. No.	Year	Production $y$ (in '000 tonnes)	$x = \text{year} - 1984$	$x^2$	$xy$
1	1981	80	-3	9	-240
2	1982	90	-2	4	-180
3	1983	92	-1	1	-92
4	1984	83	0	0	0
5	1985	94	1	1	94
6	1986	99	2	4	198
$n = 7$	1987	92	3	9	276
Total		630	0	28	56

Let the equation of trend line be  $y_e = a + bx$ .

The normal equations are:

$$\Sigma y = an + b\Sigma x \quad \dots(1)$$

$$\text{and } \Sigma xy = a\Sigma x + b\Sigma x^2 \quad \dots(2)$$

$$(1) \Rightarrow 630 = 7a + b.0 \Rightarrow a = 90$$

$$(2) \Rightarrow 56 = a.0 + 28b \Rightarrow b = 2$$

## NOTES



∴ The equation of trend is  $y_e = 90 + 2x$ , with origin 1984 and  $x$  unit = 1 year.

The slope of the straight line trend is 2. This represents the average rate of increase of  $y$  w.r.t. time. The graph of the trend values is same as that in example 5.1.

## NOTES

**Example 5.8.** Find the trend values for the following series by the method of least squares:

Year	1976	1977	1978	1979	1980	1981
Production (in crores kg)	7	10	12	14	17	24

**Solution.** Here the number of periods is equal to six. Therefore, we take  $\frac{1978 + 1979}{2} = 1978.5$  as the origin. Let  $y$  denote the variable 'production (in crores kg)'.

### Trend Line by Least Squares Method

S. No.	Year	$y$	$x = \text{year} - 1978.5$	$x^2$	$xy$
1	1976	7	-2.5	6.25	-17.5
2	1977	10	-1.5	2.25	-15
3	1978	12	-0.5	0.25	-6
4	1979	14	0.5	0.25	7
5	1980	17	1.5	2.25	25.5
$n = 6$	1981	24	2.5	6.25	60
Total		84	0	17.50	54

Let the equation of trend line be  $y_e = a + bx$ .

The normal equations are:

$$\Sigma y = an + b\Sigma x \quad \dots(1)$$

$$\text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2 \quad \dots(2)$$

$$(1) \Rightarrow 84 = 6a + b(0) \Rightarrow a = \frac{84}{6} = 14$$

$$(2) \Rightarrow 54 = a(0) + b(17.5) \Rightarrow b = \frac{54}{17.5} = 3.0857$$

∴ The equation of trend line is  $y_e = 14 + 3.0857x$ , with origin 1978.5 and  $x$  unit = 1 year.

### Trend Values

For 1976,	$x = -2.5$	∴ $y_e(1976) = 14 + (3.0857)(-2.5) = 6.2857$
For 1977,	$x = -1.5$	∴ $y_e(1977) = 14 + (3.0857)(-1.5) = 9.3714$
For 1978,	$x = -0.5$	∴ $y_e(1978) = 14 + (3.0857)(-0.5) = 12.4571$
For 1979,	$x = 0.5$	∴ $y_e(1979) = 14 + (3.0857)(0.5) = 15.5428$
For 1980,	$x = 1.5$	∴ $y_e(1980) = 14 + (3.0857)(1.5) = 18.6285$
For 1981,	$x = 2.5$	∴ $y_e(1981) = 14 + (3.0857)(2.5) = 21.7142$

**Example 5.9.** Below are given figures of production (in thousand tonnes) of a sugar factory:

Year	1976	1978	1979	1980	1981	1982	1985
Production	77	88	94	85	91	98	90

## NOTES

Fit a straight line by the least squares method and calculate the trend values.

**Solution.** We define  $x = \text{year} - 1980$  and  $y = \text{production}$ .

### Trend Line by Least Squares Method

S. No.	Year	$y$	$x = \text{year} - 1980$	$x^2$	$xy$
1	1976	77	-4	16	-308
2	1978	88	-2	4	-176
3	1979	94	-1	1	-94
4	1980	85	0	0	0
5	1981	91	1	1	91
6	1982	98	2	4	196
$n = 7$	1985	90	5	25	450
		623	1	51	159

Let the equation of the trend line be  $y_e = a + bx$ .

The normal equations are:

$$\Sigma y = an + b \Sigma x \quad \dots(1)$$

and  $\Sigma xy = a \Sigma x + b \Sigma x^2 \quad \dots(2)$

$$(1) \Rightarrow 623 = 7a + b \quad \dots(3)$$

$$(2) \Rightarrow 159 = a + 51b \quad \dots(4)$$

$$(4) \times 7 \Rightarrow 1113 = 7a + 357b \quad \dots(5)$$

$$(5) - (3) \Rightarrow 490 = 356b \Rightarrow b = \frac{490}{356} = 1.376$$

$$\therefore (4) \Rightarrow a = 159 - 51b = 159 - 51(1.376) = 88.824$$

$\therefore$  The equation of the trend line is  $y_e = 88.824 + 1.376x$  with origin = 1980 and  $x$  unit = 1 year.

### Trend values

For 1976,  $x = -4$ .  $\therefore y_e(1976) = 88.824 + 1.376(-4) = 83.32$

For 1978,  $x = -2$ .  $\therefore y_e(1978) = 88.824 + 1.376(-2) = 86.072$

For 1979,  $x = -1$ .  $\therefore y_e(1979) = 88.824 + 1.376(-1) = 87.448$

For 1980,  $x = 0$ .  $\therefore y_e(1980) = 88.824 + 1.376(0) = 88.824$

For 1981,  $x = 1$ .  $\therefore y_e(1981) = 88.824 + 1.376(1) = 90.2$

For 1982,  $x = 2$ .  $\therefore y_e(1982) = 88.824 + 1.376(2) = 91.576$

For 1985,  $x = 5$ .  $\therefore y_e(1985) = 88.824 + 1.376(5) = 95.704$

**EXERCISE 5.4****NOTES**

1. Fit a straight line trend by the method of Least Squares for the following series:

Year	1981	1982	1983	1984	1985	1986
Production	7	17	12	19	22	27

2. Below are given the production (thousand quintals) figures of a sugar factory. Fit a straight line by Least Squares method and tabulate the trend values:

Year	1972	1973	1974	1975	1976	1977	1978
Production	12	10	14	11	13	15	16

3. Find out trend values by the method of Least Squares for the following series:

Year	1980	1981	1982	1983	1984	1985
Production (in lakh units)	7	10	12	14	17	24

4. Fit a straight line trend for the following series by the method of least squares. Also, estimate the value for the year 1993:

Year	1984	1985	1986	1987	1988	1989	1990
Output	125	128	133	135	140	141	143

5. Compute secular trend by least square method from the following data:

Year	1970	1971	1972	1973	1974	1975	1976
Supply	23	25	26	24	25	29	30

6. You are given the annual profits (in ,000) for a certain firm for the years 1982-1988. Make an estimate of profit for the year 1989. You may assume linear trend in profits:

Year	1982	1983	1984	1985	1986	1987	1988
Profit (in '000 ₹)	60	72	75	65	80	85	95

7. Explain clearly what is meant by time series analysis.

The following are the figures of production (in thousand tonnes) of a sugar factory:

Year	1941	1942	1943	1944	1945
Production	80	90	92	83	94

Fit a straight line by the least squares method.

8. The sales figures of a company in lakhs of rupees for the years 1974-1981 are given below:

Year	1974	1975	1976	1977	1978	1979	1980	1981
Sales	550	560	555	585	540	525	545	585

Fit a linear trend equation and estimate the sales for the year 1973.

9. Calculate trend values from the following data by applying the method of least squares:

Year	1973	1974	1975	1976	1977	1978	1979
Sales (in crore rupees)	20	23	22	25	26	29	30

10. Fit a straight line trend by the least squares method for the following data:

Year	1951	1961	1971	1981	1991
y	34	50	67	75	85

Estimate the value of y of for the year 2001.

### Answers

- $y_e = 5.12 + 3.49x$  where origin = 1981,  $x$  unit = 1 year
- $y_e = 13 + 0.75x$ , with origin = 1975 and  $x$  unit = 1 year. Trend values are 10.75, 11.5, 12.25, 13, 13.75, 14.5, 15.25
- Trend values (in lakh units) : 6.2857, 9.3714, 12.4571, 15.5428, 18.6285, 21.7142
- $y_e = 135 + 3.1x$ , where origin = 1987 and  $x$  unit = 1 year; 153.6
- $y_e = 26 + x$ , where origin = 1973 and  $x$  unit = 1 year
- ₹ 95428.40
- $y_e = 87.8 + 2.1x$ , where origin = 1983 and  $x$  unit = 1 year
- $y_e = 555.625 + 0.4167x$ , where origin = 1977.5 and  $x$  unit = 1 year.  $y_e$  (1973) = 553.7498
- 20.071, 21.714, 23.357, 25, 26.643, 28.286, 29.929 ; estimated value for 1982 = 34.858
- $y_e = 62.2 + 1.27x$  where origin = 1971 and  $x$  unit = 1 year,  $y_e$  (2001) = 100.3

## 5.15. NON-LINEAR TREND (PARABOLIC)

There are situations where linear trend is not found suitable. Linear trend is suitable when the tendency of the actual data is to move approximately in one direction. There are number of curves representing non-linear trend. In the present section, use shall consider parabolic trends. Parabolic trends will give better trend than the straight line trends.

Let  $\{(t_i, y_i) : i = 1, 2, \dots, n\}$  be the given time series. Let  $x$  denote the deviations of the time periods  $(t_1, t_2, \dots, t_n)$  from some fixed time period, called the origin. Let  $y_e$  denote the estimated (trend) values of the variable.

Let the equation of the required parabolic trend curve be

$$y_e = a + bx + cx^2$$

where,  $a, b, c$  are constants. This trend curve will satisfy the conditions:

(i)  $\Sigma(y - y_e) = 0$

(ii)  $\Sigma(y - y_e)^2$  is least.

In order to meet these requirements, we will have to use those values of  $a, b$  and  $c$  in the trend curve equation which satisfies the following *normal equations*:

$$\Sigma y = an + b\Sigma x + c\Sigma x^2 \quad \dots(1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \dots(2)$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \quad \dots(3)$$

Here also, it is very important to mention the origin and the  $x$  unit with the trend curve equation.

## NOTES

## NOTES

There is no specific rule for choosing the origin. But if we manage to choose the origin so as to make  $\Sigma x = 0$ , then we shall be reducing the calculation involved in computing  $a$ ,  $b$  and  $c$ . In case the time periods  $t_1, t_2, \dots, t_n$  advances by equal intervals and  $\Sigma x = 0$ , then we will also have  $\Sigma x^3 = 0$ . Here, the normal equations will reduce to:

$$\Sigma y = an + b.0 + c\Sigma x^2$$

$$\Sigma xy = a.0 + b\Sigma x^2 + c.0$$

$$\Sigma x^2 y = a\Sigma x^2 + b.0 + c\Sigma x^4$$

or

$$\Sigma y = an + c\Sigma x^2 \quad \dots(1')$$

$$\Sigma xy = b\Sigma x^2 \quad \dots(2')$$

$$\Sigma x^2 y = a\Sigma x^2 + c\Sigma x^4 \quad \dots(3')$$

(2')  $\Rightarrow b = \Sigma xy / \Sigma x^2$ . The values of  $a$  and  $c$  will be obtained by solving the equations (1') and (3').

**Example 5.10.** The following table shows our urban population as percentage of total population (1921-1961):

Census year	1921	1931	1941	1951	1961
% of total population	11.4	12.1	13.9	17.3	18.0

Compute the second degree trend equation for the data given above and from the equation obtained, determine the trend value for the census year 1991.

**Solution.** Here the number of periods is five. Therefore, we take 1941 as the origin.

Let  $y$  denote the variable “% of total population”.

### Second Degree Trend Equation by Least Squares Method

S. No.	Year	$y$	$x$	$x^2$	$x^3$	$x^4$	$xy$	$x^2y$
1	1921	11.4	-20	400	-8000	160000	-228	4560
2	1931	12.1	-10	100	-1000	10000	-121	1210
3	1941	13.9	0	0	0	0	0	0
4	1951	17.3	10	100	1000	10000	173	1730
$n = 5$	1961	18.0	20	400	8000	160000	360	7200
Total		72.7	0	1000	0	340000	184	14700

Let the second degree trend equation be

$$y_e = a + bx + cx^2.$$

The normal equations are:

$$\Sigma y = an + b\Sigma x + c\Sigma x^2 \quad \dots(1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \dots(2)$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \quad \dots(3)$$

or

$$72.7 = 5a + b.0 + 1000c$$

$$184 = a.0 + 1000b + c.0$$

$$14700 = 1000a + b.0 + 340000c$$

or

$$72.7 = 5a + 1000c \quad \dots(4)$$

$$184 = 1000b \quad \dots(5)$$

$$14700 = 1000a + 340000c \quad \dots(6)$$



$$(5) \Rightarrow b = 184/1000 = 0.184$$

$$(4) \times 200 \Rightarrow 14540 = 1000a + 200000c \quad \dots(7)$$

$$(6) - (7) \Rightarrow 160 = 0 + 140000c \Rightarrow c = 0.001143$$

$$\therefore (4) \Rightarrow 72.7 = 5a + 1000(0.001143) \Rightarrow a = 14.3114$$

$\therefore$  The required equation of trend is

$$y_e = 14.3114 + 0.184x + 0.001143x^2, \text{ with origin} = 1941 \text{ and } x \text{ unit} = 1 \text{ year.}$$

$$\text{For 1991, } x = 1991 - 1941 = 50.$$

$$\therefore y_e(1991) = 14.3114 + 0.184(50) + 0.001143(50)^2 = 26.3689.$$

$\therefore$  The estimated percent of urban population for the census year 1991 = 26.3689%.

## NOTES

### EXERCISE 5.5

1. Find the equation of parabolic trend of second degree to the following data:

Year	1980	1981	1982	1983	1984	1985	1986
Outstanding loan of company 'X' (in thousand ₹)	83	60	54	21	22	13	13

2. Fit a second degree parabolic trend to the data given below:

Year	1982	1983	1984	1985	1986
Variable	7	8	10	15	20

3. The following are the production figures of an aluminium plant for the years 1990 to 2002:

Year	Production (in '000 tonnes)	Year	Production (in '000 tonnes)
1990	12	1997	21
1991	20	1998	30
1992	10	1999	35
1993	11	2000	40
1994	12	2001	37
1995	13	2002	40
1996	10		

- Find the equation of parabolic trend.
- Find the trend values for the years 1990—2002.
- Plot the original data and trend values on a graph paper.
- Estimate the production figure for the years 2003 and 2004.

### Answers

- $y_e = 30 - 12x + 2x^2$ , where origin = 1983 and  $x$  unit = 1 year.
- $y_e = 10.4286 + 3.3x + 0.7857x^2$ , where origin = 1984 and  $x$  unit = 1 year.

3. (i)  $y = 17.9 + 2.69x + 0.312x^2$  where origin = 1996 and  $x$  unit = 1 year.  
 (ii) 13.28, 12.45, 12.26, 12.71, 13.80, 15.53, 17.90, 20.91, 24.56, 28.85, 33.78, 39.35, 45.56 thousand tonnes.  
 (iv) 52.41 thousand tonnes, 59.9 thousand tonnes.

## NOTES

## 5.16. NON-LINEAR TREND (EXPONENTIAL)

In this section, we shall study the method of finding non-linear exponential trend of a given time series.

Let  $\{(t_i, y_i): i = 1, 2, \dots, n\}$  be the given time series. Let  $x$  denote the deviations of the time periods  $\{t_1, t_2, \dots, t_n\}$  from some fixed time period, called the origin. Let  $y_e$  denote the estimated (trend) values of the variable.

Let the equation of the required exponential trend curve be

$$y_e = ab^x \quad \dots(1)$$

where  $a, b$  are constants.

$$(1) \Rightarrow \log y_e = \log a + x \log b. \quad \dots(2)$$

The exponential trend curve will satisfy the conditions:

- (i)  $\Sigma(\log y - \log y_e) = 0$   
 (ii)  $\Sigma(\log y - \log y_e)^2$  is least.

In order to meet these requirements we will have to use those values of  $a$  and  $b$  in the trend curve equation which satisfies the following *normal equations*:

$$\Sigma \log y = (\log a)n + (\log b)\Sigma x \quad \dots(3)$$

$$\Sigma x \log y = (\log a) \Sigma x + (\log b) \Sigma x^2. \quad \dots(4)$$

Here also, it is very important to mention the origin and the  $x$  unit with the trend curve equation.

If origin be chosen so that  $\Sigma x = 0$ , then the above normal equations reduces to

$$\Sigma \log y = (\log a)n + (\log b).0$$

and

$$\Sigma x \log y = (\log a).0 + (\log b) \Sigma x^2.$$

$$\therefore \log a = \frac{\Sigma \log y}{n} \quad \text{and} \quad \log b = \frac{\Sigma x \log y}{\Sigma x^2}.$$

$$\therefore a = AL \left( \frac{\Sigma \log y}{n} \right) \quad \text{and} \quad b = AL \left( \frac{\Sigma x \log y}{\Sigma x^2} \right).$$

In practical problems, we prefer to choose origin in such a way that  $\Sigma x = 0$ . This will facilitate the computation of constants  $a$  and  $b$ .

**Example 5.11.** You are given the population figures of India as follows:

Census year	1911	1921	1931	1941	1951	1961	1971
Population (in crores)	25.0	25.1	27.9	31.9	36.1	43.9	54.7

Fit an exponential trend to the above data by the method of least squares and find the trend values. Also estimate the population for 1991 and 2001.

**Solution.** Here the number of periods is equal to seven, an odd number.

$\therefore$  We take 1941 (the middle most period) as the origin.



## Exponential Trend by Least Squares Method

Analysis of Time Series

S. No.	Census year	Population (in crores)	$\log y$	$x = \frac{\text{year} - 1941}{10}$	$x^2$	$x \log y$
1	1911	25.0	1.3979	-3	9	-4.1937
2	1921	25.1	1.3997	-2	4	-2.7994
3	1931	27.9	1.4456	-1	1	-1.4456
4	1941	31.9	1.5038	0	0	0
5	1951	36.1	1.5575	1	1	1.5575
6	1961	43.9	1.6425	2	4	3.2850
7	1971	54.7	1.7380	3	9	5.2140
$n = 7$			$\Sigma \log y = 10.6850$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma x \log y = 1.6178$

### NOTES

Let the equation of the exponential trend be  $y = ab^x$ .

$$\therefore \log y = \log a + x \log b \quad \dots(1)$$

The normal equations are:

$$\Sigma \log y = (\log a)n + (\log b) \Sigma x \quad \dots(2)$$

and  $\Sigma x \log y = (\log a) \Sigma x + (\log b) \Sigma x^2 \quad \dots(3)$

$$(2) \Rightarrow 10.6850 = 7 \log a + (\log b).0 \Rightarrow \log a = \frac{10.6850}{7} = 1.5264$$

$$(3) \Rightarrow 1.6178 = (\log a).0 + (\log b).28 \Rightarrow \log b = \frac{1.6178}{28} = 0.0578$$

$$\therefore (1) \Rightarrow \log y_e = 1.5264 + 0.0578x \quad \dots(4)$$

Also  $\log a = 1.5264 \Rightarrow a = \text{AL } 1.5264 = 33.60$

and  $\log b = 0.0578 \Rightarrow b = \text{AL } 0.0578 = 1.142$

$$\therefore y_e = ab^x \Rightarrow y_e = 33.6 \times (1.142)^x, \text{ where } x = \frac{\text{year} - 1941}{10}$$

This represents the exponential trend equation.

### Trend values

For 1911,  $x = -3$  and  $y_e = 33.6 \times (1.142)^{-3} = 22.5601$  crores

For 1921,  $x = -2$  and  $y_e = 33.6 \times (1.142)^{-2} = 33.6 \times (1.142)^{-3} \times 1.142$   
 $= 22.5601 \times 1.142 = 25.7636$  crores

For 1931,  $x = -1$  and  $y_e = 33.6 \times (1.142)^{-1} = 33.6 \times (1.142)^{-2} \times 1.142$   
 $= 25.7636 \times 1.142 = 29.422$  crores

For 1941,  $x = 0$  and  $y_e = 33.6 \times (1.142)^0 = 33.6$  crores

For 1951,  $x = 1$  and  $y_e = 33.6 \times (1.142)^1 = 38.3712$  crores

For 1961,  $x = 2$  and  $y_e = 33.6 \times (1.142)^2 = 33.6 \times 1.142 \times 1.142$   
 $= 38.3712 \times 1.142 = 43.8199$  crores

For 1971,  $x = 3$  and  $y_e = 33.6 \times (1.142)^3 = 33.6 \times (1.142)^2 \times 1.142$   
 $= 43.8199 \times 1.142 = 50.0423$  crores

## Estimated population for 1991 and 2001

For 1991,  $x = \frac{1991 - 1941}{10} = 5.$

$$\therefore y_e(1991) = 33.6 \times (1.142)^5 = 33.6 \times (1.142)^3 \times (1.142)^2 \\ = 50.0423 \times (1.142)^2 = 65.2634 \text{ crores.}$$

For 2001,  $x = \frac{2001 - 1941}{10} = 6.$

$$\therefore y_e(2001) = 33.6 \times (1.142)^6 \\ = 33.6 \times (1.142)^5 \times 1.142 = 65.2634 \times 1.142 \\ = 74.5408 \text{ crores.}$$

## NOTES

## EXERCISE 5.6

1. Fit an exponential trend to the following data:

Year	1998	1999	2000	2001	2002
$y$	1.6	4.5	13.8	40.2	135.0

2. Fit an exponential trend to the following data:

Year	1996	1997	1998	1999	2000
Profit (,000 ₹)	65	92	132	190	275

3. Growth of Indian merchant shipping fleet from 1968 to 1977 is given below. Fit a trend function  $y = AB^x$  where  $y$  represents shipping fleet measured in million gross registered tonnes and  $x$  is the year while  $A$  and  $B$  are constants:

Year	Shipping fleet (million tonnes)	Year	Shipping fleet (million tonnes)
1968	1.95	1973	2.89
1969	2.24	1974	3.49
1970	2.40	1975	3.87
1971	2.48	1976	5.09
1972	2.65	1977	5.48

## Answers

- $y = 13.79 (2.977)^x$ , where  $x = \text{year} - 2000$
- $y = 133 (1.43)^x$ , where  $x = \text{year} - 1998$
- $y = 3.07 (1.06)^u$ , where  $u = 2(x - 1972.5)$ .

## 5.17. SUMMARY

- A **time series** is a collection of values of a variable taken at different time periods. If  $y_1, y_2, \dots, y_n$  be the values of a variable  $y$  taken at time periods  $t_1, t_2, \dots, t_n$ , then we write this time series as  $\{(t_i, y_i); i = 1, 2, \dots, n\}$ .

- The general tendency of the values of the variable in a time series to grow or to decline over a long period of time is called **secular trend** of the times series. It indicates the general direction in which the graph of the time series appears to be going over a long period of time.
- The **seasonal variations** in a time series counts for those variations in the series which occur annually. In a time series, seasonal variations occur quite regularly. These variations play a very important role in business activities.
- The **cyclical variations** in a time series counts for the swings of graph of time series about its trend line (curve). Cyclical variations are seldom periodic and they may or may not follow same pattern after equal interval of time.
- The **irregular variations** in a time series counts for those variations which cannot be predicted before hand. This component is different from the other three components in the sense that irregular variations in a time series are very irregular.

## NOTES

---

### 5.18. REVIEW EXERCISES

---

1. Describe briefly the various characteristic movements of time series. Discuss briefly any one procedure for estimating secular trend.
2. Critically examine the different methods of measuring trend. Point out their merits and demerits.
3. Write a short note on semi-average method of estimating trend of time series.
4. Discuss the components of time series, in detail.
5. What is the time series analysis? What are the components of time series? Explain the various methods of estimating the secular trend of a time series.



## 6. INDEX NUMBERS

### NOTES

#### STRUCTURE

- 6.1. Introduction
- 6.2. Definition and Characteristics of Index Numbers
- 6.3. Uses of Constructing Index Numbers
- 6.4. Types of Index Numbers

#### I. Price Index Numbers

- 6.5. Methods
- 6.6. Simple Aggregative Method
- 6.7. Simple Average of Price Relatives Method
- 6.8. Laspeyre's Method
- 6.9. Paasche's Method
- 6.10. Dorbish and Bowley's Method
- 6.11. Fisher's Method
- 6.12. Marshall Edgeworth's Method
- 6.13. Kelly's Method
- 6.14. Weighted Average of Price Relatives Method
- 6.15. Chain Base Method

#### II. Quality Index Numbers

- 6.16. Methods
- 6.17. Index Numbers of Industrial Production

#### III. Value Index Numbers

- 6.18. Simple Aggregative Method
- 6.19. Mean of Index Numbers

#### IV. Tests of Adequacy of Index Number Formulae

- 6.20. Meaning
- 6.21. Unit Test (U.T.)
- 6.22. The Reversal Test (T.R.T.)
- 6.23. Factor Reversal Test (F.R.T.)
- 6.24. Circular Test (C.T.)

#### V. Consumer Price Index Numbers (C.P.I.)

- 6.25. Meaning
- 6.26. Significance of C.P.I.
- 6.27. Assumptions
- 6.28. Procedure
- 6.29. Methods
- 6.30. Aggregative Expenditure Method
- 6.31. Family Budget Method
- 6.32. Summary
- 6.33. Review Exercises

## 6.1. INTRODUCTION

We are generally interested in knowing as to whether the price level of a particular group of commodities is rising or falling. A teacher is interested in estimating the growth of intelligence in his students. Government may declare that the exports have increased during the current year. In all such statements, it is not possible to measure the changes in the concerned variables directly. If the exports for the current year have increased, it may not mean that exports of every item has increased. Exports of different items might have increased in different proportions, even the exports might have decreased for some of the items. We may compare the general price level of commodities in 1986 with that of price level in 1980. For this purpose, we will have to take into account the prices of all important items for both years. But, the percentage rise or fall in the prices of items is not expected to be same for each item. Had it been so, we would have immediately declared the rise or fall in the general price level of items in 1986. The change in price vary for different items. The percentage rise may be different for different commodities. It may even decrease for some items as well. Under such circumstances, we feel the necessity of some statistical device which may help us in facing such problems. The statistical devices used to measure such changes are called *Index Numbers*. Let us define 'index numbers' in a formal way.

## NOTES

## 6.2. DEFINITION AND CHARACTERISTICS OF INDEX NUMBERS

The **index numbers** are defined as specialized averages used to measure change in a variable or a group of related variables with respect to time or geographical location or some other characteristic.

In our course of discussion, we shall restrict ourselves to the study of changes in a group of related variables with respect to time only. Changes in related variables are expressed clearly by using index numbers, because these are generally expressed as percentages.

The index numbers are used to measure the change in production, prices, values, etc. in related variables over time or geographical location. The barometers are used to study changes in whether conditions, similarly the index numbers are used to study the changes in economic and business activities. That is, why, the index numbers are also called '**Economic Barometers**'.

## 6.3. USES OF CONSTRUCTING INDEX NUMBERS

1. Index numbers are used for computing real incomes from money incomes. The wages, dearness allowances, etc. are fixed on the basis of real income. The money income is divided by an appropriate consumer's price index number to get real income.

2. Index numbers are constructed to compare the changes in related variables over time. Index numbers of industrial production can be used to see the change in the production that has occurred in the current period.

3. Index numbers are used to study the changes occurred in the past. This knowledge helps in forecasting.

4. Index numbers are used to study the changes in prices, industrial production, purchasing powers of money, agricultural production, etc. of different countries. With the use of index numbers, the comparative study is also made possible for such variables.

**NOTES**

---

**6.4. TYPES OF INDEX NUMBERS**

---

There are mainly three types of index numbers:

- I. Price Index Numbers,
- II. Quantity Index Numbers,
- III. Value Index Numbers.

In our course of discussion, we shall confine mainly to 'Price Index Numbers'. Price index numbers measure the changes in prices of commodities in the current period in comparison with the prices of commodities in the base period.

**I. PRICE INDEX NUMBERS**

---

**6.5. METHODS**

---

For constructing price index numbers, the following methods are used:

- (i) Simple Aggregative Method
- (ii) Simple Average of Price Relatives Method
- (iii) Laspeyre's Method
- (iv) Paasche's Method
- (v) Drobish and Bowley's Method
- (vi) Fisher's Method
- (vii) Marshall Edgeworth's Method
- (viii) Kelly's Method
- (ix) Weighted Average of Price Relatives Method
- (x) Chain Base Method

First nine methods are fixed base methods of constructing price index number.

---

**6.6. SIMPLE AGGREGATIVE METHOD**

---

This is the simplest method of computing index number. In this method, we have

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

where 0 and 1 suffixes stand for base period and current period respectively.

$P_{01}$  = price index number for the current period

$\sum p_1$  = sum of prices of commodities per unit in the current period

$\sum p_0$  = sum of prices of commodities per unit in the base period.

In other words, this price index number is the sum of prices of commodities in the current period expressed as percentage of the sum of prices in the base period. Consider the data:



Item	Price in base period $p_0$ (in ₹)	Price in current period $P_1$ (in ₹)
A	5	6
B	8	10
C	18	27
D	112	84
E	12	15
F	6	9
Total	$\Sigma p_0 = 161$	$\Sigma p_1 = 151$

Here 
$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{151}{161} \times 100 = 93.79.$$

This index number shows that there is fall in the prices of commodities to the extent of 6.21%. It may be noted that the prices of every item has increased in the current period except for the item *D*. On the other hand, the index number is declaring a decrease in prices on an average. This is not in consistency with the definition of index numbers. In fact, this unwanted result is due to the presence of an extreme item (*D*) in the series. So, in the presence of extreme items, this method is liable to give misleading results. This is a demerit of this method.

Let us find price index number for the data given below:

Item	Unit	Price (in ₹)	
		1994 ( $p_0$ )	1996 ( $p_1$ )
Sugar	kg	6	7
Milk	litre	3	4
Ghee	kg	45	50

Here  $\Sigma p_0 = 6 + 3 + 45 = 54$

and  $\Sigma p_1 = 7 + 4 + 50 = 61$

$$\therefore P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{61}{54} \times 100 = 112.96.$$

Here we have considered the price of sugar per kg. Now we use the price of sugar per quintal, for calculating index number for the year 1996.

Item	Unit	Price (in ₹)	
		1994 ( $p_0$ )	1996 ( $p_1$ )
Sugar	quintal	600	700
Milk	litre	3	4
Ghee	kg	45	50

In this case,  $\Sigma p_0 = 600 + 3 + 45 = 648$

and  $\Sigma p_1 = 700 + 4 + 50 = 754$

$$\therefore P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{754}{648} \times 120 = 116.36.$$

The index number has changed, whereas we have not affected any change in the data except for writing the price of sugar in a different unit. This type of variation in the value of index numbers is beyond one's expectation. This is another limitation with this method.

## NOTES

## 6.7. SIMPLE AVERAGE OF PRICE RELATIVES METHOD

### NOTES

Before introducing this method of finding index number, we shall first explain the concept of 'price relative'. The price relative of a commodity in the current period with respect to base period is defined as the price of the commodity in the current period expressed as a percentage of the price in the base period. Mathematically,

$$\text{Price Relative (P)} = \frac{P_1}{P_0} \times 100.$$

For example, if the prices of a commodity be ₹ 5 and ₹ 6 in the years 1995 and 1996 respectively, then the price relative of the commodity in 1996 w.r.t. 1995 is

$$\frac{6}{5} \times 100 = 120.$$

In the simple average of price relatives method of computing index numbers, simple average of price relatives of all the items is the required index number.

Mathematically,

$$P_{01} = \frac{\sum \left( \frac{P_1}{P_0} \times 100 \right)}{n} \quad (\text{if A.M. is used})$$

i.e.,

$$P_{01} = \frac{\sum P}{n}$$

where  $P_{01}$  is the required price index number,

$$\frac{P_1}{P_0} \times 100 = \text{Price relative} = P$$

$n$  = no. of commodities under consideration.

In averaging price relatives, geometric mean is also used. In this case, the formula is

$$P_{01} = \text{Antilog} \left( \frac{\sum \log P}{n} \right)$$

It has already been observed that the index number computed by using simple aggregative method is unduly affected by the extreme items, present in the series.

We shall just show that this method of computing index number is not at all affected by the extreme items. We compute the index number for the data considered in the previous method.

Index No. by Simple A.M. of P.R. Method

Item	Price in the base period ( $P_0$ ) (in ₹)	Price in the current period ( $P_1$ ) (in ₹)	Price Relatives $P = \frac{P_1}{P_0} \times 100$
A	6	6	120
B	8	10	125
C	18	27	150
D	112	84	75
E	12	15	125
F	6	9	150
			$\sum P = 745$



$$\therefore P_{01} = \frac{\Sigma P}{n} = \frac{745}{6} = 124.17.$$

Here the index number is advocating the fact that the prices of commodities have raised on an average.

There is one more advantage of using this method. The index number, computed by averaging the price relatives is not affected by the change in measuring unit of any commodity. We illustrate this by using the data taken in the previous method:

Item	Unit	$p_0$	$p_1$	$P = \frac{p_1}{p_0} \times 100$
Sugar	kg	6	7	116.67
Milk	litre	3	4	133.33
Ghee	kg	45	50	111.11
				$\Sigma P = 361.11$

$$\therefore P_{01} = \frac{\Sigma P}{n} = \frac{361.11}{3} = 120.37.$$

Now, we consider this data once again and change the measuring units for sugar:

Item	Unit	$p_0$	$p_1$	$P = \frac{p_1}{p_0} \times 100$
Sugar	quintal	600	700	116.67
Milk	litre	3	4	133.33
Ghee	kg	45	50	111.11
				$\Sigma P = 361.11$

$$\therefore P_{01} = \frac{\Sigma P}{n} = \frac{361.11}{3} = 120.37.$$

We see that this index number is same as that for the data when the rate of sugar was expressed in kg.

Thus, the index number as calculated by this method is not affected by changing measuring units.

In averaging the price relatives, we can also make use of median, harmonic mean, etc. But, only A.M. and G.M. are generally used for this purpose.

**Example 6.1.** Calculate index number for 1994 on the basis of the prices of 1991 for the following data:

Article	A	B	C	D	E
Prices in 1991	12	25	10	5	6
Prices in 1994	15	20	12	10	15

## NOTES

**Solution.****Calculation of Inedx Nos (1991) = 100****NOTES**

Article	$P_0$	$P_1$	$P = \frac{P_1}{P_0} \times 100$
A	12	15	$\frac{15}{12} \times 100 = 125$
B	25	20	$\frac{20}{25} \times 100 = 80$
C	10	12	$\frac{12}{10} \times 100 = 120$
D	5	10	$\frac{10}{5} \times 100 = 200$
E	6	15	$\frac{15}{6} \times 100 = 250$
	$\Sigma P_0 = 58$	$\Sigma P_1 = 72$	$\Sigma P = 775$

By simple aggregative method

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{72}{58} \times 100 = 124.41.$$

By A.M. of price relatives method

$$P_{01} = \frac{\Sigma P}{n} = \frac{775}{5} = 155.$$

**Example 6.2.** From the information given below, prepare index numbers of prices for three years with average price as base:

*Rate per rupee*

Year	Wheat	Rice	Sugar
1st year	1.38 kg	1 kg	0.40 kg
2nd year	1.6 kg	0.8 kg	0.40 kg
3rd year	1 kg	0.75 kg	0.25 kg

**Solution.** Since the prices of commodities are given in the form of 'quantity prices', we shall convert these quantity prices into 'money prices'.

Price of wheat in the 1st year

$$= 2 \text{ kg per rupee}$$

 $\therefore$  Price of wheat per quintal

$$= \frac{100}{2} = ₹ 50$$

Similarly, we shall express the prices of other commodities per quintal.

Com- modity	Unit	1st year		2nd year		3rd year		Average price $P_0$
		$P_1$	$P$	$P_1$	$P$	$P_1$	$P$	
Wheat	Quintal	$\frac{100}{1.38} = 72.46$	$\frac{50}{78.33} \times 100$ = 63.83	$\frac{100}{16} = 62.5$	$\frac{62.5}{78.33} \times 100$ = 79.79	$\frac{100}{1} = 100$	$\frac{100}{78.33} \times 100$ = 127.67	$\frac{72.46 + 62.5 + 100}{3}$ = 78.33
Rice	Quintal	$\frac{100}{1} = 100$	$\frac{100}{119.44} \times 100$ = 83.72	$\frac{100}{0.8} = 125$	$\frac{125}{119.44} \times 100$ = 104.66	$\frac{100}{0.75} \times 100$ = 133.33	$\frac{133.33}{119.44} \times 100$ = 111.63	$\frac{100 + 125 + 133.33}{3}$ = 119.44
Sugar	Quintal	$\frac{100}{0.4} = 250$	$\frac{250}{300} \times 100$ = 83.33	$\frac{100}{0.4} = 250$	$\frac{250}{300} \times 100$ = 83.33	$\frac{100}{0.25} = 400$	$\frac{400}{300} = 100$ = 133.33	$\frac{250 + 250 + 400}{3}$ = 300
Total		400	230.88	437.5	267.78	633.33	372.63	497.77

NOTES

Index Numbers

## NOTES

**Index numbers by Simple Aggregative Method**

$$\text{Index no. for 1st year} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{400}{497.77} \times 100 = 83.36$$

$$\text{Index no. for 2nd year} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{437.5}{497.77} \times 100 = 87.89$$

$$\text{Index no. for 3rd year} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{633.33}{497.77} \times 100 = 127.23$$

**Index numbers by Simple A.M. of Price Relatives Method**

$$\text{Index no. for 1st year} = \frac{\Sigma P}{n} = \frac{230.88}{3} = 76.96$$

$$\text{Index no. for 2nd year} = \frac{\Sigma P}{n} = \frac{267.78}{3} = 89.26$$

$$\text{Index no. for 3rd year} = \frac{\Sigma P}{n} = \frac{372.63}{3} = 124.21.$$

**EXERCISE 6.1**

1. From the following data, construct an index number for 1996 by using the method of taking A.M. of price relatives:

Item	A	B	C	D	E	F
Price in 1995 (in ₹)	10	12	6	5	5	9
Price in 1996 (in ₹)	10	15	8	6	6	18

2. From the following data, construct price index nos. for the year 1996 by the methods:  
(i) simple A.M. of price relatives  
(ii) simple G.M. of price relatives.

Commodity	A	B	C	D	E	F
Price in 1995 (in ₹)	4	5	10	7	3	9
Price in 1996 (in ₹)	6	8	12	14	6	12

3. From the following data, construct the price index number with average price as base:

Rate per rupee			
Year	Wheat	Rice	Oil
I	10 kg	4 kg	2 kg
II	8 kg	2.5 kg	2 kg
III	5 kg	2 kg	1 kg

**Answers**

1. 133.05                      2. (i) 160.55                      (ii) 157.7  
3. 70.26, 89.16, 140.53 by using simple A.M. of price relative method

## 6.8. LASPEYRE'S METHOD

This is a method for finding weighted index numbers. In this method, base period quantities ( $q_0$ ) are used as weights. If  $P_{01}$  is the index number for the current period, then we have

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

where '0' and '1' suffixes stand for base period and current period respectively.

$\sum p_1 q_0$  = sum of products of prices of the commodities in the current period with their corresponding quantities used in the base period.

$\sum p_0 q_0$  = sum of product of prices of the commodities in the base period with their corresponding quantities used in the base period.

## 6.9. PAASCHE'S METHOD

This is a method for finding weighted index numbers. In this methods, current period quantities ( $q_1$ ) are used as weights.

If  $P_{01}$  is the required index number for the current period, then

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

where  $p_0, p_1$  represents prices per unit of commodities in the base period and current period respectively.

## 6.10. DORBISH AND BOWLEY'S METHOD

This is a method for computing weighted index numbers.

If  $P_{01}$  is the required index number for the current period, then

$$P_{01} = \frac{\left( \frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1} \right)}{2} \times 100$$

where  $p_0, p_1$  represents prices per unit of commodities in the base period and current period respectively,  $q_0, q_1$  represents number of units in the base period and current period respectively.

$$\begin{aligned} \text{We have } P_{01} &= \frac{\left( \frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1} \right)}{2} \times 100 = \frac{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 + \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100}{2} \\ &= \frac{\text{Laspeyre's index no} + \text{Paasche's index no.}}{2} \end{aligned}$$

$\therefore$  Dorbish and Bowley's index number can also be obtained by taking A.M. of Laspeyre's and Paasche's index numbers.

## NOTES



## 6.11. FISHER'S METHOD

This is a method for computing weighted index numbers.

### NOTES

If  $P_{01}$  is the required index number for the current period, then

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

where symbols  $p_0, q_0, p_1, q_1$  have their usual meaning.

$$\begin{aligned} \text{We have } P_{01} &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\left(\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100\right) \left(\frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100\right)} \\ &= \sqrt{\left(\text{Laspeyre's Index no.}\right) \left(\text{Paasche's Index no.}\right)} \end{aligned}$$

∴ Fisher's index numbers can also be obtained by taking G.M. of Laspeyre's and Paasche's index numbers. Fisher's method is considered to be the best method of computing index numbers because this method, satisfies unit test, time reversal test and factor reversal test. That is why, this method is also known as *Fisher's Ideal Method*.

## 6.12. MARSHALL EDGEWORTH'S METHOD

This is a method of computing weighted index numbers. In this method, the sum of base period quantities and current period quantities are used as weights.

If  $P_{01}$  is the required index number for the current period, then

$$P_{01} = \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)} \times 100$$

where  $p_0, q_0, p_1, q_1$  have their usual meaning.

We can also write this index numbers as

$$P_{01} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

This form is generally used for computing index numbers.

## 6.13. KELLY'S METHOD

This is a method of computing weighted index numbers. In this method, the quantities ( $q$ ) corresponding to any period can be used as weights. We can also use the average of quantities for two or more periods as weights.

If  $P_{01}$  is the required index numbers for the current period, then

$$P_{01} = \frac{\sum p_1 q}{\sum p_0 q} \times 100$$

where  $q$  represents the quantities which are to be used as weights.  $p_0, p_1$  have their usual meanings. This index number is also known as **Fixed Weights Aggregative Method**.

## 6.14. WEIGHTED AVERAGE OF PRICE RELATIVES METHOD

This is a method of computing weighted index numbers. In weighted index numbers, we give weights to every commodity in the series so that each commodity may have due influence on the index number. Till now quantity weights were used for constructing price index numbers.

In the weighted average of price relatives method, value weights ( $W$ ) are used. The values of commodities may correspond to either base period or current period or any other period.

If  $P_{01}$  is the required index number for the current period, then

$$P_{01} = \frac{\sum WP}{\sum W}, \text{ where } P = \frac{p_1}{p_0} \times 100.$$

$p_0, p_1$  have their usual meanings.

In this method, we have infact taken the weighted arithmetic mean of the price relatives. In constructing this index number, geometric mean is also used. In this case, the formula is

$$P_{01} = \text{Antilog} \left( \frac{\sum W \log P}{\sum W} \right).$$

**Example 6.3.** Construct index numbers of price for the year 1994 from the following data by applying:

1. Laspeyre's method
2. Paasche's method
3. Bowley's method
4. Fisher's method
5. Marshall Edgeworth's method

Commodity	1993		1994	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

**Solution.** Calculation of Index Nos. (1993 = 100)

Commodity	$p_0$	$q_0$	$p_1$	$q_1$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$	$p_1 q_0$
A	2	8	4	6	16	24	12	32
B	5	10	6	5	50	30	25	60
C	4	14	5	10	56	50	40	70
D	2	19	2	13	38	26	26	38
Total					160	130	103	200

$$\text{Laspeyre's price index number} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{200}{160} \times 100 = 125.$$

$$\text{Paasche's price index number} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{130}{103} \times 100 = 126.21$$

### NOTES

## Bowley's price index number

$$= \frac{\left( \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \right)}{2} \times 100 = \frac{\left( \frac{200}{160} + \frac{130}{103} \right)}{2} \times 100 = 125.607.$$

## NOTES

## Fisher's price index number

$$= \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{200}{160} \times \frac{130}{103}} \times 100 = 125.605.$$

## Marshall Edgeworth's price index number

$$= \frac{\Sigma p_1 (q_0 + q_1)}{\Sigma p_0 (q_0 + q_1)} \times 100 = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

$$= \frac{200 + 130}{160 + 103} \times 100 = 125.47.$$

**Example 6.4.** Prepare the index number for 1982 on the basis of 1962 for the following data:

Year	Commodity A		Commodity B		Commodity C	
	Price	Expenditure	Price	Expenditure	Price	Expenditure
1962	5	50	8	48	6	24
1982	4	48	7	49	5	15

**Solution.** We calculate price index number for the year 1982 by using Fisher's method.

## Calculation of Index Number

Commodity	1962			1982			$p_0 q_1$	$p_1 q_0$
	$p_0$	$p_0 q_0$	$q_0$	$p_1$	$p_1 q_1$	$q_1$		
A	5	50	10	4	48	12	60	40
B	8	48	6	7	49	7	56	42
C	6	24	4	5	15	3	18	20
Total		122			112		134	102

## Fisher's price index number

$$= \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{102}{122} \times \frac{112}{134}} \times 100 = 83.59.$$

**Example 6.5.** Calculate the weighted price index number for 2000 for the following data:

Material required	Unit	Quantity required	Price during	
			1999 (₹)	2000 (₹)
A	100 kg	500 kg	5	8
B	mt	2000 mt	9.5	14.2
C	kg	50 kg	34	42.2
D	litre	20 litres	12	24

**Solution.** Here we shall use **Kelly's method** because quantities are fixed irrespective of base and current years.

Index Numbers

### Calculation of Index Number (1999 = 100)

Material	$p_0$	$p_1$	$q$	$p_0q$	$p_1q$
A	5	8	$\frac{500}{100} = 5$	25	40
B	9.5	14.2	2000	19000	28400
C	34	42.2	50	1700	2110
D	12	24	20	240	480
Total				20965	31030

### NOTES

$$\text{Kelly's price index number} = \frac{\sum p_1q}{\sum p_0q} \times 100 = \frac{31030}{20965} \times 100 = 148.$$

**Example 6.6.** Construct an index number for the following data using weighted average (A.M. and G.M.) of price relatives method:

Commodity	Current year prices (in ₹)	Base year prices (in ₹)	Weights
A	4	5	1
B	6	5	2
C	10	8	3
D	12	10	1

**Solution.**

### Calculation of Index Numbers

Commodity	$p_0$	$p_1$	$W$	$P = \frac{p_1}{p_0} \times 100$	$\log P$	$WP$	$W \log P$
A	5	4	1	80	1.9031	80	1.9031
B	5	6	2	120	2.0792	240	4.1584
C	8	10	3	125	2.0969	375	6.2907
D	10	12	1	120	2.0792	120	2.0792
Total			7			815	14.4314

Price index no. by weighted A.M.

$$= \frac{\sum WP}{\sum W} = \frac{815}{7} = 116.43.$$

Price index no. by weighted G.M.

$$= AL \left( \frac{\sum W \log P}{\sum W} \right) = AL \left( \frac{14.4314}{7} \right) \\ = AL (2.0616) = 115.3.$$

**Example 6.7.** Prepare Index Number from the following information for the year 1980 taking the prices of 1975 as base:

**NOTES**

	Commodity			
	Wheat	Rice	Gram	Pulse
Price 1975	10	5	2	2
Price 1980	12	7	3	4

Give weights to above commodities as 4, 3, 2, 1 respectively.

**Solution.**

**Calculation of Index Number**

Commodity	$p_0$	$p_1$	$W$	$P = \frac{p_1}{p_0} \times 100$	$WP$
Wheat	10	12	4	120	480
Rice	5	7	3	140	420
Gram	2	3	2	150	300
Pulse	2	4	1	200	200
Total			10		1400

$$\therefore \text{Price index no. by weighted A.M.} = \frac{\Sigma WP}{\Sigma W} = \frac{1400}{10} = 140.$$

**EXERCISE 6.2**

1. Apply Fisher's method and calculate the price index number for 1995 from the following data:

Commodity	1994		1995	
	$p_0$	$q_0$	$p_1$	$q_1$
A	10	4	12	3
B	15	6	20	5
C	2	5	5	6
D	4	4	4	4

2. Compute Fisher's ideal price index number for 1994 for the following data:

Commodity	1993		1994	
	Price per unit	Expenditure	Price per unit	Expenditure
A	5	125	6	180
B	10	50	15	90
C	2	30	3	60
D	3	36	5	75



3. Use the data given below and calculate Fisher's ideal price index number for the year 1993 with 1990 as base:

Index Numbers

Commodity	Unit	Price (in ₹)		Quantity	
		1990	1993	1990	1993
Wheat	Quintal	90	100	20	25
Potatoes	Kilogram	1	1.20	100	130
Tomatoes	Kilogram	1	1.30	50	40

## NOTES

4. Construct Fisher's and Marshall's price index numbers by using the following data:

Commodity	Base year price	Base year quantity	Current year price	Current year quantity
A	12	100	20	120
B	4	200	4	240
C	8	120	12	120
D	20	60	24	48
E	16	80	24	52

5. From the data given below, calculate the price index number by using Fisher's ideal formula:

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	10	50	12	60
B	8	30	9	32
C	5	35	7	40

6. From the following data, find price index number for the year 2002:

Item	Price per unit		Value (2001)
	2001	2002	
A	₹ 13.75	₹ 13.75	₹ 8364
B	₹ 9.70	₹ 9.70	₹ 2207
C	₹ 6.03	₹ 8.00	₹ 876
D	₹ 466.00	₹ 433.00	₹ 701
E	₹ 1.25	₹ 1.75	₹ 534

### Answers

1. 135.4      2. 137.11      3. 111.98  
4. 139.729, 139.728      5. 121.91      6. 103.53

## 6.15. CHAIN BASE METHOD

In this method of computing index numbers, link relatives are required. The prices of commodities in the current period are expressed as the percentages of their prices in the preceding period. These are called **link relatives**.

$$\text{Link Relative (L.R.)} = \frac{\text{Price in current period}}{\text{Price in preceding period}} \times 100$$

**NOTES**

If there are more than one commodity under consideration then averages of link relatives (A.L.R.) are calculated for each period. Generally A.M. is used for averaging link relatives. These averages of link relatives (A.L.R.) for different time periods are called **chain index numbers**. The chain index number of a particular period represent the index number of that period with preceding period as the base period. This would be so except for this first period.

These chain indices can further be used to get index numbers for various periods with a particular period as the base period. These index numbers are called **chain index numbers chained to a fixed base**.

For calculating these index numbers, the following formula is used:

**C.B.I. for current period (Base fixed)**

$$= \frac{\text{A.L.R. for current period} \times \text{C.B.I. for preceding period (Base fixed)}}{100}$$

There are certain advantages of using this method. By using chain base method, comparison is possible between any two successive periods. The average of link relatives represent the index number with preceding period as the base period. This characteristic of chain base index numbers benefit businessmen to a good extent. In calculating chain base index number, some items can be introduced or withdrawn during any period. In practice, the chain base index numbers are used only in those circumstances, where the list of items changes very frequently.

**Example 6.8.** Calculate the fixed base index numbers and chain base index numbers from the following data. Are the two results same? If not, why?

Commodity	Price (in rupees)				
	1986	1987	1988	1989	1990
X	2	3	5	7	8
Y	8	10	12	4	18
Z	4	5	7	9	12

**Solution.**

**Calculation of F.B.I. (1996 = 100)**

Commodity	Price Relatives				
	1986	1987	1988	1989	1990
X	100	$\frac{3}{2} \times 100 = 150$	$\frac{5}{2} \times 100 = 250$	$\frac{7}{2} \times 100 = 350$	$\frac{8}{2} \times 100 = 400$
Y	100	$\frac{10}{8} \times 100 = 125$	$\frac{12}{8} \times 100 = 150$	$\frac{4}{8} \times 100 = 50$	$\frac{18}{8} \times 100 = 225$
Z	100	$\frac{5}{4} \times 100 = 125$	$\frac{7}{4} \times 100 = 175$	$\frac{9}{4} \times 100 = 225$	$\frac{12}{4} \times 100 = 300$
Total	300	400	575	625	925
Average of P.R. or F.B.I. (1986 = 100)	100	$\frac{400}{3} = 133.33$	$\frac{575}{3} = 191.67$	$\frac{625}{3} = 208.33$	$\frac{925}{3} = 308.33$

∴ F.B.I. for years 1987, 1988, 1989, 1990 with base 1986 are **133.33, 191.67, 208.33, 308.33** respectively.

Index Numbers

### Calculation of C.B.I. (1986 = 100)

Commodity	Link Relatives				
	1986	1987	1988	1989	1990
X	100	$\frac{3}{2} \times 100 = 150$	$\frac{5}{3} \times 100 = 166.67$	$\frac{7}{5} \times 100 = 140$	$\frac{8}{7} \times 100 = 114.29$
Y	100	$\frac{10}{8} \times 100 = 125$	$\frac{12}{10} \times 100 = 120$	$\frac{4}{12} \times 100 = 33.33$	$\frac{18}{4} \times 100 = 450$
Z	100	$\frac{5}{4} \times 100 = 125$	$\frac{7}{5} \times 100 = 140$	$\frac{9}{7} \times 100 = 128.57$	$\frac{12}{9} \times 100 = 133.33$
Total	300	400	426.67	301.9	697.62
Average of L.R.	100	$\frac{400}{3} = 133.33$	$\frac{426.67}{3} = 142.22$	$\frac{301.9}{3} = 100.643$	$\frac{697.62}{3} = 232.54$
or C.B.I.					
C.B.I. (1986 = 100)	100	$\frac{133.33 \times 100}{100} = 133.33$	$\frac{142.22 \times 133.33}{100} = 189.62$	$\frac{100.63 \times 189.62}{100} = 190.81$	$\frac{232.54 \times 190.81}{100} = 443.71$

### NOTES

∴ C.B.I. for years 1987, 1988, 1989, 1990 with base 1986 are **133.33, 189.62, 190.81, 443.71** respectively.

**Example 6.9.** The following table gives the average wholesale prices of three groups of commodities for the years 1991 to 1995. Compute chain base index numbers chained to 1991.

Group	Year				
	1991	1992	1993	1994	1995
I	4	6	8	10	12
II	16	20	24	30	36
III	8	10	16	20	24

**Solution.**

### Calculation of C.B.I. (1991 = 100)

Group	Link Relatives				
	1991	1992	1993	1994	1995
I	100	$\frac{6}{4} \times 100 = 150$	$\frac{8}{6} \times 100 = 133.33$	$\frac{10}{8} \times 100 = 125$	$\frac{12}{10} \times 100 = 120$
II	100	$\frac{20}{16} \times 100 = 125$	$\frac{24}{20} \times 100 = 120$	$\frac{30}{24} \times 100 = 125$	$\frac{36}{30} \times 100 = 120$
III	100	$\frac{10}{8} \times 100 = 125$	$\frac{16}{10} \times 100 = 160$	$\frac{20}{16} \times 100 = 125$	$\frac{24}{20} \times 100 = 120$
Total	300	400	413.33	375	360

## NOTES

Average of L.R. of C.B.I.	100	$\frac{400}{3} = 133.33$	$\frac{413.33}{3} = 137.78$	$\frac{375}{3} = 125$	$\frac{360}{3} = 120$
C.B.I. (1991 = 100)	100	$\frac{133.33 \times 100}{100} = 133.33$	$\frac{137.78 \times 133.33}{100} = 183.70$	$\frac{125 \times 183.70}{100} = 229.62$	$\frac{120 \times 229.62}{100} = 275.54$

∴ C.B.I. for years 1992, 1993, 1994, 1995 with base 1991 are 133.33, 183.70, 229.62, 275.54 respectively.

## EXERCISE 6.3

1. From the following average prices of the groups of commodities given in rupees per unit, find chain base index numbers with 1988 as the base year:

Group	1988	1989	1990	1991	1992
Ist	2	3	4	5	6
IIInd	8	10	12	15	18
IIIrd	4	5	8	10	12

2. Calculate the chain base index numbers chained to 1972 from the average prices of following commodities:

Commodity	1992	1993	1994	1995	1996
Wheat	4	6	8	10	12
Rice	16	20	24	30	36
Sugar	8	10	16	20	24

3. Compute chain base index number for 1996 with 1993 as base, by using the following data:

Commodity	Year			
	1993	1994	1995	1996
Sugar (Price per kg)	6.4	6.5	6	6.5
Gur (Price per kg)	4	3.7	4	4.5

## Answers

- 100, 133.33, 183.70, 229.62, 275.54
- 100, 133.33, 183.7, 229.63, 275.56
- 107.36.

## II. QUANTITY INDEX NUMBERS

## 6.16. METHODS

Quantity index numbers are used to show the average change in the quantities of related goods with respect to time. These index numbers are also used to measure the

level of production. In computing quantity index numbers, either prices or values are used as weights.

Let  $Q_{01}$  denotes the quantity index number for the current period. The formulae for calculating quantity index numbers are obtained by interchanging the role of 'p' and 'q' in the formulae for computing price index numbers. Various methods for computing quantity index numbers are as follows:

## NOTES

### 1. Simple Aggregative Method

$$Q_{01} = \frac{\sum q_1}{\sum q_0} \times 100.$$

### 2. Simple Average of Quantity Relative Method

$$Q_{01} = \frac{\sum Q}{n} \quad (\text{Using A.M.})$$

$$= \text{Antilog} \left( \frac{\sum \log Q}{n} \right) \quad (\text{Using G.M.})$$

where  $Q = \text{quantity relative} = \frac{q_1}{q_0} \times 100.$

### 3. Laspeyre's Method

$$Q_{01} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100.$$

### 4. Paasche's Method

$$Q_{01} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100.$$

### 5. Dorbish and Bowley's Method

$$Q_{01} = \frac{\left( \frac{\sum q_1 p_0}{\sum q_0 p_0} + \frac{\sum q_1 p_1}{\sum q_0 p_1} \right)}{2} \times 100.$$

### 6. Fisher's Ideal Method

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100.$$

### 7. Marshall Edgeworth's Method

$$Q_{01} = \frac{\sum q_1 (p_0 + p_1)}{\sum q_0 (p_0 + p_1)} \times 100.$$

### 8. Kelly's Method

$$Q_{01} = \frac{\sum q_1 p}{\sum q_0 p} \times 100.$$

### 9. Weighted Average of Quantity Relative Method

$$Q_{01} = \frac{\sum WQ}{\sum W} \quad (\text{Using A.M.})$$

$$= \text{Antilog} \left( \frac{\sum W \log Q}{\sum W} \right) \quad (\text{Using G.M.})$$

### 10. Chain Base Method

Here also, we define chain base quantity index numbers for a period as the average of link relatives (L.R.) for that particular period. These chain indices can be used to obtain quantity index numbers with a common base.



In all the above formulae, suffixes '0' and '1' stand for base period and current period respectively and

## NOTES

$p_1$  = current period price of an item

$p_0$  = base period price of an item

$q_1$  = current period quantity of an item

$q_0$  = base period quantity of an item

$Q$  = quantity relative of an item =  $\frac{q_1}{q_0} \times 100$

$W$  = value weight for an item

$p$  = price of an item in a fixed period

$n$  = no. of item under consideration.

## 6.17. INDEX NUMBERS OF INDUSTRIAL PRODUCTION

The indices of industrial production are calculated by using the methods of quantity index numbers. In the formulae for quantity index numbers, we shall take *production* in place of quantities.

**Example 6.10.** Calculate the quantity index number for 1986 by using Fisher's formula for the following data:

Commodity	1995		1996	
	Price	Quantity	Price	Quantity
A	6	70	8	120
B	8	90	10	100
C	12	140	16	280

**Solution.** Calculation of Fisher's Quantity Index No. (1995 = 100)

Commodity	$p_0$	$q_0$	$p_1$	$q_1$	$q_0 p_0$	$q_1 p_1$	$q_0 p_1$	$q_1 p_0$
A	6	70	8	120	420	960	560	720
B	8	90	10	100	720	1000	900	800
C	12	140	16	280	1680	4480	2240	3360
Total					2820	6440	3700	4880

$$\text{Fisher's quantity index number} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$$

$$= \sqrt{\frac{4880}{2820} \times \frac{6440}{3700}} \times 100 = 173.55.$$

**Example 6.11.** From the following data, construct quantity index numbers for 1986, by using the following methods:

- |                               |                                  |
|-------------------------------|----------------------------------|
| (i) Simple aggregative method | (ii) Laspeyre's method           |
| (iii) Paasche's method        | (iv) Dorbish and Bowley's method |
| (v) Fisher's method           | (vi) Marshall Edgeworth's method |

## NOTES

Commodity	1995		1996	
	Price	Value	Price	Value
A	8	80	10	110
B	10	90	12	108
C	16	256	20	340

**Solution.** Calculation of Quantity Index Nos. (1995 = 100)

Commodity	$p_0$	Value $q_0 p_0$	$q_0$	$p_1$	Value $q_1 p_1$	$q_1$	$q_1 p_0$	$q_0 p_1$
A	8	80	10	10	110	11	88	100
B	10	90	9	12	108	9	90	108
C	16	256	16	20	340	17	272	320
Total		426	35		558	37	450	528

(i)  $Q_{01}$  by simple aggregative method

$$= \frac{\Sigma q_1}{\Sigma q_0} \times 100 = \frac{37}{35} \times 100 = 105.71$$

(ii) Laspeyre's quantity index no.

$$= \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100 = \frac{450}{426} \times 100 = 105.63$$

(iii) Paasche's quantity index no.

$$= \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100 = \frac{558}{528} \times 100 = 105.68$$

(iv) Dorbish and Bowley's quantity index no.

$$= \frac{\left( \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} + \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \right)}{2} \times 100 = \frac{\left( \frac{450}{426} + \frac{558}{528} \right)}{2} \times 100 = 105.66$$

(v) Fisher's quantity index no.

$$= \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100 = \sqrt{\frac{450}{426} \times \frac{558}{528}} \times 100 = 105.66$$

(vi) Marshall Edgeworth's quantity index no.

$$= \frac{\Sigma q_1 (p_0 + p_1)}{\Sigma q_0 (p_0 + p_1)} \times 100 = \frac{\Sigma q_1 p_0 + \Sigma q_1 p_1}{\Sigma q_0 p_0 + \Sigma q_0 p_1} \times 100 = \frac{450 + 558}{426 + 528} \times 100 = 105.66$$

### III. VALUE INDEX NUMBERS

## NOTES

#### 6.18. SIMPLE AGGREGATIVE METHOD

The simple aggregative method of computing value index number ( $V_{01}$ ) is given by

$$V_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$$

where  $\sum p_1 q_1$  = sum of values of items in the current period

$\sum p_0 q_0$  = sum of values of items in the base period.

**Example 6.12.** Calculate value index number for 2000 for the following data:

Item	1998		2000	
	Price	Quantity	Price	Quantity
A	4	12	5	18
B	8	15	12	10
C	12	6	10	8
D	5	10	5	12

**Solution.** Calculation of value index number (1998 = 100)

Item	$p_0$	$q_0$	$p_1$	$q_1$	$p_0 q_0$	$p_1 q_1$
A	4	12	5	18	48	120
B	8	15	12	10	120	120
C	12	6	10	8	72	80
D	5	10	5	12	50	60
Total					290	380

$$\text{Value index number} = \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100 = \frac{380}{290} \times 100 = 131.03.$$

#### EXERCISE 6.4

1. Compute a suitable quantity index number by using the following data:

Commodity	Price in the base period	Quantity	
		Base period	Current period
A	4	7	10
B	5	8	9
C	4	10	9
D	3	12	8

2. Construct index numbers of quantity for the given data, by using the following methods:

- (i) Simple aggregative method  
(ii) Fisher's method  
(iii) Weighted average (A.M.) of quantity relatives by using base period value as weights.

## NOTES

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

3. Using Paasche's formula, compute the quantity index number and the price index number for 2000 with 1999 as base year:

Commodity	Quantity Units		Value in (₹)	
	1999	2000	1999	2000
A	100	150	500	900
B	80	100	320	500
C	60	72	150	360
D	30	33	360	297

For the above problem, also compute price index number by:

- (i) Dorbish-Bowley Method (ii) Fisher's method  
(iii) Marshall-Edworth method.

### Answers

1. 100.694 2. 66.667, 64.687, 64.375  
3. 131.02, 119.18 (i) 118.61, (ii) 118.61, (iii) 118.62

## 6.19. MEAN OF INDEX NUMBERS

If  $I_1, I_2, \dots, I_n$  are the index numbers of  $n$  groups of related items, then the index numbers of all the items of  $n$  group taken together is calculated by taking the average of these index numbers. Generally, A.M. is used for averaging the index numbers. If weights are attached with different index numbers, then weighted A.M. is to be calculated.

Let  $I$  be the index number of all the items of  $n$  groups taken together, then

$$I = \frac{I_1 + I_2 + \dots + I_n}{n} \quad \text{i.e.,} \quad I = \frac{\sum I}{n}$$

If  $W_1, W_2, \dots, W_n$  be the weights of index numbers  $I_1, I_2, \dots, I_n$  respectively, then

$$I = \frac{W_1 I_1 + W_2 I_2 + \dots + W_n I_n}{W_1 + W_2 + \dots + W_n} \quad \text{or} \quad I = \frac{\sum W I}{\sum W}$$

If G.M. is to be used for finding index number of combined group, then

$$I = AL \left( \frac{W_1 \log I_1 + W_2 \log I_2 + \dots + W_n \log I_n}{W_1 + W_2 + \dots + W_n} \right) \quad \text{or} \quad I = AL \left( \frac{\sum W \log I}{\sum W} \right)$$

**Example 6.13.** Construct the index number of business activity in India for the following data:

**NOTES**

Item	Weightage	Index
(i) Industrial Production	36	250
(ii) Mineral Production	7	135
(iii) Internal Trade	24	200
(iv) Financial Activity	20	135
(v) Exports and Imports	7	325
(vi) Shipping Activity	6	300

**Solution. Calculation of Index No. of Business Activity**

Item	Weightage W	Index I	WI
(i) Industrial Production	36	250	9000
(ii) Mineral Production	7	135	945
(iii) Internal Trade	24	200	4800
(iv) Financial Activity	20	135	2700
(v) Exports and Imports	7	325	2275
(vi) Shipping Activity	6	300	1800
Total	100		21520

$$\text{Index No. of combined group} = \frac{\sum WI}{\sum W} = \frac{21520}{100} = 215.2.$$

**Example 6.14.** A textile worker in the city of Bombay earns ₹ 350 a month. The cost of living index for a particular month is given as 136. Using the following data, find out the amount he spends on clothings and house rent.

Group	Food	Clothing	House rent	Fuel	Misc.
Expenditure	140	?	?	56	63
Group Index	180	150	100	110	80

**Solution.** Let 'a' and 'b' denote the expenditure on clothing and house rent respectively.

Group	Expenditure W	Group Index I	WI
Food	140	180	25200
Clothing	a	150	150a
House rent	b	100	100b
Fuel	56	110	6160
Misc.	63	80	5040
Total	259 + a + b = 350		36400 + 150a + 100b

$$\text{Now} \quad 259 + a + b = 350$$

$$\therefore a + b = 350 - 259 = 91$$

$$\therefore b = 91 - a.$$



$$\text{Now, cost of living index} = \frac{\Sigma WI}{\Sigma W}$$

$$\begin{aligned}\therefore 136 &= \frac{36400 + 150a + 100b}{350} \\ \therefore 47600 &= 36400 + 150a + 100(91 - a) \\ \therefore 11200 &= 150a + 9100 - 100a \\ \therefore 2100 &= 50a \\ \therefore a &= 42 \\ \therefore b &= 91 - a = 91 - 42 = 49.\end{aligned}$$

**NOTES****EXERCISE 6.5**

1. Construct index number of combined group for the following data:

Group	A	B	C	D	E
Index No.	110	95	160	170	200
Weight	4	2	1	1	2

2. Find the index number of combined group for the following data:

Group	A	B	C	D	E	F
Index No.	125	142	118.7	92	169	157
% of Weightage	25	15	10	12	13	25

3. From the following data relating to working class consumers of a city, calculate index numbers for 1993 and 1995.

Group	Weight	Group Index	
		1993	1995
Food	48	110	130
Clothing	8	120	125
Fuel	7	110	120
House rent	13	100	100
Miscellaneous	14	115	135

**Answers**

1. 136      2. 136.68      3. 110.222, 125.222

**IV. TESTS OF ADEQUACY OF INDEX NUMBER FORMULAE****6.20. MEANING**

We have studied a large number of methods of constructing index numbers. Statisticians have developed certain mathematical criterion for deciding the superiority of one method

## NOTES

over others. The following are the tests for judging the adequacy of a particular index number method :

- (i) Unit Test.
- (ii) Time Reversal Test.
- (iii) Factor Reversal Test.
- (iv) Circular Test.

## 6.21. UNIT TEST (U.T.)

An index number method is said to satisfy **unit test** if it is not changed by a change in the measuring units of some items, under consideration. All methods, except simple aggregative method, satisfies this test.

## 6.22. TIME REVERSAL TEST (T.R.T.)

An index numbers method is said to satisfy **time reversal test**, if

$$I_{01} \times I_{10} = 1$$

where  $I_{01}$  and  $I_{10}$  are the index numbers for two periods with base period and current period reversed. Here the index numbers  $I_{01}$  and  $I_{10}$  are not expressed as percentages.

The following methods of constructing index numbers satisfies this test:

- (i) Simple Aggregative Method.
- (ii) Simple G.M. of Price (or Quantity) Relatives Method.
- (iii) Fisher's Method.
- (iv) Marshall Edgeworth's Method.
- (v) Kelly's Method.

Now, we shall illustrate this test by verifying its validity for Fisher's price index number method.

$$\text{We have } P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \quad \text{and} \quad P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

where  $P_{01}$  and  $P_{10}$  are the price index numbers for the periods  $t_1$  and  $t_0$  with base periods  $t_0$  and  $t_1$  respectively.

$$\begin{aligned} \text{Now } P_{01} \times P_{10} &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} \\ &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} = \sqrt{1} = 1. \end{aligned}$$

$$\therefore P_{01} \times P_{10} = 1.$$

**Example 6.15.** Calculate price index number for the year 1996 from the following data. Use geometric mean of price relatives. Also reverse the base (1996 as base) and show whether the two results are consistent or not.

Commodity	Average price 1990 (₹)	Average Price 1996 (₹)
A	16.1	14.2
B	9.2	8.7
C	15.1	12.5
D	5.6	4.8
E	11.7	13.4
F	100	117

## NOTES

Solution.

Index No. for 1996

Commodity	$p_0$	$p_1$	$P = \frac{p_1}{p_0} \times 100$	$\log P$
A	16.1	14.2	$\frac{14.2}{16.1} \times 100 = 88.20$	1.9455
B	9.2	8.7	$\frac{8.7}{9.2} \times 100 = 94.57$	1.9757
C	15.1	12.5	$\frac{12.5}{15.1} \times 100 = 82.78$	1.9179
D	5.6	4.8	$\frac{4.8}{5.6} \times 100 = 85.71$	1.9331
E	11.7	13.4	$\frac{13.4}{11.7} \times 100 = 114.53$	2.0589
F	100	117	$\frac{117}{100} \times 100 = 117$	1.0682
$n = 6$				$\Sigma \log P = 11.8993$

$$\therefore \text{Price index no. for 1996} = \text{AL} \left( \frac{\Sigma \log P}{n} \right) = \text{AL} \left( \frac{11.8993}{6} \right) = \text{AL } 1.9832 = 96.20.$$

Index No. for 1990

Commodity	$p_0$	$p_1$	$P = \frac{p_1}{p_0} \times 100$	$\log P$
A	14.2	16.1	$\frac{16.1}{14.2} \times 100 = 113.38$	2.0547
B	8.7	9.2	$\frac{9.2}{8.7} \times 100 = 105.75$	2.0244
C	12.5	15.1	$\frac{15.1}{12.5} \times 100 = 120.80$	2.0820
D	4.8	5.6	$\frac{5.6}{4.8} \times 100 = 116.67$	2.0671
E	13.4	11.7	$\frac{11.7}{13.4} \times 100 = 87.31$	1.9410
F	117	100	$\frac{100}{117} \times 100 = 85.47$	1.9319
$n = 6$				$\Sigma \log P = 12.1011$

$$\therefore \text{Price index no. for 1990} = \text{AL} \left( \frac{\Sigma \log P}{n} \right) = \text{AL} \left( \frac{12.1011}{6} \right) = \text{AL } 2.0169 = 104.$$

Product of index numbers =  $96.20 \times 104 = 10004.8 = 10000$  (nearly)

Since the index numbers are expressed as percentages, the T.R.T. is satisfied if their products is  $(100)^2$ , which is 10000.

$\therefore$  The index numbers are consistent.

## NOTES

### 6.23. FACTOR REVERSAL TEST (F.R.T.)

An index number method is said to satisfy **factor reversal test** if the product of price index number and quantity index number, as calculated by the same method, is equal to the value index number.

In other words, if  $P_{01}$  and  $Q_{01}$  are the price index number and quantity index number for the period  $t_1$  corresponding to base period  $t_0$ , then we must have

$$P_{01} \times Q_{01} = V_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Fisher's index number method is *the only method* which satisfies this test.

Let  $P_{01}$  and  $Q_{01}$  be the Fisher's price index number and quantity index numbers respectively, then

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \quad \text{and} \quad Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

$$\begin{aligned} \text{Now} \quad P_{01} \times Q_{01} &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \\ &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \\ &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_1}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_1 q_0}} \\ &= \sqrt{\frac{\sum p_1 q_1 \times \sum p_1 q_1}{\sum p_0 q_0 \times \sum p_0 q_0}} = \frac{\sum p_1 q_1}{\sum p_0 q_0} \\ &= \text{Value index number.} \end{aligned}$$

$\therefore$  Fisher's method satisfies this test.

### 6.24. CIRCULAR TEST (C.T.)

An index number method is said to satisfy the **circular test** if  $I_{01}, I_{12}, I_{23}, \dots, I_{n-1n}$  and  $I_{n0}$  are the index numbers for the periods  $t_1, t_2, t_3, \dots, t_n, t_0$  corresponding to base periods  $t_0, t_1, t_2, \dots, t_{n-1}, t_n$  respectively, then

$$I_{01} \times I_{12} \times I_{23} \times \dots \times I_{n-1n} \times I_{n0} = 1.$$

Here, also, the index numbers have not been expressed as percentages by multiplying by 100.

If  $n = 1$ , we have  $I_{01} \times I_{10} = 1$ .

This is nothing but the condition of T.R.T. Thus, we see that the circular test is an extension of T.R.T.

If  $n = 2$ , we have

$$I_{01} \times I_{12} \times I_{20} = 1 \quad \text{or} \quad I_{01} \times I_{12} = I_{02} \quad (\because I_{02} \times I_{20} = 1)$$

The following methods satisfies circular test:

- (i) Simple Aggregative Method.
- (ii) Simple G.M. of Price (or Quantity) Relatives Method.
- (iii) Kelly's Method.

Now, we shall illustrate this test by verifying its validity to simple aggregative method for price index numbers.

Here 
$$P_{01} = \frac{\sum p_1}{\sum p_0}, P_{12} = \frac{\sum p_2}{\sum p_1}, P_{20} = \frac{\sum p_0}{\sum p_2}$$

$$\therefore P_{01} \times P_{12} \times P_{20} = \frac{\sum p_1}{\sum p_0} \times \frac{\sum p_2}{\sum p_1} \times \frac{\sum p_0}{\sum p_2} = 1.$$

$\therefore$  Simple aggregative method satisfies this test.

**Example 6.16.** Construct Fisher's Ideal Index number from the following data and show that it satisfies the factor reversal test:

Year	Article A		Article B		Article C	
	Price	Quantity	Price	Quantity	Price	Quantity
1975	16	4	4	4	2	2
1982	30	3.5	14	1.5	6	2.5

**Solution.** Let suffixes '0' and '1' refers to data for the periods 1975 and 1982 respectively.

#### Calculation of Fisher's Index Numbers

Article	$p_0$	$q_0$	$p_1$	$q_1$	$p_0q_0$	$p_1q_1$	$p_1q_0$	$p_0q_1$
A	16	4	30	3.5	64	105	120	56
B	4	4	14	1.5	16	21	56	6
C	2	2	6	2.5	4	15	12	5
Total					84	141	188	67

Now, Fisher's Ideal index number

$$= P_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100 = \sqrt{\frac{188}{84} \times \frac{141}{67}} \times 100 = 217.03$$

Verification of F.R.T.

$P_{01}$  = Fisher's price index no. for 1982 with base 1975 (= 1)

$$= \frac{217.03}{100} = 2.1703$$

$Q_{01}$  = Fisher's quantity index number for 1982 with base 1975 (= 1)

$$= \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}} = \sqrt{\frac{67}{84} \times \frac{141}{188}} = 0.7734 \quad (\text{Not as \%})$$

$V_{01}$  = Value index number of 1982 with base 1975 (= 1)

$$= \frac{\sum V_1}{\sum V_0} = \frac{\sum p_1q_1}{\sum p_0q_0} = \frac{141}{84} = 1.6786 \quad (\text{Not as \%})$$

Now,  $P_{01} \times Q_{01} = 2.1703 \times 0.7734 = 1.6785 = V_{01}$  (nearly)

$\therefore$  F.R.T. is verified.

## NOTES



**EXERCISE 6.6****NOTES**

1. Calculate Fisher's index number using the following data and check whether it satisfies the time reversal test or not.

Commodity	1991		1992	
	Quantity	Price	Quantity	Price
X	50	32	50	30
Y	35	30	40	25
Z	35	16	50	18

2. Show with the help of the following data that the time reversal test and factor reversal test are satisfied by Fisher's Ideal formula for index number construction:

Commodity	Base year		Current year	
	Price (₹)	Quantity (kg.)	Price (₹)	Quantity (kg.)
A	8	500	10	600
B	2	1000	4	800
C	6	600	8	500
D	10	300	12	400
E	4	800	2	1000

3. By using the given data show that Fisher's method of computing index numbers satisfies T.R.T. and F.R.T.

Item	1993		1995	
	Price	Value	Price	Value
A	4	12	7	21
B	60	120	65	195
C	11	44	9	36
D	27	108	30	90
E	12	72	20	100
F	25	100	20	100

**V. CONSUMER PRICE INDEX NUMBERS (C.P.I.)****6.25. MEANING**

There is no denying the fact that the rise or fall in the prices of commodities affect every family. But, this effect is not same for every family because different families consume different commodities and in different quantities. Car is not found in every house. Milk is used in almost every family but there are very few families who can afford to purchase even more than 5 litres of it, daily.

The index numbers which measures the effect of rise or fall in the prices of various goods and services, consumed by a particular group of people are called **consumer price index numbers** for that particular group of people. The consumer price index numbers help in estimating the average change in the cost of maintaining particular standard of living by a particular class of people.

## NOTES

## 6.26. SIGNIFICANCE OF C.P.I.

(i) The consumer price index numbers are used in deflating money income to real income. Money income is divided by a proper consumer price index number to obtain real income.

(ii) The consumer price index numbers are used in wage fixation and automatic increase in wages. Generally, escalator clauses are provided for automatic increase in wages in accordance with increase in consumer price index number.

(iii) The consumer price index numbers are used by the planning commission for framing rent policy, taxation policy, price policy, etc.

## 6.27. ASSUMPTIONS

The consumer price index numbers are computed under certain assumptions. These assumptions are as follows:

(i) It is assumed that the quantities of different goods and services consumed are same for base period and current period.

(ii) It is assumed that the prices of commodities are approximately same in the region covered by the consumer price index number.

(iii) It is assumed that the commodities used in preparing C.P.I. are used in equal quantities in every family in the region covered by the index number.

(iv) It is assumed that the families in the region covered by the C.P.I. are of same economic standard. Their demands are common.

These are very strong assumptions and cannot be fully met in practical life. That is why, the C.P.I. for a region will not be exactly true for every family covered by the index number.

## 6.28. PROCEDURE

The first step in computing consumer price index number is to decide the category of people for whom the index is to be computed. While fixing the domain of the index, the income and occupation of families must be taken in to consideration. Different families consume different commodities and that too in different quantities. For a particular category of people, it can be expected that their expenditure on different commodities will be almost same.

For computing index, enquiry is made about the expenditure of families on various commodities. The commodities are generally classified in the following heads:

- |                       |                |
|-----------------------|----------------|
| (a) Food              | (b) Clothing   |
| (c) Fuel and lighting | (d) House rent |
| (e) Miscellaneous.    |                |

## NOTES

After the decision about commodities is taken, the next step is to collect prices of these commodities. The price quotations must be obtained from that market, from where the concerned class of people purchase commodities. The price quotations must be absolutely free from the personal bias of the agent obtaining price quotations. The price quotations must preferably be cross checked in order to eliminate any possibility of personal bias.

All the commodities which are used by a particular class of people cannot be expected to have equal importance. For example, entertainment and house rent cannot be given equal weightage. Weights are taken in accordance with the consumption in the base period. Either base period quantities or base period expenditure on different items are generally used as weights for constructing C.P.I. The base period selected for this purpose must also be normal.

## 6.29. METHODS

There are two methods of computing consumer price index numbers.

- (i) Aggregate expenditure method.
- (ii) Family budget method.

## 6.30. AGGREGATE EXPENDITURE METHOD

In this method, generally base period quantities are used as weights.

$$\text{Consumer Price Index No.} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

where '0' and '1' suffixes stand for base period and current period respectively.

$\sum p_1 q_0$  = sum of the products of the prices of commodities in the current period with their corresponding quantities used in the base period.

$\sum p_0 q_0$  = sum of the products of the prices of commodities in the base period with their corresponding quantities used in the base period.

Sometimes, current period quantities are also used for finding consumer price index numbers.

**Example 6.17.** Calculate the cost of living index from the following data by using aggregate expenditure method.

Item	Quantity consumed in the given year	Price in base year	Price in given year
Rice	$2\frac{1}{2} \text{ Qtl.} \times 12$	12	25
Pulses	$3 \text{ kg} \times 12$	0.4	0.6
Oil	$2 \text{ kg} \times 12$	1.5	2.2
Clothing	$6 \text{ mt.} \times 12$	0.75	1
Housing		20 P.M.	30 P.M.
Miscellaneous		10 P.M.	15 P.M.

**Solution. Calculation of Cost of Living Index Number**

Index Numbers

Item	$q_1$	$p_0$	$p_1$	$p_1 q_1$	$p_0 q_1$
Rice	30	12	25	750	360
Pulses	36	0.4	0.6	21.6	14.4
Oil	24	1.5	2.2	52.8	36
Clothing	72	0.75	1.0	72	54
Housing	12	20	30	360	240
Miscellaneous	12	10	15	180	120
Total				1436.4	824.4

$$\text{Cost of living index no.} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{1436.4}{824.4} \times 100 = 174.24.$$

**NOTES****6.31. FAMILY BUDGET METHOD**

In this method, the expenditure on different commodities in the base period, are used as weights.

$$\text{Consumer Price Index No.} = \frac{\sum PW}{\sum W}$$

$$\text{where } P = \text{Price relative} = \frac{p_1}{p_0} \times 100.$$

$p_0, p_1$  refers to prices of commodities in the base period and current period respectively.

$$W = p_0 q_0.$$

$$\text{We have C.P.I.} = \frac{\sum PW}{\sum W} = \frac{\sum \left( \frac{p_1}{p_0} \times 100 \right) p_0 q_0}{\sum p_0 q_0} = \frac{\sum (p_1 \times 100) q_0}{\sum p_0 q_0} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100.$$

Therefore, the C.P.I. calculated by using both methods would be same. Family budget method is particularly used when the expenditures on various items used in the base period are given on percentage basis.

**Example 6.18.** The cost of living index for the working class families in 1988 was 168.12. The retail price indices with base 1984 = 100 and the percentages of family expenditure in 1984 are given below. Find the retail price for the rent, fuel and light group:

Group	% of Family Expenditure in 1984	Retail Price I in 1988 (1984 = 100)
Food	40	132
Rent, Fuel and Light	18	?
Clothing	9	210
Miscellaneous	33	200

**Solution.** Let 'x' be the retail price index for rent, fuel and light group:

**NOTES**

Group	% of Family Expenditure W	Retail Price Index I	IW
Food	40	132	5280
Rent, Fuel and Light	18	x	18x
Clothing	9	210	1890
Miscellaneous	33	200	6600
Total		100	13770 + 18x

$$\text{Cost of living index for 1938} = \frac{\Sigma IW}{\Sigma W}$$

$$\therefore 168.12 = \frac{13770 + 18x}{100}$$

$$\therefore 16812 = 13770 + 18x$$

$$\therefore x = \frac{16812 - 13770}{18} = 169.$$

**Example 6.19.** The group indices and corresponding weights for the working class cost of living index numbers in an industrial city for the years 1989 and 1990 are given below:

Group	Weight	Group Index for 1989	Group Index for 1990
Food	71	370	380
Clothing	3	423	504
Fuel	9	469	336
House rent	7	110	116
Miscellaneous	10	279	283

Compute the cost of living index numbers for the years 1989 and 1990. If a worker was getting ₹ 3,000 per month in 1989, do you think that he should be given some extra allowance so that he can maintain his 1989 standard of living? If so, what should be the minimum amount of this extra allowance?

**Solution.** Calculation of Cost of Living Indices for 1989 and 1990

Group	Weight W	1989		1990	
		I	IW	I	IW
Food	71	370	26270	380	26980
Clothing	3	423	1296	504	1512
Fuel	9	469	4221	336	3024
House rent	7	110	770	116	812
Miscellaneous	10	279	2790	283	2830
Total	100		35320		35158

$$\text{Cost of living index for 1989} = \frac{\Sigma IW}{\Sigma W} = \frac{35320}{100} = 353.20$$

$$\text{Cost of living index for 1990} = \frac{\Sigma IW}{\Sigma W} = \frac{35158}{100} = 351.58.$$

The worker should not be given any extra allowance, because the cost of living index has not increased in 1990.



## EXERCISE 6.7

Index Numbers

1. In the construction of a certain cost of living index number, the following group index numbers were found. Calculate the cost of living index by using weighted A.M.

Group	Index No.	Weight
Food	350	5
Fuel and Lighting	200	1
Clothing	240	1
House rent	160	1
Miscellaneous	250	2

2. The following are the group index numbers and group weights of an average working class family budget. Construct the cost of living index number by assigning the given weights:

Group	Index No.	Weight
Food	352	48
Fuel and Lighting	220	10
Clothing	230	8
House rent	160	12
Miscellaneous	190	15

3. Construct with the help of data given below the cost of living index numbers for the years 1960 and 1961, taking 1959 as the base year:

Group	Unit	Price in 1959	Price in 1960	Price in 1961
Foodgrains	per md.	16.00	18.00	20.00
Clothing	per mt	2.00	1.80	2.20
Fuel	per md.	4.00	5.00	5.50
Electricity	per unit	0.20	0.25	0.25
House rent	per room	10.00	12.00	15.00
Miscellaneous	per unit	0.50	0.60	0.75

Give weightage to the above groups in the proportion of 6, 4, 2, 2, 4 and 2 respectively.

4. From the following figures, prepare the cost of living index number by using "Aggregate Expenditure Method".

Article	Quantity Consumed in Base year	Units	Price in Base year 1971	Price in Current year 1981
Wheat	4 Qtls.	Qtl.	100	240
Rice	1 Qtl.	Qtl.	120	300
Gram	1 Qtl.	Qtl.	80	200
Pulses	2Qtls.	Qtl.	160	400
Ghee	50 kg.	kg.	20	40
Sugar	50 kg.	kg.	2	6
Fire-wood	5 Qtls.	Qtl.	16	40
House rent	1 House	House	50	100

## NOTES

## NOTES

5. Construct cost of living index for 1996 based on 1990 from the following data:

Group	Food	Housing	Clothing	Fuel	Misc.
Index No. for 1996 (Base 1990)	122	140	112	116	106
Weight	32	10	10	6	42

## Answers

1. 285                      2. 276.41                      3. 112.75, 130.75                      4. 226.05  
5. 115.72

## 6.32. SUMMARY

- The **index numbers** are defined as specialized averages used to measure change in a variable or a group of related variables with respect to time or geographical location or some other characteristic.
- The barometers are used to study changes in whether conditions, similarly the index numbers are used to study the changes in economic and business activities. That is, why, the index numbers are also called '**Economic Barometers**'.
- Index numbers are used for computing real incomes from money incomes. The wages, clearness allowances, etc. are fixed on the basis of real income.
- Index numbers are constructed to compare the changes in related variables over time.
- Index numbers are used to study the changes occurred in the past. This knowledge helps in forecasting.
- Index numbers are used to study the changes in prices, industrial production, purchasing powers of money, agricultural production, etc., of different countries.
- The **price relative** of a commodity in the current period with respect to base period is defined as the price of the commodity in the current period expressed as a percentage of the price in the base period.
- If there are more than one commodity under consideration then averages of link relatives (A.L.R.) are calculated for each period. Generally A.M. is used for averaging link relatives. These averages of link relatives (A.L.R.) for different time periods are called **chain index numbers**. The chain index number of a particular period represent the index number of that period with preceding period as the base period.
- **Quantity index numbers** are used to show the average change in the quantities of related goods with respect to time. These index numbers are also used to measure the level of production.
- The index numbers which measures the effect of rise or fall in the prices of various goods and services, consumed by a particular group of people are called **consumer price index numbers** for that particular group of people. The consumer price index numbers help in estimating the average change in the cost of maintaining particular standard of living by a particular class of people.

---

## 6.33. REVIEW EXERCISES

---

*Index Numbers*

1. "An index number is a special type of average." Discuss.
2. Write a short note on "Factor Reversal Test".
3. What is Fisher's ideal method of computing index numbers? Why is it called ideal?
4. What main points should be taken into consideration while constructing simple index nos? Explain the procedure of construction of simple index numbers taking example of five commodities.
5. Why Fisher's Ideal formula called 'Ideal'? Explain by giving an example that it satisfies time and factor reversal tests.
6. What is Index Number? What problems are involved in the construction of index numbers? Give different formulae of index numbers and state which of these is best and why?
7. What are consumer price index number? What is their significance? Discuss the steps involved in constructing a consumer price index number.

## NOTES

## NOTES

## 7. MEASURES OF CORRELATION

### STRUCTURE

- 7.1. Introduction
- 7.2. Definition
- 7.3. Correlation and Causation
- 7.4. Positive and Negative Correlation
- 7.5. Linear and Non-linear Correlation
- 7.6. Simple, Multiple and Partial Correlation

#### I. Karl Pearson's Method

- 7.7. Definition
- 7.8. Alternative Form of 'R'
- 7.9. Step Deviation Method

#### II. Spearman's Rank Correlations Method

- 7.10. Meaning
- 7.11. Case I. Non-repeated Ranks
- 7.12. Case II. Repeated Ranks
- 7.13. Summary
- 7.14. Review Exercises

### 7.1. INTRODUCTION

In practical life, we come across certain situations, where movements in one variable are accompanied by movements in other variables. For example, the expenditure of a family is very much related to the income of the concerned family. An increase in income is expected to be accompanied by an increase in the expenditure. If the data relating to a number of families is collected, then it would be found that the variables 'income' and 'expenditure' are moving in sympathy in the same direction. An increase in the day temperature may be accompanied by an increase in the sale of cold drinks. The marks in Accountancy and Mathematics papers of students in a class move in the same direction, on an average, because a student who is brilliant in one subject is expected to be so in the other subjects also.



## 7.2. DEFINITION

If the changes in the values of one variable are accompanied by changes in the values of the other variable, then the variables are said to be **correlated**. The correlated variables move in sympathy, on an average, either in the same direction or in the opposite directions. According to *L.R. Connor*, "If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in the other(s), then they are said to be correlated". In other words, variables are said to be correlated if the variations in one variable are followed by variations in the others.

## NOTES

## 7.3. CORRELATION AND CAUSATION

Two variables may be related in the sense that the changes in the values of one variable are accompanied by changes in the values of the other variable. But this cannot be interpreted in the sense that the changes in one variable has necessarily caused changes in the other variable. Their movement in sympathy may be due to mere chance. A high degree correlation between two variables may not necessarily imply the existence of a cause-effect relationship between the variables. On the other hand, if there is a cause-effect relationship between the variables, then the correlation is sure to exist between the variables under consideration. A high degree correlation between 'income' and 'expenditure' is due to the fact that expenditure is affected by the income.

Now we shall outline the reasons which may be held responsible for the existence of correlation between variables.

The correlation between variables may be due to the effect of some common cause. For example, positive correlation between the number of girls seeking admission in colleges A and B of a city may be due to the effect of increasing interest of girls towards higher education.

The correlation between variables may be due to mere chance. Consider the data regarding six students selected at random from a college.

Students	A	B	C	D	E	F
% of marks obtained in the previous exam.	42%	47%	60%	80%	55%	40%
Height (in inches)	60	62	65	70	64	59

Here the variables are moving in the same direction and a high degree of correlation is expected between the variables. We cannot expect this degree of correlation to hold good for any other sample drawn from the concerned population. In this case, the correlation has occurred just due to chance.

The correlation between variables may be due to the presence of some cause-effect relationship between the variables. For example, a high degree correlation between 'temperature' and 'sale of coffee' is due to the fact that people like taking coffee in the winter season.

The correlation between variables may also be due to the presence of interdependent relationship between the variables. For example, the presence of correlation between amount spent on entertainment of family and the total expenditure



of family is due to the fact that both variables effects each other. Similarly, the variables, 'total sale' and 'advertisement expenses' are interdependent.

## NOTES

### TYPES OF CORRELATION

Correlation is classified in the following ways:

- (i) Positive and Negative Correlation.
- (ii) Linear and Non-linear Correlation.
- (iii) Simple, Multiple and Partial Correlation.

## 7.4. POSITIVE AND NEGATIVE CORRELATION

The correlation between two variables is said to be **positive** if the variables, on an average, move in the same direction. That is, an increase (or decrease) in the value of one variable is accompanied, on an average, by an increase (or decrease) in the value of the other variable. We do not stress that the variables should move strictly in the same direction. For example, consider the data:

$x$	2	3	6	8	11
$y$	14	15	13	18	22

Here the values of  $y$  has increased corresponding to every increasing value of  $x$ , except for  $x = 6$ . The correlation between the variables  $x$  and  $y$  is positive.

The correlation between two variables is said to be **negative** if the variables, on an average, move in the opposite directions. That is, an increase (or decrease) in the value of one variable is accompanied, on an average, by a decrease (or increase) in the value of the other variable.

Here also, we do not stress that the variables should move strictly in the opposite directions. For example, consider the data:

$x$	110	107	105	95	80
$y$	8	15	14	27	36

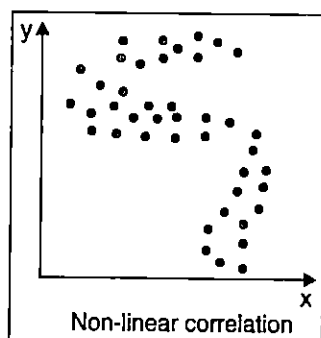
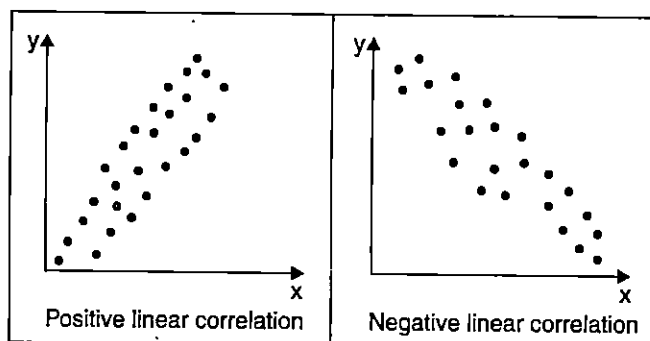
Here, a decrease in the value of  $x$  is accompanied by an increase in the value of  $y$ , except for  $x = 105$ . The correlation between  $x$  and  $y$  is negative.

Thus, we see that the correlation between two variables is positive or negative according as the movements in the variables are in same direction or in the opposite directions, on an average.

## 7.5. LINEAR AND NON-LINEAR CORRELATION

The correlation between two variables is said to be **linear** if the ratio of change in one variable to the change in the other variable is almost constant. The correlation between the 'number of students' admitted and the 'monthly fee collected' is linear in nature. Let  $x$  and  $y$  be two variables such that the ratio of change in  $x$  to the change in  $y$  is almost constant and if a scatter diagram is prepared corresponding to the variables  $x$  and  $y$ , the points in the diagrams would be almost along a line.

The extent of linear correlation is found by using Karl Pearson's method, Spearman's rank correlation method and concurrent deviation method.



## NOTES

The correlation between two variables is said to be **non-linear** if the ratio of change in one variable to the change in the other variable is not constant. The correlation between 'profit' and 'advertisement expenditure' of a company is non-linear, because if the expenditure on advertisement is doubled, the profit may not be doubled. Let  $x$  and  $y$  be two variables in which the ratio of change in  $x$  to the change in  $y$  is not constant and if a scatter diagram is drawn corresponding to the data, the points in the diagram would not be having linear tendency.

## 7.6. SIMPLE, MULTIPLE AND PARTIAL CORRELATION

The correlation is said to be **simple** if there are only two variables under consideration. The correlation between sale and profit figures of a departmental store is simple. If there are more than two variables under consideration, then the correlation is either multiple or partial. Multiple and partial coefficients of correlation are called into play when the values of one variable are influenced by more than one variable. For example, the expenditure of salaried class of people may be influenced by their monthly incomes, secondary sources of income, legacy (money etc. handed down from ancestors) etc. If we intend to find the effect of all these variables on the expenditure of families, this will be a problem of multiple correlation. In **multiple correlation**, the combined effect of a number of variables on a variable is considered. Let  $x_1, x_2, x_3$  be three variables, then  $R_{1.23}$  denotes the multiple correlation coefficient of  $x_1$  on  $x_2$  and  $x_3$ . Similarly  $R_{2.31}$  denotes the multiple correlation coefficient of  $x_2$  on  $x_3$  and  $x_1$ . In **partial correlation**, we study the relationship between any two variables, from a group of more than two variables, after eliminating the effect of other variables mathematically on the variables under consideration. Let  $x_1, x_2, x_3$  be three variables, then  $r_{12.3}$  denotes

the partial correlation coefficient between  $x_1$  and  $x_2$ . Similarly,  $r_{13.2}$  denotes the partial correlation coefficient between  $x_1$  and  $x_3$ . The methods of computing multiple and partial correlation coefficients are beyond the scope of this book. Thus, we shall be discussing the methods of computing only simple correlation coefficient.

## NOTES

### I. KARL PEARSON'S METHOD

#### 7.7. DEFINITION

Let  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  be  $n$  pairs of values of two variables  $x$  and  $y$  with respect to some characteristic (time, place, etc.). The Karl Pearson's method is used to study the presence of *linear correlation* between two variables. The Karl Pearson's coefficient of correlation, denoted by  $r(x, y)$  is defined as:

$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

or simply, 
$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2} \sqrt{\Sigma(y - \bar{y})^2}}$$

where  $\bar{x}$  and  $\bar{y}$  are the A.M.'s of  $x$ -series and  $y$ -series respectively.

This is called the *direct method* of computing Karl Pearson's coefficient of correlation.

If there is no chance of confusion, we write  $r(x, y)$ , just as  $r$ .

It can be proved mathematically that  $-1 \leq r \leq 1$ .

If the correlation between the variables is *linear*, then the value of Karl Pearson's coefficient of correlation is interpreted as follows:

Value of $r$	Degree of linear correlation between the variables
$r = +1$	Perfect positive correlation
$0.75 \leq r < 1$	High degree positive correlation
$0.50 \leq r < 0.75$	Moderate degree positive correlation
$0 < r < 0.50$	Low degree positive correlation
$r = 0$	No correlation
$-0.50 < r < 0$	Low degree negative correlation
$-0.75 \leq r < -0.50$	Moderate degree negative correlation
$-1 < r \leq -0.75$	High degree negative correlation
$r = -1$	Perfect negative correlation

**Remark 1.** The Karl Pearson's coefficient of correlation is also referred to as **product moment correlation coefficient** or as **Karl Pearson's product moment correlation coefficient**.

**Remark 2.** The Karl Pearson's coefficient of correlation,  $r$ , is also denoted by  $\rho(x, y)$  or simply by  $\rho$ . The letter  $\rho$  is the Greek letter 'rho'.

**Remark 3.** The square of Karl Pearson's coefficient of correlation is called the **coefficient of determination**.

Measures of Correlation

For example, if  $r = 0.753$ , then the coefficient of determination is  $(0.753)^2 = 0.567$ .  
The coefficient of determination always lies between 0 and 1, both inclusive.

**Remark 4.**  $r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2} \sqrt{\Sigma(y - \bar{y})^2}}$  implies

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} \sqrt{\frac{\Sigma(y - \bar{y})^2}{n}}}$$

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n \sigma_x \sigma_y}$$

NOTES

**Example 7.1.** From the data given below calculate coefficient of correlation and interpret it:

	$x$	$y$
Number of items	8	8
Mean	68	69
Sum of squares of deviations from mean	36	44

Sum of products of deviations of  $x$  and  $y$  from their respective means = 24.

**Solution.** We are given

$n = 8$ ,  $\bar{x} = 68$ ,  $\bar{y} = 69$ ,  $\Sigma(x - \bar{x})^2 = 36$ ,  $\Sigma(y - \bar{y})^2 = 44$ ,  $\Sigma(x - \bar{x})(y - \bar{y}) = 24$ .  
Coefficient of correlation,

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2} \sqrt{\Sigma(y - \bar{y})^2}} = \frac{24}{\sqrt{36} \sqrt{44}} = \frac{24}{39.7995} = +0.603.$$

$\therefore$  There is moderate degree positive linear correlation between the variables  $x$  and  $y$ .

**Example 7.2.** Two variables  $x$  and  $y$  when expressed as deviations from their respective means are as given below:

$X$	-3	-2	-1	0	+1	+2	+3
$Y$	-3	-1	0	+2	+3	+1	+2

Find the coefficient of correlation between  $x$  and  $y$ .

**Solution.** We have  $X = x - \bar{x}$  and  $Y = y - \bar{y}$ .

$$\text{Also } r(x, y) = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2} \sqrt{\Sigma(y - \bar{y})^2}} \quad \therefore \quad r(x, y) = \frac{\Sigma XY}{\sqrt{\Sigma X^2} \sqrt{\Sigma Y^2}} \quad \dots(1)$$

## NOTES

S. No.	X	Y	XY	X <sup>2</sup>	Y <sup>2</sup>
1	-3	-3	9	9	9
2	-2	-1	2	4	1
3	-1	0	0	1	0
4	0	+2	0	0	4
5	+1	+3	3	1	9
6	+2	+1	2	4	1
7	+3	+2	6	9	4
$n = 7$	$\Sigma X = 0$	$\Sigma Y = 4$	$\Sigma XY = 22$	$\Sigma X^2 = 28$	$\Sigma Y^2 = 28$

$$\therefore (1) \text{ implies } r(x, y) = \frac{22}{\sqrt{28} \sqrt{28}} = \frac{22}{28} = 0.7857.$$

**Example 7.3.** From the data given below, find the correlation coefficient between variables  $X$  and  $Y$ ;  $n = 10$ ,  $\Sigma xy = 120$ ,  $\Sigma x^2 = 90$ , S.D. of  $Y$  series = 8, where  $x$  and  $y$  denote the deviations of items of  $X$  and  $Y$  from their respective A.M.

**Solution.** We have  $n = 10$ ,  $\Sigma xy = 120$ ,  $\Sigma x^2 = 90$ ,  $\sigma_Y = 8$ .

$$\text{Also } x = X - \bar{X} \text{ and } y = Y - \bar{Y}.$$

$$\therefore \Sigma(X - \bar{X})(Y - \bar{Y}) = \Sigma xy = 120, \Sigma(X - \bar{X})^2 = \Sigma x^2 = 90$$

$$\sigma_Y = 8 \text{ implies } \sqrt{\frac{\Sigma(Y - \bar{Y})^2}{n}} = 8 \text{ or } \Sigma(Y - \bar{Y})^2 = (8)^2 \times 10 = 640.$$

$$\begin{aligned} \therefore r(X, Y) &= \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2} \sqrt{\Sigma(Y - \bar{Y})^2}} \\ &= \frac{120}{\sqrt{90} \times \sqrt{640}} = \frac{120}{3\sqrt{10} \times 8\sqrt{10}} \\ &= \frac{120}{240} = \frac{1}{2} = 0.5. \end{aligned}$$

## 7.8. ALTERNATIVE FORM OF 'R'

In the above examples, the calculations involved in **Example 5** is much more than in other examples. This is due to the fractional values of  $\bar{x}$  and  $\bar{y}$  in the data. Suppose for some data, we get  $\bar{x} = 27.374$  and  $\bar{y} = 14.873$ , then it can be well imagined that lot of time and energy would be consumed in computing the Karl Pearson's coefficient of correlation. There are very few chances to get  $\bar{x}$  and  $\bar{y}$  as whole numbers. In order to avoid the chance of facing difficulty in computing deviations of the values of variables from their respective arithmetic means, an alternative form is used which is discussed below:

$$\text{We have } r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2} \sqrt{\Sigma(y_i - \bar{y})^2}}.$$



## NOTES

$$\begin{aligned}
 \text{Now, } \Sigma(x_i - \bar{x})(y_i - \bar{y}) &= \Sigma(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) \\
 &= \Sigma x_i y_i - (\Sigma x_i) \bar{y} - \bar{x} (\Sigma y_i) + n \bar{x} \bar{y} \\
 &= \Sigma x_i y_i - \Sigma x_i \left( \frac{\Sigma y_i}{n} \right) - \left( \frac{\Sigma x_i}{n} \right) \Sigma y_i + n \left( \frac{\Sigma x_i}{n} \right) \left( \frac{\Sigma y_i}{n} \right) \\
 &= \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} = \frac{n \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n}
 \end{aligned}$$

$$\begin{aligned}
 \text{Also } \Sigma(x_i - \bar{x})^2 &= \Sigma(x_i^2 + \bar{x}^2 - 2x_i \bar{x}) = \Sigma x_i^2 + n \bar{x}^2 - 2(\Sigma x_i) \bar{x} \\
 &= \Sigma x_i^2 + n \left( \frac{\Sigma x_i}{n} \right)^2 - 2(\Sigma x_i) \left( \frac{\Sigma x_i}{n} \right) \\
 &= \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = \frac{n \Sigma x_i^2 - (\Sigma x_i)^2}{n}
 \end{aligned}$$

$$\text{Similarly, } \Sigma(y_i - \bar{y})^2 = \frac{n \Sigma y_i^2 - (\Sigma y_i)^2}{n}$$

$$\therefore r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2} \sqrt{\Sigma(y_i - \bar{y})^2}} \text{ implies}$$

$$r = \frac{n \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{\frac{n \Sigma x_i^2 - (\Sigma x_i)^2}{n}} \sqrt{\frac{n \Sigma y_i^2 - (\Sigma y_i)^2}{n}}}$$

$$\therefore r = \frac{n \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{n \Sigma x_i^2 - (\Sigma x_i)^2} \sqrt{n \Sigma y_i^2 - (\Sigma y_i)^2}}$$

For simplicity, we write

$$r = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}}$$

**Example 7.4.** Find the coefficient of correlation for the following data:

$$n = 10, \Sigma x = 50, \Sigma y = -30, \Sigma x^2 = 290, \Sigma y^2 = 300, \Sigma xy = -115.$$

$$\begin{aligned}
 \text{Solution. } r &= \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}} \\
 &= \frac{10(-115) - (50)(-30)}{\sqrt{10(290) - (50)^2} \sqrt{10(300) - (-30)^2}} \\
 &= \frac{350}{\sqrt{400} \sqrt{2100}} = \frac{35}{\sqrt{8400}} = \text{AL} \left[ \log \left( \frac{350}{\sqrt{8400}} \right) \right] \\
 &= \text{AL} \left[ \log 35 - \frac{1}{2} \log 8400 \right] = \text{AL} \left[ 1.5441 - \frac{1}{2} (3.9243) \right] \\
 &= \text{AL} (-0.4181) = \text{AL} (1.5819) = 0.3819.
 \end{aligned}$$

**Example 7.5.** Calculate the Karl Pearson's coefficient of correlation for the data given below:

$x$	4	6	8	10	11
$y$	2	3	4	6	12

## Solution.

## Calculation of 'r'

## NOTES

S. No.	x	y	xy	x <sup>2</sup>	y <sup>2</sup>
1	4	2	8	16	4
2	6	3	18	36	9
3	8	4	32	64	16
4	10	6	60	100	36
5	11	12	132	121	144
n = 5	Σx = 39	Σy = 27	Σxy = 250	Σx <sup>2</sup> = 337	Σy <sup>2</sup> = 209

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{5(250) - (39)(27)}{\sqrt{5(337) - (39)^2} \sqrt{5(209) - (27)^2}}$$

$$= \frac{197}{\sqrt{164} \sqrt{316}} = \frac{197}{227.6488} = 0.8654.$$

**Remark.** We have already found 'r' for the above data in **example 5**. The reader must have felt comfortable in using the alternative form of  $r(x, y)$ .

**Example 7.6.** Calculate the Karl Pearson's coefficient of correlation for the data given below:

(4, 2), (6, 3), (8, 4), (10, 6), (11, 12).

**Solution.** Let  $x$  and  $y$  respectively denote the first and the second variables.

## Calculation of 'r'

S. No.	x	y	xy	x <sup>2</sup>	y <sup>2</sup>
1	4	2	8	16	4
2	6	3	18	36	9
3	8	4	32	64	16
4	10	6	60	100	36
5	11	12	132	121	144
n = 5	Σx = 39	Σy = 27	Σxy = 250	Σx <sup>2</sup> = 337	Σy <sup>2</sup> = 209

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{5(250) - (39)(27)}{\sqrt{5(337) - (39)^2} \sqrt{5(209) - (27)^2}}$$

$$= \frac{197}{\sqrt{164} \sqrt{316}} = \text{AL} \left[ \log \frac{197}{\sqrt{164} \sqrt{316}} \right]$$

$$= \text{AL} \left[ \log 197 - \frac{1}{2} (\log 164 + \log 316) \right]$$

$$= \text{AL} \left[ 2.2945 - \frac{1}{2} (2.2148 + 2.4997) \right]$$

$$= \text{AL} (2.2945 - 2.3573) = \text{AL} (-0.0628)$$

$$= \text{AL} (-1 + 1 - 0.0628) = \text{AL} (\bar{1}.9372) = 0.8654.$$

**Example 7.7.** Calculate coefficient of correlation between Density of population and Death rate for the following data : Measures of Correlation

Region	Area (in sq. km.)	Population	Deaths
A	200	40,000	480
B	150	75,000	1,200
C	120	72,000	1,080
D	80	20,000	270

## NOTES

**Solution.** Let the variables  $x$  and  $y$  denote 'density of population' and 'death rate' respectively.

We have

$$\text{Density of population}^* = \frac{\text{Population}}{\text{Area}} \quad \text{and} \quad \text{Death rate}^* = \frac{\text{No. of deaths}}{\text{Population}} \times 100.$$

$$\therefore \text{ For region A, } x = \frac{40000}{200} = 200, y = \frac{480}{40000} \times 100 = 1.2.$$

$$\text{For region B, } x = \frac{75000}{150} = 500, y = \frac{1200}{75000} \times 100 = 1.6.$$

$$\text{For region C, } x = \frac{72000}{120} = 600, y = \frac{1080}{72000} \times 100 = 1.5.$$

$$\text{For region D, } x = \frac{20000}{80} = 250, y = \frac{270}{20000} \times 100 = 1.35.$$

### Correlation between $x$ and $y$

S.No.	$x$	$y$	$xy$	$x^2$	$y^2$
1	200	1.2	240	40000	1.44
2	500	1.6	800	250000	2.56
3	600	1.5	900	360000	2.25
4	250	1.35	337.5	62500	1.8225
$n = 4$	$\Sigma x = 1550$	$\Sigma y = 5.65$	$\Sigma xy = 2277.5$	$\Sigma x^2 = 712500$	$\Sigma y^2 = 8.0725$

$$\begin{aligned}
 r &= \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \sqrt{n\Sigma y^2 - (\Sigma y)^2}} \\
 &= \frac{4(2277.5) - (1550)(5.65)}{\sqrt{4(712500) - (1550)^2} \sqrt{4(8.0725) - (5.65)^2}} \\
 &= \frac{352.5}{\sqrt{447500} \sqrt{0.3675}} = \frac{352.5}{405.532} = 0.8692.
 \end{aligned}$$

**Example 7.8.** In two sets of variables of  $X$  and  $Y$  with 50 observations of each, the following data were observed:

$$\bar{X} = 10, \text{ S.D. of } X = 3, \bar{Y} = 6, \text{ S.D. of } Y = 2, r_{XY} = +0.3.$$

However, on subsequent verification it was found that one pair with value of  $X (= 10)$  and value of  $Y (= 6)$  was inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of correlation coefficient affected?

## NOTES

**Solution.** We have  $n = 50$ ,  $\bar{X} = 10$ ,  $\sigma_X = 3$ ,  $\bar{Y} = 6$ ,  $\sigma_Y = 2$ ,  $r_{XY} = 0.3$ .

$$\bar{X} = \frac{\Sigma X}{n} \Rightarrow 10 = \frac{\Sigma X}{50} \Rightarrow \Sigma X = 500$$

$$\bar{Y} = \frac{\Sigma Y}{n} \Rightarrow 6 = \frac{\Sigma Y}{50} \Rightarrow \Sigma Y = 300$$

$$\sigma_X = 3 \Rightarrow \sqrt{\frac{\Sigma X^2}{n} - \bar{X}^2} = 3 \Rightarrow \frac{\Sigma X^2}{50} - (10)^2 = 9$$

$$\Rightarrow \Sigma X^2 = 109 \times 50 = 5450$$

$$\sigma_Y = 2 \Rightarrow \sqrt{\frac{\Sigma Y^2}{n} - \bar{Y}^2} = 2 \Rightarrow \frac{\Sigma Y^2}{50} - (6)^2 = 4$$

$$\Rightarrow \Sigma Y^2 = 40 \times 50 = 2000.$$

Also 
$$r_{XY} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2} \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$\therefore 0.3 = \frac{50\Sigma XY - (500)(300)}{\sqrt{50 \times 5450 - (500)^2} \sqrt{50 \times 2000 - (300)^2}}$$

$$\therefore = \frac{50 \Sigma XY - 150000}{150 \times 100}$$

$$\Rightarrow 0.3 \times 15000 = 50 \Sigma XY - 150000.$$

$$\Rightarrow 50 \Sigma XY = 4500 + 150000 \Rightarrow \Sigma XY = 3090.$$

After dropping the incorrect pair ( $X = 10$ ,  $Y = 6$ ), we have 49 pairs of values. Now we find correct values of  $\Sigma X$ ,  $\Sigma Y$ ,  $\Sigma X^2$ ,  $\Sigma Y^2$  and  $\Sigma XY$ .

**Corrected sums**

$$\Sigma X = 500 - 10 = 490, \quad \Sigma Y = 300 - 6 = 294,$$

$$\Sigma X^2 = 5450 - (10)^2 = 5350, \quad \Sigma Y^2 = 2000 - (6)^2 = 1964,$$

$$\Sigma XY = 3090 - (10 \times 6) = 3030.$$

$$\therefore \text{Correct } r_{XY} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2} \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{49(3030) - (490)(294)}{\sqrt{49(5350) - (490)^2} \sqrt{49(1964) - (294)^2}}$$

$$= \frac{4410}{\sqrt{22050} \sqrt{9800}} = \frac{4410}{14700} = 0.3.$$

**EXERCISE 7.3**

1. Find the coefficient of correlation for the following data:

$x$	2	10	8	6	8
$y$	4	6	7	10	6

## NOTES

2. Find the coefficient of correlation for the following data:

$x$	2	3	4	5	6
$y$	4	3	2	8	10

3. Calculate the coefficient of correlation between  $x$  and  $y$  for the following data:

$x$	2	4	5	6	3	6	8	10
$y$	5	6	6	8	4	8	12	15

4. Find Karl Pearson's coefficient of correlation between  $x$  and  $y$  for the following data:

$x$	3	4	8	9	6	2	1
$y$	5	3	7	7	6	9	2

5. Find the coefficient of correlation for the following data:

$x$	1	2	3	4	5	6	7	8	9	10
$y$	10	9	8	8	6	12	4	3	18	1

6. Calculate the coefficient of correlation between  $X$  and  $Y$  for the following data:

$X$	1	2	3	4	5	6	7	8	9
$Y$	9	8	10	12	11	13	14	16	15

7. Calculate the coefficient of correlation for the following data:

$x$	10	7	12	12	9	16	12	18	8	12	14	16
$y$	6	4	7	8	10	7	10	15	5	6	11	13

8. With the following data in 6 cities, calculate the coefficient of correlation by Pearson's method between the density of population and the death rate.

City	Area in square kilometres	Population (in thousands)	No. of deaths
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840
E	120	72	1224
F	80	24	312

9. Coefficient of correlation between variables  $x$  and  $y$  for 20 pairs is 0.3; means of  $x$  and  $y$  are respectively 15 and 20, standard deviations are 4 and 5 respectively. After calculations, it was found that one pair with values (27, 35) was taken as (17, 30). Find the correct coefficient of correlation between  $x$  and  $y$ .

## Answers

1.  $r = 0.2859$       2.  $r = 0.7825$       3.  $r = 0.9623$   
 4.  $r = 0.4078$       5.  $r = -0.1840$       6.  $r = 0.95$   
 7.  $r = 0.748$       8.  $r = 0.9876$       9.  $r = 0.521$



## 7.9. STEP DEVIATION METHOD

### NOTES

When the values of  $x$  and  $y$  are numerically high, as in **Example 12** of Article 10.15, the step deviation method is used.

Deviations of values of variables  $x$  and  $y$  are calculated from some chosen arbitrary numbers, called  $A$  and  $B$ . Let  $h$  be a *positive* common factor of all the deviations  $(x - A)$  of items in the  $x$ -series. The definition of  $h$  is valid, since at least one common factor "1" exist for all the deviations. Similarly let  $k$  be a *positive* factor of all the deviations  $(y - B)$  of items in the  $y$ -series.

$$\text{Let } u = \frac{x - A}{h} \quad \text{and} \quad v = \frac{y - B}{k}$$

$\therefore$  The variables  $u$  and  $v$  are obtained by changing origin and scale of the variables  $x$  and  $y$  respectively.

Since correlation coefficient is independent of change of origin and scale, we have

$$r(x, y) = r(u, v).$$

$$\therefore r(x, y) = \frac{\Sigma(u - \bar{u})(v - \bar{v})}{\sqrt{\Sigma(u - \bar{u})^2} \sqrt{\Sigma(v - \bar{v})^2}}$$

On simplification, we get

$$r(x, y) = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

The values of  $u$  and  $v$  are called the **step deviations** of the values of  $x$  and  $y$  respectively. In the above form:

$\Sigma u$  is the sum of step deviations of the items of  $x$ -series.

$\Sigma v$  is the sum of step deviations of the items of  $y$ -series.

$\Sigma uv$  is the sum of the products of the step deviations of items of  $x$ -series with the corresponding step deviations of items of  $y$ -series.

$\Sigma u^2$  is the sum of the squares of the step deviations of items of  $x$ -series.

$\Sigma v^2$  is the sum of the squares of the step deviations of items of  $y$ -series.

In practical problems, the choice of common factors  $h$  and  $k$  would not create any problem. Even if we do not care to compute step deviations, by dividing the deviations of values of  $x$  and  $y$  by some common factor, the formula would still work. Suppose we have taken deviations ( $u$ ) of the items of  $x$ -series from  $A$ ,

$$\text{i.e., } u = x - A = \frac{x - A}{1}$$

We can consider the values of  $u$  as the step deviations of the items of  $x$ -series, taking '1' as the common factor. Similar argument would also work for  $y$ -series.

Therefore, in solving problems, we first calculate deviations of items of  $x$ -series and  $y$ -series from some convenient and suitable assumed means  $A$  and  $B$  respectively. These deviations of  $x$ -series and  $y$ -series are then divided by positive common factors, if at all desired. If we do not bother to divide these deviations by common factors, then these deviations would be thought of as *step deviations* of items of given series with '1' as the common factor for both series.

Thus if  $u = x - A$  and  $v = y - B$ , then, we have

$$r(x, y) = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

**Example 7.9.** Find the correlation coefficient between 'height of father' and 'height of son', for the following data:

Measures of Correlation

Height of father (in inches)	65	66	67	67	68	69	70	72
Height of son (in inches)	67	68	65	68	72	72	69	71

NOTES

**Solution.** Let  $x$  and  $y$  denote the variables 'height of father' and 'height of son' respectively.

#### Calculation of 'r'

S. No.	$x$	$y$	$u = x - A$ $A = 68$	$v = y - B$ $B = 69$	$uv$	$u^2$	$v^2$
1	65	67	-3	-2	6	9	4
2	66	68	-2	-1	2	4	1
3	67	65	-1	-4	4	1	16
4	67	68	-1	-1	1	1	1
5	68	72	0	3	0	0	9
6	69	72	1	3	3	1	9
7	70	69	2	0	0	4	0
8	72	71	4	2	8	6	4
$n = 8$			$\Sigma u = 0$	$\Sigma v = 0$	$\Sigma uv = 24$	$\Sigma u^2 = 36$	$\Sigma v^2 = 44$

Now

$$\begin{aligned}
 r &= \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \sqrt{n\Sigma v^2 - (\Sigma v)^2}} \\
 &= \frac{8(24) - 0 \times 0}{\sqrt{8(36) - 0^2} \sqrt{8(44) - 0^2}} = \frac{8(24)}{\sqrt{8} \sqrt{36} \sqrt{8} \sqrt{44}} \\
 &= \frac{24}{6 \times \sqrt{44}} = \frac{4}{\sqrt{4} \times \sqrt{11}} = \frac{2}{\sqrt{11}} \\
 &= AL \left\{ \log 2 - \frac{1}{2} \log 11 \right\} = AL \left\{ 0.3010 - \frac{1}{2} (1.0414) \right\} \\
 &= AL \{ 0.3010 - 0.5207 \} = AL \{ -0.2197 \} \\
 &= AL \{ -1 + 1 - 0.2197 \} = AL \{ \bar{1}.7803 \} = 0.6030.
 \end{aligned}$$

$\therefore$

$$r = 0.6030.$$

It shows that there is moderate degree positive linear correlation between the variables.

**Example 7.10.** Psychology test of intelligence and of arithmetical ability were applied to 10 children. Here is a record of ungrouped data showing intelligence and arithmetic ratios. Calculate Karl Pearson's coefficient of correlation:

Child	A	B	C	D	E	F	G	H	I	J
I.R.	105	104	102	101	100	99	98	96	93	92
A.R.	101	103	100	98	95	96	104	92	97	94

**Solution.** Let  $x$  and  $y$  denote the variables I.R. and A.R. respectively.

**NOTES**

Child	$x$	$y$	$u = x - A$ $A = 100$	$v = y - B$ $B = 96$	$uv$	$u^2$	$v^2$
A	105	101	5	5	25	25	25
B	104	103	4	7	28	16	49
C	102	100	2	4	8	4	16
D	101	98	1	2	2	1	4
E	100	95	0	-1	0	0	1
F	99	96	-1	0	0	1	0
G	98	104	-2	8	-16	4	64
H	96	92	-4	-4	16	16	16
I	93	97	-7	1	-7	49	1
J	92	94	-8	-2	16	64	4
$n = 10$			$\Sigma u = 10$	$\Sigma v = -20$	$\Sigma uv = 72$	$\Sigma u^2 = 180$	$\Sigma v^2 = 180$

Now

$$\begin{aligned}
 r &= \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \sqrt{n\Sigma v^2 - (\Sigma v)^2}} \\
 &= \frac{10(72) - (-10)(20)}{\sqrt{10(180) - (-10)^2} \sqrt{10(180) - (20)^2}} \\
 &= \frac{720 + 200}{\sqrt{1800 - 100} \sqrt{1800 - 400}} = \frac{920}{\sqrt{1700} \sqrt{1400}} \\
 &= AL \left\{ \log 920 - \frac{1}{2} (\log 1700 + \log 1400) \right\} \\
 &= AL \left\{ 2.9638 - \frac{1}{2} (3.2304 + \log 3.1461) \right\} \\
 &= AL \{-0.2244\} = AL \{1.7756\} = 0.5965. \\
 \therefore r &= 0.5965.
 \end{aligned}$$

It shows that there is moderate degree positive linear correlation between the variables.

**Example 7.11. Given:**

No. of pairs of observations	$= 10$
Sum of deviations of $x$	$= -170$
Sum of deviations of $y$	$= -20$
Sum of squares of deviations of $x$	$= 8288$
Sum of squares of deviations of $y$	$= 2264$
Sum of product of deviations of $x$ and $y$	$= 3044$

Find out coefficient of correlation when the arbitrary means of  $x$  and  $y$  are 82 and 68 respectively.

**Solution.** Let  $u = x - 82$ ,  $v = y - 68$ .

$\therefore$  We are given

$$\begin{aligned}
 \Sigma u &= -170 & \Sigma v &= -20, & \Sigma u^2 &= 8288, \\
 \Sigma v^2 &= 2264, & \Sigma uv &= 3044.
 \end{aligned}$$

Let 'r' be the coefficient of correlation between the variables x and y.

$$\begin{aligned}
 \therefore r &= \frac{n\sum uv - (\sum u)(\sum v)}{\sqrt{n\sum u^2 - (\sum u)^2} \sqrt{n\sum v^2 - (\sum v)^2}} \\
 &= \frac{10(3044) - (-170)(-20)}{\sqrt{10(8288) - (-170)^2} \sqrt{10(2264) - (-20)^2}} \\
 &= \frac{30440 - 3400}{\sqrt{82880 - 28900} \sqrt{22640 - 400}} = \frac{27040}{\sqrt{53980} \sqrt{22240}} \\
 &= AL \left\{ \log 27040 - \frac{1}{2} (\log 53980 + \log 22240) \right\} \\
 &= AL \left\{ 4.4320 - \frac{1}{2} (4.7322 + 4.3472) \right\} = AL \left\{ 4.4320 - \frac{1}{2} (9.0794) \right\} \\
 &= AL \{ 4.4320 - 4.5397 \} = AL \{ -0.1077 \} = AL \{ \bar{1}.8923 \} = 0.7803. \\
 \therefore r &= 0.7803.
 \end{aligned}$$

## NOTES

**Example 7.12.** From the following table giving the distribution of students and also regular players among them according to age group, find out correlation coefficient between 'age' and 'playing habit':

Age	15-16	16-17	17-18	18-19	19-20	20-21
No. of students	200	270	340	360	400	300
No. of regular players	150	162	170	180	180	120

**Solution.** We are to find the degree of correlation between the variables 'age' and 'playing habit.' The numbers of students in each age group is not same. So, first of all we shall express the number of regular players in each age group as the percentage of students in the corresponding age group. Let x and y denote the variables 'age' and 'percentage of regular players' respectively.

### Calculation of 'r'

Age	Mid-pts. of age groups x	No. of students	No. of regular players	% of regular players y	u = x - A A = 17.5	v = y - B B = 50	uv	u <sup>2</sup>	v <sup>2</sup>
15-16	15.5	200	150	75	-2	25	-50	4	625
16-17	16.5	270	162	60	-1	10	-10	1	100
17-18	17.5	340	170	50	0	0	0	0	0
18-19	18.5	360	180	50	1	0	0	1	0
19-20	19.5	400	180	45	2	-5	-10	4	25
20-21	20.5	300	120	40	3	-10	-30	9	100
n = 6					$\sum u = 3$	$\sum v = 20$	$\sum uv = -100$	$\sum u^2 = 19$	$\sum v^2 = 850$

## NOTES

Now 
$$r = \frac{n\sum uv - (\sum u)(\sum v)}{\sqrt{n\sum u^2 - (\sum u)^2} \sqrt{n\sum v^2 - (\sum v)^2}}$$

$$= \frac{6(-100) - (3)(20)}{\sqrt{6(19) - (3)^2} \sqrt{6(850) - (20)^2}}$$

$$= \frac{-660}{\sqrt{105} \sqrt{4700}} = \frac{-660}{702.4956} = -0.9395.$$

It shows that there is high degree negative linear correlation between the variables.

## EXERCISE 7.4

1. The following table gives the value of iron ore exported and value of steel imported in India during 1970-71 to 1976-77. Find the value of correlation coefficient between exports and imports.

Year	1970-71	1971-72	1972-73	1973-74	1974-75	1975-76	1976-77
Export of iron ore ('000 ₹)	42	44	58	55	89	98	66
Import of steel ('000 ₹)	56	49	53	58	65	76	58

2. Find the coefficient of correlation between income and expenditure of a wage-earner and comment on the result.

Month	Jan.	Feb.	Mar.	Apr.	May	June	July
Income (₹)	46	54	56	56	58	60	62
Expenditure (₹)	36	40	44	54	42	58	54

3. The following table gives the distribution of the total population and those who are wholly or partially blind among them. Find out if there is any relation between 'age' and 'blindness'.

Age	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of persons ('000)	100	60	40	36	24	11	6	3
No. of blinds	55	40	40	40	36	22	18	15

4. Find the correlation coefficient between age and playing habit of the following students:

Age (in years)	No. of students	Regular players
15	250	200
16	200	150
17	150	90
18	120	48
19	100	30
20	80	12



5. Calculate the coefficient of correlation and its probable error between the heights of fathers and sons for the following data: Measures of Correlation

Height of Father (in inches)	65	66	67	68	69	70	71
Height of Son (in inches)	67	68	66	69	72	72	69

## NOTES

6. Calculate the coefficient of correlation for the following data:

$x$	100	200	300	400	500	600	700
$y$	30	50	60	80	100	110	130

7. Calculate Karl Pearson's coefficient of correlation for the following data:

- (i) Sum of deviations of  $x = 5$
- (ii) Sum of deviations of  $y = 4$
- (iii) Sum of squares of deviations of  $x = 40$
- (iv) Sum of squares of deviations of  $y = 50$
- (v) Sum of products of deviations of  $x$  and  $y = 32$
- (vi) No. of pairs of observations = 10

8. Calculate correlation coefficient for the following data:

$$n = 10, \Sigma x = 140, \Sigma y = 150, \Sigma(x - 10)^2 = 180, \Sigma(y - 15)^2 = 215, \Sigma(x - 10)(y - 15) = 60.$$

(Hint. Let  $u = x - 10$ ,  $v = y - 15$ .)

$$\therefore \Sigma u^2 = 180, \Sigma v^2 = 215, \Sigma uv = 60.$$

$$\text{Now } \Sigma u = \Sigma(x - 10) = \Sigma x - n(10) = 140 - 10 \times 10 = 40 \text{ etc.}$$

### Answers

- 1.  $r = 0.9042$
- 2.  $r = 0.769$
- 3.  $r = 0.8982$
- 4.  $r = -0.9276$
- 5.  $r = 0.668$ , P.E. = 0.1412
- 6.  $r = 0.9972$
- 7.  $r = 0.7042$
- 8. 0.915.

## II. SPEARMAN'S RANK CORRELATION METHOD

### 7.10. MEANING

In practical life, we come across problems of estimating correlation between variables, which are not quantitative in nature. Suppose, we are interested in deciding if there is any correlation between the variables 'honesty' and 'smartness' among a group of students. Here the variables 'honesty' and 'smartness' are not capable of quantitative measurement. These variables are qualitative in nature. Ranking is possible in case of qualitative variables.

Spearman's rank correlation method is used for studying linear correlation between variables which are not necessarily quantitative in nature. This method works for both quantitative as well as qualitative variables.

Let  $n$  pairs of values of variables  $x$  and  $y$  be given. The first step is to express the values of the variables in ranks. In case of qualitative variables, the data would be given in the desired form. For quantitative variables, the ranks are allotted according to the magnitude of the values of the variables. Generally the I rank is allotted to the item with highest value. If the highest value of the first variable is allotted I rank, then the same method is to be adopted for finding the ranks of the values of the other variable. In allotting ranks, difficulty arises when the values of two or more items in a series are equal. We shall consider this case separately.

## 7.11. CASE I. NON-REPEATED RANKS

### NOTES

Let  $R_1$  and  $R_2$  represent the ranks of the items corresponding to the variables  $x$  and  $y$  respectively.

The coefficient of rank correlation ( $r_k$ ) is given by the formula:

$$r_k = 1 - \frac{6\sum D^2}{n(n^2 - 1)},$$

where  $n$  is the number of pairs and  $D$  denotes the difference between ranks i.e., ( $R_1 - R_2$ ) of the corresponding values of the variables.

**Example 7.13.** Two judges in a beauty competition rank the 12 entries as follows :

$x$	1	2	3	4	5	6	7	8	9	10	11	12
$y$	12	9	6	10	3	5	4	7	8	2	11	1

What degree of agreement is there between the judges?

**Solution.** Here the ranks are denoted by  $x$  and  $y$ , therefore,  $D = x - y$ .

Calculation of ' $r_k$ '

S. No.	$x$	$y$	$D = x - y$	$D^2$
1	1	12	-11	121
2	2	9	-7	49
3	3	6	-3	9
4	4	10	-6	36
5	5	3	2	4
6	6	5	1	1
7	7	4	3	9
8	8	7	1	1
9	9	8	1	1
10	10	2	8	64
11	11	11	0	0
12	12	1	11	121
$n = 12$				$\sum D^2 = 416$

Coefficient of rank correlation,

$$r_k = 1 - \frac{6\sum D^2}{n(n^2 - 1)} = 1 - \frac{6(416)}{12(12^2 - 1)} = 1 - 1.4545 = -0.4545.$$

It shows that there is low degree negative linear correlation between the variables. This means that the judges are not agreeing, though the degree of disagreement is low.

**Example 7.14.** Ten competitors in a beauty contest are ranked by three judges in the following order:

Ist judge	1	5	4	8	9	6	10	7	3	2
IInd judge	4	8	7	6	5	9	10	3	2	1
IIIrd judge	6	7	8	1	5	10	9	2	3	4

Use the rank correlation coefficient to discuss which pair of judges has the nearest approach to common taste in beauty.

**Solution.** Let  $R_1$ ,  $R_2$  and  $R_3$  denote the variables 'ranks by Ist judge', ranks by IInd judge' and 'ranks by IIIrd judge' respectively. Let  $r_{12}$ ,  $r_{23}$  and  $r_{13}$  stand for the coefficients of rank correlation between the variables  $R_1$  and  $R_2$ ,  $R_2$  and  $R_3$ ,  $R_1$  and  $R_3$  respectively.

### Calculation of $r_{12}$ , $r_{23}$ and $r_{13}$

S. No.	$R_1$	$R_2$	$R_3$	$D_{12} = R_1 - R_2$	$D_{23} = R_2 - R_3$	$D_{13} = R_1 - R_3$	$D_{12}^2$	$D_{23}^2$	$D_{13}^2$
1	1	4	6	-3	-2	-5	9	4	25
2	5	8	7	-3	1	-2	9	1	4
3	4	7	8	-3	-1	-4	9	1	16
4	8	6	1	2	5	7	4	25	49
5	9	5	5	4	0	4	16	0	16
6	6	9	10	-3	-1	-4	9	1	16
7	10	10	9	0	1	1	0	1	1
8	7	3	2	4	1	5	16	1	25
9	3	2	3	1	-1	0	1	1	0
10	2	1	4	1	-3	-2	1	9	4
$n = 10$							$\Sigma D_{12}^2 = 74$	$\Sigma D_{23}^2 = 44$	$\Sigma D_{13}^2 = 156$

$$\text{We have } r_{12} = 1 - \frac{6\Sigma D_{12}^2}{n(n^2 - 1)} = 1 - \frac{6(74)}{10(10^2 - 1)} = 0.5515.$$

$$r_{23} = 1 - \frac{6\Sigma D_{23}^2}{n(n^2 - 1)} = 1 - \frac{6(44)}{10(10^2 - 1)} = 0.7333.$$

$$r_{13} = 1 - \frac{6\Sigma D_{13}^2}{n(n^2 - 1)} = 1 - \frac{6(156)}{10(10^2 - 1)} = 0.0545.$$

By comparing the rank correlation coefficients, we find that  $r_{23}$  is the greatest (and positive) and so we conclude that the IInd judge and IIIrd judge have the nearest approach to common taste in beauty.

**Example 7.15.** The ranks of 16 students in tests in 'Mathematics' and 'Statistics' were as follows. The two numbers within the brackets denoting the ranks of the same student in Mathematics and Statistics respectively.

(1, 1), (2, 10), (3, 3), (4, 4), (5, 5), (6, 7), (7, 2), (8, 6), (9, 8),

(10, 11), (11, 15), (12, 9), (13, 14), (14, 12), (15, 16), (16, 13).

(i) Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and Statistics.

(ii) What does the value of the coefficient obtained indicates?

(iii) If you had found out Karl Pearson's coefficient of correlation between the ranks of these 16 students, would your result be the same as obtained in (i) or different?

**Solution.** Let  $R_1$  and  $R_2$  denote the ranks in 'Mathematics' and Statistics respectively.

### NOTES

## NOTES

Calculation of  $r_s$ 

S. No.	$R_1$	$R_2$	$D = R_1 - R_2$	$D^2$
$n = 16$				
1	1	1	0	0
2	2	10	-8	64
3	3	3	0	0
4	4	4	0	0
5	5	5	0	0
6	6	7	-1	1
7	7	2	5	25
8	8	6	2	4
9	9	8	1	1
10	10	11	-1	1
11	11	15	-4	16
12	12	9	3	9
13	13	14	-1	1
14	14	12	2	4
15	15	16	-1	1
16	16	13	3	9
$\Sigma D^2 = 136$				

Coefficient of rank correlation,

$$r_s = 1 - \frac{6 \Sigma D^2}{n(n^2 - 1)} = 1 - \frac{6(136)}{16((16)^2 - 1)} = 1 - 0.2 = 0.8.$$

- (ii) The value of  $r_s = 0.8$  shows that there is high degree positive linear correlation between the variables ranks in Mathematics and Statistics.
- (iii) Let  $x$  and  $y$  denote the ranks in 'Mathematics' and 'Statistics' respectively i.e.,  $x = R_1$  and  $y = R_2$

Calculation of  $r$ 

S. No.	$x$	$y$	$xy$	$x^2$	$y^2$
$n = 16$					
1	1	1	1	1	1
2	2	10	20	4	100
3	3	3	9	9	9
4	4	4	16	16	16
5	5	5	25	25	25
6	6	7	42	36	49
7	7	2	14	49	4
8	8	6	48	64	36
9	9	8	72	81	64
10	10	11	110	100	121
11	11	15	165	121	225
12	12	9	108	144	81
13	13	14	182	169	196
14	14	12	168	196	144
15	15	16	240	225	256
16	16	13	208	256	169
$\Sigma x = 136$					
$\Sigma y = 136$					
$\Sigma xy = 1428$					
$\Sigma x^2 = 1496$					
$\Sigma y^2 = 1496$					

∴ Karl Pearson's coefficient of correlation,

$$\begin{aligned}
 r &= \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2} \sqrt{n\sum y^2 - (\sum y)^2}} \\
 &= \frac{16(1428) - (136)(136)}{\sqrt{16(1496) - (136)^2} \sqrt{16(1496) - (136)^2}} \\
 &= \frac{4352}{\sqrt{5440} \sqrt{5440}} = \frac{4352}{5440} = 0.8.
 \end{aligned}$$

This coefficient is same as the rank correlation coefficient.

**Remark.** If the non-repeated ranks are given in the data, then the Karl Pearson's coefficient of correlation and Spearman's coefficient are always equal.

## NOTES

### 7.12. CASE II. REPEATED RANKS

Here we shall consider the case, when the values of two or more items in a series are equal. In such cases, we allot equal ranks to all the items with equal values. Suppose that the values of three items in a series are equal at the fourth place, then each item

with equal value would be allotted rank  $\frac{4+5+6}{3} = 5$ . Similarly, if there happen to be

two items in a series with equal values at the seventh place, then each item with equal value would be allotted rank  $\frac{7+8}{2} = 7.5$ .

In case of repeated ranks, the coefficient of rank correlation is given by the formula,

$$r_k = 1 - \frac{6 \left\{ \sum D^2 + \frac{1}{12} (m^3 - m) + \dots \right\}}{n(n^2 - 1)}$$

where  $n$  is the number of pairs and  $D$  denote the difference between ranks ( $R_1 - R_2$ ) of

the corresponding values of the variables. In  $\frac{1}{12} (m^3 - m)$ ,  $m$  is number of items whose

ranks are equal. The term  $\frac{1}{12} (m^3 - m)$  is to be added for each group of items with equal ranks. Now, we shall illustrate this method by taking some examples.

**Example 7.16.** Following are the marks obtained by ten students in Hindi and English. Calculate coefficient of correlation by method of rank differences.

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks in Hindi	45	56	39	54	45	40	56	60	30	36
Marks in English	40	36	30	44	36	32	45	42	20	36

**Solution.** Let  $R_1$  and  $R_2$  denote the ranks of the variables 'marks in Hindi' and 'marks in English' respectively. The first rank is allotted to the greatest item in each series.

### Calculation of $\tau_R$

<i>x</i>	48	33	40	9	16	16	4	20	9	6	19
<i>y</i>	13	13	24	6	15	4	20	9	9	6	19

**Solution.** Let  $R_1$  and  $R_2$  denote the ranks of the variables  $x$  and  $y$  respectively. The first rank is allotted to the greatest item in each series.

### Calculation of $r_k$

S. No.	$x$	$y$	$R_1$	$R_2$	$D = R_1 - R_2$	$D^2$
1	48	18	3	5.5	-2.5	6.25
2	38	18	5	5.5	-0.5	0.25
3	40	24	4	1	3	9
4	9	6	10	8.5	1.5	2.25
5	16	15	8	4	4	16
6	16	4	8	10	-2	4
7	65	20	1	2	-1	1
8	24	9	6	7	-1	1
9	16	6	8	8.5	-0.5	0.25
10	57	19	2	3	-1	1



Now, the coefficient of rank correlation is

$$r_k = 1 - \frac{6 \left\{ \Sigma D^2 + \frac{1}{12} (m^3 - m) + \dots \right\}}{n(n^2 - 1)}$$

Here the items 16, 13, 6 are repeated thrice, twice, twice respectively. Therefore, we shall add the correcting factor  $\frac{1}{12} (m^3 - m)$  three times in the values of  $\Sigma D^2$ , with the values of  $m$  as 3, 2, 2.

$$\begin{aligned} \therefore r_k &= 1 - \frac{6 \left\{ 41 + \frac{1}{12} (3^3 - 3) + \frac{1}{12} (2^3 - 2) + \frac{1}{12} (2^3 - 2) \right\}}{10(10^2 - 1)} \\ &= 1 - \frac{6 \left\{ 41 + 2 + \frac{1}{2} + \frac{1}{2} \right\}}{990} = 1 - \frac{44}{165} = 0.7333. \end{aligned}$$

It shows that there is a moderate degree positive linear correlation between the variables.

**Example 7.18.** The coefficient of rank correlation of the marks obtained by 10 students in Auditing and Accounting was found to be 0.5. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

**Solution.** We have

$$\begin{aligned} \text{Incorrect } r_k &= 0.5 \\ n &= 10 \end{aligned}$$

Incorrect difference of ranks (D) = 3

Correct difference of rank (D) = 7

$$\text{We know that } r_k = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$$

$$\therefore \text{Incorrect } r_k = 1 - \frac{6(\text{incorrect } \Sigma D^2)}{n(n^2 - 1)}$$

$$\therefore 0.5 = 1 - \frac{6(\text{incorrect } \Sigma D^2)}{10(10^2 - 1)}$$

$$\therefore \text{Incorrect } \Sigma D^2 = 82.5.$$

$$\begin{aligned} \text{Now } \text{Correct } \Sigma D^2 &= \text{incorrect } \Sigma D^2 - (\text{incorrect } D^2) + (\text{correct } D^2) \\ &= 82.5 - (3)^2 + (7)^2 = 82.5 - 9 + 49 = 122.5. \end{aligned}$$

$$\begin{aligned} \text{Correct } r_k &= 1 - \frac{6(\text{correct } \Sigma D^2)}{n(n^2 - 1)} = 1 - \frac{6(122.5)}{10(10^2 - 1)} \\ &= 1 - 0.7424 = 0.2575. \end{aligned}$$

## Merits

1. This method is applicable to both qualitative and quantitative variables.
2. Only this method is applicable when ranks are given.
3. This method involves less calculation work as compared to Karl Pearson's method.

## NOTES

**Demerits**

This method is applicable only when the correlation between the variables is linear.

**NOTES****EXERCISE 7.6**

1. From the following data, calculate Spearman's Rank Correlation coefficient.

S. No.	1	2	3	4	5	6	7	8	9	10
Rank Difference	-2	-4	-1	+3	+2	0	-2	+3	+3	2

2. Ten students were examined in Economics and Statistics. The ranks obtained by the students are as follows:

Economics	1	2	3	4	5	6	7	8	9	10
Statistics	2	4	1	5	3	9	7	10	6	8

Calculate the coefficient of rank correlation.

3. Ten students got following percentage of marks in Mathematics and Accountancy papers.

Mathematics	81	36	98	25	75	82	92	62	65	39
Accountancy	84	51	91	60	68	62	86	58	35	49

Find the rank correlation coefficient.

4. Calculate the coefficient of rank correlation for the following data of marks of eight students in Statistics and Accountancy:

Marks in Statistics	52	63	45	36	72	65	45	25
Marks in Accountancy	62	53	51	25	79	43	60	30

5. Ten competitors in an intelligence test are ranked by three examiners in the following order:

Ist Examiner	9	3	7	5	1	6	2	4	10	8
IInd Examiner	9	1	10	4	3	8	5	2	7	6
IIIrd Examiner	6	3	8	7	2	4	1	5	9	10

Calculate the appropriate rank correlation to help you answer the following questions:

- (i) Which pair of judges agree the most?  
 (ii) Which pair of judges disagree the most?
6. An office has 12 clerks. The long serving clerks feel that they should have a seniority increment based on length of service. An assessment of their efficiency by their departmental manager and the personnel department produces a ranking of efficiency. This is shown below together with a ranking of their length of service. Do the data support the claim of clerks for a seniority increment?

Ranking according to length of service	1	2	3	4	5	6	7	8	9	10	11	12
Ranking according to efficiency	2	3	5	1	9	10	11	12	8	7	6	4

7. Find the coefficient of correlation between  $x$  and  $y$  by the method of rank differences:

*Measures of Correlation*

$x$	42	48	35	50	50	57	45	40	50	39
$y$	90	110	95	95	95	120	115	128	116	130

### Answers

- |                        |                      |                     |
|------------------------|----------------------|---------------------|
| 1. $r_k = 0.6364$      | 2. $r_k = 0.7575$    | 3. $r_k = 0.7575$   |
| 4. $r_k = 0.643$       | 5. (i) Ist and IIIrd | (ii) IInd and IIIrd |
| 6. $r_k = 0.3776$ , No | 7. $r_k = -0.0556$ . |                     |

### NOTES

## 7.13. SUMMARY

- Two variables may be related in the sense that the changes in the values of one variable are accompanied by changes in the values of the other variable. But this cannot be interpreted in the sense that the changes in one variable has necessarily caused changes in the other variable. Their movement in sympathy may be due to mere chance. A high degree correlation between two variables may not necessarily imply the existence of a cause-effect relationship between the variables. On the other hand, if there is a cause-effect relationship between the variables, then the correlation is sure to exist between the variables under consideration.
- The correlation between two variables is said to be **positive** if the variables, on an average, move in the same direction. That is, an increase (or decrease) in the value of one variable is accompanied, on an average, by an increase (or decrease) in the value of the other variable.
- The correlation between two variables is said to be **linear** if the ratio of change in one variable to the change in the other variable is almost constant. The correlation between the 'number of students' admitted and the 'monthly fee collected' is linear in nature.
- The correlation is said to be **simple** if there are only two variables under consideration. In **multiple correlation**, the combined effect of a number of variables on a variable is considered. Let  $x_1, x_2, x_3$  be three variables, then  $R_{1.23}$  denotes the multiple correlation coefficient of  $x_1$  on  $x_2$  and  $x_3$ . Similarly  $R_{2.31}$  denotes the multiple correlation coefficient of  $x_2$  on  $x_3$  and  $x_1$ . In **partial correlation**, we study the relationship between any two variables, from a group of more than two variables, after eliminating the effect of other variables mathematically on the variables under consideration.

## 7.14. REVIEW EXERCISES

- Explain the meaning of the term 'Correlation'. Does it always signify cause and effect relationship?
- What is correlation? Distinguish between positive and negative correlation.
- If the 'r' between the annual values of exports during the last ten years and the annual number of children born during the same period is + 0.8. What inference, if any, would you draw?

## NOTES

4. What is a scatter diagram?
5. Explain the meaning of the term 'correlation'. Name the different measures of correlation and discuss their uses.
6. Define correlation and discuss its significance in statistical analysis.
7. Explain different methods of computing correlation.
8. What do you understand by correlation? Explain its various types in detail.
9. What is coefficient of concurrent deviation? How is it determined?
10. Elucidate the main features of Karl Pearson's coefficient of correlation.
11. What is correlation?
12. "If two variables are independent the correlation between them is zero, but the converse is not always true." Comment.

## 8. REGRESSION ANALYSIS

### NOTES

#### STRUCTURE

- 8.1. Introduction
- 8.2. Meaning
- 8.3. Uses of Regression Analysis
- 8.4. Types of Regression
- 8.5. Regression Lines
- 8.6. Regression Equations
- 8.7. Step Deviation Method
- 8.8. Regression Lines for Grouped Data
- 8.9. Properties of Regression Coefficients and Regression Lines
- 8.10. Summary
- 8.11. Review Exercises

### 8.1. INTRODUCTION

In the discussion of correlation, we estimated the degree of relationship between variables. The coefficient of correlation  $r$ , ( $-1 \leq r \leq 1$ ) measured the *degree* of relationship between variables. A numerically high value of ' $r$ ' resulted because of closeness of relation between the variables, under consideration. The coefficient of correlation is unable to depict the *nature* of relationship between the variable. For a given data, regarding the corresponding values of two related variables, the coefficient of correlation cannot give the *estimated value* of a variable, corresponding to a certain value of the other related variable. For example, the coefficient of correlation between 'height' and 'weight' of a group of students of a university cannot help to give the estimated weight (resp. height) of a student with given height (resp. weight). This type of assignment is dealt with the tools of *regression analysis*.

### 8.2. MEANING

The literal meaning of the word 'regression' is 'stepping back towards the average'. British biometrician *Sir Francis Galton* (1822–1911) studies the heights of many persons and concluded that the offspring of abnormally tall or short parents tend to *regress* to the average population height. In statistics, *regression analysis* is concerned with the measure of average relationship between variables. Here we shall deal with

## NOTES

the derivation of appropriate functional relationships between variables. Regression explains the nature of relationship between variables.

There are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable* (or *regressed variable* or *predicted variable* or *explained variable*). The variable which influences the value of dependent variable is called *independent variable* (or *regressor* or *predictor* or *explanator*). Prediction is possible in regression analysis, because here we study the average relationship between related variables.

---

### 8.3. USES OF REGRESSION ANALYSIS

---

The tools of regression analysis are definitely more important and useful than those of correlation analysis. Some of the important uses of regression analysis are as follows:

(i) Regression analysis helps in establishing relationship between dependent variable and independent variables. The independent variables may be more than one. Such relationships are very useful in further studies of the variables, under consideration.

(ii) Regression analysis is very useful for prediction. Once a relation is established between dependent variable and independent variables, the value of dependent variable can be predicted for given values of the independent variables. This is very useful for predicting sale, profit, investment, income, population, etc.

(iii) Regression analysis is specially used in Economics for estimating demand function, production function, consumption function, supply function, etc. A very important branch of Economics, called *Econometrics*, is based on the techniques of regression analysis.

(iv) The coefficient of correlation between two variables can be found easily by using the regression lines between the variables.

---

### 8.4. TYPES OF REGRESSION

---

If there are only two variables under consideration, then the regression is called **simple regression**. For example, the study of regression between 'income' and 'expenditure' for a group of family would be termed as simple regression. If there are more than two variables under consideration then the regression is called **multiple regression**. In this text, we shall restrict ourselves to the study of only simple regression. The regression is called **partial regression** if there are more than two variables under consideration and relation between only two variables is established after excluding the effect of other variables. The simple regression is called **linear regression** if the point on the scatter diagram of variables lies almost along a line otherwise it is termed as **non-linear regression** or **curvilinear regression**.

---

### 8.5. REGRESSION LINES

---

Let the variables under consideration be denoted by ' $x$ ' and ' $y$ '. The line used to estimate the value of  $y$  for a given value of  $x$  is called the *regression line of  $y$  on  $x$* . Similarly, the



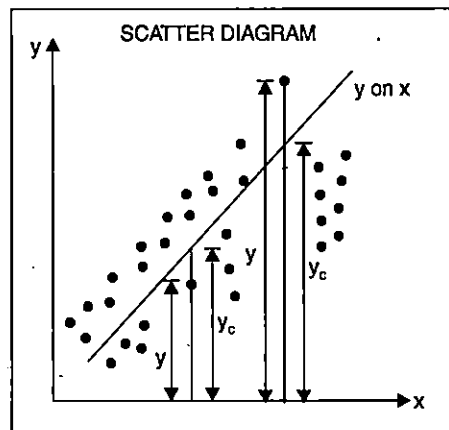
line used to estimate the value of  $x$  for a given value of  $y$  is called the *regression line of  $x$  on  $y$* . In regression line of  $y$  on  $x$  ( $x$  on  $y$ ), the variable  $y$  is considered as the dependent (independent) variable whereas  $x$  is considered as the independent (dependent) variable. The position of regression lines depends upon the given pairs of value of the variables. Regression lines are also known as *estimating lines*. We shall see that in case of perfect correlation between the variables, the regression lines will be coincident. The angle between the regression lines will increase for  $0^\circ$  to  $90^\circ$  as the correlation coefficient numerically decreases from 1 to 0. If for a particular pair of variables,  $r = 0$ , then the regression lines will be perpendicular to each other. The regression lines will be determined by using the *principle of least squares*.

## NOTES

### 8.6. REGRESSION EQUATIONS

We have already noted that for two variables  $x$  and  $y$ , there can be two regression lines. If the intention is to depict the change in  $y$  for a given change in  $x$ , then the regression line of  $y$  on  $x$  is to be used. Similar argument also works for regression line of  $x$  on  $y$ .

(i) **Regression equation of  $y$  on  $x$ .** The regression equation of  $y$  on  $x$  is estimated by using the 'principle of least squares'. This principle will ensure that the sum of the squares of the *vertical* deviations of actual values of  $y$  from estimated values for all possible values of  $x$  is minimum.



Mathematically,  $\Sigma(y - y_c)^2$  is least, where  $y$  and  $y_c$  are the corresponding actual and computed values of  $y$  for a particular value of  $x$ .

Let  $n$  pairs of values  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  of two variables  $x$  and  $y$  be given.

Let the regression equation of  $y$  on  $x$  be  $y = a + bx$ . ... (1)

By using derivatives, it can be proved that the constants  $a$  and  $b$  are found by using the *normal equations*:

$$\Sigma y = an + b\Sigma x \quad \dots (2)$$

and  $\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \dots (3)$

Dividing (2) by  $n$ , we get

$$\frac{\Sigma y}{n} = a + b \frac{\Sigma x}{n}$$

$\Rightarrow$

$$\bar{y} = a + b \bar{x} \quad \dots (4)$$

## NOTES

Subtracting (4) from (1), we get

$$y - \bar{y} = b(x - \bar{x}) \quad \dots(5)$$

Multiplying (2) by  $\Sigma x$  and (3) by  $n$  and subtracting, we get

$$(\Sigma x)(\Sigma y) - n\Sigma xy = b(\Sigma x)^2 - bn\Sigma x^2$$

$$\Rightarrow n\Sigma xy - (\Sigma x)(\Sigma y) = b(n\Sigma x^2 - (\Sigma x)^2)$$

$$\therefore b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

The constant  $b$  is denoted by  $b_{yx}$  and is called **regression coefficient** of  $y$  on  $x$ .

$$\therefore (5) \Rightarrow y - \bar{y} = b_{yx}(x - \bar{x}), \text{ where } b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

**Remark.**  $b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$  implies

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \sqrt{n\Sigma y^2 - (\Sigma y)^2}} \times \frac{\sqrt{n\Sigma y^2 - (\Sigma y)^2}}{n} = r \times \frac{\sqrt{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2}}{\sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}} = r \frac{\sigma_y}{\sigma_x}$$

$$\therefore b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

Thus we see that the regression equation of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ ,

where  $\bar{x} = \frac{\Sigma x}{n}$ ,  $\bar{y} = \frac{\Sigma y}{n}$ ,  $b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$ , which is also equal to  $r \frac{\sigma_y}{\sigma_x}$ .

**Example 8.1.** Find the regression equation of  $y$  on  $x$  when we know :

$$\bar{x} = 68.2, \bar{y} = 9.9, \frac{\sigma_y}{\sigma_x} = 0.44, r = 0.76.$$

**Solution.** We have  $\bar{x} = 68.2$ ,  $\bar{y} = 9.9$ ,  $\frac{\sigma_y}{\sigma_x} = 0.44$ ,  $r = 0.76$ .

The regression equation of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ .

$$\Rightarrow y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \quad \Rightarrow y - 9.9 = (0.76)(0.44)(x - 68.2)$$

$$\Rightarrow y - 9.9 = 0.3344(x - 68.2) \quad \Rightarrow y = 0.3344x + 9.9 - (0.3344)(68.2)$$

$$\Rightarrow y = 0.3344x - 12.9061.$$

**Example 8.2.**  $x$  and  $y$  are correlated variables. Ten observations of values of  $(x, y)$  have the following results:

$$\Sigma x = 55, \Sigma y = 55, \Sigma xy = 350, \Sigma x^2 = 385.$$

Predict the value of  $y$  when the value of  $x$  is 6.

**Solution.** To predict the value of  $y$  for a given value of  $x$ , we shall require the equation of regression line of  $y$  on  $x$ .

The equation of regression line of  $y$  on  $x$  is

$$y - \bar{y} = b_{yx}(x - \bar{x}). \quad \dots(1)$$

We have

$$\Sigma x = 55, \Sigma y = 55, \Sigma xy = 350, \Sigma x^2 = 385, n = 10$$

Now

$$\bar{x} = \frac{\Sigma x}{n} = \frac{55}{10} = 5.5, \bar{y} = \frac{\Sigma y}{n} = \frac{55}{10} = 5.5$$

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{10(350) - (55)(55)}{10(385) - (55)^2} = \frac{475}{825} = 0.5758.$$

$$\therefore (1) \Rightarrow y - 5.5 = 0.5758(x - 5.5)$$

$$\Rightarrow y = 0.5758x + 2.3331.$$

This is the equation of regression line of  $y$  on  $x$ .

When  $x = 6$ , the predicted value of  $y$

$$= 0.5758(6) + 2.3331 = 5.7879.$$

**Example 8.3.** For the following data, find the regression line of  $y$  on  $x$ :

$x$	1	2	3	4	5	8	10
$y$	9	8	10	12	14	16	15

**Solution.**

**Regression line of  $y$  on  $x$**

S. No.	$x$	$y$	$xy$	$x^2$
1	1	9	9	1
2	2	8	16	4
3	3	10	30	9
4	4	12	48	16
5	5	14	70	25
6	8	16	128	64
7	10	15	150	100
$n = 7$	$\Sigma x = 33$	$\Sigma y = 84$	$\Sigma xy = 451$	$\Sigma x^2 = 219$

The regression line of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{33}{7} = 4.714, \bar{y} = \frac{\Sigma y}{n} = \frac{84}{7} = 12$$

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(451) - (33)(84)}{7(219) - (33)^2} = \frac{385}{444} = 0.867.$$

$\therefore$  The equation of regression line of  $y$  on  $x$  is

$$y - 12 = 0.867(x - 4.714)$$

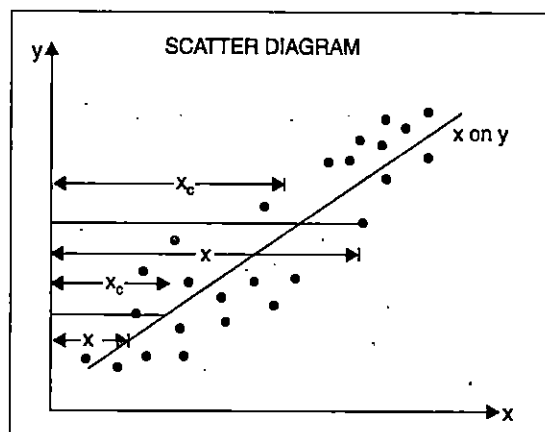
or 
$$y = 0.867x + 12 - (0.867)(4.714)$$

or 
$$y = 0.867x - 7.913.$$

(ii) **Regression equation of  $x$  on  $y$ .** The regression equation of  $x$  on  $y$  is also estimated by using the 'principle of least squares'. This principle will ensure that the sum of the squares of the *horizontal* deviations of actual values of  $x$  from estimated values for all possible values of  $y$  is minimum. Mathematically,  $\Sigma(x - \bar{x}_e)^2$  is least, where  $x$  and  $\bar{x}_e$  are the corresponding actual and computed values of  $x$  for a particular value of  $y$ .

## NOTES

## NOTES



Let  $n$  pairs of values  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  of two variables  $x$  and  $y$  be given.

Let the regression equation of  $x$  on  $y$  be  $x = a + by$  ... (1)

By using derivatives, it can be proved that the constants  $a$  and  $b$  are found by using the *normal equations*:

$$\Sigma x = ax + b\Sigma y \quad \dots (2)$$

and  $\Sigma xy = a\Sigma y + b\Sigma y^2 \quad \dots (3)$

Dividing (2) by  $n$ , we get

$$\frac{\Sigma x}{n} = a + b \frac{\Sigma y}{n}$$

$$\Rightarrow \bar{x} = a + b\bar{y} \quad \dots (4)$$

Subtracting (4) from (1), we get

$$x - \bar{x} = b(y - \bar{y}) \quad \dots (5)$$

Multiplying (2) by  $\Sigma y$  and (3) by  $n$  and subtracting, we get

$$(\Sigma x)(\Sigma y) - n\Sigma xy = b(\Sigma y)^2 - bn\Sigma y^2$$

$$\Rightarrow n\Sigma xy - (\Sigma x)(\Sigma y) = b(n\Sigma y^2 - (\Sigma y)^2)$$

$$\therefore b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$$

The constant  $b$  is denoted by  $b_{xy}$  and is called **regression coefficient** of  $x$  on  $y$ .

$$\therefore (5) \Rightarrow x - \bar{x} = b_{xy}(y - \bar{y}), \text{ where } b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$$

**Remark.**  $b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$  implies

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \sqrt{n\Sigma y^2 - (\Sigma y)^2}} \times \frac{\sqrt{n\Sigma x^2 - (\Sigma x)^2}}{n} = r \times \frac{\sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}}{\sqrt{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2}} = r \frac{\sigma_x}{\sigma_y}$$

$$\therefore b_{xy} = r \frac{\sigma_y}{\sigma_x}$$

Thus, we see that the regression equation of  $x$  on  $y$  is  $x - \bar{x} = b_{xy}(y - \bar{y})$ ,

where  $\bar{x} = \frac{\Sigma x}{n}$ ,  $\bar{y} = \frac{\Sigma y}{n}$ ,  $b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$ , which is also equal to  $r \frac{\sigma_x}{\sigma_y}$ .

**Example 8.4.** Find the regression coefficient  $b_{xy}$  between  $x$  and  $y$  for the following data:

$$\Sigma x = 30, \Sigma y = 42, \Sigma xy = 199, \Sigma x^2 = 184, \Sigma y^2 = 318, n = 6.$$

**Solution.** 
$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{6(199) - (30)(42)}{6(318) - (42)^2} = \frac{-66}{144} = -0.4583.$$

**NOTES**

**Example 8.5.** For observations of pairs  $(x, y)$  of the variables  $x$  and  $y$ , the following results are obtained:

$$\Sigma x = 110, \Sigma y = 70, \Sigma x^2 = 2500, \Sigma y^2 = 2000, \Sigma xy = 100, n = 20.$$

Find the equation of the regression line of  $x$  on  $y$ . Estimate the value of  $x$  when  $y = 4$ .

**Solution.** We have

$$\Sigma x = 110, \Sigma y = 70, \Sigma x^2 = 2500, \Sigma y^2 = 2000, \Sigma xy = 100, n = 20.$$

The equation of regression line of  $x$  on  $y$  is

$$x - \bar{x} = b_{xy}(y - \bar{y}). \quad \dots(1)$$

Now 
$$\bar{x} = \frac{\Sigma x}{n} = \frac{110}{20} = 5.5, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{70}{20} = 3.5$$

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{20(100) - (110)(70)}{20(2000) - (70)^2} = \frac{-5700}{35100} = -0.1624$$

$$\therefore (1) \Rightarrow x - 5.5 = -0.1624(y - 3.5)$$

$$\Rightarrow x = -0.1624y + (0.1624)(3.5) + 5.5$$

$$\Rightarrow x = -0.1624y + 6.0684.$$

This is the regression equation of  $x$  on  $y$ .

When  $y = 4$ , the estimated value of  $y$

$$= -0.1624(4) + 6.0684 = 5.4188.$$

**Example 8.6.** Find the equations of the line of regression of  $y$  on  $x$  and  $x$  on  $y$  for the data:

$x$	5	2	1	4	3
$y$	5	8	4	2	10

**Solution.**

**Regression Equations**

S. No.	$x$	$y$	$xy$	$x^2$	$y^2$
1	5	5	25	25	25
2	2	8	16	4	64
3	1	4	4	1	16
4	4	2	8	16	4
5	3	10	30	9	100
$n = 5$	$\Sigma x = 15$	$\Sigma y = 29$	$\Sigma xy = 83$	$\Sigma x^2 = 55$	$\Sigma y^2 = 209$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{15}{5} = 3, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{29}{5} = 5.8$$

The regression equation of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ .

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{5(83) - (15)(29)}{5(55) - (15)^2} = \frac{-20}{50} = -0.4$$

## NOTES

∴ The equation is

$$y - 5.8 = -0.4(x - 3)$$

or

$$y = -0.4x + (0.4)3 + 5.8$$

or

$$y = -0.4x + 7.$$

The regression equation of  $x$  on  $y$  is  $x - \bar{x} = b_{xy}(y - \bar{y})$ .

$$b_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum y^2 - (\sum y)^2} = \frac{5(83) - (15)(29)}{5(209) - (29)^2} = \frac{-20}{204} = -0.098$$

∴ The equation is

$$x - 3 = -0.098(y - 5.8)$$

or

$$x = 0.098y + (0.098)(5.8) + 3$$

or

$$x = -0.098y + 3.5684.$$

**Example 8.7.** You are given below the following information about advertisement and sales:

	Advt. Expenditure (x) (in crore rupees)	Sales (y) (in crore rupees)
Mean	20	120
S.D.	5	25

Coefficient of correlation = 0.8.

(i) Calculate the regression equations.

(ii) Find the likely sales when advertisement expenditure is ₹ 25 crores.

(iii) What should be advertisement budget if the company wants to attain sales target of ₹ 150 crores?

**Solution.** We have

$$\bar{x} = 20, \bar{y} = 120, \sigma_x = 5, \sigma_y = 25, r = 0.8$$

(i) The regression equation of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ .

$$\Rightarrow y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \Rightarrow y - 120 = (0.8) \frac{25}{5} (x - 20)$$

$$\Rightarrow y - 120 = 4(x - 20) \Rightarrow y = 4x + 40.$$

The regression equation of  $x$  on  $y$  is  $x - \bar{x} = b_{xy}(y - \bar{y})$ .

$$\Rightarrow x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y}) \Rightarrow x - 20 = (0.8) \frac{5}{25} (y - 120)$$

$$\Rightarrow x - 20 = 0.16(y - 120) \Rightarrow x = 0.16y + 0.8.$$

(ii) We are to estimate the value of  $y$  for a given value of  $x$ .

∴ We use regression equation of  $y$  on  $x$  which is  $y = 4x + 40$ .

∴ When  $x = 25$ , the estimated value of  $y = 4(25) + 40 = 140$

∴ Estimated sales = ₹ 140 crores.

(iii) We are to estimate the value of  $x$  for a given value of  $y$ .

∴ We use regression equation of  $x$  on  $y$  which is  $x = 0.16y + 0.8$

∴ When  $y = 150$ , the estimated value of  $x = (0.16)(150) + 0.8 = 24.8$

∴ Estimated advt. expenditure = ₹ 24.8 crores.



**Example 8.8. Given:**

	<i>x</i> -series	<i>y</i> -series
Mean	5	4
S.D.	1.224	1.414

**NOTES**

Sum of products of deviations from means of *x* and *y* series = 6

Number of items = 8.

(i) Obtain the regression equations.

(ii) Estimate the value of *x* when *y* = 5.

**Solution.** (i) We have  $\bar{x} = 5$ ,  $\bar{y} = 4$ ,  $\sigma_x = 1.224$ ,  $\sigma_y = 1.414$ ,  $\Sigma(x - \bar{x})(y - \bar{y}) = 6$  and  $n = 8$ .

$$\begin{aligned} \text{Now } r &= \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2} \sqrt{\Sigma(y - \bar{y})^2}} = \frac{6}{n \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} \sqrt{\frac{\Sigma(y - \bar{y})^2}{n}}} \\ &= \frac{6}{8\sigma_x\sigma_y} = \frac{3}{4(1.224)(1.414)} = 0.433. \end{aligned}$$

Regression equation of *y* on *x* is

$$y - \bar{y} = b_{yx}(x - \bar{x}).$$

$$\Rightarrow y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \quad \Rightarrow y - 4 = 0.433 \times \frac{1.414}{1.224} (x - 5)$$

$$\Rightarrow y - 4 = 0.5(x - 5) \quad \Rightarrow y = 0.5x + 4 - 2.5$$

$$\Rightarrow y = 0.5x + 1.5.$$

Regression equation of *x* on *y* is

$$x - \bar{x} = b_{xy}(y - \bar{y}).$$

$$\Rightarrow x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y}) \Rightarrow x - 5 = 0.433 \times \frac{1.224}{1.414} (y - 4)$$

$$\Rightarrow x - 5 = 0.375(y - 4) \Rightarrow x = 0.375y + 5 - (0.375) \times 4$$

$$\Rightarrow x = 0.375y + 3.5.$$

(ii) When *y* = 5, the estimated value of *x* = (0.375) 5 + 3.5 = 5.375.

(By using regression equation of *x* on *y*)

**Example 8.9. You are given the following data:**

$\Sigma x = 300$ ,  $\bar{x} = 50$ ,  $\Sigma y = 240$ , variance of *x* = 2.56, variance of *y* = 1.96, coefficient of correlation between *x* and *y* = + 0.6.

Find :

(i) Two regression coefficients

(ii) Two regression equations.

**Solution.** We have

$$\Sigma x = 300, \quad \bar{x} = 50, \quad \Sigma y = 240, \quad \sigma_x^2 = 2.56 \quad \sigma_y^2 = 2.25, \quad r = + 0.6.$$

$$\bar{x} = \frac{\Sigma x}{n} \quad \Rightarrow 50 = \frac{300}{n} \quad \Rightarrow n = \frac{300}{50} = 6$$

$$\sigma_x^2 = 2.56 \Rightarrow \sigma_x = +\sqrt{2.56} = 1.6$$

$$\sigma_y^2 = 1.96 \Rightarrow \sigma_y = +\sqrt{1.16} = 1.4$$

**NOTES**

$$(i) \quad b_{yx} = r \frac{\sigma_y}{\sigma_x} = 0.6 \times \frac{1.4}{1.6} = 0.525$$

and 
$$b_{xy} = r \frac{\sigma_x}{\sigma_y} = 0.6 \times \frac{1.6}{1.4} = 0.686$$

(ii) Regression equation of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ .

$$\Rightarrow y - 40 = 0.525(x - 50) \quad \left( \bar{y} = \frac{\Sigma y}{n} = \frac{240}{6} = 40 \right)$$

$$\Rightarrow y = 0.525x + 40 - (0.525) 50$$

$$\Rightarrow y = 0.525x + 13.75.$$

Regression equation of  $x$  on  $y$  is  $x - \bar{x} = b_{xy}(y - \bar{y})$ .

$$\Rightarrow x - 50 = 0.686(y - 40)$$

$$\Rightarrow x = 0.686y + 50 - (0.686)(40)$$

$$\Rightarrow x = 0.686y + 22.56.$$

**Example 8.10.** From the following data, find the regression equations:

$x$	6	2	10	4	8
$y$	9	11	5	8	7

**Solution.**

**Estimation of Regression Equations**

S. No.	$x$	$y$	$xy$	$x^2$	$y^2$
1	6	9	54	36	81
2	2	11	22	4	121
3	10	5	50	100	25
4	4	8	32	16	64
5	8	7	56	64	49
$n = 5$	$\Sigma x = 30$	$\Sigma y = 40$	$\Sigma xy = 214$	$\Sigma x^2 = 220$	$\Sigma y^2 = 340$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{30}{5} = 6, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{40}{5} = 8.$$

The regression equation of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ .

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{5(214) - (30)(40)}{5(220) - (30)^2} = \frac{-130}{200} = -0.65.$$

$\therefore$  The equation is

$$y - 8 = -0.65(x - 6) \quad \text{or} \quad y = -0.65x + (0.65)6 + 8$$

or

$$y = 11.9 - 0.65x.$$

The regression equation of  $x$  on  $y$  is  $x - \bar{x} = b_{xy}(y - \bar{y})$ .

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{5(214) - (30)(40)}{5(340) - (40)^2} = \frac{-130}{100} = -1.3.$$

∴ The equation is

$$x - 6 = -1.3(y - 8) \quad \text{or} \quad x = -1.3y + (1.3)8 + 6$$

or  $x = 16.4 - 1.3y.$

**Example 8.11.** In order to find the correlation coefficient between two variables  $X$  and  $Y$  from 12 pairs of observations, the following calculation were made:

$$\Sigma X = 30, \Sigma X^2 = 670, \Sigma Y = 5, \Sigma Y^2 = 285, \Sigma XY = 344.$$

On subsequent verification, it was discovered that the pair ( $X = 11, Y = 4$ ) was copied wrongly, the correct values being  $X = 10$  and  $Y = 14$ . After making necessary correction, find the:

## NOTES

- (i) regression coefficients
- (ii) regression equations and
- (iii) correlation coefficient.

**Solution.** We have  $\Sigma X = 30, \Sigma X^2 = 670, \Sigma Y = 5, \Sigma Y^2 = 285, \Sigma XY = 344.$

Incorrect pair = ( $X = 11, Y = 4$ )

Correct pair = ( $X = 10, Y = 14$ )

**Corrected sums**

$$\Sigma X = 30 - 11 + 10 = 29$$

$$\Sigma Y = 5 - 4 + 14 = 15$$

$$\Sigma X^2 = 670 - (11)^2 + (10)^2 = 649$$

$$\Sigma Y^2 = 285 - (4)^2 + (14)^2 = 465$$

$$\Sigma XY = 344 - (11 \times 4) + (10 \times 14) = 440.$$

(i) **Regression coefficients**

$$b_{YX} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{12(440) - (29)(15)}{12(649) - (29)^2} = \frac{4845}{6947} = 0.6974$$

$$b_{XY} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma Y^2 - (\Sigma Y)^2} = \frac{12(440) - (29)(15)}{12(465) - (15)^2} = \frac{4845}{5355} = 0.9048.$$

(ii) **Regression equations**

$$\bar{X} = \frac{\Sigma X}{n} = \frac{29}{12} = 2.42, \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{15}{12} = 1.25$$

Regression equation of  $Y$  on  $X$  is  $Y - \bar{Y} = b_{YX}(X - \bar{X}).$

$$\Rightarrow Y - 1.25 = 0.6974(X - 2.42)$$

$$\Rightarrow Y = 0.6974X + 1.25 - (0.6974)(2.42)$$

$$\Rightarrow Y = 0.6974X - 0.438.$$

Regression equation of  $X$  on  $Y$  is  $X - \bar{X} = b_{XY}(Y - \bar{Y}).$

$$\Rightarrow X - 2.42 = 0.9048(Y - 1.25)$$

$$\Rightarrow X = 0.9048Y + 2.42 - (0.9048)(1.25)$$

$$\Rightarrow X = 0.9048Y + 1.289.$$

$$(iii) \quad b_{YX} \cdot b_{XY} = \left( r \frac{\sigma_Y}{\sigma_X} \right) \left( r \frac{\sigma_X}{\sigma_Y} \right) = r^2$$

$$\therefore r = \pm \sqrt{b_{YX} \cdot b_{XY}}$$

## NOTES

$$\therefore r = \pm \sqrt{0.6974 \times 0.9048} = \pm 0.7944$$

$$b_{YX} > 0 \Rightarrow r > 0.$$

$$\left( \because \frac{\sigma_y}{\sigma_x} > 0 \right)$$

$$\therefore r = +0.7944.$$

## EXERCISE 8.1

1. Find  $b_{yx}$  from the following data:

$$\Sigma x = 30, \Sigma y = 42, \Sigma xy = 199, \Sigma x^2 = 184, \Sigma y^2 = 318, n = 6.$$

2. Find  $b_{yx}$  from the following data:

$x$	1	2	3	4	5
$y$	6	8	7	6	8

3. The following results were worked out from scores in Statistics and Mathematics in a certain examination:

	Scores in Statistics ( $x$ )	Scores in Mathematics ( $y$ )
A.M.	39.5	47.5
S.D.	10.8	17.8

Karl Pearson's coefficient of correlation is 0.42. Find both regression lines. Estimate the value of  $y$  when  $x = 50$  and  $x$  when  $y = 30$ .

4. From the following data find the yield of wheat in kg per unit area when the rain fall is 9 inches:

	Mean	S.D.
Yield of wheat per unit area (in kg)	10	8
Annual rainfall (in inches)	8	2

Coefficient of correlation = 0.5.

5. You are given below the following information about advertisement expenditure and sales:

	Advt. Expenditure ( $x$ ) (in crore rupees)	Sales ( $y$ ) (in crore rupees)
Mean	10	90
S.D.	3	12

Coefficient of correlation = + 0.8.

- Calculate two regression equations.
  - Find the likely sales when advertisement expenditure is ₹ 30 crores.
  - What should be the advertisement budget if the company wants to attain sales target of ₹ 150 crores?
6. Find the equations of regression lines for the following pairs ( $x, y$ ) for variables  $x$  and  $y$ :  
(1, 2), (2, 5), (3, 3), (4, 8), (5, 7).

## NOTES

7. For the following data, determine the regression lines:

$x$	6	2	10	4	8
$y$	9	11	5	8	7

From these regression lines, estimate the value of:

- (i)  $y$  when  $x = 5$  and (ii)  $x$  when  $y = 10$ .

8. Find the regression lines for the following pairs  $(x, y)$  for variables  $x$  and  $y$ :

$(1, 6), (5, 1), (3, 0), (2, 0), (1, 1), (1, 2), (7, 1), (3, 5)$ .

9. By using the following data regarding pairs  $(x, y)$  for variables  $x$  and  $y$ , find the most likely value of  $y$ , when  $x = 6.2$ :

$(1, 9), (2, 8), (3, 10), (4, 12), (5, 11), (6, 13), (7, 14), (8, 16), (9, 15)$ .

10. A computer while calculating the correlation coefficient between two variables  $x$  and  $y$  obtained the following constants:

$$n = 25, \Sigma x = 127, \Sigma y = 100, \Sigma x^2 = 650, \Sigma y^2 = 450, \Sigma xy = 516.$$

It was however, later discovered at the time of checking that it copied down two pairs of

observations as: 

$x$	$y$
8	12
6	8

 while the correct values were: 

$x$	$y$
8	10
6	10

 . After making

the necessary corrections, find the:

- (i) regression coefficients (ii) regression equations and  
(iii) correlation coefficient.

### Answers

1.  $-0.3235$                       2.  $0.2$
3.  $y = 0.6922x + 20.1581, x = 0.2548y + 27.397, 54.77, 35.04$                       4.  $12 \text{ kg}$
5. (i)  $x = 0.2y - 8, y = 3.2x + 58,$                       (ii) ₹ 154 crores                      (iii) ₹ 22 crores
6. Regression line of  $y$  on  $x: y = 1.1 + 1.3x;$   
Regression line of  $x$  on  $y: x = 0.5 + 0.5y$
7.  $y = 11.9 - 0.65x, x = 16.4 - 1.3y$                       (i)  $8.65$                       (ii)  $3.4$
8.  $y = 2.8745 - 0.3042x, x = 3.4306 - 0.2778y$                       9.  $13.14$
10. (i)  $b_{yx} = 0.826, b_{xy} = 0.095$   
(ii) Regression equation of  $y$  on  $x: y = 0.826x - 1.96$   
Regression equation of  $x$  on  $y: x = 0.095y + 4.7$   
(iii)  $r = 0.28$

## 8.7. STEP DEVIATION METHOD

When the values of  $x$  and  $y$  are numerically high, the step deviation method is used.

Deviations of values of variables  $x$  and  $y$  are calculated from some chosen arbitrary numbers, called  $A$  and  $B$ . Let  $h$  be a positive common factor of all deviations  $(x - A)$  of items in the  $x$ -series. Similarly let  $k$  be a positive factor of all deviations  $(y - B)$  of items in the  $y$ -series. The step deviations are:

$$u = \frac{x - A}{h}, v = \frac{y - B}{k}.$$

In practical problems, if we do not bother to divide the deviations by common factors, then these deviations would be thought of as step deviations of items of given series with '1' as the common factor for both series.

The equation of regression line of  $y$  on  $x$  in terms of step deviations is

$$y - \bar{y} = b_{yx}(x - \bar{x}),$$

**NOTES**

where

$$\bar{x} = A + \left(\frac{\Sigma u}{n}\right)h, \quad \bar{y} = B + \left(\frac{\Sigma v}{n}\right)k$$

and

$$b_{yx} = b_{vu} \cdot \frac{k}{h} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \cdot \frac{k}{h}.$$

The equation of regression line of  $x$  on  $y$  in terms of step deviations is

$$x - \bar{x} = b_{xy}(y - \bar{y}),$$

where

$$\bar{x} = A + \left(\frac{\Sigma u}{n}\right)h, \quad \bar{y} = B + \left(\frac{\Sigma v}{n}\right)k$$

and

$$b_{xy} = b_{uv} \cdot \frac{h}{k} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2} \cdot \frac{h}{k}.$$

**Remark.** In particular if  $u = x - A$  and  $v = y - B$  i.e., when  $h = 1, k = 1$ , we have

$$\bar{x} = A + \frac{\Sigma u}{n}, \quad \bar{y} = B + \frac{\Sigma v}{n},$$

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \quad \text{and} \quad b_{xy} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2}.$$

**Example 8.12.** An investigation into the demand for television sets in 7 towns has resulted in the following data:

Population $x$ (in thousand)	11	14	14	17	17	21	25
No. of T.V. sets demanded, $y$	15	27	27	30	34	38	46

Calculate the regression equation of  $y$  on  $x$  and estimate the demand for T.V. sets for a town with a population of 30 thousands.

**Solution.** Computation of Regression Equation of  $y$  on  $x$

S. No.	$x$	$y$	$u = x - A$ $A = 17$	$v = y - B$ $B = 27$	$uv$	$u^2$
1	11	15	-6	-12	72	36
2	14	27	-3	0	0	9
3	14	27	-3	0	0	9
4	17	30	0	3	0	0
5	17	34	0	7	0	0
6	21	38	4	11	44	16
7	25	46	8	19	152	64
$n = 7$			$\Sigma u = 0$	$\Sigma v = 28$	$\Sigma uv = 268$	$\Sigma u^2 = 134$

Regression equation of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ .

$$\text{We have } \bar{x} = A + \frac{\Sigma u}{n} = 17 + \frac{0}{7} = 17$$

$$\bar{y} = B + \frac{\Sigma v}{n} = 27 + \frac{28}{7} = 31$$



$$b_{yx} = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum u^2 - (\sum u)^2} = \frac{7(268) - (0)(28)}{7(134) - (0)^2} = \frac{268}{134} = 2$$

∴ The required equation is

$$y - 31 = 2(x - 17) \text{ or } y = 2x - 3.$$

When population is 30 thousand i.e.,  $x = 30$ , the estimated value of

$$y = 2(30) - 3 = 57.$$

∴ The estimated demand for T.V. sets is 57.

**Example 8.13.** Obtain the two regression equations from the following data:

$x$	25	28	35	32	31	36	29	38	34	32
$y$	43	46	49	41	36	32	31	30	33	39

Also find the value of  $y$  when  $x$  is equal to 30.

**Solution.** Computation of Regression Equations

S. No.	$x$	$y$	$u = x - A$ $A = 32$	$v = y - B$ $B = 38$	$uv$	$u^2$	$v^2$
1	25	43	-7	5	-35	49	25
2	28	46	-4	8	-32	16	64
3	35	49	3	11	33	9	121
4	32	41	0	3	0	0	9
5	31	36	-1	-2	2	1	4
6	36	32	4	-6	-24	16	36
7	29	31	-3	-7	21	9	49
8	38	30	6	-8	-48	36	64
9	34	33	2	-5	-10	4	25
10	32	39	0	1	0	0	1
$n = 10$			$\sum u = 0$	$\sum v = 0$	$\sum uv = -93$	$\sum u^2 = 140$	$\sum v^2 = 398$

**Regression equation of 'y on x'**

The regression equation of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ .

$$\bar{x} = A + \frac{\sum u}{n} = 32 + \frac{0}{10} = 32$$

$$\bar{y} = B + \frac{\sum v}{n} = 38 + \frac{0}{10} = 38$$

$$b_{yx} = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum u^2 - (\sum u)^2} = \frac{10(-93) - (0)(0)}{10(140) - (0)^2} = -\frac{93}{140} = -0.6643.$$

∴ The required equation is  $y - 38 = -0.6643(x - 32)$ .

$$\Rightarrow y = -0.6643x + 38 + (0.6643)(32)$$

$$\Rightarrow y = -0.6643x + 59.2576.$$

When  $x = 30$ , the estimated value of  $y = (-0.6643)(30) + 59.2576 = 39.3286$ .

**Regression equation of 'x on y'**

The regression equation of  $x$  on  $y$  is  $x - \bar{x} = b_{xy}(y - \bar{y})$ .

$$\bar{x} = 32, \bar{y} = 38$$

$$b_{xy} = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum v^2 - (\sum v)^2} = \frac{10(-93) - (0)(0)}{10(398) - (0)^2} = -\frac{93}{398} = -0.2337.$$

## NOTES

$\therefore$  The required equation is  $x - 32 = -0.2337(y - 38)$ .

$$\Rightarrow x = -0.2337y + 32 + (0.2337)(38)$$

$$\Rightarrow x = -0.2337y + 40.8806.$$

## NOTES

**Example 8.14.** Following are the heights of fathers and sons in inches:

Height of father	65	66	67	68	69	71	73	67
Height of son	67	68	64	72	70	69	70	68

Find the two lines of regression and estimate the height of the son when the height of the father is 67.5 inches.

**Solution.** Let the variables 'height of father' and 'height of son' be denoted by  $x$  and  $y$  respectively.

### Computation of Regression Equations

S. No.	$x$	$y$	$u = x - A$ $A = 68$	$v = y - B$ $B = 68$	$uv$	$u^2$	$v^2$
1	65	67	-3	-1	3	9	1
2	66	68	-2	0	0	4	0
3	67	64	-1	-4	4	1	16
4	68	72	0	4	0	0	16
5	69	70	1	2	2	1	4
6	71	69	3	1	3	9	1
7	73	70	5	2	10	25	4
8	67	68	-1	0	0	1	0
$n = 8$			$\Sigma u = 2$	$\Sigma v = 4$	$\Sigma uv = 22$	$\Sigma u^2 = 50$	$\Sigma v^2 = 42$

### Regression equation of 'y on x'

The regression equation of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ .

$$\bar{x} = A + \frac{\Sigma u}{n} = 68 + \frac{2}{8} = 68.25 \text{ inches}$$

$$\bar{y} = B + \frac{\Sigma v}{n} = 68 + \frac{4}{8} = 68.5 \text{ inches}$$

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} = \frac{8(22) - (2)(4)}{8(50) - (2)^2} = \frac{168}{396} = 0.4242$$

$\therefore$  The required equation is

$$y - 68.5 = 0.4242(x - 68.25).$$

$$\Rightarrow y = 0.4242x + 68.5 - (0.4242)(68.25)$$

$$\Rightarrow y = 0.4242x + 39.4835.$$

### Regression equation of 'x on y'

The regression equation of  $x$  on  $y$  is  $x - \bar{x} = b_{xy}(y - \bar{y})$ .

$$\bar{x} = 68.25 \text{ inches, } \bar{y} = 68.5 \text{ inches}$$

$$b_{xy} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2} = \frac{8(22) - (2)(4)}{8(42) - (4)^2} = \frac{168}{320} = 0.525$$

$\therefore$  The required equation is

$$x - 68.25 = 0.525(y - 68.5).$$

$$\Rightarrow x = 0.525y + 68.25 - (0.525)(68.5)$$

$$\Rightarrow x = 0.525y + 32.2875.$$

To find the estimated value of height of son ( $y$ ) for a given value of height of father ( $x$ ), we require regression equation of  $y$  on  $x$  i.e.,

$$y = 0.4242x + 39.4835.$$

$\therefore$  When  $x = 67.5$  inches, the estimated value of

$$y = (0.4242)(67.5) + 39.4835 = 68.117 \text{ inches.}$$

**Example 8.15.** Students of a class have obtained marks as given below in Paper I and Paper II of Statistics:

Paper I	45	55	56	58	60	65	68	70	75	80	85
Paper II	56	50	48	60	62	64	65	70	74	82	90

Find the means, coefficient of correlation, regression coefficients and regression equations.

**Solution.** Let the variables 'marks in Paper I' and 'marks in Paper II' be denoted by  $x$  and  $y$  respectively.

#### Computation of $\bar{x}, \bar{y}, r$

S. No.	$x$	$y$	$u = x - A$ $A = 60$	$v = y - B$ $B = 70$	$uv$	$u^2$	$v^2$
1	45	56	-15	-14	210	225	196
2	55	50	-5	-20	100	25	400
3	56	48	-4	-22	88	16	484
4	58	60	-2	-10	20	4	100
5	60	62	0	-8	0	0	64
6	65	64	5	-6	-30	25	36
7	68	65	8	-5	-40	64	25
8	70	70	10	0	0	100	0
9	75	74	15	4	60	225	16
10	80	82	20	12	240	400	144
11	85	90	25	20	500	625	400
$n = 11$			$\Sigma u = 57$	$\Sigma v = -49$	$\Sigma uv = 1148$	$\Sigma u^2 = 1709$	$\Sigma v^2 = 1865$

$$\text{Means } \bar{x} = A + \frac{\Sigma u}{n} = 60 + \frac{57}{11} = 65.1818.$$

$$\bar{y} = B + \frac{\Sigma v}{n} = 70 + \frac{(-49)}{11} = 65.5455.$$

#### Coefficient of correlation

$$\begin{aligned}
 r &= \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \sqrt{n\Sigma v^2 - (\Sigma v)^2}} \\
 &= \frac{11(1148) - (57)(-49)}{\sqrt{11(1709) - (57)^2} \sqrt{11(1865) - (-49)^2}} \\
 &= \frac{15421}{\sqrt{15550} \sqrt{18114}} = \frac{15421}{124.70 \times 134.59} = 0.9188.
 \end{aligned}$$

#### NOTES



## Regression coefficients

$$b_{yx} = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum u^2 - (\sum u)^2} = \frac{11(1148) - (57)(-49)}{11(1709) - (57)^2} = \frac{15421}{15550} = 0.9917$$

## NOTES

$$b_{xy} = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum v^2 - (\sum v)^2} = \frac{11(1148) - (57)(-49)}{11(1865) - (-49)^2} = \frac{15421}{18114} = 0.8513.$$

## Regression equations

Regression equation of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ .

$$\Rightarrow y - 65.5455 = 0.9917(x - 65.1818)$$

$$\Rightarrow y = 0.9917x + 65.5455 - (0.9917)(65.1818)$$

$$\Rightarrow y = 0.9917x + 0.9047.$$

Regression equation of  $x$  on  $y$  is  $x - \bar{x} = b_{xy}(y - \bar{y})$ .

$$\Rightarrow x - 65.1818 = 0.8513(y - 65.5455)$$

$$\Rightarrow x = 0.8513y + 65.1818 - (0.8513)(65.5455)$$

$$\Rightarrow x = 0.8513y + 9.3829.$$

## EXERCISE 8.2

1. The following data relates to 'advertising expenditure' (in lakhs of rupees) and 'sales' (in crores of rupees)

Advertising expenditure (in lakhs of rupees)	10	12	15	23	20
Sales (in crores of rupees)	14	17	23	25	21

Estimate (i) the sales corresponding to advertising expenditure of ₹ 30 lakhs and (ii) the advertising expenditure for a sales target of 35 crores.

2. Find two lines of regression from the following data:

Age of husband	25	22	28	26	35	20	22	40	20	18
Age of wife	18	15	20	17	22	14	16	21	15	14

Hence estimate (i) the age of husband when the age of wife is 19 years and (ii) the age of wife when the age of husband is 30 years.

3. Find the regression equation of  $y$  on  $x$  for the following data:

$x$	78	89	97	69	59	79	68	61
$y$	125	137	156	112	107	136	124	108

4. Find the two regression equations for the following series. What is the most likely value of  $x$  when  $y = 20$  and most likely value of  $y$  when  $x = 22$ ?

$x$	35	25	29	31	27	24	33	36
$y$	23	27	26	21	24	20	29	30

5. Find the regression equations for the following data:

$x$	23	26	39	31	36	21	30	39
$y$	45	48	45	42	31	39	38	32

Also find:

- (i) the value of  $y$  when  $x = 30$   
 (ii) correlation coefficient between  $x$  and  $y$ .
6. Obtain the lines of regression and show them on the graph paper for the following data:

$x$	65	66	67	67	68	69	71	71
$y$	67	68	64	68	70	70	69	68

### Answers

- (i) 29.9666 crore rupees (ii) 31.75 lakh rupees
- If  $x$  and  $y$  respectively represent 'age of husband' and 'age of wife' then the regression equations are  $x = 2.23y - 12.76$  and  $y = 0.385x + 7.34$   
 (i) 30 years nearly (ii) 19 years nearly
- $y = 1.212x + 34.725$
- Regression equation of  $x$  on  $y$ :  $x = 0.543y + 16.425$   
 Regression equation of  $y$  on  $x$ :  $y = 0.352x + 14.44$   
 $x = 27.285$  when  $y = 20$ ,  $y = 22.184$  when  $x = 22$
- Regression equation of  $y$  on  $x$ :  $y = 0.4x + 27.75$   
 Regression equation of  $x$  on  $y$ :  $x = 0.511y + 10.185$   
 (i) 39.75 (ii) 0.452
- Regression equation of  $y$  on  $x$ :  $y = 0.2353x + 52$   
 Regression equation of  $x$  on  $y$ :  $x = 0.4615y + 36.618$ .

## 8.8. REGRESSION LINES FOR GROUPED DATA

In case of grouped data if either  $x$  or  $y$  or both variables represent classes, then their respective mid-points are taken as their representatives.

$$\text{In this case, if } u = \frac{x - A}{h}, \quad v = \frac{y - B}{k},$$

then the regression line of  $y$  on  $x$  is  $y - \bar{y} = b_{yx}(x - \bar{x})$ ,

$$\text{where } \bar{x} = A + \left( \frac{\sum fu}{N} \right) h,$$

$$\bar{y} = B + \left( \frac{\sum fv}{N} \right) k$$

$$\text{and } b_{yx} = \frac{N \sum fuv - (\sum fu)(\sum fv)}{N \sum fu^2 - (\sum fu)^2} \cdot \frac{k}{h}$$

### NOTES

The regression line of  $x$  on  $y$  is

$$x - \bar{x} = b_{xy}(y - \bar{y}),$$

## NOTES

where

$$\bar{x} = A + \left( \frac{\Sigma fu}{N} \right) h,$$

$$\bar{y} = B + \left( \frac{\Sigma fv}{N} \right) k$$

and

$$b_{xy} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma f v^2 - (\Sigma fv)^2} \cdot \frac{h}{k}$$

**Example 8.16.** Compute regression lines corresponding to the marks obtained by 25 students in Economics and Statistics:

	Marks in Statistics			
Marks in Economics	30-40	40-50	50-60	60-70
30-40	3	1	1	0
40-50	2	6	1	2
50-60	1	2	2	1
60-70	0	1	1	1

**Solution.** Let  $x$  and  $y$  denote the variables 'marks in Statistics' and 'marks in Economics' respectively.

Class of  $x$  : 30-40 40-50 50-60 60-70

Mid-point ( $x$ ) : 35 45 55 65

Deviation from

$A = 45$  : -10 0 10 20

Step deviation by  $h = 10$

$\left( u = \frac{x - 45}{10} \right)$  : -1 0 1 2

Class of  $y$  : 30-40 40-50 50-60 60-70

Mid-point ( $y$ ) : 35 45 55 65

Deviation from

$B = 45$  : -10 0 10 20

Step deviation by  $k = 10$

$\left( v = \frac{y - 45}{10} \right)$  : -1 0 1 2



# Regression Table

Regression Analysis

x \ y	u	30-40	40-50	50-60	60-70	f	fu	fu <sup>2</sup>	fuv
		-1	0	1	2				
30-40	-1	3	0	-1		5	-5	5	2
40-50	0	3	1	1	0	11	0	0	0
50-60	1	2	6	1	2	6	6	6	3
60-70	2	1	2	2	1	3	6	12	6
		0	1	1	1				
f		6	10	5	4	N = 25	Σfu = 7	Σfu <sup>2</sup> = 23	Σfuv = 11
fu		-6	0	5	8	Σfu = 7			
fu <sup>2</sup>		6	0	5	16	Σfu <sup>2</sup> = 27			
fuv		2	0	3	6	Σfuv = 11			

## NOTES

Now  $\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h = 45 + \left(\frac{7}{25}\right)10 = 47.8$

$\bar{y} = B + \left(\frac{\Sigma fv}{N}\right)k = 45 + \left(\frac{7}{25}\right)10 = 47.8$

$b_{yx} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma fu^2 - (\Sigma fu)^2} \cdot \frac{k}{h} = \frac{25(11) - (7)(7)}{25(27) - (7)^2} \cdot \frac{10}{10} = \frac{2260}{6260} = 0.361$

$b_{xy} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma fv^2 - (\Sigma fv)^2} \cdot \frac{h}{k} = \frac{25(11) - (7)(7)}{25(23) - (7)^2} \cdot \frac{10}{10} = \frac{2260}{5260} = 0.429$

The regression line of y on x is  $y - \bar{y} = b_{yx}(x - \bar{x})$

or  $y - 47.8 = 0.361(x - 47.8)$   
 or  $y = 0.361x + 47.8 - (0.361)(47.8)$   
 or  $y = 0.361x + 30.544.$

The regression line of x on y is  $x - \bar{x} = b_{xy}(y - \bar{y})$

or  $x - 47.8 = 0.429(y - 47.8)$   
 or  $x = 0.429y + 47.8 - (0.429)(47.8)$   
 or  $x = 0.429y + 27.294.$

**EXERCISE 8.3****NOTES**

1. The following table gives the frequency, according to groups of marks obtained by 67 students in an intelligence test. Compute the regression lines between the variables age ( $x$ ) and marks ( $y$ ):

Test marks	Age (in years)			
	18	19	20	21
200—250	4	4	2	1
250—300	3	5	4	2
300—350	2	6	8	5
350—400	1	4	6	10

2. Following is the distribution of students according to their height and weight:

Height (in inches)	Weight (in lbs)			
	90—100	100—110	110—120	120—130
50—55	4	7	5	2
55—60	6	10	7	4
60—65	6	12	10	7
65—70	3	8	6	3

Obtain (i) the coefficients of regression and (ii) the regression equations.

**Answers**

1.  $y = 21.5134x - 109.7157$ ,  $x = 0.008y + 17.1727$   
 2. (i)  $b_{yx} = 0.152$ ,  $b_{xy} = 0.041$ , (ii)  $y = 0.152x + 99.93$ ,  $x = 0.041y + 55.88$ .

## 8.9. PROPERTIES OF REGRESSION COEFFICIENTS AND REGRESSION LINES

(i) We have  $b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$  and  $b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$ .

$\sigma_x$  and  $\sigma_y$  are always non-negative.

$\therefore$  The signs of  $b_{yx}$  and  $b_{xy}$  are same as that of  $r$ .

$\therefore$  The signs of regression coefficients and correlation coefficient are same.

Thus  $b_{yx}$ ,  $b_{xy}$  and  $r$  are all either positive or negative.

(ii)  $b_{yx} \cdot b_{xy} = r \cdot \frac{\sigma_y}{\sigma_x} \cdot r \cdot \frac{\sigma_x}{\sigma_y} = r^2$ .

Now  $0 \leq r^2 \leq 1$  because  $-1 \leq r \leq 1$ .

$\therefore 0 \leq b_{yx} b_{xy} \leq 1$ .

$\therefore$  The product of regression coefficients is non-negative and cannot exceed one.

$$(iii) b_{yx} \cdot b_{xy} = r \frac{\sigma_y}{\sigma_x} \cdot r \frac{\sigma_x}{\sigma_y} = r^2$$

$$\therefore r = \pm \sqrt{b_{yx} b_{xy}}.$$

The sign of  $r$  is taken as that of regression coefficients.

$$(iv) \text{ The regression line of } y \text{ on } x \text{ is } y - \bar{y} = b_{yx} (x - \bar{x}).$$

$$\Rightarrow y = b_{yx} x + (\bar{y} - b_{yx} \bar{x})$$

$\therefore$  When  $y$  is kept on the left side, then the coefficient of  $x$  on the right side gives the regression coefficient of  $y$  on  $x$ .

For example, let  $4x + 7y - 9 = 0$  be the regression line of  $y$  on  $x$ .

$$\text{We write this as } y = -\frac{4}{7}x + \frac{9}{7}.$$

$$\therefore \text{ Regression coefficient of } y \text{ on } x = \text{coefficient of } x = -\frac{4}{7}.$$

$$\text{The regression line of } x \text{ on } y \text{ is } x - \bar{x} = b_{xy} (y - \bar{y}).$$

$$\Rightarrow x = b_{xy} y + (\bar{x} - b_{xy} \bar{y})$$

$\therefore$  When  $x$  is kept on the left side, then the coefficient of  $y$  on the right side gives the regression coefficient of  $x$  on  $y$ .

For example, let  $5x + 9y - 8 = 0$  be the regression line of  $x$  on  $y$ .

$$\text{We write this as } x = -\frac{9}{5}y + \frac{8}{5}.$$

$$\therefore \text{ Regression coefficient of } x \text{ on } y = \text{coefficient of } y = -\frac{9}{5}.$$

$$(v) \text{ The regression line of } y \text{ on } x \text{ is } y - \bar{y} = b_{yx} (x - \bar{x}).$$

This equation is satisfied by the point  $(\bar{x}, \bar{y})$ . This point also lies on the regression line of  $x$  on  $y$ :

$$x - \bar{x} = b_{xy} (y - \bar{y})$$

$\therefore$  The point  $(\bar{x}, \bar{y})$  is common to both regression lines. In other words, if the correlation between the variables is not perfect, then the regression lines intersect at  $(\bar{x}, \bar{y})$ .

#### (vi) Angle between the lines of regression

$$\text{The regression line of } y \text{ on } x \text{ is } y - \bar{y} = b_{yx} (x - \bar{x}).$$

$$\Rightarrow y = b_{yx} x + (\bar{y} - b_{yx} \bar{x}) \quad \therefore \text{ Slope} = b_{yx} = m_1 \text{ (say)}$$

$$\text{The regression line of } x \text{ on } y \text{ is } x - \bar{x} = b_{xy} (y - \bar{y}).$$

$$\Rightarrow y = \frac{1}{b_{xy}} x + \left( \bar{y} - \frac{1}{b_{xy}} \bar{x} \right) \quad \therefore \text{ Slope} = \frac{1}{b_{xy}} = m_2 \text{ (say)}$$

Let  $\theta$  be the acute angle between the regression lines.

$$\therefore \tan \theta = \left| \frac{m_1 - m_2}{1 + m_1 m_2} \right| = \left| \frac{b_{yx} - \frac{1}{b_{xy}}}{1 + b_{yx} \cdot \frac{1}{b_{xy}}} \right| = \left| \frac{b_{yx} b_{xy} - 1}{b_{yx} + b_{xy}} \right|$$

## NOTES

## NOTES

$$= \left| \frac{r \frac{\sigma_y}{\sigma_x} \cdot r \frac{\sigma_x}{\sigma_y} - 1}{r \frac{\sigma_x}{\sigma_y} + r \frac{\sigma_y}{\sigma_x}} \right| = \left| \frac{r^2 - 1}{r \left( \frac{\sigma_x^2 + \sigma_y^2}{\sigma_x \sigma_y} \right)} \right|$$

$$= \frac{|r^2 - 1| \sigma_x \sigma_y}{|r| (\sigma_x^2 + \sigma_y^2)} = \frac{(1 - r^2) \sigma_x \sigma_y}{|r| (\sigma_x^2 + \sigma_y^2)}$$

$$\therefore \tan \theta = \frac{(1 - r^2) \sigma_x \sigma_y}{|r| (\sigma_x^2 + \sigma_y^2)}$$

**Particular cases:**

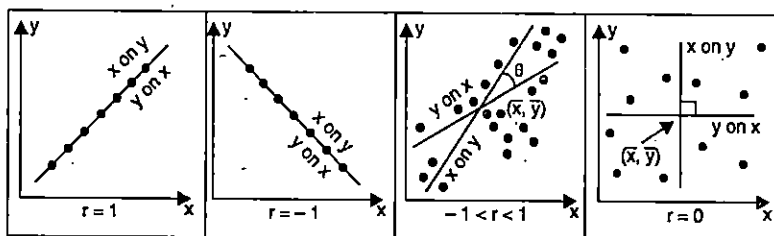
(i)  $r = 0$ . In this case,  $\tan \theta$  is not defined.

$\therefore \theta = 90^\circ$  i.e., the regression lines are *perpendicular* to each other.

(ii)  $r = 1$  (or  $-1$ ). In this case,  $\tan \theta = 0$ .

$\therefore$  The regression lines are *coincident*, because the point  $(\bar{x}, \bar{y})$  is on both the regression lines.

Thus, we see that if the variables are not correlated, then the regression lines are perpendicular to each other and if the variables are perfectly correlated, then the regression lines are coincident. The closeness of regression lines measure the degree of linear correlation between the variables.



**Example 8.17.** Find the mean of the variable  $X$  and  $Y$  and correlation coefficient from the following informations:

Regression equation of  $Y$  on  $X$ :  $2Y - X - 50 = 0$ .

Regression equation of  $X$  on  $Y$ :  $3Y - 2X - 10 = 0$ .

**Solution.** Regression equation of  $Y$  on  $X$  is  $2Y - X - 50 = 0$  ... (1)

Regression equation of  $X$  on  $Y$  is  $3Y - 2X - 10 = 0$  ... (2)

Multiplying (1) by 2, we get  $4Y - 2X - 100 = 0$  ... (3)

Subtracting (3) from (2), we get  $-Y + 90 = 0 \Rightarrow Y = 90$

$\therefore (1) \Rightarrow 2(90) - X - 50 = 0 \Rightarrow X = 130$

$\therefore \bar{X} = 130, \bar{Y} = 90.$

(1)  $\Rightarrow Y = \frac{1}{2}X + 25 \Rightarrow b_{yx} = \frac{1}{2}$

(2)  $\Rightarrow X = \frac{3}{2}Y - 5 \Rightarrow b_{xy} = \frac{3}{2}$

Now  $r = +\sqrt{b_{yx} \cdot b_{xy}}$  (Regression coefficients are +ve)

$$= \sqrt{\frac{1}{2} \times \frac{3}{2}} = \frac{\sqrt{3}}{2} = 0.866.$$



**Example 8.18.** Out of the following two regression lines, find the line of regression of  $y$  on  $x$ :

$$4x - 5y + 33 = 0 \quad \text{and} \quad 20x - 9y - 107 = 0.$$

**Solution.** The regression lines are

$$4x - 5y + 33 = 0 \quad \dots(1)$$

$$20x - 9y - 107 = 0. \quad \dots(2)$$

Let (1) be the regression line of  $y$  on  $x$ .

$\therefore$  (2) is the regression line of  $x$  on  $y$ .

$$(1) \Rightarrow y = \frac{4}{5}x + \frac{33}{5} \quad \therefore b_{yx} = \frac{4}{5}$$

$$(2) \Rightarrow x = \frac{9}{20}y + \frac{107}{20} \quad \therefore b_{xy} = \frac{9}{20}$$

$b_{yx}$  and  $b_{xy}$  are of same signs.

$$\text{Also} \quad b_{yx} \cdot b_{xy} = \frac{4}{5} \times \frac{9}{20} = \frac{36}{100} \leq 1.$$

$\therefore$  Our choice of regression lines is correct.

$\therefore$  Regression line of  $y$  on  $x$  is  $4x - 5y + 33 = 0$ .

**Example 8.19.** Find the regression coefficients  $b_{yx}$  and  $b_{xy}$  when the lines of regression are

$$4x + 3y + 7 = 0 \quad \text{and} \quad 3x + 4y + 8 = 0.$$

**Solution.** The regression lines are

$$4x + 3y + 7 = 0 \quad \dots(1) \quad \text{and} \quad 3x + 4y + 8 = 0 \quad \dots(2)$$

Let (1) be the regression line of  $y$  on  $x$ .

$\therefore$  (2) is the regression line of  $x$  on  $y$ .

$$(1) \Rightarrow y = -\frac{4}{3}x - \frac{7}{3} \quad \therefore b_{yx} = -\frac{4}{3}$$

$$(2) \Rightarrow x = -\frac{4}{3}y - \frac{8}{3} \quad \therefore b_{xy} = -\frac{4}{3}$$

$$b_{yx} \cdot b_{xy} = \left(-\frac{4}{3}\right)\left(-\frac{4}{3}\right) = \frac{16}{9} > 1$$

This is impossible, because  $0 \leq b_{yx} \cdot b_{xy} \leq 1$ .

$\therefore$  Our supposition is wrong.

(1) is the regression line of  $x$  on  $y$  and (2) is the regression line  $y$  on  $x$ .

$$(1) \Rightarrow x = -\frac{3}{4}y - \frac{7}{4}$$

$$\text{and} \quad (2) \Rightarrow y = -\frac{3}{4}x - 2$$

$$\therefore b_{xy} = -\frac{3}{4} \quad \text{and} \quad b_{yx} = -\frac{3}{4}$$

**Example 8.20.** The equations of regression lines are given by  $4x - 5y + 35 = 0$  and  $5x - 2y = 20$ .

Also, variance of  $x$ -series is 9. Find :

- (i) mean values of  $x$  and  $y$  variables
- (ii) correlation coefficient between  $x$  and  $y$  variables
- (iii) standard deviation of  $y$  series.

## NOTES

## NOTES

**Solution.** The equations of regression lines are

$$4x - 5y + 35 = 0 \quad \dots(1)$$

$$5x - 2y - 20 = 0 \quad \dots(2)$$

$$(i) (1) \times 2 \Rightarrow 5x - 10y + 70 = 0 \quad \dots(3)$$

$$(2) \times 5 \Rightarrow 25x - 10y - 100 = 0 \quad \dots(4)$$

$$(3) - (4) \Rightarrow -17x + 170 = 0 \Rightarrow x = 10$$

$$\therefore (1) \Rightarrow 4(10) - 5y + 35 = 0$$

$$\Rightarrow 5y = 40 + 35 = 75 \Rightarrow y = 15$$

$\therefore (10, 15)$  is on both lines.

$$\therefore \bar{x} = 10, \bar{y} = 15.$$

(ii) Let (1) be the regression line of  $y$  on  $x$ .

$\therefore$  (2) is the regression line of  $x$  on  $y$ .

$$(1) \Rightarrow y = \frac{4}{5}x + 7 \quad \therefore b_{yx} = \frac{4}{5}$$

$$(2) \Rightarrow x = \frac{2}{5}y + 4 \quad \therefore b_{xy} = \frac{2}{5}$$

$\therefore b_{yx}$  and  $b_{xy}$  are of same sign.

$$\text{Also } b_{yx} \cdot b_{xy} = \left(\frac{4}{5}\right)\left(\frac{2}{5}\right) = \frac{8}{25} \leq 1$$

$\therefore$  Our choice of regression lines is correct.

$$\therefore b_{yx} = \frac{4}{5} \text{ and } b_{xy} = \frac{2}{5}$$

$$\therefore r = +\sqrt{b_{yx} \cdot b_{xy}} = +\sqrt{\frac{4}{5} \times \frac{2}{5}} = \frac{2\sqrt{2}}{5} = 0.5657.$$

$$(iii) b_{yx} = \frac{4}{5} \Rightarrow r \frac{\sigma_y}{\sigma_x} = \frac{4}{5}$$

$$\Rightarrow \frac{2\sqrt{2}}{5} \cdot \frac{\sigma_y}{3} = \frac{4}{5} \quad (\because \text{var. } x = 9 \Rightarrow \sigma_x = \sqrt{9} = 3)$$

$$\Rightarrow \sigma_y = \frac{4 \times 3}{2\sqrt{2}} = 2\sqrt{2} = 4.2426.$$

**Example 8.21.** Equations of two regression lines are:

$$4x + 3y + 7 = 0 \quad \text{and} \quad 3x + 4y + 8 = 0$$

Find:

(i) mean of  $x$ , mean of  $y$ ;

(ii) regression coefficients  $b_{yx}$  and  $b_{xy}$  and

(iii) correlation coefficient between  $x$  and  $y$ .

**Solution.** The equations of regression lines are:

$$4x + 3y + 7 = 0 \quad \dots(1)$$

$$\text{and} \quad 3x + 4y + 8 = 0 \quad \dots(2)$$

$$(i) (1) \times 4 \Rightarrow 16x + 12y + 28 = 0 \quad \dots(3)$$

$$(2) \times 3 \Rightarrow 9x + 12y + 24 = 0 \quad \dots(4)$$

$$(3) - (4) \Rightarrow 7x + 4 = 0 \Rightarrow x = -4/7.$$



$$\therefore (1) \Rightarrow 4(-4/7) + 3y + 7 = 0$$

$$\Rightarrow y = \frac{-7 + (16/7)}{3} = -\frac{11}{7}$$

$\therefore (-4/7, -11/7)$  is on both lines.

Since regression lines intersect at  $(\bar{x}, \bar{y})$ , we have

$$\bar{x} = -4/7, \bar{y} = -11/7.$$

(ii) Let (1) be the regression line of  $y$  on  $x$ .

$\therefore$  (2) is the regression line of  $x$  on  $y$ .

$$(1) \Rightarrow y = -\frac{4}{3}x - \frac{7}{3} \quad \therefore b_{yx} = -\frac{4}{3}$$

$$(2) \Rightarrow x = -\frac{4}{3}y - \frac{8}{3} \quad \therefore b_{xy} = -\frac{4}{3}$$

$$\therefore b_{yx} \cdot b_{xy} = \left(-\frac{4}{3}\right)\left(-\frac{4}{3}\right) = \frac{16}{9} > 1$$

This is impossible, because  $0 \leq b_{yx} \cdot b_{xy} \leq 1$ .

$\therefore$  Our supposition is wrong.

$\therefore$  (1) is the regression line of  $x$  on  $y$  and (2) is the regression line of  $y$  on  $x$ .

$$(1) \Rightarrow x = -\frac{3}{4}y - \frac{7}{4} \quad \therefore b_{xy} = -\frac{3}{4}$$

$$(2) \Rightarrow y = -\frac{3}{4}x - 2 \quad \therefore b_{yx} = -\frac{3}{4}$$

$$(iii) \quad r = -\sqrt{b_{yx}b_{xy}} = -\sqrt{\left(-\frac{3}{4}\right)\left(-\frac{3}{4}\right)} = -\frac{3}{4}$$

#### EXERCISE 8.4

- In a problem of regression analysis, the two regression coefficients are found to be  $-0.6$  and  $-1.4$ . What is the correlation coefficient?
- Out of the following two regression lines, find the regression line of  $x$  on  $y$ :
  - $13x - 10y + 11 = 0$ ,  $2x - y - 1 = 0$
  - $x + 4y + 11 = 0$ ,  $4x + y - 7 = 0$ .
- Find the correlation coefficient when the lines of regression are
 
$$2x - 9y + 6 = 0, x - 2y + 1 = 0.$$
- The equations of two regression lines obtained in a correlation analysis are as follows:
 
$$3x + 13y = 19, x + 3y = 5.$$
 Obtain (i) the means of  $x$  and  $y$   
 (ii) the regression coefficients  $b_{yx}$  and  $b_{xy}$   
 (iii) the correlation coefficient.
- In a partially destroyed record, the following data are available:  
 Variance of  $x = 25$ .  
 Regression equation of  $x$  on  $y$ :  $5x - y = 22$ .  
 Regression equation of  $y$  on  $x$ :  $64x - 45y = 24$ .  
 Find:
  - Mean values of  $x$  and  $y$ .
  - Coefficient of correlation between  $x$  and  $y$
  - Standard deviation of  $y$ .

#### NOTES

## NOTES

6. The regression equations of a bivariate distribution are:

Regression equation of  $y$  on  $x$ :  $4y = 9x + 15$

Regression equation of  $x$  on  $y$ :  $25x = 6y + 7$ .

Find:

- (i) Coefficient of correlation.
  - (ii) The ratio of means of  $x$  and  $y$ .
  - (iii) The ratio of S.D. of  $x$  and  $y$ .
7. In a partially destroyed laboratory record of an analysis of correlation data, the following results are legible:
- Variance of  $x = 9$
- Regression equations:  $4x - 5y + 33 = 0$   
 $20x - 9y = 107$ .
- On the basis of above information, find:
- (i) the mean values of  $x$  and  $y$ .
  - (ii) standard deviation of  $y$ -series, and
  - (iii) the coefficient of correlation.
8. The equations of regression lines are given to be  $3x + 2y - 26 = 0$  and  $6x + y - 31 = 0$ . Find the values of  $\bar{x}$ ,  $\bar{y}$  and  $r$ .

## Answers

1.  $-0.9165$
2. (i)  $2x - y - 1 = 0$  (ii)  $4x + y - 7 = 0$  3.  $r = 0.6667$
4. (i)  $\bar{x} = 2, \bar{y} = 1$  (ii)  $b_{yx} = -\frac{3}{13}, b_{xy} = -3$  (iii)  $r = -\frac{3}{\sqrt{13}}$
5. (i)  $\bar{x} = 6, \bar{y} = 8$  (ii)  $r = 8/15$  (iii)  $\sigma_y = 40/3$
6. (i)  $r = 0.7348$  (ii)  $\bar{x}/\bar{y} = 59/219$  (iii)  $\sigma_x/\sigma_y = \sqrt{8}/\sqrt{75}$
7. (i)  $\bar{x} = 13, \bar{y} = 17$  (ii)  $\sigma_y = 4$ , (iii)  $r = 0.6$
8.  $\bar{x} = 4, \bar{y} = 7, r = -0.5$ .

## 8.10. SUMMARY

- In statistics, *regression analysis* is concerned with the measure of average relationship between variables. Here we shall deal with the derivation of appropriate functional relationships between variables. Regression explains the nature of relationship between variables.
- There are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable* (or *regressed variable* or *predicted variable* or *explained variable*). The variable which influences the value of dependent variable is called *independent variable* (or *regressor* or *predictor* or *explanator*).
- If there are only two variables under consideration, then the regression is called **simple regression**. If there are more than two variables under consideration then the regression is called **multiple regression**. The regression is called **partial regression** if there are more than two variables under consideration and relation between only two variables is established after excluding the effect of other variables. The simple regression is called **linear regression** if the

point on the scatter diagram of variables lies almost along a line otherwise it is termed as **non-linear regression** or **curvilinear regression**.

## 8.11. REVIEW EXERCISES

1. What are regression coefficients? Show that  $r^2 = b_{yx} \cdot b_{xy}$ .
2. Point out the role of regression analysis in business with the help of few examples.
3. What is regression? Why are there, in general two regression lines? Under what conditions can there be only one regression line?
3. What is regression analysis? Explain its use in business problems with suitable examples.
4. What do you mean by regression coefficients? What are the uses of regression analysis?

## NOTES

## NOTES

## 9. PROBABILITY

### STRUCTURE

- 9.1. Introduction
- 9.2. Random Experiment
- 9.3. Sample Space
- 9.4. Tree Diagram
- 9.5. Event
- 9.6. Algebra of Events
- 9.7. Equality Likely Outcomes
- 9.8. Exhaustive Outcomes
- 9.9. Three Approaches of Probability
- 9.10. Classical Approach of Probability
- 9.11. 'Odds In Favour' and 'Odds Against' an Event
- 9.12. Mutually Exclusive Events
- 9.13. Addition Theorem (For Mutually Exclusive Events)
- 9.14. Addition Theorem (General)
- 9.15. Conditional Probability
- 9.16. Independent Events
- 9.17. Dependent Events
- 9.18. Independent Experiments
- 9.19. Multiplication Theorem
- 9.20. Total Probability Rule
- I. Baye's Theorem**
- 9.21. Motivation
- 9.22. Criticism of Classical Approach of Probability
- 9.23. Empirical Approach of Probability
- 9.24. Subjective Approach of Probability
- 9.25. Summary
- 9.26. Review Exercises

### 9.1. INTRODUCTION

The words 'Probability' and 'Chance' are quite familiar to everyone. Many a times, we come across statements like, "Probably it may rain today", "Chances of his visit to the university are very few", "It is possible that he may pass the examination with good marks". In the above statements, the words probably, chance, possible, etc. convey the



sense of uncertainty about the occurrence of some event. Ordinarily, it may appear that there cannot be any exact measurement for these uncertainties, but in Statistics, we do have methods for calculating the degree of certainty of events in numerical value, provided certain conditions are satisfied.

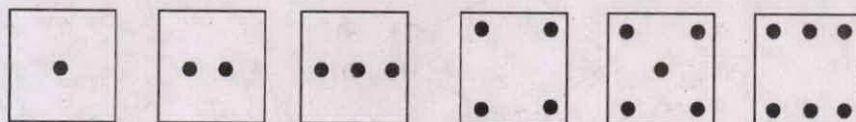
## NOTES

### 9.2. RANDOM EXPERIMENT

When we perform experiments in science and engineering, repeatedly under very nearly identical conditions, we get almost the same result. Such experiments are called **deterministic experiments**.

There also exist experiments in which the results may not be essentially the same even if the experiment is performed under very nearly identical conditions. Such experiments are called **random experiments**. If we toss a coin, we may get 'head' or 'tail'. This is a random experiment. Throwing of a die is also a random experiment as any of the six faces of the die may come up. In this experiment, there are six possibilities (1 or 2 or 3 or 4 or 5 or 6).

**Remark 1.** A *die* is a small cube used in gambling. On its six faces, dots are marked as:



Numbers on a die

Plural of the word die is *dice*. The outcome of throwing a die is the number of dots on its upper most face.

**Remark 2.** A *pack of cards* consists of four suits called *Spades*, *Hearts*, *Diamonds* and *Clubs*. Each suit consists of 13 cards, of which nine cards are numbered from 2 to 10, an ace, a king, a queen and a jack (or knave). Spades and clubs are black faced cards, while hearts and diamonds are red faced cards. The kings, queens and jacks are called *face cards*.

### 9.3. SAMPLE SPACE

The **sample space** of a random experiment is defined as the set of all possible outcomes of the experiment. The possible outcomes are called **sample points**. The sample space is generally denoted by the letter *S*. We list the sample space of some random experiments:

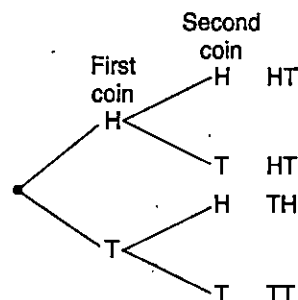
Random Experiment	Sample Space
1. Throwing of a fair die	$S = \{1, 2, 3, 4, 5, 6\}$
2. Tossing of an unbiased coin	$S = \{H, T\}$
3. Tossing of two unbiased coins	$S = \{HH, HT, TH, TT\}$
4. A family of two children	$S = \{BB, BG, GB, GG\}$

### 9.4. TREE DIAGRAM

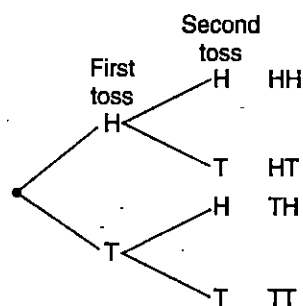
A **Tree diagram** is a device used to enumerate all the logical possibilities of a sequence of steps where each step can occur in a finite number of ways. A tree diagram is constructed from left to right and the number of branches at any point corresponds to the number of ways the next step can occur.

**Illustration 1.** The tree diagram of the sample space of the toss of two coins is shown in the figure.

## NOTES



**Illustration 2.** The tree diagram of the sample space of the two tosses of a coin is shown in the figure.



## 9.5. EVENT

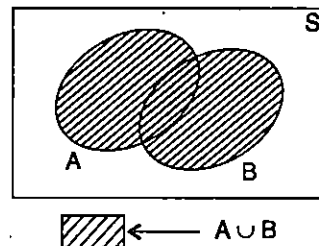
An **event** is defined as a subset of the sample space. An event is called an **elementary (or simple) event** if it contains only one sample point. In the experiment of rolling a die, the event A of getting '3' is a simple event. We write  $A = \{3\}$ . An event is called an **impossible event** if it can never occur. In the above example, the event  $B = \{7\}$  of getting '7' is an impossible event. On the other hand, an event which is sure to occur is called a **certain event**. In the above example of rolling a die, the event C of getting a number less than 7 is a certain event.

In the throwing of two dice, the cases favourable to getting sum 7 are 6 viz. (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) and (6, 1).

## 9.6. ALGEBRA OF EVENTS

We know that the events of a random experiment are sets, being subsets of the sample space. Thus, we can use set operations to form new events.

Let A and B be any two events associated with a random experiment.





## NOTES

The event of occurrence of either A or B or both is written as 'A or B' and is denoted by the subset  $A \cup B$  of the sample space. The event of occurrence of both A and B is written as 'A and B' and is denoted by the subset  $A \cap B$  of the sample space. For simplicity, the event  $A \cap B$  is also denoted by 'AB'.

The event of non-occurrence of event A is written as 'not A' and is denoted by the subset  $A'$ , which is the complement of set A. The event  $A'$  is called the **complementary event** of the event A.

**Illustration.** Let a die be tossed.

$$\therefore S = \{1, 2, 3, 4, 5, 6\}$$

Let A = event of getting an even number

and B = event of getting a number less than 5

$$\therefore A = \{2, 4, 6\} \text{ and } B = \{1, 2, 3, 4\}$$

Here

$A \cup B$  = event of occurrence of either A or B or both  
 $= \{1, 2, 3, 4, 6\}$  (See figure)

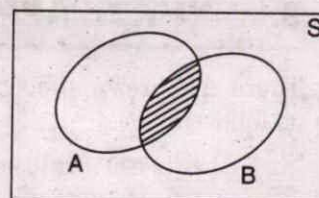
$A \cap B$  = event of occurrence of both A and B  
 $= \{2, 4\}$  (See figure)

Also,  $A'$  = event of non-occurrence of

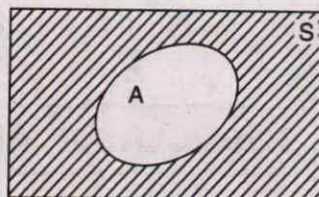
$$A = \{1, 3, 5\}$$

and  $B'$  = event of non-occurrence of

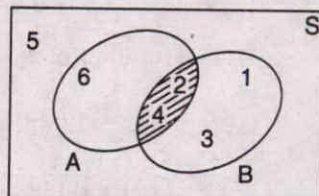
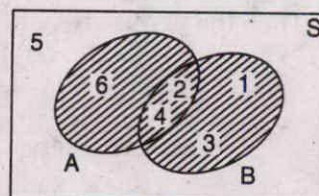
$$B = \{5, 6\}.$$



$\leftarrow A \cap B$



$\leftarrow A'$



## 9.7. EQUALITY LIKELY OUTCOMES

The outcomes of a random experiment are called **equally likely**, if all of these have equal preferences. In the experiment of tossing a unbiased coin, the outcomes 'Head' and 'Tail' are equally likely.

In our discussion, we shall always assume the outcomes of a random experiment to be equally likely.

## 9.8. EXHAUSTIVE OUTCOMES

The outcomes of a random experiment are called **exhaustive**, if these cover all the possible outcomes of the experiment. In the experiment of rolling a die, the outcomes 1, 2, 3, 4, 5, 6 are exhaustive.

## NOTES

## 9.9. THREE APPROACHES OF PROBABILITY

There are three approaches of discussing probability of events. These approaches are as follows:

1. Classical approach
2. Empirical approach
3. Subjective approach

We shall first discuss classical approach of probability.

## 9.10. CLASSICAL APPROACH OF PROBABILITY

Suppose in a random experiment, there are  $n$  exhaustive, equally likely outcomes. Let  $A$  be an event and there are  $m$  outcomes (cases) favourable to the happening of it. Then the **probability**  $P(A)$  of the happening of the event  $A$  is defined as

$$P(A) = \frac{\text{Total no. of cases favourable to the happening of } A}{\text{Total no. of exhaustive, equally likely cases}} = \frac{m}{n}$$

It may be observed from this definition, that  $0 \leq m \leq n$ .

$$\therefore 0 \leq \frac{m}{n} \leq 1 \quad \text{or} \quad 0 \leq P(A) \leq 1.$$

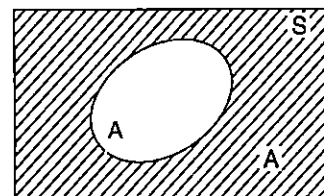
The number of cases favourable to the non-happening of the event  $A$  is  $n - m$ .

$$\therefore P(\text{not } A) = \frac{n - m}{n} = \frac{n}{n} - \frac{m}{n} = 1 - \frac{m}{n} = 1 - P(A).$$

$$\therefore P(A) + P(\text{not } A) = 1, \text{ i.e. } P(A) + P(\bar{A}) = 1.$$

If  $A$  is a *sure event*, the  $P(A) = \frac{n}{n} = 1$  and if  $A$  happen to be an *impossible event*, then

$$P(A) = \frac{0}{n} = 0.$$



## 9.11. 'ODDS IN FAVOUR' AND 'ODDS AGAINST' AN EVENT

The ratio of cases in favour of  $A$  and cases against  $A$  is called the **odds in favour** of  $A$ . Similarly, the ratio of cases against  $A$  and cases in favour of  $A$  is called the **odds against**  $A$ .

If odds in favour of  $A$  are  $m : n$ , then (i) odds against  $A$  are  $n : m$  and (ii) probability of  $A = \frac{m}{m + n}$ .

If odds against  $A$  are  $m : n$ , then (i) odds in favour of  $A$  are  $n : m$  and (ii) probability of  $A = \frac{n}{m + n}$ .



**Illustration.** Let  $P(A) = \frac{3}{5}$ .

Let total number of cases be  $5\lambda$ .

$\therefore P(A) = \frac{3\lambda}{5\lambda}$  and this implies that cases in favour of A are  $3\lambda$  and cases against

A are  $5\lambda - 3\lambda = 2\lambda$ .

$\therefore$  odds in favour of A are  $3\lambda : 2\lambda$  or  $3 : 2$  and odds against A are  $2\lambda : 3\lambda$  or  $2 : 3$ .

**Example 9.1.** Find the probability of getting the sum 10 in a single throw of two dice.

**Solution.** Here  $S = \{(1, 1), (1, 2), (1, 3), \dots, (6, 5), (6, 6)\}$ .

$\therefore$  No. of possible outcomes  $= 6 \times 6 = 36$ .

Let A be the event of getting sum 10.

$\therefore A = \{(4, 6), (5, 5), (6, 4)\}$

$$P(A) = \frac{3}{36} = \frac{1}{12}.$$

**Remark.** In the above experiment, the sample point  $(a, b)$  means that 'a' is on the first die and 'b' is on the second die.

**Example 9.2.** Find the probability of getting sum 10 in two throws of a die.

**Solution.** Here  $S = \{(1, 1), (1, 2), (1, 3), \dots, (6, 5), (6, 6)\}$

$\therefore$  No. of possible outcomes  $= 6 \times 6 = 36$

Let A be the event of getting sum 10

$\therefore A = \{(4, 6), (5, 5), (6, 4)\}$

$$P(A) = \frac{3}{36} = \frac{1}{12}.$$

**Remark.** In the above experiment, the sample point  $(a, b)$  means that 'a' occurred in the first toss and 'b' occurred in the second toss.

**Example 9.3.** From a bag containing 4 red and 5 green balls, a ball is drawn at random. What is the probability that it is a red ball?

**Solution.** Total no. of balls  $= 4 + 5 = 9$

No. of red balls  $= 4$

$\therefore$  Prob. of getting a red ball  $= \frac{\text{Total no. of red ball}}{\text{Total no. of balls}} = \frac{4}{9}$ .

**Example 9.4.** Three coins are tossed. Find the probability of getting at least two heads.

**Solution.** Let S be the sample space.

$\therefore S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$

Let A be the event of getting at least two heads.

$\therefore A = \{HHH, HHT, HTH, THH\}$

$\therefore P(\text{at least two heads})$

$$= P(A) = \frac{\text{No. of cases favourable to A}}{\text{Total no. of exhaustive, equally likely cases}} = \frac{4}{8} = \frac{1}{2}.$$

**Example 9.5.** Find the probability of getting a 'King' or a 'Queen' in a single draw from a well-shuffled pack of playing cards.

**Solution.** Let A be the event of getting a king or a queen in the draw.

$\therefore$  No. of favourable cases for the happening of the event  $A = 4 + 4 = 8$

## NOTES

$$\text{Total no. of cases} = 52$$

$$\therefore P(\text{King or Queen}) = P(A) = \frac{8}{52} = \frac{2}{13}$$

**NOTES**

**Example 9.6.** Two unbiased dice are thrown. Find the probability that the total of the numbers on the dice is greater than 8.

**Solution.** Let S be the sample space.

$$\therefore S = \{(1, 1), (1, 2), (1, 3), \dots, (6, 5), (6, 6)\}$$

There are  $6 \times 6 = 36$  exhaustive, equally likely outcomes.

Let A be the event of getting total greater than 8.

$$\therefore A = \{(3, 6), (4, 5), (5, 4), (6, 3), (4, 6), (5, 5), (6, 4), (5, 6), (6, 5), (6, 6)\}$$

$$\therefore P(\text{sum} > 8) = P(A) = \frac{\text{No. of cases favourable to A}}{\text{Total No. of cases}} = \frac{10}{36} = \frac{5}{18}$$

**Example 9.7.** Three unbiased dice are thrown simultaneously. Find the probability of getting (i) sum not greater than 5, (ii) sum at least 15, (iii) sum equal to 8.

**Solution.** Let S be the sample space.

$$\therefore S = \{(1, 1, 1), (1, 1, 2), \dots, (6, 6, 5), (6, 6, 6)\}$$

There are  $6 \times 6 \times 6 = 216$  exhaustive equally likely outcomes.

(i) Let A = event that sum is not greater than 5.

$$\therefore A = \{(1, 1, 1), (1, 1, 2), (1, 2, 1), (2, 1, 1), (1, 1, 3), (1, 2, 2), (1, 3, 1), (2, 1, 2), (2, 2, 1), (3, 1, 1)\}$$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{10}{216} = \frac{5}{108}$$

(ii) Let B = event that sum is at least 15

$$\therefore B = \{(3, 6, 6), (4, 5, 6), (4, 6, 5), (5, 4, 6), (5, 5, 5), (5, 6, 4), (6, 3, 6), (6, 4, 5), (6, 5, 4), (6, 6, 3), (4, 6, 6), (5, 5, 6), (5, 6, 5), (6, 4, 6), (6, 5, 5), (6, 6, 4), (5, 6, 6), (6, 5, 6), (6, 6, 5), (6, 6, 6)\}$$

$$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{20}{216} = \frac{5}{54}$$

(iii) Let C = event of getting sum 8

$$\therefore C = \{(1, 1, 6), (1, 2, 5), (1, 3, 4), (1, 4, 3), (1, 5, 2), (1, 6, 1), (2, 1, 5), (2, 2, 4), (2, 3, 3), (2, 4, 2), (2, 5, 1), (3, 1, 4), (3, 2, 3), (3, 3, 2), (3, 4, 1), (4, 1, 3), (4, 2, 2), (4, 3, 1), (5, 1, 2), (5, 2, 1), (6, 1, 1)\}$$

$$\therefore P(C) = \frac{n(C)}{n(S)} = \frac{21}{216} = \frac{7}{72}$$

**Example 9.8.** Find the probability that in a random arrangement of the letters of the word DAUGHTER, the letter D occupies the first place.

**Solution.** The word DAUGHTER contains 8 letters and all are different.

$$\therefore \text{Total no. of possible arrangements} = 8! = 40320$$

Let A be the event of getting a word with D at the first place.



∴ Favourable cases to the event A

$$= 1 \times \text{no. of ways of arranging 7 letters (except D).}$$

$$= 1 \times 7! = 5040$$

∴ P(D occupies first place)

$$P(A) = \frac{5040}{40320} = \frac{1}{8}.$$

**Example 9.9.** Find the probability that in a random arrangement of letters of the word **MATHEMATICS**, the consonants occur together.

**Solution.** The word **MATHEMATICS**, contains 2 M's, 2 A's, 2 T's, 1 H, 1 E, 1 I, 1 C and 1 S.

∴ Total no. of exhaustive cases

$$= \frac{11!}{2!2!2!} = 4989600$$

For finding the no. of favourable cases to the event under consideration, we shall consider all consonants **M, T, H, M, T, C, S** as one block. So in arranging 2 A's, 1 E's, 1 I's, 1 (**MTHMTCS**), all the consonants will occur together. The consonants **MTHMTCS** can arrange themselves in

$$\frac{7!}{2!2!} = 1260 \text{ ways}$$

$$\therefore \text{No. of favourable cases} = \frac{5!}{2!1!1!1!} \times 1260 = 75600$$

$$\therefore P(\text{consonants are together}) = \frac{75600}{4989600} = 0.01515.$$

**Example 9.10.** A bag contains 10 red and 8 black balls. Two balls are drawn at random. Find the probability that:

(i) both balls are red.

(ii) one ball is red and the other is black.

**Solution.** (i) Total number of balls = 10 + 8 = 18

$$(ii) P(\text{both red balls}) = \frac{\text{No. of selections of 2 out of 10 red balls}}{\text{No. of selections of 2 out of 18 balls}}$$

$$= \frac{{}^{10}C_2}{{}^{18}C_2} = \frac{10 \times 9}{1 \times 2} + \frac{18 \times 17}{1 \times 2} = \frac{10 \times 9}{18 \times 17} = \frac{5}{17}.$$

(ii) P(one red ball and one black ball)

$$= \frac{\left( \begin{smallmatrix} \text{No. of selections of} \\ 1 \text{ out of 10 red balls} \end{smallmatrix} \right) \left( \begin{smallmatrix} \text{No. of selections of} \\ 1 \text{ out of 8 black balls} \end{smallmatrix} \right)}{\text{No. of selections of 2 out of 18 balls.}}$$

$$= \frac{{}^{10}C_1 \times {}^8C_1}{{}^{18}C_2}$$

$$= (10 \times 8) + \frac{18 \times 17}{2} = \frac{10 \times 8}{9 \times 17} = \frac{80}{153}.$$

## NOTES

**Example 9.11.** From a group of 8 children (5 boys and 3 girls) three children are selected at random. Calculate the probabilities that the selected group contains:

(i) no girl

(ii) only one girl

(iii) only one particular girl.

## NOTES

**Solution.** No. of girls = 3

No. of boys = 5

Let  $A_0, A_1, A_2$  and  $A_3$  be the events that there are no girl, one girl, two girls and three girls in the random selection of three children.

$$(i) \text{ Required probability} = P(A_0) = \frac{{}^5C_3}{{}^8C_3} = \frac{5}{28}.$$

$$(ii) \text{ Required probability} = P(A_1) = \frac{{}^3C_1 \times {}^5C_2}{{}^8C_3} = \frac{15}{28}.$$

(iii) The particular girl can be selected in only one way. The remaining two children in the selection must be boys.

$$\therefore \text{ Required probability} = \frac{1 \times {}^5C_2}{{}^8C_3} = \frac{5}{28}.$$

## EXERCISE 9.1

- Find the probability of getting an odd number in a single throw of a fair die.
- A single letter is selected at random from the word "PROBABILITY". What is the probability that it is a vowel?
- Find the probability of getting at most two heads in a single toss of three coins.
- From a pack of 52 playing cards, one card is drawn at random. What is the probability that it will be a queen of club or king of diamond?
- Two unbiased dice are thrown. Find the probability that the total of the numbers on the dice is 8.
- Tickets are numbered from 1 to 100. These are well-shuffled and a ticket is drawn at random. What is the probability that the drawn ticket has a number which is greater than 75?
- The following table gives a distribution of marks:

Marks	No. of students
0—10	7
10—20	10
20—30	15
30—40	28
40—50	23
50—60	45
60—70	12
70—80	7
80—90	2
90—100	1

An individual is taken at random from the above group. Find the probability that:

- his marks are below 50
- his marks are below 70.
- his marks are between 30 and 60
- his marks are above 70.



8. Find the probability that in a random arrangement of the letters of the word **VOWEL**, the letter **V** occupies the first place.
9. Find the probability that in a random arrangement of letters the word **BHARAT**, the two **A** occupies the first two places.
10. Find the probability that in a random arrangement of the letters of the word **STATISTICS**, the three **T** are in the beginning.
11. Find the probability that in a random arrangement of the letters of the word **STATISTICS**, the three **T** are together.
12. Four cards are drawn from a pack of playing cards. Find the probability that none is a king.
13. A bag contains 6 white, 4 red and 10 black balls. Two balls are drawn at random. Find the probability that both balls are black.
14. A bag contains 7 white, 5 red and 8 black balls. Two balls are drawn at random. Calculate the probability that none is white.

**NOTES****Answers**

- |  |  |                     |                   |
|--|--|---------------------|-------------------|
| 1. $\frac{1}{2}$                             | 2. $\frac{4}{11}$                                  | 3. $\frac{7}{8}$    | 4. $\frac{1}{26}$ |
| 5. $\frac{5}{36}$                            | 6. $\frac{1}{4}$                                   |                     |                   |
| 7. (i) 0.5533                                | (ii) 0.9333  | (iii) 0.64          | (iv) 0.0667       |
| 8. 0.2                                       | 9. 0.0667  | 10. 0.0083          | 11. 0.0667        |
| 12. $\frac{{}^{48}C_4}{{}^{52}C_4} = 0.7187$ | 13. $\frac{{}^{10}C_2}{{}^{20}C_2} = \frac{9}{38}$ | 14. $\frac{39}{95}$ |                   |

**9.12. MUTUALLY EXCLUSIVE EVENTS**

Two events associated with a random experiment are said to be **mutually exclusive** if both cannot occur together in the same trial. In the experiment of throwing a die, the events  $A = \{1, 4\}$  and  $B = \{2, 5, 6\}$  are mutually exclusive events. In the same experiment, the events  $A = \{1, 4\}$  and  $C = \{2, 4, 5, 6\}$  are not mutually exclusive because if 4 appear on the die, then it is favourable to both events  $A$  and  $C$ . The definition of mutually exclusive events can also be extended to more than two events. We say that more than two events are mutually exclusive if the happening of one of these rules out the happening of all other events. The events  $A = \{1, 2\}$ ,  $B = \{3\}$  and  $C = \{6\}$ , are mutually exclusive in connection with the experiment of throwing a single die.

**9.13. ADDITION THEOREM (FOR MUTUALLY EXCLUSIVE EVENTS)**

If  $A$  and  $B$  are two mutually exclusive events associated with a random experiment, then

$$P(A \text{ or } B) = P(A) + P(B).$$

**Proof.** Let  $n$  be the total number of exhaustive, equally like cases of the experiment.

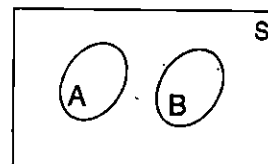
Let  $m_1$  and  $m_2$  be the number of cases favourable to the happening of the events A and B respectively.

## NOTES

$$P(A) = \frac{m_1}{n}$$

and

$$P(B) = \frac{m_2}{n}$$



Since the events are given to be mutually exclusive, therefore there cannot be any sample point common to both events A and B.

$\therefore$  The event A or B can happen in exactly  $m_1 + m_2$  ways.

$$\therefore P(A \text{ or } B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B).$$

Hence,  $P(A \text{ or } B) = P(A) + P(B)$ .

This theorem can also be extended to even more than two events.

If  $A_1, A_2, \dots, A_k$  are m.e. events, then

$$P(A_1 \text{ or } A_2 \text{ or } \dots \text{ or } A_k) = P(A_1) + P(A_2) + \dots + P(A_k).$$

**Example 9.12.** A box contains 4 red balls, 4 green balls and 7 white balls. What is the probability that a ball drawn is either red or white?

**Solution.** Total no. of balls =  $4 + 4 + 7 = 15$

Let A = event of drawing a red ball

B = event of drawing a white ball.

4 Red
4 Green
7 White

The events A and B are m.e. because a ball cannot be both red and white.

$$P(A) = \frac{\text{No. of red balls}}{\text{Total no. of balls}} = \frac{4}{15}$$

$$P(B) = \frac{\text{No. of white balls}}{\text{Total no. of balls}} = \frac{7}{15}$$

Now A or B is the event of drawing either a red ball or a white ball. By addition theorem, the required probability

$$P(A \text{ or } B) = P(A) + P(B) = \frac{4}{15} + \frac{7}{15} = \frac{11}{15}$$

**Example 9.13.** In a single throw of 2 dice, determine the probability of getting total 7 or 11.

**Solution.** Here  $S = \{(1, 1), (1, 2), \dots, (6, 5), (6, 6)\}$ .

Let A and B be the events of getting total 7 and 11 respectively.

$$\therefore A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$$

and

$$B = \{(5, 6), (6, 5)\}$$

The events A and B are mutually exclusive and

$$P(A) = \frac{6}{36} = \frac{1}{6} \quad \text{and} \quad P(B) = \frac{2}{36} = \frac{1}{18}$$

$\therefore$  By addition theorem,

$$P(\text{total is 7 or 9}) = P(A \cup B) = P(A) + P(B)$$

$$= \frac{1}{6} + \frac{1}{18} = \frac{3+1}{18} = \frac{2}{9}$$



**Example 9.14.** In a single throw of two dice, find the probability that neither a doublet nor a total of 9 will appear.

**Solution.** Here  $S = \{(1, 1), (1, 2), \dots, (6, 5), (6, 6)\}$ .

$\therefore$  Number of possible outcomes  $= 6 \times 6 = 36$ .

Let  $E$  = event that doublet is occurred

and  $F$  = event that sum is 9.

$\therefore E = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$

and  $F = \{(3, 6), (4, 5), (5, 4), (6, 3)\}$ .

$\therefore P(E) = \frac{6}{36} = \frac{1}{6}$  and  $P(F) = \frac{4}{36} = \frac{1}{9}$ .

$P(\text{neither a doublet nor a total of 9})$

$$= P(E^c \cap F^c) = P((E \cup F)^c) = 1 - P(E \cup F) \quad \dots(1)$$

The events  $E$  and  $F$  are *m.e.*

$\therefore$  By addition theorem,

$$P(E \cup F) = P(E) + P(F) = \frac{1}{6} + \frac{1}{9} = \frac{5}{18}$$

$$\therefore (1) \Rightarrow P(E^c \cap F^c) = 1 - \frac{5}{18} = \frac{13}{18}.$$

### EXERCISE 9.2

1. A and B are mutually exclusive events for which  $P(A) = 0.3$ ,  $P(B) = p$  and  $P(A \cup B) = 0.5$ . Find the value of  $p$ .
2. Find the probability that a card drawn from a pack of playing cards is either a 'queen' or a 'king'.
3. A and B are two mutually exclusive events of an experiment. If  $P(\text{'not } A) = 0.65$ ,  $P(A \cup B) = 0.65$  and  $P(B) = p$ , find the value of  $p$ .  
[Hint. Use  $P(A \cup B) = (1 - P(\text{'not } A)) + P(B)$ .]
4. From a set of 17 cards, numbered 1, 2, 3, ..., 16, 17, one is drawn at random. Show that the chance that its number is divisible by 3 or 7 is  $7/17$ .
5. Find the probability of getting the sum 9 or 11 in a single throw of two dice.
6. In a single throw of three dice, find the probability of getting a total of 17 or 18.

### Answers

- |                   |                   |        |                  |
|-------------------|-------------------|--------|------------------|
| 1. 0.2            | 2. $\frac{2}{13}$ | 3. 0.3 | 5. $\frac{1}{6}$ |
| 6. $\frac{1}{54}$ |                   |        |                  |

## 9.14. ADDITION THEOREM (GENERAL)

If A and B are two events not necessarily mutually exclusive, associated with a random experiments, then

$$P(A \text{ or } B) = P(A) + P(B) - P(AB).$$

### NOTES

## NOTES

**Proof.** Let  $n$  be the total number of exhaustive equally likely cases of the experiment.

Let  $m_1$  and  $m_2$  be the number of cases favourable to the happening of the events  $A$  and  $B$  respectively.

$$\therefore P(A) = \frac{m_1}{n} \quad \text{and} \quad P(B) = \frac{m_2}{n}$$

Since the events are given to be not necessarily non-mutually exclusive, there may be some sample points common to both events  $A$  and  $B$ .

Let  $m_3$  be number of these common sample points.  $m_3$  will be zero in case  $A$  and  $B$  are mutually exclusive.

$$\therefore P(AB) = \frac{m_3}{n}$$

The  $m_3$  sample points which are common to both events  $A$  and  $B$ , are included in the events  $A$  and  $B$  separately.

$$\begin{aligned} \therefore \text{Number of sample points in the event } A \text{ or } B \\ = m_1 + m_2 - m_3 \end{aligned}$$

$m_3$  is subtracted from  $m_1 + m_2$  to avoid counting of common sample points twice.

$$\therefore P(A \text{ or } B) = \frac{m_1 + m_2 - m_3}{n} = \frac{m_1}{n} + \frac{m_2}{n} - \frac{m_3}{n} = P(A) + P(B) - P(AB).$$

Hence,  $P(A \text{ or } B) = P(A) + P(B) - P(AB)$ .

**Corollary 1.** If events  $A$  and  $B$  happen to be mutually exclusive events, then  $P(AB) = 0$  and in this case *addition theorem* implies

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(AB) = P(A) + P(B) - 0 = P(A) + P(B) \\ \therefore P(A \text{ or } B) &= P(A) + P(B). \end{aligned}$$

This is the same as the **addition theorem** for mutually exclusive events.

**Corollary 2.** If  $A, B, C$  are three events associated with a random experiment, then

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) - P(BC) - P(CA) - P(AB) + P(ABC).$$

**Example 9.15.** Find the probability that a card drawn from a pack of playing cards is either a 'queen' or a 'spade'.

**Solution.** Total number of cases (cards) = 52

Let  $A$  = event of drawing a 'queen'

$B$  = event of drawing a 'spade'

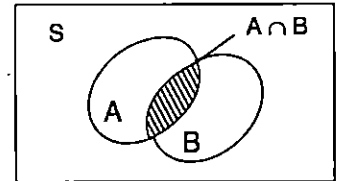
$$\therefore P(A) = \frac{4}{52} \quad \text{and} \quad P(B) = \frac{13}{52}$$

Here the events are not mutually exclusive as drawing of the card 'queen of spade' is common to both events.

$$\therefore P(AB) = \frac{1}{52}$$

By *addition theorem*, the probability of getting either a queen or a spade is

$$P(A \text{ or } B) = P(A) + P(B) - P(AB) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$





**Example 9.16.** One number is drawn from numbers 1 to 150. Find the probability that it is divisible by either 3 or 5.

**Solution.** Here  $S = \{1, 2, 3, \dots, 149, 150\}$ .

Let  $A =$  event that the number is divisible by 3

$$\therefore A = \{3, 6, 9, \dots, 147, 150\}$$

$$\therefore P(A) = \frac{50}{150}$$

Let  $B =$  event that the number is divisible by 5.

$$\therefore B = \{5, 10, 15, \dots, 145, 150\}$$

$$\therefore P(B) = \frac{30}{150}$$

The events  $A$  and  $B$  are not *m.e.* because the sample points 15, 30, 45, ..., 150 are common to both.

$$\therefore AB = \{15, 30, 45, \dots, 135, 150\}$$

$$\therefore P(AB) = \frac{10}{150}$$

By *addition theorem*, the required probability of getting a multiple of either 3 or 5 is

$$P(A \text{ or } B) = P(A) + P(B) - P(AB) = \frac{50}{150} + \frac{30}{150} - \frac{10}{150} = \frac{70}{150} = \frac{7}{15}$$

**Example 9.17.** A student applies for a job in two firms  $X$  and  $Y$ . The probability of his being selected in firm  $X$  is 0.7 and being rejected in the firm  $Y$  is 0.5. The probability of at least one of his application being rejected is 0.6. What is the probability that he will be selected in one of the firms?

**Solution.** Let  $A =$  event of getting selected in  $X$

$B =$  event of getting selected in  $Y$

$$\therefore P(A) = 0.7, P(B) = 1 - 0.5 = 0.5$$

$$P(AB) = 1 - P(\text{rejecting in at least one firm}) = 1 - 0.6 = 0.4$$

$$\therefore P(\text{selected in one of the firm})$$

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

$$= 0.7 + 0.5 - 0.4 = 0.8.$$

### EXERCISE 9.3

- Find the probability that a card drawn from a pack of playing cards is either a 'king' or a 'club'.
- A drawer contains 50 bolts and 150 nuts. Half of the bolts and half of the nuts are rusted. If one item is chosen at random, what is the probability that it is rusted or is a bolt?
- From 30 tickets marked with first 30 numerals, one is drawn at random. Find the probability that it is:
  - a multiple of 5 or 7
  - a multiple of 3 or 7.
- A construction company is bidding for two contracts  $A$  and  $B$ . The probability that the company will get contract  $A$  is  $3/5$ , the probability that the company will get contract  $B$  is  $1/3$  and the probability that the company will get both the contracts is  $1/8$ . What is the probability that the company will get contract  $A$  or  $B$ ?

### NOTES

## NOTES

5. The probability that a contractor will get a plumbing contract is  $\frac{2}{3}$  and the probability that he will not get an electric contract is  $\frac{5}{9}$ . The probability of getting at least one contract is  $\frac{4}{5}$ . What is the probability that he will get both contracts?

$$\left[ \text{Hint. } \frac{4}{5} = \frac{2}{3} + \left(1 - \frac{5}{9}\right) - P(AB). \right]$$

6. Find the probability of getting at least one five, in a single throw of two dice.

## Answers

- |                     |                    |                      |                      |
|---------------------|--------------------|----------------------|----------------------|
| 1. $\frac{4}{13}$   | 2. $\frac{5}{8}$   | 3. (i) $\frac{1}{3}$ | (ii) $\frac{13}{30}$ |
| 4. $\frac{97}{120}$ | 5. $\frac{14}{45}$ | 6. $\frac{11}{36}$   |                      |

## 9.15. CONDITIONAL PROBABILITY

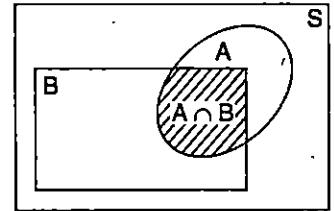
Let us consider the random experiment of throwing a die. Let  $A$  be the event of getting an odd number on the die.

$$\therefore S = \{1, 2, 3, 4, 5, 6\} \text{ and } A = \{1, 3, 5\}.$$

$$\therefore P(A) = \frac{3}{6} = \frac{1}{2}.$$

Let  $B = \{2, 3, 4, 5, 6\}$ .

If, after the die is thrown, we are given the information that the event  $B$  has occurred, then the probability of event  $A$  will no more be  $\frac{1}{2}$ , because in this case, the favourable cases are three and the total number of possible outcomes will be five and not six. The probability of event  $A$ , with the condition that event  $B$  has happened will be  $\frac{3}{5}$ . This conditional probability is denoted as  $P(A/B)$ . Let us define the concept of conditional probability in a formal manner.



Let  $A$  and  $B$  be any two events associated with a random experiment. The probability of occurrence of event  $A$  when the event  $B$  has already occurred is called the **conditional probability** of  $A$  when  $B$  is given and is denoted as  $P(A/B)$ . The conditional probability  $P(A/B)$  is meaningful only when  $P(B) \neq 0$ , i.e. when  $B$  is not an impossible event.

By definition,

$P(A/B)$  = Probability of occurrence of event  $A$  when the event  $B$  has already occurred.

$$= \frac{\text{no. of cases favourable to } B \text{ which are also favourable to } A}{\text{no. of cases favourable to } B}$$

$$\therefore P(A/B) = \frac{\text{no. of cases favourable to } A \cap B}{\text{no. of cases favourable to } B}$$

$$\text{Also, } P(A/B) = \frac{\text{no. of cases in the sample space}}{\text{no. of cases favourable to } B} \cdot \frac{\text{no. of cases favourable to } A \cap B}{\text{no. of cases in the sample space}}$$

$$\therefore P(A/B) = \frac{P(A \cap B)}{P(B)}.$$



**Remark 1.** If  $P(A) \neq 0$ , the  $P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}$ .

**Remark 2.** If A and B are *m.e.* events, then

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0}{P(B)} = 0 \quad \text{and} \quad P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{0}{P(A)} = 0.$$

$\therefore$  If A and B are *m.e.* events, then A/B and B/A are impossible events.

For an illustration, let us consider the random experiment of throwing two coins.

$$\therefore S = \{HH, HT, TH, TT\}.$$

Let  $A = \{HH, HT\}$ ,  $B = \{HH, TH\}$ ,  $C = \{HH, HT, TH\}$  and  $D = \{TT\}$ .

$$\therefore P(A) = \frac{2}{4} = \frac{1}{2}, \quad P(B) = \frac{2}{4} = \frac{1}{2}, \quad P(C) = \frac{3}{4}, \quad P(D) = \frac{1}{4}.$$

A/B is the event of getting A with the condition that B has occurred.

$$\therefore P(A/B) = \frac{n(A \cap B)}{n(B)} = \frac{n\{HH\}}{n\{HH, TH\}} = \frac{1}{2}.$$

$$\text{Similarly, } P(A/C) = \frac{n\{HH, HT\}}{n\{HH, HT, TH\}} = \frac{2}{3} \quad \text{and} \quad P(B/C) = \frac{n\{HH, TH\}}{n\{HH, HT, TH\}} = \frac{2}{3}.$$

We observe that  $P(A/C) \neq P(A)$ ,  $P(B/C) \neq P(B)$ .

The events A and D are *m.e.* and we have

$$P(A/D) = \frac{n(A \cap D)}{n(D)} = \frac{0}{1} = 0 \quad \text{and} \quad P(D/A) = \frac{n(D \cap A)}{n(A)} = \frac{0}{2} = 0.$$

**Example 9.18.** If  $P(E) = 0.40$ ,  $P(F) = 0.35$  and  $P(E \cup F) = 0.55$ , find  $P(E/F)$ .

**Solution.** We have  $P(E) = 0.40$ ,  $P(F) = 0.35$ ,  $P(E \cup F) = 0.55$ .

By addition theorem,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

$$\therefore 0.55 = 0.40 + 0.35 - P(E \cap F)$$

$$\therefore P(E \cap F) = 0.75 - 0.55 = 0.20.$$

$$\text{Required probability, } P(E/F) = \frac{P(E \cap F)}{P(F)} = \frac{0.20}{0.35} = \frac{4}{7}.$$

**Example 9.19.** A coin is tossed twice and the four possible outcomes are assumed to be equally likely. If E is the event "both head and tail have occurred", and F the event "at most one tail is observed", find  $P(E)$ ,  $P(F)$ ,  $P(E/F)$  and  $P(F/E)$ .

**Solution.** We have  $S = \{HH, HT, TH, TT\}$

$$E = \{HT, TH\} \quad \text{and} \quad F = \{HH, HT, TH\}.$$

$$\therefore E \cap F = \{HT, TH\}.$$

$$\therefore P(E) = \frac{n(E)}{n(S)} = \frac{2}{4} = \frac{1}{2}, \quad P(F) = \frac{n(F)}{n(S)} = \frac{3}{4},$$

$$P(E/F) = \frac{n(E \cap F)}{n(F)} = \frac{2}{3} \quad \text{and} \quad P(F/E) = \frac{n(E \cap F)}{n(E)} = \frac{2}{2} = 1.$$

**Example 9.20.** A die is thrown twice and the sum of the number appearing is observed to be 6. What is the conditional probability that the number 4 has appeared at least once?

**Solution.** Let S be the sample space of the experiment.

$$\therefore S = \{(1, 1), (1, 2), \dots, (6, 5), (6, 6)\}.$$

## NOTES

## NOTES

Let  $A =$  event of getting sum 6  
 and  $B =$  event of getting 4 at least once.  
 $\therefore A = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$   
 and  $B = \{(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (1, 4), (2, 4), (3, 4), (5, 4), (6, 4)\}.$

$$\therefore P(A) = \frac{5}{36} \quad \text{and} \quad P(B) = \frac{11}{36}$$

$$\text{Also } A \cap B = \{(4, 2), (2, 4)\} \quad \therefore P(A \cap B) = \frac{2}{36}$$

Now, required probability

= Probability of getting 4 on at least one die given that sum is 6

$$= P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{2/36}{5/36} = \frac{2}{5}$$

Alternatively,

$$P(B/A) = \frac{\text{no. of cases favourable to 'A} \cap \text{B'}}{\text{no. of cases favourable to A}} = \frac{2}{5}$$

**Remark 1.** In practical problems, it would be easier to use the formulae:

$$(i) P(A/B) = \frac{\text{no. of favourable to 'A} \cap \text{B'}}{\text{no. of cases favourable to B}}$$

$$(ii) P(B/A) = \frac{\text{no. of cases favourable to 'B} \cap \text{A' i.e. 'A} \cap \text{B'}}{\text{no. of cases favourable to A}}$$

**Remark 2.** The event  $A \cap B$  is same as  $B \cap A$  and each consists of sample points which are common to both A and B.

**Example 9.21.** One card is drawn from a well shuffled pack of 52 cards. If E is the event "the card drawn is either a king or an ace" and F is the event "the card drawn is either an ace or a jack", then find the probability of the conditional event E/F.

**Solution.** There are 4 kings and 4 aces in the pack.

$$\therefore P(E) = \frac{4+4}{52} = \frac{2}{13}$$

There are 4 aces and 4 jacks in the pack.

$$\therefore P(F) = \frac{4+4}{52} = \frac{2}{13}$$

The event  $E \cap F$  contain 4 aces.

$$\therefore P(E \cap F) = \frac{4}{52} = \frac{1}{13}$$

$$\therefore \text{Required probability} = P(E/F) = \frac{P(E \cap F)}{P(F)} = \frac{1/13}{2/13} = \frac{1}{2}$$

**Example 9.22.** Three fair coins are tossed. Find the probability that they are all tails, if:

- (i) at least one of the coins show tail      (ii) two coins show tail  
 (iii) at least two coins show head      (iv) at most one coin show head.

**Solution.** Here  $S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$

Let  $A =$  event of getting all tails  $\therefore A = \{TTT\}$

(i) Let  $B =$  event that at least one of the coins show tail

$\therefore B = \{HHT, HTH, THH, HTT, THT, TTH, TTT\}$



$$\therefore A \cap B = \{TTT\}$$

$$\therefore \text{Required probability} = P(A/B) = \frac{n(A \cap B)}{n(B)} = \frac{1}{7}$$

(ii) Let  $B = \text{event that two coins show tail}$

$$\therefore B = \{HTT, THT, TTH\}$$

$$\therefore A \cap B = \phi$$

$$\therefore \text{Required probability} = \frac{n(A \cap B)}{n(B)} = \frac{0}{3} = 0.$$

(iii) Let  $B = \text{event that at least two coins show head}$

$$\therefore B = \{HHH, HHT, HTH, THH\}$$

$$\therefore A \cap B = \phi$$

$$\therefore \text{Required probability} = P(A/B) = \frac{n(A \cap B)}{n(B)} = \frac{0}{4} = 0.$$

(iv) Let  $B = \text{event that at most one coin show head.}$

$$\therefore B = \{HTT, THT, TTH, TTT\}$$

$$\therefore A \cap B = \{TTT\}$$

$$\therefore \text{Required probability} = P(A/B) = \frac{n(A \cap B)}{n(B)} = \frac{1}{4}.$$

**Remark.** The value of  $P(A/B)$  is equal to  $\frac{n(A \cap B)}{n(B)}$  which is also equal to  $\frac{P(A \cap B)}{P(B)}$ .

### EXERCISE 9.4

- If  $P(\text{not } A) = 0.7$ ,  $P(B) = 0.7$  and  $P(B/A) = 0.5$ , then find  $P(A/B)$  and  $P(A \cup B)$ .
- For two events  $A$  and  $B$ ,  $P(A) = 0.5$ ,  $P(B) = 0.6$  and  $P(A \cap B) = 0.8$ . Find the conditional probabilities  $P(A/B)$  and  $P(B/A)$ .
- A die is thrown. Find that probability that the number obtained is greater than 2 if:
  - no other information is given,
  - it is given that the number obtained is less than 5.
- A pair of fair dice is thrown. Find the probability that the sum is 10 or greater if 5 appears on the first die.
- A pair of fair dice is thrown. If the two numbers appearing are different, find the probability that the sum is 4 or less.
- The probability that a person stopping at a petrol pump will ask to have his tyres checked is 0.12, the probability that he will ask to have his oil checked is 0.29 and the probability that he will ask to have both of them checked is 0.07.
  - What is the probability that a person who has oil checked will also have tyre checked?
  - What is the probability that a person stopping at the petrol pump will have either tyres or oil checked?

**[Hint:** Let  $A$  and  $B$  be the events of getting 'tyres checked' and 'oil checked' respectively.

$$\therefore P(A) = 0.12, P(B) = 0.29, P(A \cap B) = 0.07.$$

$$(ii) \text{ Required probability} = P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$(iii) \text{ Required probability} = P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

### NOTES

## NOTES

1. 0.2143, 0.85

2. 0.5, 0.6

3. (i)  $\frac{2}{3}$

(ii)  $\frac{1}{2}$

4.  $\frac{1}{3}$

5.  $\frac{2}{15}$

6. (i)  $\frac{7}{29}$

(ii)  $\frac{17}{50}$

**9.16. INDEPENDENT EVENTS**

Let A and B be two events associated with a random experiment. We know that

$$P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

$$\therefore P(A \cap B) = P(A) P(B/A).$$

In general  $P(B/A)$  may or may not be equal to  $P(B)$ . When  $P(B/A)$  and  $P(B)$  are equal, then the events A and B are of special importance.

Two events associated with a random experiment are said to be **independent events** if the occurrence or non-occurrence of one event does not affect the probability of the occurrence of the other event. For example, the events A and B are independent events when  $P(A/B) = P(A)$  and  $P(B/A) = P(B)$ .

**Theorem.** Let A and B be events associated with a random experiment. The events A and B are independent if and only if

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

**Proof.** Let A and B be independent events.

$$\begin{aligned} \therefore P(A \cap B) &= \frac{P(A \cap B)}{P(B)} \times P(B) = P(A/B) P(B) \\ &= P(A) P(B) \end{aligned}$$

$$[\because P(A/B) = P(A)]$$

$$\therefore P(A \cap B) = P(A) P(B).$$

Conversely, let  $P(A \cap B) = P(A) P(B)$ .

$$\therefore P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) P(B)}{P(B)} = P(A)$$

$$\text{and } P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{P(A) P(B)}{P(A)} = P(B).$$

$$\therefore P(A/B) = P(A) \text{ and } P(B/A) = P(B).$$

$\therefore$  A and B are independent events.

**Remark 1.**  $P(A \cap B) = P(A) P(B)$  is the necessary and sufficient condition for the events A and B to be independent.

**Remark 2.** Let A and B be events associated with a random experiment.

(i) Let A and B be *m.e.*  $\therefore P(A \cap B) = 0$

$\therefore P(A \cap B) \neq P(A) P(B)$  i.e. A and B are not independent events.

$\therefore$  **Mutually exclusive events cannot be independent.**

(ii) Let A and B be independent.

$\therefore P(A \cap B) = P(A) P(B)$  i.e.  $P(A \cap B) \neq 0$ .

$\therefore$  A and B are not *m.e.* events.

$\therefore$  **Independent events cannot be mutually exclusive.**

**Important observation.** If A and B be any two events associated with a random experiment, then their physical description is not sufficient to decide if A and B are independent events or not. A and B are declared to be independent events only when we have  $P(AB) = P(A) P(B)$ .



## 9.17. DEPENDENT EVENTS

Let A and B be two events associated with a random experiment. If A and B are not independent events, then these are called **dependent events**.

∴ In case of dependent events, we have  $P(A \cap B) = P(A) P(B/A)$ .

**Theorem.** Let A and B be events associated with a random experiment.

If A and B are independent, then show that the events (i)  $\bar{A}$ , B (ii) A,  $\bar{B}$  (iii)  $\bar{A}$ ,  $\bar{B}$  are also independent.

**Proof.** The events A and B are independent.

$$\therefore P(A \cap B) = P(A) P(B) \quad \dots(1)$$

$$(i) (A \cap B) \cap (\bar{A} \cap B) = (A \cap \bar{A}) \cap (B \cap B) = \phi \cap B = \phi$$

$$\text{and } (A \cap B) \cup (\bar{A} \cap B) = (A \cup \bar{A}) \cap B = S \cap B = B.$$

∴ The events  $A \cap B$  and  $\bar{A} \cap B$  are *m.e.* and their union is B.

∴ By *addition theorem*, we have

$$P(B) = P(A \cap B) + P(\bar{A} \cap B).$$

$$\Rightarrow P(\bar{A} \cap B) = P(B) - P(A \cap B) = P(B) - P(A) P(B) \quad [\text{Using (1)}]$$

$$= (1 - P(A)) P(B) = P(\bar{A}) P(B).$$

∴  $P(\bar{A} \cap B) = P(\bar{A}) P(B)$  i.e.  $\bar{A}$  and B are independent.

$$(ii) (A \cap B) \cap (A \cap \bar{B}) = (A \cap A) \cap (B \cap \bar{B}) = A \cap \phi = \phi$$

$$\text{and } (A \cap B) \cup (A \cap \bar{B}) = A \cap (B \cup \bar{B}) = A \cap S = A$$

∴ The events  $A \cap B$  and  $A \cap \bar{B}$  are *m.e.* and their union is A.

∴ By *addition theorem*, we have

$$P(A) = P(A \cap B) + P(A \cap \bar{B}).$$

$$\Rightarrow P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - P(A) P(B) \quad [\text{Using (1)}]$$

$$= P(A)(1 - P(B)) = P(A) P(\bar{B}).$$

∴  $P(A \cap \bar{B}) = P(A) P(\bar{B})$  i.e. A and  $\bar{B}$  are independent.

$$(iii) (\bar{A} \cap B) \cap (\bar{A} \cap \bar{B}) = (\bar{A} \cap \bar{A}) \cap (B \cap \bar{B}) = \bar{A} \cap \phi = \phi$$

$$\text{and } (\bar{A} \cap B) \cup (\bar{A} \cap \bar{B}) = \bar{A} \cap (B \cup \bar{B}) = \bar{A} \cap S = \bar{A}.$$

∴ The events  $\bar{A} \cap B$  and  $\bar{A} \cap \bar{B}$  are *m.e.* and their union is  $\bar{A}$ .

∴ By *addition theorem*, we have

$$P(\bar{A}) = P(\bar{A} \cap B) + P(\bar{A} \cap \bar{B}) \quad \dots(1)$$

$$\Rightarrow P(\bar{A} \cap \bar{B}) = P(\bar{A}) - P(\bar{A} \cap B) = P(\bar{A}) - P(\bar{A}) P(B)$$

$$[\text{Using part (i)}]$$

$$= P(\bar{A}) (1 - P(B)) = P(\bar{A}) P(\bar{B}).$$

∴  $P(\bar{A} \cap \bar{B}) = P(\bar{A}) P(\bar{B})$  i.e.  $\bar{A}$  and  $\bar{B}$  are independent.

## NOTES



**Example 9.23.** If  $A$  and  $B$  are independent events such that  $P(A \cup B) = 0.6$  and  $P(A) = 0.2$ , find  $P(B)$ .

**Solution.** We have  $P(A \cup B) = 0.6$  and  $P(A) = 0.2$ .

By *addition theorem*, we have

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

$$\therefore P(A \cup B) = P(A) + P(B) - P(A)P(B) \quad (\because A \text{ and } B \text{ are independent})$$

$$\Rightarrow 0.6 = 0.2 + P(B) - (0.2)P(B)$$

$$\Rightarrow 0.4 = P(B)(1 - 0.2)$$

$$\Rightarrow (0.8)P(B) = 0.4 \Rightarrow P(B) = \frac{0.4}{0.8} = \frac{1}{2} = 0.5.$$

**Example 9.24.** A coin is tossed thrice and all the eight outcomes are assumed equally likely. In which of the following cases are the events  $E$  and  $F$  independent?

(i)  $E$ : the first throw results in head.

$F$ : the last throw results in tail.

(ii)  $E$ : the number of heads is two.

$F$ : the last throw results in head.

(iii)  $E$ : the number of heads is odd.

$F$ : the number of tails is odd.

**Solution.** Let  $S$  be the sample space.

$$\therefore S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

(i) Here  $E = \{HHH, HHT, HTH, HTT\}$  and  $F = \{HHT, HTT, THT, TTT\}$ .

$$\therefore P(E) = \frac{4}{8} = \frac{1}{2} \quad \text{and} \quad P(F) = \frac{4}{8} = \frac{1}{2}.$$

There are 2 cases favourable to the event  $E \cap F$ , namely HHT and HTT.

$$\therefore P(E \cap F) = \frac{2}{8} = \frac{1}{4}.$$

$$\text{Also, } P(E)P(F) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = P(E \cap F).$$

$\therefore$  The events  $E$  and  $F$  are **independent**.

(ii) Here  $E = \{HHT, HTH, THH\}$  and  $F = \{HHH, HTH, THH, TTH\}$ .

$$\therefore P(E) = \frac{3}{8} \quad \text{and} \quad P(F) = \frac{4}{8} = \frac{1}{2}$$

There are 2 cases favourable to the event  $E \cap F$ , namely HTH and THH.

$$\therefore P(E \cap F) = \frac{2}{8} = \frac{1}{4}.$$

$$\text{Also, } P(E)P(F) = \frac{3}{8} \times \frac{1}{2} = \frac{3}{16} \neq P(E \cap F).$$

$\therefore$  The events  $E$  and  $F$  are **not independent**.

(iii) Here  $E = \{HHH, HTT, THT, TTH\}$  and  $F = \{HHT, HTH, THH, TTT\}$ .

$$\therefore P(E) = \frac{4}{8} = \frac{1}{2} \quad \text{and} \quad P(F) = \frac{4}{8} = \frac{1}{2}.$$

There is no case favourable to the event  $E \cap F$ .

## NOTES

$$\therefore P(E \cap F) = \frac{0}{8} = 0.$$

$$\text{Also, } P(E)P(F) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \neq P(E \cap F).$$

$\therefore$  The events E and F are **not independent**.

### EXERCISE 9.5

- Two events A and B are such that  $P(A) = 0.6$ ,  $P(B) = 0.2$  and  $P(A \cap B) = 0.8$ . Does this imply that A and B are independent?
  - The events A and B are given to be independent. Find  $P(B)$  if it is given that  $P(A) = 0.35$  and  $P(A \cup B) = 0.60$ .
- If  $P(\text{not } B) = 0.65$ ,  $P(A \cup B) = 0.85$  and A and B are independent events, find  $P(A)$ .
  - If  $P(\text{not } A) = 0.4$ ,  $P(A \cup B) = 0.75$  and A, B are given to be independent events, find the value of  $P(B)$ .
- A coin is tossed twice and all possible outcomes are assumed to be equally likely. A is the event : both head and tail have occurred and B is the event : at least one tail has occurred". Show that A and B are not independent.
- One card is drawn from a pack of 52 cards so that each card is equally likely to be selected. A is the event : "the card is a heart" and B is the event : "the card is a king." Show that A and B are independent.
- A die is thrown and the 6 possible outcomes are assumed to be equally likely. If E is the event : "the number appearing is a multiple of 3", and F is the event : "the number appearing is even". Show that the events E and F are independent.

### Answers

- No.
  - $\frac{5}{13}$
- $\frac{10}{13}$
  - $\frac{3}{8}$

## 9.18. INDEPENDENT EXPERIMENTS

Two random experiments are said to be **independent experiments** if the occurrence or non-occurrence of an event in one experiment does not in any way affect the probability of occurrence of any event in the other experiment. For example, two tosses of a coin are independent experiments.

## 9.19. MULTIPLICATION THEOREM

If A and B be events associated with independent experiments  $E_1$  and  $E_2$  respectively, then prove that

$$P(AB) = P(A)P(B).$$

**Proof.** Since the random experiments  $E_1$  and  $E_2$  are independent, the sample spaces of the experiments are not affected by the events.

Let  $n_1$  and  $n_2$  be the numbers of exhaustive, equally likely cases in the first and second experiment respectively.

Let  $m_1$  be the number of cases favourable to the happening of the event A out of  $n_1$  cases of the first experiment.

### NOTES

## NOTES

$$P(A) = \frac{m_1}{n_1}$$

Let  $m_2$  be the number of cases favourable to the happening of the event B out of  $n_2$  cases of the second experiment.

$$P(B) = \frac{m_2}{n_2}$$

By the **Fundamental principle of events**, the number of cases favourable to the happening of the event AB in this specified order is  $m_1 m_2$ . Also the number of exhaustive, equally likely cases in the combined experiment is  $n_1 n_2$ .

$$P(AB) = \frac{m_1 m_2}{n_1 n_2} = \frac{m_1}{n_1} \cdot \frac{m_2}{n_2} = P(A)P(B).$$

$$P(AB) = P(A) P(B).$$

The theorem can also be extended to even more than two events.

Let  $A_1, A_2, \dots, A_k$  be  $k$  events associated with random experiments  $E_1, E_2, \dots, E_k$  with probabilities  $p_1, p_2, \dots, p_k$  respectively, then

$$P(A_1 A_2 \dots A_k) = P(A_1) P(A_2) \dots P(A_k) = p_1 p_2 \dots p_k.$$

Also

$$P(\text{not } A_1) = P(\bar{A}_1) = 1 - p_1$$

$$P(\text{not } A_2) = P(\bar{A}_2) = 1 - p_2$$

$$P(\text{not } A_k) = P(\bar{A}_k) = 1 - p_k.$$

Since  $A_1, A_2, \dots, A_k$  are associated with independent experiments, therefore, the events  $\bar{A}_1, \bar{A}_2, \dots, \bar{A}_k$  are also associated with independent experiments.

Now  $P(\text{event of happening of at least one of } A_1, A_2, \dots, A_k)$

$$= 1 - P(\text{event of happening of none of } A_1, A_2, \dots, A_k)$$

$$= 1 - P(\bar{A}_1 \bar{A}_2 \dots \bar{A}_k) = 1 - P(\bar{A}_1) P(\bar{A}_2) \dots P(\bar{A}_k)$$

(By Multiplication theorem)

$$= 1 - (1 - p_1)(1 - p_2) \dots (1 - p_k).$$

**Remark.** If A and B are events associated with experiments which are not independent, then the probability of the event 'AB' is found by using the result:

$$P(AB) = P(A) P(B/A).$$

This result can also be extended to more than two experiments.

**Example 9.25.** A and B appeared for an interview for two posts. Probability of A's rejection is  $2/5$  and that of B's selection is  $4/7$ . Find the probability that one of them is selected.

**Solution.** The random experiments 'interview of A' and 'interview of B' are experiment.

Let

E = event that A is selected

and

F = event that B is selected.

$\therefore$

$$P(\bar{E}) = \frac{2}{5} \quad \text{and} \quad P(F) = \frac{4}{7}.$$



Also,  $P(E) = 1 - P(\bar{E}) = 1 - \frac{2}{5} = \frac{3}{5}$

and  $P(\bar{F}) = 1 - P(F) = 1 - \frac{4}{7} = \frac{3}{7}$

Required probability = P(only one is selected)

$$= P(E \bar{F} \cup \bar{E} F) = P(E \bar{F}) + P(\bar{E} F)$$

(Using addition theorem)

$$= P(E) P(\bar{F}) + P(\bar{E}) P(F)$$

(Using multiplication theorem)

$$= \frac{3}{5} \times \frac{3}{7} + \frac{2}{5} \times \frac{4}{7} = \frac{17}{35}$$

**Example 9.26.** The odds in favour of one student passing a test are 3 : 7. The odds against another student passing it are 3 : 5. What is the probability that both pass the test?

**Solution.** Let A = event that first pass the test.

$$\therefore P(A) = \frac{3}{3+7} = \frac{3}{10}$$

Let B = event the second pass the test.

$$\therefore P(B) = \frac{5}{3+5} = \frac{5}{8}$$

The random experiments of results of students are independent.

$$\therefore P(\text{both pass the test}) = P(AB) = P(A)P(B) = \frac{3}{10} \times \frac{5}{8} = \frac{3}{16}$$

**Example 9.27.** A speaks truth in 60% of the cases and B in 90% of the cases. In what percentage of cases, are they likely to contradict each other in stating the same fact?

**Solution.** The random experiments of speeches of A and B are independent.

Let E = event of A speaking truth

and F = event of B speaking truth.

$$\therefore P(E) = \frac{60}{100} = \frac{6}{10} \quad \text{and} \quad P(F) = \frac{90}{100} = \frac{9}{10}$$

Probability of A and B contradicting each other =  $P(E \bar{F} \text{ or } \bar{E} F)$

$$= P(E \bar{F}) + P(\bar{E} F) = P(E) P(\bar{F}) + P(\bar{E}) P(F)$$

$$= P(E)(1 - P(F)) + (1 - P(E))P(F)$$

$$= \frac{6}{10} \left(1 - \frac{9}{10}\right) + \left(1 - \frac{6}{10}\right) \frac{9}{10} = \frac{6}{10} \times \frac{1}{10} + \frac{4}{10} \times \frac{9}{10} = \frac{42}{100}$$

$\therefore$  A and B are likely to contradict each other in 42% cases.

**Example 9.28.** A problem in statistics is given to five students A, B, C, D and E. Their chances of solving the problem are  $\frac{1}{2}$ ,  $\frac{1}{3}$ ,  $\frac{1}{4}$ ,  $\frac{1}{5}$  and  $\frac{1}{6}$  respectively. What is the probability that the problem will be solved?

**Solution.** The random experiments of trying the problem by the given students are independent.

## NOTES

## NOTES

Let  $A_1$  = event that A fails to solve the problem  
 $A_2$  = event that B fails to solve the problem  
 $A_3$  = event that C fails to solve the problem  
 $A_4$  = event that D fails to solve the problem  
 $A_5$  = event that E fails to solve the problem.

$$\begin{aligned}\therefore P(A_1) &= 1 - \frac{1}{2} = \frac{1}{2} & P(A_2) &= 1 - \frac{1}{3} = \frac{2}{3} \\ P(A_3) &= 1 - \frac{1}{4} = \frac{3}{4} & P(A_4) &= 1 - \frac{1}{5} = \frac{4}{5} \\ P(A_5) &= 1 - \frac{1}{6} = \frac{5}{6}\end{aligned}$$

$$\begin{aligned}\text{Now, } P(\text{event that the problem is solved by at least one student}) \\ &= 1 - P(\text{event that the problem is not solved by any of the five students}) \\ &= 1 - P(A_1 A_2 A_3 A_4 A_5) \\ &= 1 - P(A_1) P(A_2) P(A_3) P(A_4) P(A_5) \text{ (By Multiplication Theorem)} \\ &= 1 - \left( \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \times \frac{4}{5} \times \frac{5}{6} \right) = 1 - \frac{1}{6} = \frac{5}{6}.\end{aligned}$$

**Example 9.29.** Three bags contains 7 white 8 red, 9 white 6 red and 5 white 7 red balls respectively. One ball, at random, is drawn from each bag. Find the probability that all of them are of the same colour.

**Solution.** The three random experiments of drawing balls from given bags are independent.

7 White 8 Red	9 White 6 Red	5 White 7 Red
Bag I	Bag II	Bag III

Let  $W_i$  and  $R_i$  be the events of drawing white ball and red ball respectively from the  $i$ th bag,  $i = 1, 2, 3$ .

$$\begin{aligned}\text{Required probability} &= P(\text{all balls of same colour}) \\ &= P(W_1 W_2 W_3 \text{ or } R_1 R_2 R_3) \\ &= P(W_1 W_2 W_3) + P(R_1 R_2 R_3) \\ &= P(W_1)P(W_2)P(W_3) + P(R_1)P(R_2)P(R_3) \\ &= \frac{7}{7+8} \times \frac{9}{9+6} \times \frac{5}{5+7} + \frac{8}{7+8} \times \frac{6}{9+6} \times \frac{7}{5+7} \\ &= \frac{7}{15} \times \frac{9}{15} \times \frac{5}{12} + \frac{8}{15} \times \frac{6}{15} \times \frac{7}{12} = \frac{651}{2700} = \frac{217}{900}.\end{aligned}$$

**Example 9.30.** A bag has 4 red and 5 black balls, a second bag has 3 red and 7 black balls. One ball is drawn from the first and two from the second. Find the probability that out of three balls, two are black and one is red.

**Solution.** The random experiments of 'drawing one ball from first bag' and 'drawing two balls from second bag' are independent.

4 Red 5 Black	3 Red 7 Black
Bag I	Bag II

We are to find the probability two black balls and one red ball.



∴ Required probability

$$= P(F_r S_{bb} \text{ or } F_b S_{rb}),$$

where  $F_r$  is the event of drawing red ball from the first bag, etc.

$$= P(F_r S_{bb}) + P(F_b S_{rb}) = P(F_r) P(S_{bb}) + P(F_b) P(S_{rb})$$

$$= \frac{4}{4+5} \times \frac{{}^7C_2}{{}^{3+7}C_2} + \frac{5}{4+5} \times \frac{{}^3C_1 {}^7C_1}{{}^{3+7}C_2}$$

$$= \frac{4}{9} \times \frac{\frac{7 \times 6}{1 \times 2}}{\frac{10 \times 9}{1 \times 2}} + \frac{5}{9} \times \frac{3 \times 7}{10 \times 9} = \frac{4}{9} \times \frac{7}{15} + \frac{5}{9} \times \frac{7}{15} = \frac{7}{15}.$$

**Example 9.31.** In a hockey match, the probability of winning of Indian team against Pakistani team is  $1/4$ . Three matches are played. Find the probability that:

## NOTES

(i) India loses all the matches.

(ii) India wins at least one match.

(iii) India wins two matches.

**Solution.** The random experiments of matches between Indian team and Pakistani team are independent.

Let  $A_i$  to the event of winning of Indian team in the  $i$ th match,  $i = 1, 2, 3$ .

$$\therefore P(A_i) = \frac{1}{4} \quad \text{and} \quad P(\bar{A}_i) = 1 - \frac{1}{4} = \frac{3}{4}$$

(i)  $P(\text{India losing all the matches})$

$$= P(\bar{A}_1 \bar{A}_2 \bar{A}_3) = P(\bar{A}_1) P(\bar{A}_2) P(\bar{A}_3) = \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{64}.$$

(ii)  $P(\text{India winning at least one match})$

$$= 1 - P(\text{India losing all the matches})$$

$$= 1 - \frac{27}{64} = \frac{37}{64}.$$

[Using part (i)]

(iii)  $P(\text{India winning two matches})$

$$= P(A_1 A_2 \bar{A}_3 \text{ or } A_1 \bar{A}_2 A_3 \text{ or } \bar{A}_1 A_2 A_3)$$

$$= P(A_1 A_2 \bar{A}_3) + P(A_1 \bar{A}_2 A_3) + P(\bar{A}_1 A_2 A_3)$$

$$= P(A_1) P(A_2) P(\bar{A}_3) + P(A_1) P(\bar{A}_2) P(A_3) + P(\bar{A}_1) P(A_2) P(A_3)$$

$$= \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} + \frac{1}{4} \times \frac{3}{4} \times \frac{1}{4} + \frac{3}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{9}{64}.$$

**Example 9.32.** Three groups of children consists of respectively 3 girls and 1 boy, 2 girls and 2 boys, 1 girl and 3 boys. One child is selected at random from each group. Find the chance that three selected children comprise 1 girl and 2 boys.

**Solution.** The random experiments of drawing one child from each group are independent.

3 Girls 1 Boy	2 Girls 2 Boys	1 Girl 3 Boys
Group I	Group II	Group III

## NOTES

Let  $G_i$  and  $B_i$  be the events of selecting a girl and a boy respectively from the  $i$ th group,  $i = 1, 2, 3$ .

$\therefore$  P(1 girl and 2 boys)

$$\begin{aligned}
 &= P(G_1 B_2 B_3 \text{ or } B_1 G_2 B_3 \text{ or } B_1 B_2 G_3) \\
 &= P(G_1 B_2 B_3) + P(B_1 G_2 B_3) + P(B_1 B_2 G_3) \\
 &= P(G_1)P(B_2)P(B_3) + P(B_1)P(G_2)P(B_3) + P(B_1)P(B_2)P(G_3) \\
 &= \left(\frac{3}{4} \times \frac{2}{4} \times \frac{3}{4}\right) + \left(\frac{1}{4} \times \frac{2}{4} \times \frac{3}{4}\right) + \left(\frac{1}{4} \times \frac{2}{4} \times \frac{1}{4}\right) = \frac{18+6+2}{64} = \frac{13}{12}
 \end{aligned}$$

**Example 9.33.** A bag contains 6 white ball and 4 black balls. Two balls are drawn at random one by one without replacement. Find the probability that both balls are white.

**Solution.** Since the first ball is not replaced before the second draw, the random experiments of drawing balls are not independent.

Let  $W_1$  = event that first ball is white

$W_2$  = event that second ball is white

P(both balls are white)

$$\begin{aligned}
 &= P(W_1 W_2) = P(W_1)P(W_2/W_1) \\
 &= \frac{6}{6+4} \times \frac{6-1}{(6-1)+4} = \frac{6}{10} \times \frac{5}{9} = \frac{1}{3}
 \end{aligned}$$

**Example 9.34.** From a pack of playing cards, two cards are drawn one by one without replacement. Find the probability that:

(i) first is king and second is queen

(ii) one is king and other in queen.

**Solution.** Since the first card is not replaced before the second draw, the random experiments of drawing cards are not independent.

Let  $K_i$  and  $Q_i$  be the events of drawing a king and a queen respectively in the  $i$ th draw,  $i = 1, 2$ .

(i) P(first is king and second is queen)

$$\begin{aligned}
 &= P(K_1 Q_2) = P(K_1) P(Q_2/K_1) \\
 &= \frac{4}{52} \times \frac{4}{51} = \frac{4}{663}
 \end{aligned}$$

(ii) P(one king and one queen)

$$\begin{aligned}
 &= P(K_1 Q_2 \text{ or } Q_1 K_2) = P(K_1 Q_2) + P(Q_1 K_2) \\
 &= P(K_1) P(Q_2/K_1) + P(Q_1) P(K_2/Q_1) \\
 &= \frac{4}{52} \times \frac{4}{51} + \frac{4}{52} \times \frac{4}{51} = \frac{8}{663}
 \end{aligned}$$

**Example 9.35.** From a well-shuffled pack of playing cards, two cards are drawn at random one by one. Find the probability that they are both kings if the first card is :  
(i) replaced, (ii) not replaced before the second draw.

**Solution.** Let  $K_1$  and  $K_2$  be the events of getting kings in the first draw and second draw respectively before the second draw.

(i) Since the first card is replaced, the random experiments are independent.

$$\therefore P(\text{both kings}) = P(K_1 K_2) = P(K_1) P(K_2) = \frac{4}{52} \times \frac{4}{52} = \frac{1}{169}$$



(ii) Since the first card is not replaced, the random experiments are not independent.

$$\therefore P(\text{both kings}) = P(K_1 K_2) = P(K_1) P(K_2/K_1) = \frac{4}{52} \times \frac{3}{51} = \frac{1}{221}.$$

**Example 9.36.** A bag contains 8 red and 5 white balls. Two successive drawings of three balls are made such that (i) balls are replaced before second trial, (ii) balls are not replaced before second trial. Find the probability that 1st drawing will give 3 white and the 2nd 3 red balls.

## NOTES

**Solution.** Let  $W_1$  and  $R_2$  be the events of getting 3 white balls in the first draw and 3 red balls in the second draw respectively.

8 Red 5 White
------------------

(i) Since the balls of first draw are replaced, the random experiments are independent.

$$\therefore P(W_1 R_2) = P(W_1) P(R_2) = \frac{{}^5C_3}{{}^{13}C_3} \times \frac{{}^8C_3}{{}^{13}C_3} = \frac{10}{286} \times \frac{56}{286} = 0.0068$$

(ii) Since the balls of first draw are not replaced, the random experiments are not independent.

$$\therefore P(W_1 R_2) = P(W_1) P(R_2/W_1) = \frac{{}^5C_3}{{}^{13}C_3} \times \frac{{}^8C_3}{{}^{10}C_3} = \frac{10}{286} \times \frac{56}{120} = 0.0163.$$

**Example 9.37.** In each of a set of games, it is 2 to 1 in favour of the winner of the previous game. What is the chance that the player who wins the first game shall win at least three of the next four games?

**Solution.** Let  $W_i$  be the event that the winner of the first game wins the  $i$ th game,  $i = 2, 3, 4, 5$ .

$\therefore P(\text{Winner of the first game wins at least 3 out of the next 4 games})$

$$= P(W_2 W_3 W_4 \bar{W}_5 \text{ or } W_2 W_3 \bar{W}_4 W_5 \text{ or } W_2 \bar{W}_3 W_4 W_5 \text{ or } \bar{W}_2 W_3 W_4 W_5 \text{ or } W_2 W_3 W_4 W_5)$$

$$= P(W_2 W_3 W_4 \bar{W}_5) + P(W_2 W_3 \bar{W}_4 W_5) + P(W_2 \bar{W}_3 W_4 W_5) + P(\bar{W}_2 W_3 W_4 W_5) + P(W_2 W_2 W_3 W_4)$$

$$= P(W_2) P(W_3/W_2) P(W_4/W_2 W_3) P(\bar{W}_5/W_2 W_3 W_4) + P(W_2) P(W_3/W_2) P(\bar{W}_4/W_2 W_3) P(W_5/W_2 W_3 \bar{W}_4) + P(W_2) P(\bar{W}_3/W_2) P(W_4/W_2 \bar{W}_3) P(W_5/W_2 \bar{W}_3 W_4) + P(\bar{W}_2) P(W_3/\bar{W}_2) P(W_4/\bar{W}_2 W_3) P(W_5/\bar{W}_2 W_3 W_4) + P(W_2) P(W_3/W_2) P(W_4/W_2 W_3) P(W_5/W_2 W_3 W_4)$$

$$= \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} + \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} + \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3}$$

$$+ \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{2}{3} + \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3}$$

$$= \frac{8+4+4+4+16}{81} = \frac{36}{81}.$$

## EXERCISE 9.6

## NOTES

1. Two cards are drawn from a pack of cards in succession (with replacement). Find the probability that the first card is spade and the second is a black king.
2. A husband and a wife appear in an interview for two vacancies for the same post. The probability of husband's selection is  $\frac{2}{5}$  and that of wife is  $\frac{4}{5}$ . What is the probability that both of them will be selected?
3. A man wants to marry a girl having qualities : white complexion—the probability of getting such a girl is one in twenty, handsome dowry — the probability of getting this is one in thirty. Find the probability of his getting married to a white complexioned girl who may also bring handsome dowry.
4. (i) A problem in statistics is given to three students Ram, Shyam and Radheyshyam whose chances of solving it are 0.3, 0.5 and 0.6 respectively. Find the probability that the problem will be solved.  
 (ii) A problem in statistics is given to three students, A, B and C whose chances of solving it are  $\frac{1}{2}$ ,  $\frac{1}{3}$  and  $\frac{1}{4}$  respectively. Find the probability that the problem will be solved.  
 (iii) A problem in statistics is given to four students A, B, C and D whose chances of solving it are  $\frac{1}{2}$ ,  $\frac{1}{3}$ ,  $\frac{1}{4}$ ,  $\frac{1}{4}$  respectively. Find the probability that the problem is solved.
5. The probability of A winning a race is  $\frac{1}{5}$  and the probability of B winning the race is  $\frac{1}{6}$ . Find the probability that none will win the race.
6. (i) The odds in favour of 'A' solving a problem are 7 : 6, and the odds against 'B' solving the same problem are 11 : 8. What is the probability that the problem will be solved, if both try the problem?  
 (ii) A can solve 90% of problems given in a book and B can solve 70%. What is the probability that at least one of them will solve the problem, selected at random?
7. The odds in favour of first speaking the truth are 3 : 2. The odds in favour of second speaking the truth are 5 : 3. In what percentage of cases are they likely to contradict each other on an identical point?
8. What is the probability of throwing 6 with a die at least once in 3 attempts?
9. A can solve 75% of problems and B can solve 70%. What is the probability that at least one of them will solve the problem, selected at random.
10. Find the probability of drawing a heart on each of the two consecutive draws of one card from a well-shuffled pack of playing cards, if the card is not replaced after the first draw.
11. Find the probability of drawing a king, a queen and a knave in that order from a pack of playing cards in three consecutive draws of one card. The first two cards drawn are replaced.
12. A bag contains 10 red and 6 black balls, 4 balls are drawn successively one by one and are not replaced. What is the probability that these are alternatively of different colours?
13. A bag contains 13 balls numbered from 1 to 13. Suppose an even number is considered as a success. Two balls are drawn one by one without replacement. Find the probability of getting one success.
14. A student is trying to seek admission in either of the two colleges. The probability that he is admitted in first college is  $\frac{3}{5}$  and that in second college is  $\frac{1}{3}$ . Find the probability that he is admitted at least one of the colleges.

15. A bag contains 2 white balls and 3 black balls. Four persons, A, B, C, D in the order named each draws one ball and does not replace it. The first to draw a white ball receive ₹ 50. Determine their expectations.

Probability

[Hint. Let A, B, C, D themselves denote the probability of their winning.

$$\begin{aligned}\therefore P(A) &= \frac{2}{5} \\ P(B) &= \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10} \\ P(C) &= \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5} \\ P(D) &= \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{2}{2} = \frac{1}{10}\end{aligned}$$

$\therefore$  Their respective expectations are ₹  $\left(50 \times \frac{2}{5}\right)$ , ₹  $\left(50 \times \frac{3}{10}\right)$ , ₹  $\left(50 \times \frac{1}{5}\right)$ , ₹  $\left(50 \times \frac{1}{10}\right)$ .

NOTES

### Answers

1.  $\frac{1}{104}$
2.  $\frac{8}{25}$
3.  $\frac{1}{600}$
4. (i) 0.86 -
- (ii)  $\frac{3}{4}$
- (iii)  $\frac{13}{16}$
5.  $\frac{4}{5} \times \frac{5}{6} = \frac{2}{3}$
6. (i)  $\frac{181}{247}$
- (ii)  $\frac{97}{100}$
7.  $\left[\left(\frac{3}{5} \times \frac{3}{8}\right) + \left(\frac{2}{5} \times \frac{5}{8}\right)\right] 100\% = 47.5\%$
8.  $1 - \left(\frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}\right) = \frac{91}{216}$
9.  $\frac{37}{40}$
10.  $\frac{13}{52} \times \frac{12}{51} = \frac{1}{17}$
11. 0.000455
12.  $\frac{45}{364}$
13.  $\frac{6}{13} \times \frac{7}{12} + \frac{7}{13} \times \frac{6}{12} = \frac{7}{13}$
14.  $\frac{11}{15}$
15. ₹ 20, ₹ 15, ₹ 10, ₹ 5.

## 9.20. TOTAL PROBABILITY RULE

Let  $E_1, E_2, \dots, E_n$  be  $n$  mutually exclusive and exhaustive events, with non-zero probabilities, of a random experiment. If  $A$  be any arbitrary event of the sample space of the above random experiment with  $P(A) > 0$ , then

$$P(A) = P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + \dots + P(E_n)P(A/E_n).$$

**Proof.** Let  $S$  be the sample space of the random experiment.

Since  $E_1, E_2, \dots, E_n$  are exhaustive, we have

$$S = E_1 \cup E_2 \cup \dots \cup E_n.$$

Now

$$A = S \cap A = (E_1 \cup E_2 \cup \dots \cup E_n) \cap A$$

$\Rightarrow$

$$A = (E_1 \cap A) \cup (E_2 \cap A) \cup \dots \cup (E_n \cap A) \quad \dots(1)$$

Since  $E_1, E_2, \dots, E_n$  are mutually exclusive, we have

$$E_i \cap E_j = \phi \text{ for } i \neq j.$$

$$\text{Now } (E_i \cap A) \cap (E_j \cap A) = (E_i \cap E_j) \cap A = \phi \cap A = \phi$$



## NOTES

$\therefore E_1 \cap A, E_2 \cap A, \dots, E_n \cap A$  are also mutually exclusive.

By **addition theorem**, (1) implies

$$P(A) = P(E_1 \cap A) + P(E_2 \cap A) + \dots + P(E_n \cap A)$$

$$\Rightarrow P(A) = P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + \dots + P(E_n)P(A/E_n).$$

**Remark.** In practical problems, it is found convenient to write as follows:

$$P(A) = P(E_1 A \text{ or } E_2 A \text{ or } \dots \text{ or } E_n A) = P(E_1 A) + P(E_2 A) + \dots + P(E_n A)$$

$$\therefore P(A) = P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + \dots + P(E_n)P(A/E_n).$$

**Corollary.** In particular if  $n = 2$ , we have

$$P(A) = P(E_1)P(A/E_1) + P(E_2)P(A/E_2).$$

**Example 9.38.** A bag contains 4 red and 3 black balls. A second bag contains 2 red and 4 black balls. One bag is selected at random. From the selected bag, one ball is drawn. Find the probability that it is a red ball.

**Solution.** Let  $E_1$  and  $E_2$  be the events of selecting first bag and second bag respectively.

$$\therefore P(E_1) = \frac{1}{2}, P(E_2) = \frac{1}{2}$$

Let  $R$  be the event of drawing a red ball.

$$\therefore P(R/E_1) = P(\text{Red ball is drawn from first bag}) = \frac{4}{7}$$

$$\text{Similarly, } P(R/E_2) = \frac{2}{6} = \frac{1}{3}$$

Now,  $P(\text{selecting a red ball})$

$$= P(R) = P(E_1 R \text{ or } E_2 R) = P(E_1 R) + P(E_2 R)$$

$$= P(E_1)P(R/E_1) + P(E_2)P(R/E_2)$$

$$= \frac{1}{2} \times \frac{4}{7} + \frac{1}{2} \times \frac{1}{3} = \frac{2}{7} + \frac{1}{6} = \frac{19}{42}$$

**Example 9.39.** Two urns contains 2 red, 4 white and 3 red, 7 white balls. One of the urns is chosen at random and a ball is drawn from it. Find the probability that the ball drawn is (i) red (ii) white.

**Solution.** Let  $E_1$  and  $E_2$  be the events of choosing the first urn and the second urn respectively.

$$\therefore P(E_1) = \frac{1}{2}, P(E_2) = \frac{1}{2}$$

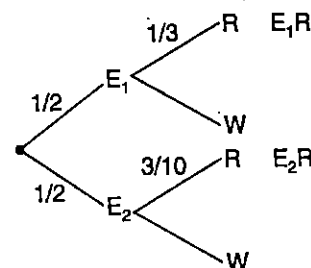
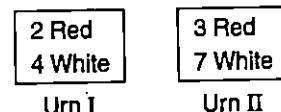
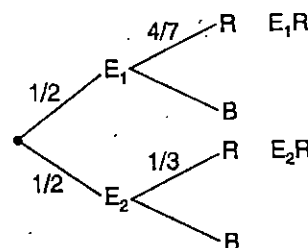
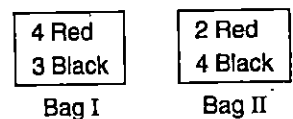
(i) Let  $R$  be the event of drawing a red ball.

$$\therefore P(R/E_1) = \frac{2}{2+4} = \frac{1}{3}$$

$$P(R/E_2) = \frac{3}{3+7} = \frac{3}{10}$$

Now  $P(\text{drawing a red ball}) = P(R)$

$$= P(E_1 R \text{ or } E_2 R) = P(E_1 R) + P(E_2 R)$$



$$= P(E_1)P(R/E_1) + P(E_2)P(R/E_2)$$

$$= \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{3}{10} = \frac{19}{60}$$

(ii) Let W be the event of drawing a white ball.

$$\therefore P(W/E_1) = \frac{4}{2+4} = \frac{2}{3}$$

$$P(W/E_2) = \frac{7}{3+7} = \frac{7}{10}$$

Now P(drawing a white ball)

$$= P(W) = P(E_1W \text{ or } E_2W) = P(E_1W) + P(E_2W)$$

$$= P(E_1)P(W/E_1) + P(E_2)P(W/E_2)$$

$$= \frac{1}{2} \times \frac{2}{3} + \frac{1}{2} \times \frac{7}{10} = \frac{41}{60}$$

**Example 9.40.** Suppose that 5 men out of 100 men and 25 women out of 1000 women are good orator. Assuming that there are equal number of men and women, find the probability, of choosing an orator.

**Solution.** Let  $E_1$  and  $E_2$  be the events of choosing a man and a woman respectively

$\therefore P(E_1) = \frac{1}{2}$  and  $P(E_2) = \frac{1}{2}$ , because there are equal number of men and women.

Let A be the event of choosing an orator

$\therefore P(A/E_1)$  = probability that a man is an orator

$$= \frac{5}{100} = \frac{1}{20}$$

$P(A/E_2)$  = probability that a woman is an orator

$$= \frac{25}{1000} = \frac{1}{40}$$

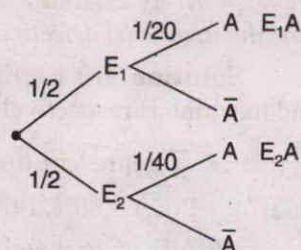
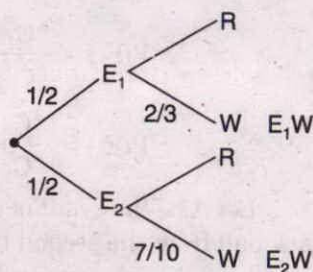
$\therefore P(\text{orator is chosen}) = P(A) = P(E_1A \text{ or } E_2A)$

$$= P(E_1A) + P(E_2A) = P(E_1)P(A/E_1) + P(E_2)P(A/E_2)$$

$$= \frac{1}{2} \times \frac{1}{20} + \frac{1}{2} \times \frac{1}{40} = \frac{3}{80}$$

**Example 9.41.** There are two bags. The first bag contains 5 white and 3 black balls and the second bag contains 3 white and 5 black balls. Two balls are drawn at random from the first bag and are put into the second bag, without noticing their colours. Then two balls are drawn from the second bag. Find the probability that these balls are white and black.

**Solution.** Let  $E_1$ ,  $E_2$  and  $E_3$  be the events of transferring 2 white, 1 white and 1 black, 2 black balls respectively from the first bag to second bag.



5 White
3 Black
Bag I

3 White
5 Black
Bag II

$$P(E_1) = \frac{{}^5C_2}{{}^8C_2} = \frac{10}{28} = \frac{5}{14}$$

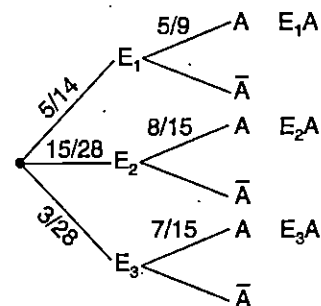
## NOTES

$$P(E_2) = \frac{{}^5C_1 \times {}^3C_1}{{}^8C_2} = \frac{5 \times 3}{28} = \frac{15}{28}$$

$$P(E_3) = \frac{{}^3C_2}{{}^8C_2} = \frac{3}{28}$$

Let A be the event of drawing one white and one black ball from the second bag.

$$\begin{aligned} \therefore P(A) &= P(E_1 A \text{ or } E_2 A \text{ or } E_3 A) \\ &= P(E_1 A) + P(E_2 A) + P(E_3 A) \\ &= P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + P(E_3)P(A/E_3) \\ &= \frac{5}{14} \times \frac{{}^5C_1 \times {}^5C_1}{{}^{10}C_2} + \frac{15}{28} \times \frac{{}^4C_1 \times {}^6C_1}{{}^{10}C_2} + \frac{3}{28} \times \frac{{}^3C_1 \times {}^7C_1}{{}^{10}C_2} \\ &= \frac{5}{14} \times \frac{5}{9} + \frac{15}{28} \times \frac{8}{15} + \frac{3}{28} \times \frac{7}{15} = \frac{673}{1260} \end{aligned}$$



**Example 9.42.** Two machines A and B produce respectively 60% and 40% of the total numbers of items of a factory. The percentages of defective output of these machines are respectively 2% and 5%. If an item is selected at random, what is the probability that the item is (i) defective (ii) non-defective?

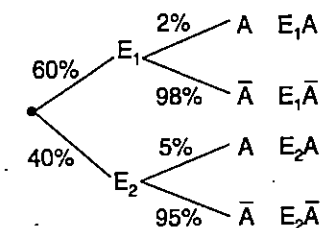
**Solution.** Let  $E_1$ ,  $E_2$  be the events of drawing an item produced by machine A and machine B respectively. Let A be the event of selecting a defective item.

$\therefore \bar{A}$  represent the event of selecting a non-defective item. We have

$$P(E_1) = 60\%, P(E_2) = 40\%.$$

$P(A/E_1)$  = probability that a defective item is produced by A = 2%.

$P(A/E_2)$  = probability that a defective item is produced by B = 5%.



(i) P(selected item is defective)

$$\begin{aligned} &= P(A) = P(E_1 A \text{ or } E_2 A) = P(E_1 A) + P(E_2 A) \\ &= P(E_1)P(A/E_1) + P(E_2)P(A/E_2) = (60\%)(2\%) + (40\%)(5\%) \\ &= \frac{60}{100} \times \frac{2}{100} + \frac{40}{100} \times \frac{5}{100} = \frac{320}{10000} = 0.032. \end{aligned}$$

(ii) P(selected item is non-defective)

$$\begin{aligned} &= P(\bar{A}) = P(E_1 \bar{A} \text{ or } E_2 \bar{A}) = P(E_1 \bar{A}) + P(E_2 \bar{A}) \\ &= P(E_1)P(\bar{A}/E_1) + P(E_2)P(\bar{A}/E_2) = (60\%)(98\%) + (40\%)(95\%) \\ &= \frac{60}{100} \times \frac{98}{100} + \frac{40}{100} \times \frac{95}{100} = \frac{9680}{10000} = 0.968. \end{aligned}$$

## EXERCISE 9.7

1. A bag contains 3 white and 2 black balls and another bag contains 2 white and 4 black balls. One bag is chosen at random. From the selected bag, one ball is drawn. Find the probability that the ball drawn is white.

## NOTES

2. Find the probability of drawing a one-rupee coin from a purse with two compartments one of which contains 3 fifty-paise coins and 2 one-rupee coins and other contains 2 fifty-paise coins and 3 one-rupee coins.
3. An unbiased coin is tossed. If the result is a head, a pair of unbiased dice is rolled and the sum of the numbers obtained is noted. If the result is a tail, a card from a well shuffled pack of eleven cards numbered 2, 3, 4, ....., 12 is picked and the number on the card is noted. What is the probability that the noted number is either 7 or 8?
4. In a bolt factory, machines A, B and C manufacture 25%, 35% and 40% of the total bolts. Of their outputs 5%, 4% and 2% are respectively defective bolts. A bolt is drawn at random from the output. What is the probability that the bolt drawn is defective?
5. We are given three boxes as follows:  
Box I has 10 light bulbs of which 4 are defective.  
Box II has 6 light bulbs of which 1 is defective.  
Box III has 8 light bulbs of which 3 are defective.  
We select a box at random and then draw a bulb at random. What is the probability that the bulb is defective?
6. A bag contains 6 white and 7 black balls, and another bag contains 4 white and 5 black balls. A ball is taken from the first bag and without seeing its colour is put in the second bag. A ball is taken from the latter. Find the probability that the ball drawn is white.
7. Bag A contains 5 white and 6 black balls. Bag B contains 4 white and 3 black balls. A ball is transferred from bag A to the bag B and then a ball is taken out of the second bag. Find the probability of this ball being black.
8. A bag contains 3 white and 5 black balls and a second bag contains 5 white and 3 black balls. One ball is transferred from first bag to the second and then a ball is drawn from the second bag. Find the probability that the ball drawn white.
9. An urn contains 10 white and 3 black balls, while another urn contains 3 white and 5 black balls. Two balls are drawn from the first urn and put in to the second urn and then a ball is drawn from the latter. What is the probability of drawing a white ball?

## Answers

- |                      |                    |                      |                    |
|----------------------|--------------------|----------------------|--------------------|
| 1. $\frac{7}{15}$    | 2. $\frac{1}{2}$   | 3. $\frac{193}{792}$ | 4. 0.345           |
| 5. $\frac{113}{360}$ | 6. $\frac{29}{65}$ | 7. $\frac{39}{88}$   | 8. $\frac{43}{72}$ |
| 9. $\frac{59}{130}$  |                    |                      |                    |

## I. BAYES' THEOREM

**9.21. MOTIVATION**

Let there be two or more urns, each containing some white balls and red balls. Suppose an urn is chosen at random and a ball is drawn from that chosen urn. By using *addition theorem* and *multiplication theorem*, we can find the probability of drawing a white ball (or red ball) from the urn chosen.

But in case, we are given that the ball drawn is white and we are interested in finding the probability of the event that the ball was drawn from the 1st urn or II<sup>nd</sup> urn, etc., then the situation is not the same as in the previous case. Now the probability of the drawn urn will depend upon the colour of the drawn ball.

To tackle this type of problems, Bayes' theorem is used. This theorem was enunciated by British mathematician Thomas Bayes in 1763.

Let  $E_1, E_2, \dots, E_n$  be  $n$  mutually exclusive and exhaustive events, with non-zero probabilities, of a random experiment. If  $A$  be any arbitrary event of the sample space of the above experiment with  $P(A) > 0$ , then

## NOTES

$$P(E_i/A) = \frac{P(E_i)P(A/E_i)}{\sum_{j=1}^n P(E_j)P(A/E_j)}, \quad 1 \leq i \leq n.$$

**Proof.** Let  $S$  be the sample space of the random experiment.

$$\therefore S = E_1 \cup E_2 \cup \dots \cup E_n \quad (\because E_1, E_2, \dots, E_n \text{ are exhaustive})$$

$$\begin{aligned} \text{Now } A &= S \cap A = (E_1 \cup E_2 \cup \dots \cup E_n) \cap A \\ &= (E_1 \cap A) \cup (E_2 \cap A) \cup \dots \cup (E_n \cap A). \end{aligned}$$

$$\begin{aligned} \therefore P(A) &= P(E_1 \cap A) + P(E_2 \cap A) + \dots + P(E_n \cap A)^* \\ &= P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + \dots + P(E_n)P(A/E_n) \end{aligned}$$

$$\text{or } P(A) = \sum_{j=1}^n P(E_j)P(A/E_j) \quad \dots(1)$$

$$\text{Now, } P(E_i/A) = \frac{P(E_i \cap A)}{P(A)}, \quad 1 \leq i \leq n$$

$$\therefore P(E_i/A) = \frac{P(E_i)P(A/E_i)}{\sum_{j=1}^n P(E_j)P(A/E_j)}, \quad 1 \leq i \leq n. \quad [\text{Using (1)}]$$

**Remark 1.** If  $n = 2$ , then

$$P(E_1/A) = \frac{P(E_1)P(A/E_1)}{P(E_1)P(A/E_1) + P(E_2)P(A/E_2)}$$

$$\text{and } P(E_2/A) = \frac{P(E_2)P(A/E_2)}{P(E_1)P(A/E_1) + P(E_2)P(A/E_2)}$$

**Example 9.43.** In 1988, there will be three candidates for the position of principal - A, B and C. The chances of their selection are in the proportion 4 : 2 : 3 respectively. The probability that A, if selected, will introduce co-education in the college is 0.3. The probability of B and C doing the same are respectively 0.5 and 0.8. What is the probability that there will be co-education in the college in 1988? Also find the probability that the co-education in the college was introduced by the principal B.

**Solution.** Let  $E_1, E_2, E_3$  be the events of selection of A, B, C as principal respectively. Let  $A$  be the event of introduction of co-education in the college.

$$\therefore P(E_1) = \frac{4}{4+2+3} = \frac{4}{9}, \quad P(E_2) = \frac{2}{4+2+3} = \frac{2}{9}$$

$$\text{and } P(E_3) = \frac{3}{4+2+3} = \frac{3}{9}.$$

$$\text{Also, } P(A/E_1) = \frac{3}{10}, \quad P(A/E_2) = \frac{5}{10}, \quad P(A/E_3) = \frac{8}{10}.$$

Now,  $P(\text{co-education is introduced in the college})$

$$\begin{aligned} &= P(A) = P(E_1 A \text{ or } E_2 A \text{ or } E_3 A) = P(E_1 A) + P(E_2 A) + P(E_3 A) \\ &= P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + P(E_3)P(A/E_3) \\ &= \left(\frac{4}{9} \times \frac{3}{10}\right) + \left(\frac{2}{9} \times \frac{5}{10}\right) + \left(\frac{3}{9} \times \frac{8}{10}\right) = \frac{46}{90} = \frac{23}{45}. \end{aligned}$$



By Bayes' theorem,

P(Co-education was introduced by the principal B)

$$= P(E_2/A) = \frac{P(E_2)P(A/E_2)}{P(A)} = \frac{\frac{2}{9} \times \frac{5}{10}}{\frac{23}{45}} = \frac{5}{23}$$

Probability

NOTES

**Example 9.44.** A manufacturing firm produces steel pipes in three plants with daily production volume of 500, 1000 and 2000 units respectively. According to past experience, it is known that the fraction of defective outputs produced by the three plants are respectively 0.005, 0.008, 0.010. If a pipe is selected from a day's total production and found to be defective, find out the probability that it came from the first plant.

**Solution.** Let  $E_1$ ,  $E_2$  and  $E_3$  be the events of drawing a pipe produced by first plant, second plant and third plant respectively. Let  $A$  be the event of drawing a defective pipe.

$$\therefore P(E_1) = \frac{500}{500 + 1000 + 2000} = \frac{1}{7}$$

$$P(E_2) = \frac{1000}{500 + 1000 + 2000} = \frac{2}{7} \quad \text{and} \quad P(E_3) = \frac{2000}{500 + 1000 + 2000} = \frac{4}{7}$$

Also  $P(A/E_1) = 0.005$ ,  $P(A/E_2) = 0.008$  and  $P(A/E_3) = 0.010$ .

The events  $E_1$ ,  $E_2$ ,  $E_3$  are mutually exclusive and exhaustive.

By Bayes' theorem,  $P(\text{Plant I produced the defective pipe}) = P(E_1/A)$

$$\begin{aligned} &= \frac{P(E_1)P(A/E_1)}{P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + P(E_3)P(A/E_3)} \\ &= \frac{\frac{1}{7}(0.005)}{\frac{1}{7}(0.005) + \frac{2}{7}(0.008) + \frac{4}{7}(0.010)} \\ &= \frac{0.005}{0.005 + 0.016 + 0.040} = \frac{0.005}{0.061} = \frac{5}{61} \end{aligned}$$

**Example 9.45.** An insurance company insured 2000 scooter drivers, 4000 car drivers and 6000 truck drivers. The probability of an accident involving a scooter driver, car driver and a truck driver is 0.01, 0.03 and 0.15 respectively. One of the insured drivers meets with an accident. What is the probability that he is a car driver?

**Solution.** Let  $E_1$ ,  $E_2$ ,  $E_3$  be the events of drawing scooter driver, car driver, truck driver respectively.

Total number of drivers = 2000 + 4000 + 6000 = 12000

$$\therefore P(E_1) = \frac{2000}{12000} = \frac{1}{6}, \quad P(E_2) = \frac{4000}{12000} = \frac{1}{3} \quad \text{and} \quad P(E_3) = \frac{6000}{12000} = \frac{1}{2}$$

Let  $A$  be the event of getting an accident.

$$\therefore P(A/E_1) = 0.01, \quad P(A/E_2) = 0.03 \quad \text{and} \quad P(A/E_3) = 0.15$$

The events  $E_1$ ,  $E_2$  and  $E_3$  are mutually exclusive and exhaustive.

∴ By Bayes' theorem

$$P(\text{accident involved car driver}) = P(E_2/A)$$

$$= \frac{P(E_2)P(A/E_2)}{P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + P(E_3)P(A/E_3)}$$

$$= \frac{\frac{1}{3} \times 0.03}{\frac{1}{6} + 0.01 + \frac{1}{3} \times 0.03 + \frac{1}{2} \times 0.15} = \frac{0.01}{0.526} = \frac{0.06}{0.52} = \frac{3}{26}$$

## NOTES

### EXERCISE 9.8

1. A bag contains 4 black and 1 white balls and another bag contains 5 black and 4 white balls. One bag is chosen and a ball is drawn. If the ball drawn is black, find the probability that it is drawn from the first bag.
2. Two urns contain 4 white, 6 blue and 4 white, 5 blue balls. One of the urns is selected at random and a ball is drawn. If the ball drawn is white, find the probability that it is drawn from the second urn.
3. Assume that a factory has two machines. Past records shows that machine I produces 60% of the items of output and machine II produces 40% of the items. Further, 2% of the items produced by machine I were defective and only 1% produced by machine II were defective. If a defective item is drawn at random, what is the probability that it was produced by machine I?
4. In a factory, machines A, B and C produces 40%, 40% and 20% respectively. Of the total of their output 1%, 1% and 3% are defective. An item is drawn at random from the total production and found to be defective. Find the probability that this item is produced by the machine C.
5. A manufacturing firm produces pipes in two plants with daily production volume of 5000 and 7000 units respectively. According to past experience, it is known that the fraction of defective outputs produced by the plants are 0.01 and 0.02 respectively. If a pipe is selected at random from a day's total production and found to be defective, find out the probability that it came from the second plant.

### Answers

- |          |          |        |        |
|----------|----------|--------|--------|
| 1. 36/61 | 2. 10/19 | 3. 3/4 | 4. 3/7 |
| 5. 14/19 |          |        |        |

## 9.22. CRITICISM OF CLASSICAL APPROACH OF PROBABILITY

Though the classical approach of measuring probability seems to be quite simple and straight forward, but this approach is subjected to certain points of criticism.

In defining probability of an event, we assume that all the possible outcomes are equally likely. This means that all the possible outcomes of an experiment have equal chances of being occurred. In other words, the probability of occurrence of each outcome is equal. Thus, in classical approach, we define probability of an event in terms of probabilities of various outcomes of the experiment. Thus, this definition of probability is circular in nature.



The classical approach is based on abstract reasoning and is suitable under ideal conditions. For example, we say that in the experiment of throwing a die, the probability of getting 2 is  $1/6$ . But this will be true if the die is unbiased i.e. it is perfect. In practice, perfection is not achieved. Thus, this approach is not realistic in nature.

## NOTES

### 9.23. EMPIRICAL APPROACH OF PROBABILITY

This approach is based upon repetitive experiments under uniform conditions. Suppose a coin is perfectly balanced and we toss it 100 times. In 100 tosses, we may get head 56 times. Again if we toss this coin 1000 times, we may get head 519 times. Again if we toss this coin 10,000 times, we may get head 5085 times. In these experiments, we see that the ratio  $56/100$ ,  $519/1000$ ,  $5085/10000$  is tending toward  $1/2$ , which should be the probability of getting head in any toss of the coin. In empirical approach, the probability of an event is defined in terms of a ratio of the type explained above.

If an experiment is repeated  $n$  times under uniform conditions and an event  $E$  occurs ' $m$ ' times, then the probability of the event  $E$  is defined as

$$P(E) = \lim_{n \rightarrow \infty} \frac{m}{n}$$

In the definition, ' $\lim_{n \rightarrow \infty}$ ' emphasizes the fact that  $n$  must be very large. In the

real mathematical sense, we cannot measure  $\lim_{n \rightarrow \infty} \frac{m}{n}$ , because we cannot repeat any

experiment infinitely many times. Thus, in this approach, we content ourselves by assuming that  $n$  takes large and practically possible values. If we toss a perfect coin two times, it is not expected that we shall definitely get one head and one tail. But if the same coin is tossed 5000 times, we may get about 2500 heads. Thus, the probability defined in this approach is a long run concept, because to find the probability of an event, we have to repeat our experiment a large number of times. In any experiment, we shall have  $m \leq n$ .

$$\therefore P(E) = \lim_{n \rightarrow \infty} \frac{m}{n} \text{ implies } 0 \leq P(E) \leq 1.$$

The empirical approach of probability is based on experience.

**Limitations.** The computation of empirical probability requires repetition of experiment, a very large number of times. This restricts the suitability of this approach. In many cases, the experiment may not be repeated a large number of times. If  $E$  be the event that a particular student secures 70% marks or more in all the examinations given to him in a particular year, then this experiment cannot be repeated many number of times. This approach is also not applicable to experiments which are not expected to occur frequently in future.

## NOTES

---

## 9.24. SUBJECTIVE APPROACH OF PROBABILITY

---

The classical and empirical approaches of probability are *objective* in nature. In the subjective approach, the probability of an event is considered as a measure of one's confidence in the occurrence of that particular event. The probability of the event that the student 'A' will pass the examination cannot be calculated by any of the above discussed objective approaches. The events of his passing and failing are not equally likely cases. Had these cases been equally likely, we could have used classical approach and said that the probability is  $\frac{1}{2}$ . In this case, the experiment is such that it cannot be repeated under uniform conditions. Thus, the empirical approach also fails to comment upon the probability of this event. In such cases, the subjective approach is found useful. In subjective approach, the probability of an event represent the degree of faith which a rational person reposes in the occurrence of that certain event. The degree of faith will depend upon his judgement, personal outlook, etc. In this approach, the probability of a event, differ from person to person and that is why it is called subjective approach. In this approach, the probability of an event also suffer from personal bias of its estimator.

---

## 9.25. SUMMARY

---

- When we perform experiments in science and engineering, repeatedly under very nearly identical conditions, we get almost the same result. Such experiments are called **deterministic experiments**.  
There also exist experiments in which the results may not be essentially the same even if the experiment is performed under very nearly identical conditions. Such experiments are called **random experiments**.
- The **sample space** of a random experiment is defined as the set of all possible outcomes of the experiment. The possible outcomes are called **sample points**.
- A **Tree diagram** is a device used to enumerate all the logical possibilities of a sequence of steps where each step can occur in a finite number of ways.
- An **event** is defined as a subset of the sample space. An event is called an **elementary (or simple) event** if it contains only one sample point. In the experiment of rolling a die, the event A of getting '3' is a simple event. We write  $A = \{3\}$ . An event is called an **impossible event** if it can never occur.

---

## 9.26. REVIEW EXERCISES

---

1. Explain the fundamental concepts of 'Probability'.
2. Define 'probability'.
3. Write the fundamental concepts of probability calculation.
4. Define 'probability' and explain its importance in Statistics.
5. Explain the term 'Mutually Exclusive Events' by taking some examples.
6. What is conditional probability? Explain with the help of an example.
7. Define probability and explain the Addition law of probability giving suitable examples.



8. Explain what do you understand by the term 'probability'. State and prove the addition and multiplication theorems of probability.
9. Explain short notes on any two:
  - (i) Dependent and independent events
  - (ii) Mutually exclusive and equally likely events
  - (iii) Simple and compound events.
10. Explain the Multiplication Theorem of Probability with suitable example.
11. Explain Bayes' theorem with the help of an example.
12. Define probability in different ways. Giving their merits and demerits by examples. State which is the best.
13. Discuss in detail the Classical and Empirical approaches to probability.
14. Explain the various approaches to probability.

**NOTES**



# 10. PROBABILITY DISTRIBUTIONS

## (Binomial, Poisson, Normal Distributions)

## NOTES

## STRUCTURE

- 10.1. Introduction
- 10.2. Empirical Distribution

**I. Binomial Distribution**

- 10.3. Introduction
- 10.4. Conditions
- 10.5. Binomial Variable
- 10.6. Binomial Probability Function
- 10.7. Binomial Frequency Distribution

**II. Property of Binomial Distribution**

- 10.8. The Shape of B.D.
- 10.9. The Limiting Case of B.D.
- 10.10. Mean of B.D.
- 10.11. Variance and S.D. of B.D.
- 10.12.  $\gamma_1$  And  $\gamma_2$  of B.D.
- 10.13. Recurrence Formula for B.D.
- 10.14. Fitting of a Binomial Distribution

**III. Poisson Distribution**

- 10.15. Introduction
- 10.16. Conditions
- 10.17. Poisson Variable
- 10.18. Poisson Probability Function
- 10.19. Poisson Frequency Distribution

**IV. Property of Poisson Distribution**

- 10.20. The Shape of P.D.
- 10.21. Special Usefulness of P.D.
- 10.22. Mean of P.D.
- 10.23. Variance and S.D. of P.D.
- 10.24.  $\gamma_1$  and  $\gamma_2$  of P.D.
- 10.25. Recurrence Formula for P.D.
- 10.26. Fitting of a Poisson Distribution

**V. Normal Distribution**

- 10.27. Introduction
- 10.28. Probability Function of Continuous Random Variable
- 10.29. Normal Distribution
- 10.30. Definition
- 10.31. Standard Normal Distribution
- 10.32. Area Under Normal Curve
- 10.33. Table of Area Under Standard Normal Curve
- 10.34. Properties of Normal Distribution
- 10.35. Fitting of a Normal Distribution
- 10.36. Summary
- 10.37. Review Exercises

## 10.1. INTRODUCTION

We know that a real valued function defined on the sample space of a random experiment is called a *random variable*. A random variable is either discrete or continuous.

Let  $x$  be a discrete random variable assuming values  $x_1, x_2, x_3, \dots, x_n$  corresponding to the various outcomes of a random experiment. If the probability of occurrence of  $x = x_i$  is  $P(x_i) = p_i, 1 \leq i \leq n$  such that  $p_1 + p_2 + p_3 + \dots + p_n = 1$ , then the function,  $P(x_i) = p_i, 1 \leq i \leq n$  is called the *probability function* of the random variable  $x$  and the set  $\{P(x_1), P(x_2), P(x_3), \dots, P(x_n)\}$  is called the *probability distribution* of  $x$ .

## NOTES

## 10.2. EMPIRICAL DISTRIBUTION

Let  $x$  be a discrete random variable assuming values  $x_1, x_2, \dots, x_n$  corresponding to various outcomes of a random experiment. Let this random experiment be repeated  $N$  times. Let the random variable  $x$  take values  $x_1, x_2, \dots, x_n$  with respective frequencies  $f_1, f_2, \dots, f_n$  where  $f_1 + f_2 + \dots + f_n = N$ .

The distribution

$x$	$x_1$	$x_2$	....	$x_n$
$f$	$f_1$	$f_2$	....	$f_n$

is called an **empirical**

**distribution.**

**Illustration.** Let the random experiment be of tossing of two coins.

$\therefore S = \{HH, HT, TH, TT\}$ .

Let  $x$  be random variable "square of number of tails," then  $x$  takes the values  $0^2 = 0, 1^2 = 1, 2^2 = 4$ . Let this random experiment be repeated 100 times and let the observed frequencies be as follows:

HH	HT	TH	TT
↓	↓	↓	↓
24	27	23	26

$\therefore$  The empirical distribution corresponding to above experiment is

$x$	0 (HH)	1 (HT, TH)	4 (TT)
$f$	24	50 (= 27 + 23)	26

Now we shall consider three very important types of probability distributions.

### I. BINOMIAL DISTRIBUTION

## 10.3. INTRODUCTION

The binomial distribution is a particular type of probability distribution. This was discovered by **James Bernoulli (1654–1705)** in the year 1700. This distribution mainly deals with attributes. An attribute is either present or absent with respect to

elements of a population. For example, if a coin is tossed, we get either *head* or *tail*. The workers of a factory may be classified as *skilled* and *unskilled*. An item of a population of articles produced in a firm may be either defective or non-defective.

## NOTES

### 10.4. CONDITIONS

The following conditions are essential for the applicability of binomial distribution:

(i) **The random experiment is performed for a finite and fixed number of trials.** If in an experiment, a coin is tossed repeatedly or a ball is drawn from an urn repeatedly, then each toss or draw is called a **trial**. For example, if a coin is tossed 6 times, then this experiment has 6 trials. The number of trials in an experiment is generally denoted by ' $n$ '.

(ii) **The trials are independent.** By this we mean that the result of a particular trial is not going to effect the result of any other trial. For example, if a coin is tossed or a die is thrown, the trials would be independent. If from a pack of playing cards, some draws of one card are made without replacing the cards, then the trials would not be independent. But if the card drawn is replaced before the next draw, then the trials would be independent.

(iii) **Each trial must result in either "success" or "failure".** In other words, in every trial, there should be only two possible outcomes i.e., *success* or *failure*. For example, if a coin is tossed, then either *head* or *tail* is observed. Similarly, if an item is examined, it is either *defective* or *non-defective*.

(iv) **The probability of success in each trial is same.** In other words, this condition requires that the probability of *success* should not change in different trials. For example, if a sample of two items is drawn, then the probability of exactly one being defective will be constant in different trials provided the items are replaced before the next draw.

### 10.5. BINOMIAL VARIABLE

A random variable which counts the number of successes in a random experiment with trials satisfying above four conditions is called a **Binomial variable**.

For example, if a coin is tossed 5 times and the event of getting head is *success*, then the possible values of the binomial variable are 0, 1, 2, 3, 4, 5. This is so, because, the minimum number of successes is 0 and maximum number is 5.

### 10.6. BINOMIAL PROBABILITY FUNCTION

When a fair coin is tossed, we have only two possibilities: head and tail. Let us call the occurrence of head as 'success'. Therefore, the occurrence of tail would be a 'failure'. Let this coin be tossed 5 times. Suppose we are interested in finding the probability of getting 4 heads and 1 tail i.e., of getting 4 successes. If S and F denote 'success' and 'failure' in a trial respectively, then there are  ${}^5C_4 = 5$  ways of having 4 successes.

These are: SSSSF, SSSFS, SSFSS, SFSSS, FSSSS.

The probability of getting 4 successes in each case is  $\left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)$ , because the trials are independent.

## NOTES

$\therefore$  By using *addition theorem*, the required probability of having 4 successes is  ${}^5C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)$ , which is equal to  $\frac{5}{32}$ . Now we shall generalise this method of finding the probabilities for different values of *binomial variables*.

Let a random experiment satisfying the conditions of **binomial distribution** be performed. Let the number of trials in the experiment be  $n$ . Let  $p$  denote the probability of *success* in any trial.

$\therefore$  Probability of failure,  $q = 1 - p$

Let  $x$  denote the binomial variable corresponding to this experiment.

$\therefore$  The possible values of  $x$  are  $0, 1, 2, \dots, n$ .

If there are  $r$  successes in  $n$  trials, then there would be  $n - r$  failures. One of the ways in which  $r$  successes may occur is

$\begin{array}{cc} \text{SS} \dots \text{SS} & \text{FF} \dots \text{F} \\ \underbrace{\hspace{2cm}} & \underbrace{\hspace{2cm}} \\ r \text{ times} & n - r \text{ times} \end{array}$

where S and F denote success and failure in trials.

Now,  $P(\text{SS} \dots \text{SFF} \dots \text{F}) = P(S)P(S) \dots P(S)P(F)P(F) \dots P(F)$

( $\because$  the trials are independent)

$$= p.p \dots p.q.q \dots q = p^r q^{n-r}.$$

We know that  ${}^nC_r$  is the number of combinations of  $n$  things taking  $r$  at a time. Therefore, the number of ways in which  $r$  successes can occur in  $n$  trials is equal to the number of ways of choosing  $r$  trials (for successes) out of total  $n$  trials i.e., it is  ${}^nC_r$ . Therefore, there are  ${}^nC_r$  ways in which we get  $r$  successes and  $n - r$  failures and the probability of occurrence of each of these ways is  $p^r q^{n-r}$ . Hence the probability of  $r$  successes in  $n$  trials in any order is

$$P(x=r) = p^r q^{n-r} + p^r q^{n-r} + \dots {}^nC_r \text{ terms} \quad (\text{By addition theorem})$$

or 
$$P(x=r) = {}^nC_r p^r q^{n-r}, \quad 0 \leq r \leq n.$$

This is called the **binomial probability function**. The corresponding **binomial distribution** is

$x$	0	1	2	.....	$n$
$P(x)$	${}^nC_0 p^0 q^n$	${}^nC_1 p^1 q^{n-1}$	${}^nC_2 p^2 q^{n-2}$	.....	${}^nC_n p^n q^0$

The probabilities of 0 success, 1 success, 2 successes, .....,  $n$  successes are respectively the 1st, 2nd, 3rd, .....,  $(n+1)$ th terms in the binomial expansion of  $(q+p)^n$ . This is why, it is called **binomial distribution**.

## 10.7. BINOMIAL FREQUENCY DISTRIBUTION

If a random experiment, satisfying the requirements of binomial distribution, is repeated  $N$  times, then the expected frequency of getting  $r$  ( $0 \leq r \leq n$ ) successes is given by

$$N.P(x=r) = N.{}^nC_r p^r q^{n-r}, \quad 0 \leq r \leq n.$$

The frequencies of getting 0 success, 1 success, 2 successes, .....,  $n$  successes are respectively the 1st, 2nd, 3rd, .....,  $(n + 1)$ th terms in the expansion of  $N(q + p)^n$ .

**Example 10.1.** An unbiased coin is tossed six times. Find the probability of obtaining:

## NOTES

- (i) no head
- (ii) all heads
- (iii) at least one head i.e., one or more heads
- (iv) exactly 4 heads
- (v) less than 3 heads
- (vi) more than 4 heads
- (vii) more than 4 and less than 6 heads
- (viii) more than 6 heads.

**Solution.** Let  $p$  be the probability of success i.e., of getting head in the toss of the coin.

$$\therefore n = 6, p = \frac{1}{2} \text{ and } q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}.$$

Let  $x$  be the binomial variable 'no. of successes'.

By **Binomial distribution**,  $P(x = r) = {}^nC_r p^r q^{n-r}$ ,  $0 \leq r \leq n$ .

$$\therefore P(x = r) = {}^6C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{6-r} = {}^6C_r \left(\frac{1}{2}\right)^6 = {}^6C_r \frac{1}{64}, \quad 0 \leq r \leq n.$$

$$(i) P(\text{no head}) = P(x = 0) = {}^6C_0 \frac{1}{64} = \frac{1}{64}$$

$$(ii) P(\text{all heads}) = P(x = 6) = {}^6C_6 \frac{1}{64} = \frac{1}{64}$$

$$(iii) P(\text{at least one head}) = 1 - P(\text{no head}) = 1 - \frac{1}{64} = \frac{63}{64} \quad [\text{Using part (i)}]$$

$$(iv) P(\text{exactly 4 heads}) = P(x = 4) = {}^6C_4 \frac{1}{64} = \frac{15}{64}$$

$$(v) P(\text{less than 3 heads}) = P(x < 3) = P(x = 0 \text{ or } 1 \text{ or } 2)$$

$$= P(x = 0) + P(x = 1) + P(x = 2)$$

$$= {}^6C_0 \frac{1}{64} + {}^6C_1 \frac{1}{64} + {}^6C_2 \frac{1}{64}$$

$$= (1 + 6 + 15) \frac{1}{64} = \frac{22}{64} = \frac{11}{32}$$

$$(vi) P(\text{more than 4 heads})$$

$$= P(x > 4) = P(x = 5 \text{ or } 6) = P(x = 5) + P(x = 6)$$

$$= {}^6C_5 \frac{1}{64} + {}^6C_6 \frac{1}{64} = (6 + 1) \frac{1}{64} = \frac{7}{64}$$

$$(vii) P(\text{more than 4 heads and less than 6 heads})$$

$$= P(4 < x < 6) = P(x = 5) = {}^6C_5 \frac{1}{64} = \frac{6}{64} = \frac{3}{32}.$$

$$(viii) P(\text{more than 6 heads}) = P(x > 6) = 0. \quad (\because \text{The event is impossible.})$$

**Example 10.2.** A die is thrown 4 times. Getting a number greater than 2 is a success. Find the probability of getting (i) exactly 1 success (ii) less than 3 successes (iii) more than 3 successes.

**Solution.** Let  $p$  be the probability of success i.e., of getting number greater than 2 in the throw of a die.



$$\therefore n = 4, p = \frac{4}{6} = \frac{2}{3} \text{ and } q = 1 - p = 1 - \frac{2}{3} = \frac{1}{3}.$$

Let  $x$  be the binomial variable "no. of successes".

By **Binomial distribution**,  $P(x = r) = {}^nC_r p^r q^{n-r}$ ,  $0 \leq r \leq n$ .

$$\therefore P(x = r) = {}^4C_r \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{4-r}, 0 \leq r \leq 4.$$

$$(i) P(\text{exactly 1 success}) = P(x = 1)$$

$$= {}^4C_1 \left(\frac{2}{3}\right)^1 \left(\frac{1}{3}\right)^{4-1} = 4 \times \frac{2}{3} \times \frac{1}{27} = \frac{8}{81}.$$

$$(ii) P(\text{less than 3 successes}) = P(x < 3)$$

$$= P(x = 0 \text{ or } x = 1 \text{ or } x = 2)$$

$$= P(x = 0) + P(x = 1) + P(x = 2)$$

(Addition theorem for m.e. events)

$$= {}^4C_0 \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^{4-0} + {}^4C_1 \left(\frac{2}{3}\right)^1 \left(\frac{1}{3}\right)^{4-1} + {}^4C_2 \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^{4-2}$$

$$= \left(1 \times 1 \times \frac{1}{81}\right) + \left(4 \times \frac{2}{3} \times \frac{1}{27}\right) + \left(6 \times \frac{4}{9} \times \frac{1}{9}\right)$$

$$= \frac{1 + 8 + 24}{81} = \frac{33}{81}.$$

$$(iii) P(\text{more than 3 successes}) = P(x > 3)$$

$$= P(x = 4) = {}^4C_4 \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^0 = 1 \times \frac{16}{81} \times 1 = \frac{16}{81}.$$

**Example 10.3.** There are 20% chances for a worker of an industry to suffer from an occupational disease. 50 workers were selected at random and examined for the occupational disease. Find the probability that (i) only one worker is found suffering from the disease; (ii) more than 3 are suffering from the disease; (iii) none is suffering from the disease.

**Solution.** Let  $p$  be the probability of success i.e., a worker is suffering from disease.

$$\therefore n = 50, p = \frac{20}{100} = \frac{1}{5} \text{ and } q = 1 - p = 1 - \frac{1}{5} = \frac{4}{5}.$$

Let  $x$  be the binomial variable, "no. of successes".

By **Binomial distribution**,  $P(x = r) = {}^nC_r p^r q^{n-r}$ ,  $0 \leq r \leq n$ .

$$\therefore P(x = r) = {}^{50}C_r \left(\frac{1}{5}\right)^r \left(\frac{4}{5}\right)^{50-r}, 0 \leq r \leq 50.$$

$$(i) P(\text{only one is suffering}) = P(x = 1)$$

$$= {}^{50}C_1 \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^{49} = 50 \times \frac{1}{5} \times \left(\frac{4}{5}\right)^{49} = 10 \left(\frac{4}{5}\right)^{49}.$$

$$(ii) P(\text{more than 3 are suffering}) = P(x > 3) = 1 - P(x \leq 3)$$

$$= 1 - P(x = 0 \text{ or } x = 1 \text{ or } x = 2 \text{ or } x = 3)$$

$$= 1 - \{P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)\}$$

## NOTES

## NOTES

$$\begin{aligned}
&= 1 - \left\{ {}^{50}C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{50} + {}^{50}C_1 \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^{49} + {}^{50}C_2 \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^{48} + {}^{50}C_3 \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^{47} \right\} \\
&= 1 - \left\{ 1 \times 1 \times \left(\frac{4}{5}\right)^{50} + 50 \times \frac{1}{5} \times \left(\frac{4}{5}\right)^{49} \right. \\
&\quad \left. + \left(\frac{50 \times 49}{1 \times 2}\right) \times \left(\frac{1}{5}\right)^2 \times \left(\frac{4}{5}\right)^{48} + \left(\frac{50 \times 49 \times 48}{1 \times 2 \times 3}\right) \times \left(\frac{1}{5}\right)^3 \times \left(\frac{4}{5}\right)^{47} \right\} \\
&= 1 - \frac{4^{47}}{5^{50}} \{64 + (50 \times 16) + (1225 \times 4) + (19600 \times 1)\} = 1 - \left(\frac{4^{47}}{5^{50}} \times 25364\right).
\end{aligned}$$

(iii)  $P(\text{none is suffering}) = P(x = 0)$

$$= {}^{50}C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{50} = 1 \times 1 \times \left(\frac{4}{5}\right)^{50} = \left(\frac{4}{5}\right)^{50}.$$

**Example 10.4.** There are 64 beds in a garden and 3 seeds of a particular type of flower are sown in each bed. The probability of a flower being white is  $1/4$ . Find the number of beds with 3, 2, 1 and 0 white flowers.

**Solution.** Let  $p$  be the probability of success i.e., the flower is white.

$$\therefore n = 3, N = 64, p = \frac{1}{4}, q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}.$$

Let  $x$  be the binomial variable, 'no. of successes'.

By Binomial distribution,  $P(x = r) = {}^nC_r p^r q^{n-r}$ ,  $0 \leq r \leq n$ .

$$\therefore P(x = r) = {}^3C_r \left(\frac{1}{4}\right)^r \left(\frac{3}{4}\right)^{3-r}, 0 \leq r \leq 3.$$

$$\therefore P(3 \text{ white flowers}) = P(x = 3) = {}^3C_3 \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^0 = \frac{1}{64}$$

$$P(2 \text{ white flowers}) = P(x = 2) = {}^3C_2 \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^1 = \frac{9}{64}$$

$$P(1 \text{ white flower}) = P(x = 1) = {}^3C_1 \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^2 = \frac{27}{64}$$

$$P(\text{no white flower}) = P(x = 0) = {}^3C_0 \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^3 = \frac{27}{64}$$

$\therefore$  Number of beds with 3 white flowers

$$= 64 \times P(x = 3) = 64 \times \frac{1}{64} = 1$$

Number of beds with 2 white flowers

$$= 64 \times P(x = 2) = 64 \times \frac{9}{64} = 9$$

Number of beds with 1 white flower

$$= 64 \times P(x = 1) = 64 \times \frac{27}{64} = 27$$

Number of beds with no white flower

$$= 64 \times P(x = 0) = 64 \times \frac{27}{64} = 27.$$



**Example 10.5.** In an experiment, a fair die is thrown 6 times. The event of occurring number greater than 4 is the 'success' of the experiment. Find the Binomial distribution of the experiment. If the same experiment is repeated 3645 times, find also the Binomial frequency distribution.

**Solution.** Let  $p$  be the probability of success i.e., of getting number greater than 4

$$\therefore n = 6, p = \frac{2}{6} = \frac{1}{3} \text{ and } q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}.$$

Let  $x$  be the binomial variable, "no. of successes".

By **Binomial distribution**,  $P(x = r) = {}^nC_r p^r q^{n-r}$ ,  $0 \leq r \leq n$ .

$$\therefore P(x = r) = {}^6C_r \left(\frac{1}{3}\right)^r \left(\frac{2}{3}\right)^{6-r}, 0 \leq r \leq 6.$$

$$\text{Now } P(x = 0) = {}^6C_0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^6 = 1 \times 1 \times \frac{64}{729} = \frac{64}{729}$$

$$P(x = 1) = {}^6C_1 \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^5 = 6 \times \frac{1}{3} \times \frac{32}{243} = \frac{192}{729}$$

$$P(x = 2) = {}^6C_2 \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^4 = 15 \times \frac{1}{9} \times \frac{16}{81} = \frac{240}{729}$$

$$P(x = 3) = {}^6C_3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^3 = 20 \times \frac{1}{27} \times \frac{8}{27} = \frac{160}{729}$$

$$P(x = 4) = {}^6C_4 \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^2 = 15 \times \frac{1}{81} \times \frac{4}{9} = \frac{60}{729}$$

$$P(x = 5) = {}^6C_5 \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right)^1 = 6 \times \frac{1}{243} \times \frac{2}{3} = \frac{12}{729}$$

$$P(x = 6) = {}^6C_6 \left(\frac{1}{3}\right)^6 \left(\frac{2}{3}\right)^0 = 1 \times \frac{1}{729} \times 1 = \frac{1}{729}$$

$\therefore$  The required distributions are as follows:

Binomial Distribution		Binomial Frequency Distribution	
$x$	$P(x)$	$x$	$N \cdot P(x) = 3645 \times P(x)$
0	$\frac{64}{729}$	0	$3645 \times \frac{64}{729} = 320$
1	$\frac{192}{729}$	1	$3645 \times \frac{192}{729} = 960$
2	$\frac{240}{729}$	2	$3645 \times \frac{240}{729} = 1200$
3	$\frac{160}{729}$	3	$3645 \times \frac{160}{729} = 800$
4	$\frac{60}{729}$	4	$3645 \times \frac{60}{729} = 300$
5	$\frac{12}{729}$	5	$3645 \times \frac{12}{729} = 60$
6	$\frac{1}{729}$	6	$3645 \times \frac{1}{729} = 5$

## NOTES

## EXERCISE 10.1

## NOTES

1. If a coin is tossed six times, what is the probability of obtaining four or more heads?
2. An unbiased coin is tossed 8 times. Find the probability of obtaining (i) exactly 2 heads (ii) more than 2 heads (iii) all heads.
3. The incidence of occupational disease in a factory is such that the workers have a 25% chances of suffering from it. What is the probability that out of 6 workmen, 4 or more contact the disease?
4. Out of 800 families with 4 children each, what percentage would be expected to have (a) 2 boys and 2 girls (b) at least one boy (c) no girl (d) at most 2 girls? Assume equal probabilities for boys and girls.
5. In a certain town, 20% of population is literate, and assume that 200 investigators takes a sample of 10 individuals to see whether they are literate. How many investigators would you expect to report that 3 persons or less are literate in their samples?
6. Eight coins are tossed at a time, 256 times. Find the expected frequencies of 0, 1, 2, 3 successes (getting head).
7. During war, 1 ship out of 9 was sunk on an average in making a certain voyage. What was the probability that exactly 3 out of a convoy of 6 ships would arrive safely?
8. The probability of a man hitting a target is  $\frac{1}{3}$ . How many least number of times, must he fire so that the probability of hitting the target at least once is more than 90%?
9. An unbiased coin is tossed 10 times. Find the probability of getting exactly 5 heads?
10. The probability that a student will be graduate is 0.4. Determine the probability that out of 5 students (i) none (ii) one (iii) at least one (iv) all will be graduate.

## Answers

- |  |   |                      |
|--|---|----------------------|
| 1. $\frac{11}{32}$   | 2. $\frac{7}{64}, \frac{219}{256}, \frac{1}{256}$                           | 3. $\frac{77}{2048}$ |
| 4. 37.5%, 93.75%, 6.25%, 68.75%                                      | 5. 176  |                      |
| 6. 1, 8, 28, 56  | 7. ${}^6C_3 \left(\frac{8}{9}\right)^3 \left(\frac{1}{9}\right)^3 = 0.0193$ |                      |
| 8. $1 - \left(\frac{2}{3}\right)^n > \frac{9}{10} \Rightarrow n = 6$ | 9. 0.2461   |                      |
| 10. (i) 0.07776  | (ii) 0.2592   |                      |
| (iii) 0.92224  | (iv) 0.01024  |                      |

## II. PROPERTIES OF BINOMIAL DISTRIBUTION

## 10.8. THE SHAPE OF B.D.

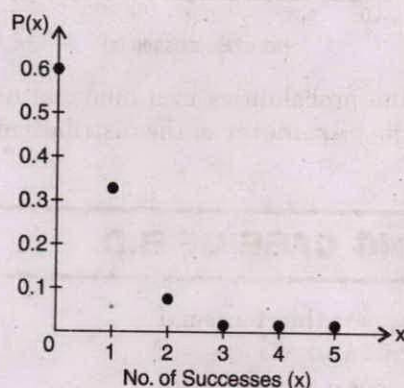
The shape of the binomial distribution depends upon the probability of success ( $p$ ) and the number of trials in the experiment. If  $p = q = \frac{1}{2}$ , then the distribution will be symmetrical for every value of  $n$ . If  $p \neq q$ , then the distribution would be asymmetrical i.e., skewed. The magnitude of skewness varies as the difference between  $p$  and  $q$ . We illustrate this by taking  $p = 0.1$ ,  $p = 0.5$ ,  $p = 0.9$  and assuming that there are 5 trials in the experiment.

**Case I.  $p = 0.1, n = 5$** 

We have  $P(x=r) = {}^5C_r \left(\frac{1}{10}\right)^r \left(\frac{9}{10}\right)^{5-r}, 0 \leq r \leq 5$ .

The B.D. is

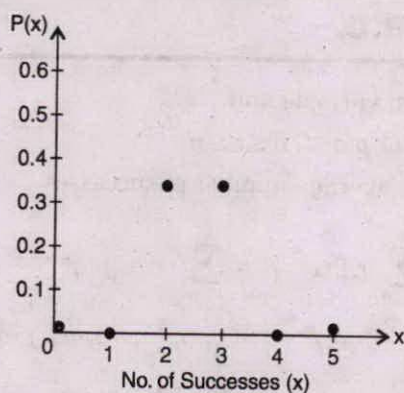
$x$	0	1	2	3	4	5
$P(x)$	0.59049	0.32805	0.07290	0.00810	0.00045	0.00001

**Case II.  $p = 0.5, n = 5$** 

We have  $P(x=r) = {}^5C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{5-r}, 0 \leq r \leq 5$ .

The B.D. is

$x$	0	1	2	3	4	5
$P(x)$	0.03125	0.15625	0.31250	0.31250	0.15625	0.03125

**Case III.  $p = 0.9, n = 5$** 

We have  $P(x=r) = {}^5C_r \left(\frac{9}{10}\right)^r \left(\frac{1}{10}\right)^{5-r}, 0 \leq r \leq 5$ .

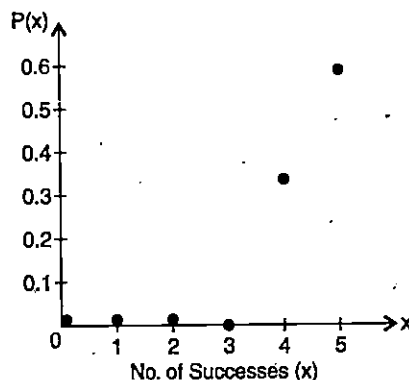
The B.D. is

$x$	0	1	2	3	4	5
$P(x)$	0.00001	0.00045	0.00810	0.07290	0.32805	0.59049

**NOTES**



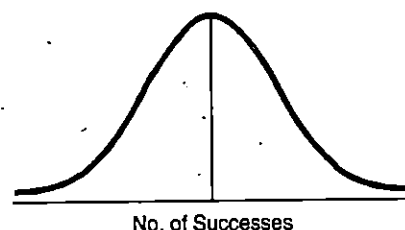
## NOTES



Thus, we see that the probabilities in a binomial distribution depends upon  $n$  and  $p$ . These are called the **parameter** of the distribution.

### 10.9. THE LIMITING CASE OF B.D.

As number of trials ( $n$ ) in the binomial distribution increases, the number of successes also increases. If neither  $p$  nor  $q$  is very small; then as  $n$  approaches infinity, the skewness in the distribution disappears and it becomes continuous. We shall see that such a continuous, bell shaped distribution is called a **normal distribution**. Thus, the normal distribution is a limiting case of binomial distribution as  $n$  approaches infinity.



### 10.10. MEAN OF B.D.

Let  $x$  be a binomial random variable and

$$P(x=r) = {}^nC_r p^r q^{n-r}, \quad 0 \leq r \leq n.$$

The *mean* of  $x$  is the average number of successes.

$$\begin{aligned} \therefore \text{Mean, } \mu &= \sum_{r=0}^n r \cdot P(x=r) = \sum_{r=0}^n r \cdot {}^nC_r p^r q^{n-r} \\ &= 0 \cdot {}^nC_0 p^0 q^n + 1 \cdot {}^nC_1 p^1 q^{n-1} + 2 \cdot {}^nC_2 p^2 q^{n-2} + \dots + n \cdot {}^nC_n p^n q^0 \\ &= 0 + n \cdot p q^{n-1} + 2 \cdot \frac{n(n-1)}{12} p^2 q^{n-2} + \dots + n \cdot 1 \cdot p^n \\ &= np \left\{ q^{n-1} + \frac{n-1}{1} p q^{n-2} + \dots + p^{n-1} \right\} \\ &= np \{ {}^{n-1}C_0 p^0 q^{n-1} + {}^{n-1}C_1 p^1 q^{n-2} + \dots + {}^{n-1}C_{n-1} p^{n-1} q^0 \} \\ &= np (q+p)^{n-1} = np(1)^{n-1} = np. \end{aligned}$$

$$\therefore \text{Mean of } x = np.$$

## 10.11. VARIANCE AND S.D. OF B.D.

### NOTES

Let  $x$  be a binomial random variable and

$$P(x = r) = {}^nC_r p^r q^{n-r}, \quad 0 \leq r \leq n.$$

The variance and standard deviation of  $x$  measures the dispersion of the binomial distribution and are given by

$$\text{Variance} = \sum_{r=0}^n r^2 \cdot P(x = r) - \mu^2$$

and

$$\text{S.D.} = \sqrt{\sum_{r=0}^n r^2 \cdot P(x = r) - \mu^2}.$$

$$\begin{aligned} \text{Now } \sum_{r=0}^n r^2 \cdot P(x = r) &= \sum_{r=0}^n r^2 \cdot {}^nC_r p^r q^{n-r} \\ &= 0 \cdot {}^nC_0 p^0 q^n + 1^2 \cdot {}^nC_1 p^1 q^{n-1} + 2^2 \cdot {}^nC_2 p^2 q^{n-2} + 3^2 \cdot {}^nC_3 p^3 q^{n-3} + \dots + n^2 \cdot {}^nC_n p^n q^0 \\ &= 0 + 1 \cdot \frac{n}{1} p q^{n-1} + 2^2 \cdot \frac{n(n-1)}{1 \cdot 2} p^2 q^{n-2} + \frac{3^2 \cdot n(n-1)(n-2)}{1 \cdot 2 \cdot 3} p^3 q^{n-3} + \dots + n^2 \cdot 1 \cdot p^n \cdot 1 \\ &= np \left\{ q^{n-1} + \frac{2(n-1)}{1} p q^{n-2} + \frac{3(n-1)(n-2)}{1 \times 2} p^2 q^{n-3} + \dots + n p^{n-1} \right\} \\ &= np \left\{ \left( q^{n-1} + \frac{n-1}{1} p q^{n-2} + \frac{(n-1)(n-2)}{1 \times 2} p^2 q^{n-3} + \dots + p^{n-1} \right) \right. \\ &\quad \left. + \left( \frac{n-1}{1} p q^{n-2} + \frac{2(n-1)(n-2)}{1 \times 2} p^2 q^{n-3} + \dots + (n-1) p^{n-1} \right) \right\} \\ &= np \{ (q + p)^{n-1} + (n-1) p (q^{n-2} + (n-2) p q^{n-3} + \dots + p^{n-2}) \} \\ &= np \{ 1 + (n-1) p (q + p)^{n-2} \} = np \{ 1 + (n-1) p \cdot 1 \} \\ &= np \{ 1 + np - p \} = np + n^2 p^2 - np^2. \end{aligned}$$

$$\begin{aligned} \therefore \text{Variance} &= \sum_{r=0}^n r^2 \cdot P(x = r) - \mu^2 = (np + n^2 p^2 - np^2) - (np)^2 = np - np^2 \\ &= np(1 - p) = npq. \end{aligned}$$

$$\text{Also, } \text{S.D.} = \sqrt{\text{variance}} = \sqrt{npq}.$$

## 10.12. $\gamma_1$ AND $\gamma_2$ OF B.D.

The values of  $\gamma_1$  and  $\gamma_2$  for the binomial probability function

$$P(x = r) = {}^nC_r p^r q^{n-r}, \quad 0 \leq r \leq n$$

are given by

$$\gamma_1 = \frac{1 - 2p}{\sqrt{npq}} \quad \text{and} \quad \gamma_2 = \frac{1 - 6pq}{npq}.$$

### 10.13. RECURRENCE FORMULA FOR B.D.

#### NOTES

Let  $x$  be a binomial random variable and

$$P(x=r) = {}^nC_r p^r q^{n-r}, 0 \leq r \leq n.$$

$$\text{For } 0 \leq k < n, P(k) = {}^nC_k p^k q^{n-k} \text{ and } P(k+1) = {}^nC_{k+1} p^{k+1} q^{n-(k+1)}$$

$$\text{Dividing, we get } \frac{P(k+1)}{P(k)} = \frac{{}^nC_{k+1} p^{k+1} q^{n-k-1}}{{}^nC_k p^k q^{n-k}}$$

$$= \frac{n!}{(k+1)!(n-(k+1))!} \cdot \frac{k!(n-k)!}{n!} \cdot \frac{p}{q} = \frac{n-k}{k+1} \cdot \frac{p}{q}$$

$$\therefore P(k+1) = \frac{n-k}{k+1} \cdot \frac{p}{q} P(k) \text{ for } 0 \leq k < n.$$

This is the required recurrence formula.

**Example 10.6.** The mean and S.D. of a binomial distribution are 20 and 4 respectively. Calculate  $n$ ,  $p$  and  $q$ .

**Solution.** Let the binomial distribution be

$$P(x=r) = {}^nC_r p^r q^{n-r}, 0 \leq r \leq n.$$

$$\therefore \text{Mean} = np \text{ and S.D.} = \sqrt{npq}.$$

We are given mean = 20, S.D. = 4.

$$\therefore np = 20, \sqrt{npq} = 4.$$

$$\Rightarrow \sqrt{20q} = 4 \Rightarrow 20q = 16 \Rightarrow q = \frac{4}{5}$$

$$\therefore p = 1 - q = 1 - \frac{4}{5} = \frac{1}{5}.$$

$$np = 20 \text{ implies } n \times \frac{1}{5} = 20 \text{ i.e., } n = 100$$

$$\therefore n = 100, p = \frac{1}{5}, q = \frac{4}{5}.$$

**Example 10.7.** If the sum of the mean and the variance of a binomial distribution of 5 trials is  $9/5$ , then find the binomial distribution.

**Solution.** Let the binomial distribution be

$$P(x=r) = {}^nC_r p^r q^{n-r}, 0 \leq r \leq n.$$

$$\therefore \text{Mean} = np \text{ and variance} = npq.$$

By the given condition,

$$np + npq = \frac{9}{5} \text{ and } n = 5.$$

$$\Rightarrow 5p + 5p(1-p) = \frac{9}{5} \Rightarrow 5p + 5p - 5p^2 = \frac{9}{5}$$

$$\Rightarrow 25p^2 - 50p + 9 = 0 \therefore p = \frac{1}{5}$$

$$\therefore q = 1 - p = 1 - \frac{1}{5} = \frac{4}{5}.$$

$\therefore$  The binomial distribution is

$$P(x=r) = {}^5C_r \left(\frac{1}{5}\right)^r \left(\frac{4}{5}\right)^{5-r}, 0 \leq r \leq 5.$$



**Example 10.8.** Is the following statement correct? "The mean and variance of a binomial distribution are respectively 6 and 9". Probability Distributions

**Solution.** Let the binomial distribution be

$$P(x = r) = {}^nC_r p^r q^{n-r}, 0 \leq r \leq n.$$

Now mean = 6  $\Rightarrow np = 6$

variance = 9  $\Rightarrow npq = 9$

$\therefore 6q = 9 \Rightarrow q = \frac{3}{2}$

This is impossible, because probability of an event can never be greater than 1.

$\therefore$  The given statement is not correct.

NOTES

### EXERCISE 10.2

1. Determine the binomial distribution whose mean is 5 and standard deviation is  $\sqrt{2.5}$ .
2. Determine the probability of 3 successes in a binomial distribution whose mean and variance are respectively 2 and  $\frac{3}{2}$ .
3. For a binomial distribution, the mean is 6 and the standard deviation is  $\sqrt{2}$ . Find the probability of getting 7 successes.
4. Is there any inconsistency in the statement. "The mean of a Binomial Distribution is 80 and S.D. is 8." If no inconsistency is found, what shall be the values of  $p$ ,  $q$  and  $n$ ?

#### Answers

1.  $P(x = r) = {}^{10}C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{10-r}, 0 \leq r \leq 10$

2. 0.2076

3. 0.2341

4.  $\frac{1}{5}, \frac{4}{5}, 400$

## 10.14. FITTING OF A BINOMIAL DISTRIBUTION

Let  $x$  be a binomial random variable of an experiment. Let probability of  $x$  successes be given by

$$P(x = r) = {}^nC_r p^r q^{n-r}, 0 \leq r \leq n.$$

By fitting of a binomial distribution, we mean to find out the theoretical frequencies of the values of the binomial random variable  $x = 0, 1, 2, \dots, n$ , when the experiment of  $n$  trials is repeated for, say,  $N$  times. The theoretical frequencies are given by

$$N \cdot P(x = r) = N \cdot {}^nC_r p^r q^{n-r}, 0 \leq r \leq n.$$

The recurrence formula can also be made use of, if desired.

**Example 10.9.** The following data are the number of seeds germinating out of 10 on damp filter for 80 sets of seeds. Fit a binomial distribution to this data.

## NOTES

$x$	0	1	2	3	4	5	6	7	8	9	10
$f$	6	20	28	12	8	6	0	0	0	0	0

**Solution.****Calculation of Expected Frequencies**

$x$	Observed frequency ( $f$ )	$fx$
0	6	0
1	20	20
2	28	56
3	12	36
4	8	32
5	6	30
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0
Total	80	174

Hence  $n = 10$ ,  $N = 80$ 

$$\text{Mean, } \bar{x} = \frac{\sum fx}{N} = \frac{174}{80} = 2.175$$

Let  $p$  denote the probability of success in a trial.

$$\therefore \text{Mean} = np \Rightarrow 10p = 2.175$$

$$\therefore p = 0.2175 \text{ and } q = 1 - p = 0.7825$$

 $\therefore$  The binomial distribution is

$$P(x = r) = {}^{10}C_r (0.2175)^r (0.7825)^{10-r}, 0 \leq r \leq 10.$$

 $\therefore$  For  $x = 0$ , expected frequency =  $80 \times P(0)$ 

$$= 80 \times {}^{10}C_0 (0.2175)^0 (0.7825)^{10} = 6.8854 \approx 7$$

For  $x = 1$ , expected frequency =  $80 \times P(1)$ 

$$= 80 \times {}^{10}C_1 (0.2175)^1 (0.7825)^9 = 19.1385 \approx 19$$

For  $x = 2$ , expected frequency =  $80 \times P(2)$ 

$$= 80 \times {}^{10}C_2 (0.2175)^2 (0.7825)^8 = 23.9382 \approx 24$$

For  $x = 3$ , expected frequency =  $80 \times P(3)$ 

$$= 80 \times {}^{10}C_3 (0.2175)^3 (0.7825)^7 = 17.7427 \approx 18$$

For  $x = 4$ , expected frequency =  $80 \times P(4)$ 

$$= 80 \times {}^{10}C_4 (0.2175)^4 (0.7825)^6 = 8.6302 \approx 8^*$$

\*We have approximated 8.6302 to 8 in order to keep the sum of all expected frequencies equal to 80.



For  $x = 5$ , expected frequency  $= 80 \times P(5)$

$$= 80 \times {}^{10}C_5 (0.2175)^5 (0.7825)^5 = 2.8768 \approx 3$$

For  $x = 6$ , expected frequency  $= 80 \times P(6)$

$$= 80 \times {}^{10}C_6 (0.2175)^6 (0.7825)^4 = 0.6636 \approx 1$$

For  $x = 7$ , expected frequency  $= 80 \times P(7)$

$$= 80 \times {}^{10}C_7 (0.2175)^7 (0.7825)^3 = 0.1046 \approx 0$$

For  $x = 8$ , expected frequency  $= 80 \times P(8)$

$$= 80 \times {}^{10}C_8 (0.2175)^8 (0.7825)^2 = 0.0108 \approx 0$$

For  $x = 9$ , expected frequency  $= 80 \times P(9)$

$$= 80 \times {}^{10}C_9 (0.2175)^9 (0.7825)^1 = 0.0006 \approx 0$$

For  $x = 10$ , expected frequency  $= 80 \times P(10)$

$$= 80 \times {}^{10}C_{10} (0.2175)^{10} (0.7825)^0 = 0.000016 \approx 0.$$

$\therefore$  The expected frequencies are as given in the following table:

$x$	0	1	2	3	4	5	6	7	8	9	10
Exp. freq.	7	19	24	18	8	3	1	0	0	0	0

**Example 10.10.** Five dice are thrown 96 times. The number of times 4 or 5 or 6 was actually thrown in the experiment is given in the following table:

No. of dice (Each showing 4 or 5 or 6)	0	1	2	3	4	5
Observed frequency	1	10	24	35	18	8

Fit a binomial distribution assuming:

- the dice are perfect.
- the nature of dice is not known.

**Solution.** Let  $x$  be the binomial random variable\* of the experiment. The possible values of  $x$  are 0, 1, 2, 3, 4, 5. Let  $p$  be the probability of getting 4 or 5 or 6 in a single throw.

Here  $n = 5$ ,  $N = 96$

$\therefore$  The expected frequency of getting  $r$  successes

$$= N {}^nC_r p^r q^{n-r} = 96 {}^5C_r p^r q^{5-r}, 0 \leq r \leq 5.$$

(i) In this case, the dice are perfect.

$$\therefore p = \frac{3}{6} = \frac{1}{2}$$

Also  $q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$

$\therefore$  For  $x = 0$ , the expected frequency  $= 96 \times {}^5C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 = 3$

For  $x = 1$ , the expected frequency  $= 96 \times {}^5C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^4 = 15$

For  $x = 2$ , the expected frequency  $= 96 \times {}^5C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = 30$

For  $x = 3$ , the expected frequency  $= 96 \times {}^5C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = 30$

For  $x = 4$ , the expected frequency  $= 96 \times {}^5C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^1 = 15$

For  $x = 5$ , the expected frequency  $= 96 \times {}^5C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^0 = 3.$

\* The experiment of throwing 5 dice can be considered as throwing of one die 5 times.

## NOTES

## NOTES

(ii) In this case, the value of  $p$  will be estimated by using the observed frequencies. We have

$$\text{Mean} = \frac{\sum fx}{N} = \frac{(1 \times 0) + (10 \times 1) + (24 \times 2) + (35 \times 3) + (18 \times 4) + (8 \times 5)}{96} = \frac{275}{96}$$

Also, mean =  $np$

$$\therefore 5p = \frac{275}{96} \quad \text{or} \quad p = 0.5729$$

$$\therefore q = 1 - p = 1 - 0.5729 = 0.4271.$$

$\therefore$  Expected frequency for  $x = r$

$$= 96 {}^5C_r (0.5729)^r (0.4271)^{5-r}, 0 \leq r \leq 5.$$

For  $x = 0$ , the expected frequency

$$= 96 \times {}^5C_0 (0.5729)^0 (0.4271)^5 = 1.3643 \approx 1$$

For  $x = 1$ , the expected frequency

$$= 96 \times {}^5C_1 (0.5729)^1 (0.4271)^4 = 9.1503 \approx 9$$

For  $x = 2$ , the expected frequency

$$= 96 \times {}^5C_2 (0.5729)^2 (0.4271)^3 = 24.5479 \approx 25$$

For  $x = 3$ , the expected frequency

$$= 96 \times {}^5C_3 (0.5729)^3 (0.4271)^2 = 32.9281 \approx 33$$

For  $x = 4$ , the expected frequency

$$= 96 \times {}^5C_4 (0.5729)^4 (0.4271)^1 = 22.0844 \approx 22$$

For  $x = 5$ , the expected frequency

$$= 96 \times {}^5C_5 (0.5729)^5 (0.4271)^0 = 5.9247 \approx 6.$$

## EXERCISE 10.3

- 8 unbiased coins were tossed 256 times. Fit a binomial distribution.
- 7 coins are tossed and the number of the heads noted. The experiment is repeated 128 times and the following distribution is obtained:

No. of heads	0	1	2	3	4	5	6	7
Frequency	7	16	19	35	30	13	7	1

Fit a binomial distribution assuming:

- the coins are unbiased.
- the nature of coins is not known.

- Four dice are thrown 162 times. The number of times 5 or 6 was actually thrown in the experiment is given in the following table:

No. of dice (Each showing 5 or 6)	0	1	2	3	4
Observed frequency	22	50	50	35	5

Fit a binomial distribution assuming:

- the dice are perfect.
- the nature of dice is not known.



4. Ten coins were tossed 1024 times and the following frequencies are observed.

No. of heads	0	1	2	3	4	5	6	7	8	9	10
Frequency	2	10	38	106	188	257	226	128	59	7	3

Compare these frequencies with the expected frequencies.

### Answers

- 1, 8, 28, 56, 70, 56, 28, 8, 1.
- (i) 1, 7, 21, 35, 35, 21, 7, 1. (ii) 1, 8, 23, 36, 34, 19, 6, 1.
- (i) 32, 64, 48, 16, 2 (ii) 18, 52, 58, 29, 5.
- $1 \left( = 1024^{10}C_0 (0.5135)^0 (0.4865)^{10} \right)$ , 8, 38, 107, 200, 251, 221, 133, 52, 12, 1.

## III. POISSON DISTRIBUTION

### 10.15. INTRODUCTION

The Poisson distribution is also a discrete probability distribution. This was discovered by French mathematician **Simon Denis Poisson** (1781 – 1840) in the year 1837. This distribution deals with the evaluation of probabilities of *rare* events such as “no. of car accidents on road”, “no. of earthquakes in a year”, “no. of misprints in a book”, etc.

### 10.16. CONDITIONS

The Poisson distribution is derived as a limiting case of binomial distribution. So, the conditions for the applicability of the Poisson distribution are same as those for the applicability of Binomial distribution. Here the additional requirement is that the probability of ‘success’ is quite near to zero.

### 10.17. POISSON VARIABLE

A random variable which counts the number of successes in a random experiment with trials satisfying above conditions is called a **Poisson variable**. If the probability of an article being defective is  $1/500$  and the event of getting a defective article is *success* and samples of 10 articles are checked for defective articles, then the possible values of the Poisson variable are 0, 1, 2, ..... 10.

### 10.18. POISSON PROBABILITY FUNCTION

Let a random experiment satisfying the conditions of Poisson Distribution be performed. Let the number of trials in the experiment be  $n$ , which is very large. Let  $p$  denote the probability of *success* in any trial. We assume that  $p$  is very small, i.e., we are dealing with a rare event. Let  $x$  denote the Poisson variable corresponding to this experiment.

∴ The possible values of  $x$  are 0, 1, 2, .....  $n$ .

## NOTES

The Poisson distribution is obtained as a limiting case of the corresponding binomial distribution of the experiment.

It can be proved mathematically that the probability of  $r$  successes is given by

## NOTES

$$P(x = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

This is called the **Poisson probability function**. The corresponding **Poisson distribution** is

$x$	0	1	2	3 .....
$P(x)$	$\frac{e^{-m} m^0}{0!}$	$\frac{e^{-m} m^1}{1!}$	$\frac{e^{-m} m^2}{2!}$	$\frac{e^{-m} m^3}{3!} \dots\dots$

The constant  $m$  is the product of  $n$  and  $p$  and is called the **parameter** of the Poisson distribution.

## 10.19. POISSON FREQUENCY DISTRIBUTION

If a random experiment, satisfying the requirements of Poisson distribution, is repeated  $N$  times, then the expected frequency of getting  $r$  successes is given by

$$N \cdot P(x = r) = N \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

**Remark 1.** The distribution to be used in solving a problem is generally given in the problem. If it is not given, then the student should make use of Poisson distribution only when the event in the problem is of rare nature i.e., the probability of happening of event is quite near to zero.

**Remark 2.** The values of  $e^{-m}$  required in any particular problem is generally given with the problem itself, otherwise, the value of  $e^{-m}$  can be found out by using the table given at the end. In the examination hall, generally the table of  $e^{-m}$  is available for students. If at all the value of  $e^{-m}$  is neither given with the problem nor the table of  $e^{-m}$  is supplied in the examination hall, then the students are advised to retain their final result in terms of  $e^{-m}$ .

**Example 10.11.** A company makes electric toys. The probability that an electric toy is defective is 0.01. What is the probability that a shipment of 300 toys will contain exactly 5 defective toys?

**Solution.** Let  $x$  be the Poisson variable, 'no. of defective toys'

By Poisson distribution,  $P(x = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$

Here  $n = 300, \quad p = 0.01$

$\therefore m = np = 300 \times 0.01 = 3$

$\therefore P(x = r) = \frac{e^{-3} 3^r}{r!}, \quad r = 0, 1, 2, \dots, 300$

$\therefore P(5 \text{ defective toys}) = P(x = 5)$

$$= \frac{e^{-3} (3)^5}{5!} = \frac{0.04979 \times 243}{120} = 0.1008.$$

**Example 10.12.** 10% of the tools produced in a certain factory turns out to be defective. Find the probability that in a sample of 10 tools chosen at random, (i) exactly two (ii) more than two will be defective by using Poisson approximation to binomial distribution.

## NOTES

**Solution.** Let  $x$  be the Poisson variable, "no. of defective tools in the sample".

By Poisson distribution,  $P(x = r) = \frac{e^{-m} m^r}{r!}$ ,  $r = 0, 1, 2, \dots$

Here  $n = 10, p = \frac{10}{100} = \frac{1}{10} \therefore m = np = 10 \times \frac{1}{10} = 1$

$$\therefore P(x = r) = \frac{e^{-1}(1)^r}{r!}, \quad r = 0, 1, 2, \dots, 10.$$

(i)  $P(\text{exactly 2 defectives}) = P(x = 2)$

$$= \frac{e^{-1}}{2!} = \frac{0.36788}{2} = 0.18394.$$

(ii)  $P(\text{more than 2 defectives}) = P(x > 2) = 1 - P(x \leq 2)$

$$= 1 - P(x = 0 \text{ or } x = 1 \text{ or } x = 2)$$

$$= 1 - \{P(x = 0) + P(x = 1) + P(x = 2)\}$$

$$= 1 - \left\{ \frac{e^{-1}}{0!} + \frac{e^{-1}}{1!} + \frac{e^{-1}}{2!} \right\} = 1 - e^{-1} \left\{ 1 + 1 + \frac{1}{2} \right\}$$

$$= 1 - (0.36788)(2.5) = 0.0803.$$

**Example 10.13.** A telephone exchange receives on an average 4 calls per minute. Find the probability on the basis of Poisson distribution ( $m = 4$ ), of:

(i) 2 or less calls per minute,

(ii) upto 4 calls per minute,

(iii) more than 4 calls per minute.

**Solution.** Let  $x$  be the Poisson variable "no. of calls per minute".

By Poisson distribution,

$$P(x = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

Here  $m =$  average number of successes i.e., calls per minute  $= 4$

$$\therefore P(x = r) = \frac{e^{-4}(4)^r}{r!}, \quad r = 0, 1, 2, \dots$$

(i)  $P(2 \text{ or less calls per minute}) = P(x \leq 2)$

$$= P(x = 0 \text{ or } x = 1 \text{ or } x = 2)$$

$$= P(x = 0) + P(x = 1) + P(x = 2)$$

$$= \frac{e^{-4}(4)^0}{0!} + \frac{e^{-4}(4)^1}{1!} + \frac{e^{-4}(4)^2}{2!}$$

$$= e^{-4} \{1 + 4 + 8\} = 0.01832 \times 13 = 0.2382$$



## NOTES

$$(ii) P(\text{upto 4 calls per minute}) = P(x \leq 4)$$

$$= P(x = 0 \text{ or } x = 1 \text{ or } x = 2 \text{ or } x = 3 \text{ or } x = 4) \\ = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4)$$

$$= \frac{e^{-4}(4)^0}{0!} + \frac{e^{-4}(4)^1}{1!} + \frac{e^{-4}(4)^2}{2!} + \frac{e^{-4}(4)^3}{3!} + \frac{e^{-4}(4)^4}{4!}$$

$$= e^{-4} \left\{ 1 + 4 + 8 + \frac{64}{6} + \frac{256}{24} \right\}$$

$$= 0.01832 \times 34.3333 = 0.6289$$

$$(iii) P(\text{more than 4 calls per minute}) = P(x > 4)$$

$$= 1 - P(x \leq 4) = 1 - 0.6289 = 0.3711. \quad [\text{By part (ii)}]$$

**Example 10.14.** A manufacturer of bulbs knows that on an average 5% of his production is defective. He sells bulbs in boxes of 100 pieces and guarantees that not more than 4 bulbs will be defective in a box. What is the probability that a box will meet the guarantee ( $e^{-5} = 0.0067$ )?

**Solution.** Let  $x$  be the Poisson variable, 'no. of defective bulbs per box',  
By Poisson distribution,

$$P(x = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

Here  $n = 100, p = \frac{5}{100}$ .

$$\therefore m = np = 100 \times \frac{5}{100} = 5$$

$$\therefore P(x = r) = \frac{e^{-5} 5^r}{r!}, \quad r = 0, 1, 2, \dots, 100.$$

$$P(\text{box will meet the guarantee}) = P(x \leq 4)$$

$$= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4)$$

$$= \frac{e^{-5}(5)^0}{0!} + \frac{e^{-5}(5)^1}{1!} + \frac{e^{-5}(5)^2}{2!} + \frac{e^{-5}(5)^3}{3!} + \frac{e^{-5}(5)^4}{4!}$$

$$= e^{-5} \left[ \frac{1}{1} + \frac{5}{1} + \frac{25}{2} + \frac{125}{6} + \frac{625}{24} \right] = 0.0067 \times 65.37499 = 0.438.$$

**Example 10.15.** A car-hire firm has two cars, which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused ( $e^{-1.5} = 0.2231$ ).

**Solution.** Let  $x$  be the Poisson variable, 'no. of demands per day'.

$\therefore$  By Poisson distribution,

$$P(x = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

Here  $m$  = average number of demands per day = 1.5

$$\therefore P(x = r) = \frac{e^{-1.5}(1.5)^r}{r!}, \quad r = 0, 1, 2, \dots$$

Now, proportion of days on which neither car is used

$$= P(x = 0) = \frac{e^{-1.5}(1.5)^0}{0!} = 0.2231$$

Also, proportion of days on which some demand is refused

$$\begin{aligned} &= P(x > 2) = 1 - P(x \leq 2) = 1 - P(x = 0 \text{ or } x = 1 \text{ or } x = 2) \\ &= 1 - \{P(x = 0) + P(x = 1) + P(x = 2)\} \\ &= 1 - \left\{ \frac{e^{-1.5}(1.5)^0}{0!} + \frac{e^{-1.5}(1.5)^1}{1!} + \frac{e^{-1.5}(1.5)^2}{2!} \right\} \\ &= 1 - e^{-1.5} \left\{ 1 + 1.5 + \frac{2.25}{2} \right\} = 1 - (0.2231)(3.625) = 0.1913. \end{aligned}$$

## NOTES

**Example 10.16.** 250 passengers have made reservations for a flight from Delhi to Mumbai. If the probability that a passenger, who has reservation, will not turn up is 0.016. Find the probability that at the most 3 passengers will not turn up.

(given  $e^{-4} = 0.0183$ )

**Solution.** Let  $x$  be the random variable 'no. of passengers not turning up'.

$\therefore$  By Poisson distribution,

$$P(x = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

Here

$$n = 250, p = 0.016$$

$\therefore$

$$m = np = 250 \times \frac{16}{1000} = 4$$

$\therefore$

$$P(x = r) = \frac{e^{-4}(4)^r}{r!}, \quad r = 0, 1, 2, \dots, 250$$

$\therefore$  Prob. that at most 3 passengers will not turn up

$$\begin{aligned} &= P(x \leq 3) = P(x = 0 \text{ or } x = 1 \text{ or } x = 2 \text{ or } x = 3) \\ &= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) \\ &= \frac{e^{-4}(4)^0}{0!} + \frac{e^{-4}(4)^1}{1!} + \frac{e^{-4}(4)^2}{2!} + \frac{e^{-4}(4)^3}{3!} \\ &= e^{-4} \left[ 1 + 4 + \frac{16}{2} + \frac{64}{6} \right] = 0.0183 \times 23.67 = 0.433. \end{aligned}$$

### EXERCISE 10.4

- Find the probability that at most 5 defective bolts will be found in a box of 200 bolts if it is known that 2% of such bolts are expected to be defective (you may take the distribution to be of Poisson type). ( $e^{-4} = 0.0183$ ).
- A telephone exchange receives on an average 3 calls per minute. Find the probability on the basis of Poisson distribution ( $m = 3$ ), of:
  - exactly 1 call per minute
  - exactly 3 calls per minute
  - less than 3 calls per minute
  - more than 1 call per minute.
- Assuming that the probability of a total accident in a factory during a year is  $1/1200$ , calculate the probability that in a factory employing 300 workers, there will be at least two total accidents in a year. ( $e^{-0.25} = 0.7788$ ).

## NOTES

4. The probabilities of a Poisson variate taking the values 3 and 4 are equal. Calculate the probabilities of the variate taking the values 0 and 1.
5. Find the probability that at most 5 defective articles will be found in a box of 200 articles, if experience shows that 2% of such articles are defective.
6. Assume that the probability of an individual coal miner killed in a mine accident during a year is  $1/2500$ . Calculate the probability that in a mine employing 2000 miners, there will be (i) no fatal accident and (ii) at least one fatal accident in a year.
7. The probability that a man aged 60 years will die within a year is 0.01125. What is the probability that out of 12 such men, at least 11 will reach their 61st birth day?
8. An office switch board receives telephone calls at the rate of 3 per minute on an average. What is the probability of receiving (i) no call in one minute interval and (ii) at the most 3 calls in one minute interval?
9. 2% bulbs, manufactured by a company are defective. Find the probability that in a sample of 2000 bulbs: (i) less than 2 bulbs are defective (ii) more than 3 bulbs are defective. (Use  $e^{-4} = 0.0183$ )
10. In a town 10 accidents took place in a span of 50 days. Assuming that the number of accidents per day follows the Poisson distribution, find the probability that there will be three or more accidents in a day. (use  $e^{-0.2} = 0.8187$ )

## Answers

1.  $e^{-4} \left\{ \frac{4^0}{0!} + \frac{4^1}{1!} + \frac{4^2}{2!} + \frac{4^3}{3!} + \frac{4^4}{4!} + \frac{4^5}{5!} \right\} = 0.7845$  2. 0.1494, 0.2241, 0.4232, 0.8008
3.  $1 - e^{-0.25} \left( 1 + \frac{0.25}{1!} \right) = 0.0265$
4.  $e^{-4} = 0.01832$ ,  $4e^{-4} = 0.07328$  5. 0.7853
6. (i) 0.4493 (ii) 0.5507
7.  $P(\text{at least 11 will survive}) = P(\text{at most one die}) = P(x \leq 1) = e^{-0.135} (1 + 0.135) = 0.9916$
8. (i) 0.0498 (ii) 0.6473
9. (i) 0.092 (ii) 0.567 10. 0.0012.

## IV. PROPERTIES OF POISSON DISTRIBUTION

**10.20. THE SHAPE OF P.D.**

The shape of the Poisson distribution depends upon the parameter  $m$ , the average number of successes per unit. As value of  $m$  increases, the graph of Poisson distribution would get closer to a symmetrical continuous curve.

**10.21. SPECIAL USEFULNESS OF P.D.**

The Poisson distribution is specially used when there are events which do not occur as outcomes of a definite number of trials in an experiment, rather occur randomly in nature. This distribution is used when the event under consideration is rare and casual. In finding probabilities by Poisson distribution, we require only the measure of average chance of occurrence ( $m$ ) based on past experience or a small sample drawn for the purpose.



## 10.22. MEAN OF P.D.

Let  $x$  be a Poisson random variable and

$$P(x=r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

The mean of  $x$  is the average numbers of successes.

$$\begin{aligned}\therefore \text{Mean } (\mu) &= \sum_{r=0}^{\infty} r \cdot P(x=r) = \sum_{r=0}^{\infty} r \cdot \frac{e^{-m} m^r}{r!} \\&= 0 \cdot \frac{e^{-m} m^0}{0!} + 1 \cdot \frac{e^{-m} m^1}{1!} + 2 \cdot \frac{e^{-m} m^2}{2!} + 3 \cdot \frac{e^{-m} m^3}{3!} + \dots \\&= 0 + me^{-m} \left( \frac{1}{1!} + \frac{2m}{2!} + \frac{3m^2}{3!} + \dots \right) = me^{-m} \left( 1 + \frac{m}{1!} + \frac{m^2}{2!} + \dots \right) \\&= me^{-m} \cdot e^m = me^0 = m \cdot 1 = m.\end{aligned}$$

$\therefore$  Mean of  $x = m$ .

## 10.23. VARIANCE AND S.D. OF P.D.

Let  $x$  be a Poisson random variable and

$$P(x=r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

The variance and standard deviation of  $x$  measures the dispersion of the Poisson distribution and are given by

$$\text{variance} = \sum_{r=0}^{\infty} r^2 \cdot P(x=r) - \mu^2 \quad \text{and} \quad \text{S.D.} = \sqrt{\sum_{r=0}^{\infty} r^2 \cdot P(x=r) - \mu^2}.$$

$$\begin{aligned}\text{Now } \sum_{r=0}^{\infty} r^2 \cdot P(x=r) &= \sum_{r=0}^{\infty} r^2 \cdot \frac{e^{-m} m^r}{r!} \\&= 0^2 \cdot \frac{e^{-m} m^0}{0!} + 1^2 \cdot \frac{e^{-m} m^1}{1!} + 2^2 \cdot \frac{e^{-m} m^2}{2!} + 3^2 \cdot \frac{e^{-m} m^3}{3!} + 4^2 \cdot \frac{e^{-m} m^4}{4!} + \dots \\&= 0 + me^{-m} \left( \frac{1}{1!} + \frac{2m}{1!} + \frac{3m^2}{2!} + \frac{4m^3}{3!} + \dots \right) \\&= me^{-m} \left\{ \left( 1 + \frac{m}{1!} + \frac{m^2}{2!} + \frac{m^3}{3!} + \dots \right) + \left( \frac{m}{1!} + \frac{2m^2}{2!} + \frac{3m^3}{3!} + \dots \right) \right\} \\&= me^{-m} \left\{ e^m + m \left( 1 + \frac{m}{1!} + \frac{m^2}{2!} + \dots \right) \right\} = me^{-m} \{ e^m + me^m \} \\&= me^{-m} e^m (1+m) = me^0 (1+m) = m(1+m) = m + m^2.\end{aligned}$$

$$\therefore \text{Variance} = \sum_{r=0}^{\infty} r^2 \cdot P(x=r) - \mu^2 = (m + m^2) - m^2 = m.$$

$$\text{Also, } \text{S.D.} = \sqrt{\text{variance}} = \sqrt{m}.$$

NOTES

## 10.24. $\gamma_1$ AND $\gamma_2$ OF P.D.

The values of  $\gamma_1$  and  $\gamma_2$  for the Poisson probability function

$$P(x=r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

NOTES

are given by

$$\gamma_1 = \frac{1}{\sqrt{m}} \quad \text{and} \quad \gamma_2 = \frac{1}{m}.$$

## 10.25. RECURRENCE FORMULA FOR P.D.

Let  $x$  be a Poisson variable and  $P(x=r) = \frac{e^{-m} m^r}{r!}, r = 0, 1, 2, \dots$

$$\text{For } k \geq 0, \quad P(k) = \frac{e^{-m} m^k}{k!} \quad \text{and} \quad P(k+1) = \frac{e^{-m} m^{k+1}}{(k+1)!}$$

$$\text{Dividing, we get} \quad \frac{P(k+1)}{P(k)} = \frac{e^{-m} m^{k+1}}{(k+1)!} \cdot \frac{k!}{e^{-m} m^k} = \frac{m}{k+1}$$

$$\therefore \quad P(k+1) = \frac{m}{k+1} P(k), \quad k = 0, 1, 2, \dots$$

This is the required recurrence formula.

**Example 10.17.** Criticise the following statement, "The mean and standard deviation of a Poisson distribution are 5 and 2 respectively".

**Solution.** Let  $x$  be a Poisson variable and

$$P(x=r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

By the given condition,

$$\text{Mean} = 5, \text{ S.D.} = 2$$

$$\therefore \quad \text{Variance} = (2)^2 = 4.$$

Now, in Poisson distribution,

$$\text{mean} = \text{variance} = m$$

$$\therefore \quad 5 = 4. \text{ This is impossible.}$$

$\therefore$  The given statement is incorrect.

**Example 10.18.** If  $x$  is a Poisson random variable such that:

$$P(x=2) = 9P(x=4) + 90P(x=6),$$

then find mean, standard deviation and  $\gamma_1$ .

$$\text{Solution. We have} \quad P(x=r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \dots$$

where  $m$  is the average no. of occurrence of  $x$ .

By the given condition,

$$P(x=2) = 9 P(x=4) + 90 P(x=6).$$



$$\therefore \frac{e^{-m}m^2}{2!} = 9 \frac{e^{-m}m^4}{4!} + 90 \frac{e^{-m}m^6}{6!}$$

$$\text{or} \quad \frac{m^2}{2} = \frac{9m^4}{24} + \frac{90m^6}{720} \quad \text{or} \quad \frac{1}{2} = \frac{3m^2}{8} + \frac{m^4}{8} \quad (\because m \neq 0)$$

$$\text{or} \quad 4 = 3m^2 + m^4 \quad \text{or} \quad (m^2 + 4)(m^2 - 1) = 0$$

$$\Rightarrow \quad m^2 = -4, 1 \quad \Rightarrow \quad m^2 = 1 \Rightarrow m = 1 \quad (\because m^2 > 0)$$

$$\therefore \text{Mean} = 1, \text{standard deviation} = \sqrt{1} = 1 \quad \text{and} \quad \gamma_1 = 1/\sqrt{1} = 1.$$

**NOTES****EXERCISE 10.5**

1. The mean of a Poisson distribution is  $\sqrt{8}$ , find the value of its S.D.
2. Criticise the following statement: "The mean and variance of a Poisson distribution are 4 and 2.1 respectively."
3. Comment on the following statement: "The mean and variance of a Poisson distribution are equal only if the average occurrence of the Poisson variable is  $\leq 4$ ."
4. The standard deviation of a Poisson distribution is 3. Find the probability of getting 3 successes.

**Answers**

1. 1.6818
4. 0.0149.

**10.26. FITTING OF A POISSON DISTRIBUTION**

Let  $x$  be a Poisson random variable. Let probability of  $x$  successes be given by

$$P(x=r) = \frac{e^{-m}m^r}{r!}, \quad r = 0, 1, 2, \dots$$

By fitting of a Poisson distribution, we mean to find the theoretical frequencies of the values of the Poisson random variable  $x = 0, 1, 2, \dots$  when the experiment is repeated  $N$  times. The theoretical frequencies are given by

$$N \cdot P(x=r) = N \cdot \frac{e^{-m}m^r}{r!} \quad \text{for} \quad r = 0, 1, 2, \dots$$

The use of recurrence formula would be found very useful.

**Example 10.19.** The distribution of typing mistakes committed by a typist is given below:

No. of mistakes per page	0	1	2	3	4	5
No. of pages	142	156	69	27	5	1

Assuming a Poisson model, find out the expected frequencies.

## NOTES

**Solution. Calculation of Expected Frequencies**

No. of mistakes per page $x$	No. of pages $f$	$fx$
0	142	0
1	156	156
2	69	138
3	27	81
4	5	20
5	1	5
Total	$N = 400$	400

$\therefore m = \text{average of no. of mistakes per page}$

$$= \frac{\sum fx}{N} = \frac{400}{400} = 1.$$

$\therefore$  The Poisson distribution is

$$P(x=r) = \frac{e^{-m} m^r}{r!} = \frac{e^{-1}(1)^r}{r!} = \frac{e^{-1}}{r!}, \quad r = 0, 1, 2, \dots$$

For  $x = 0$ , expected frequency =  $400 \cdot P(0)$

$$= 400 \cdot \frac{e^{-1}}{0!} = 400(0.36788) = 147.152 \approx 147$$

For  $x = 1$ , expected frequency =  $400 \cdot P(1)$

$$= 400 \cdot \frac{e^{-1}}{1!} = 400(0.36788) = 147.152 \approx 147$$

For  $x = 2$ , expected frequency =  $400 \cdot P(2)$

$$= 400 \cdot \frac{e^{-1}}{2!} = 400(0.36788)/2 = 73.576 \approx 74$$

For  $x = 3$ , expected frequency =  $400 \cdot P(3)$

$$= 400 \cdot \frac{e^{-1}}{3!} = 400(0.36788)/6 = 24.525 \approx 25$$

For  $x = 4$ , expected frequency =  $400 \cdot P(4)$

$$= 400 \cdot \frac{e^{-1}}{4!} = 400(0.36788)/24 = 6.131 \approx 6$$

For  $x = 5$ , expected frequency =  $400 \cdot P(5)$

$$= 400 \cdot \frac{e^{-1}}{5!} = 400(0.36788)/120 = 1.2263 \approx 1.$$

**Example 10.20.** Letters were received in an office on each of 100 days. Assuming the following data to form a random sample from a Poisson distribution, find the expected frequencies, correct to the nearest unit, taking  $e^{-4} = 0.0183$ .

No. of letters	0	1	2	3	4	5	6	7	8	9	10
Frequency	1	4	15	22	21	20	8	6	2	0	1

**Solution.****Calculation of Expected Frequencies**

Probability Distributions

No. of letters $x$	Frequency $f$	$fx$
0	1	0
1	4	4
2	15	30
3	22	66
4	21	84
5	20	100
6	8	48
7	6	42
8	2	16
9	0	0
10	1	10
Total	$N = 100$	400

$\therefore m = \text{average no. of letters per day}$

$$= \frac{\sum fx}{N} = \frac{400}{100} = 4.$$

$\therefore$  The Poisson distribution is

$$P(x=r) = \frac{e^{-m} m^r}{r!} = \frac{e^{-4} 4^r}{r!}, \quad r = 0, 1, 2, \dots$$

By recurrence formula,

$$P(k+1) = \frac{4}{k+1} P(k), \quad k = 0, 1, 2, \dots$$

Let  $f(x)$  denote the expected frequency of  $x$ .

$$\therefore f(k+1) = N \cdot P(k+1) = 100 \cdot \frac{4}{k+1} P(k) = \frac{4}{k+1} 100 \cdot P(k) = \frac{4}{k+1} f(k)$$

$$\therefore f(k+1) = \frac{4}{k+1} f(k), \quad k = 0, 1, 2, \dots$$

Now  $f(0) = 100 P(0) = 100 \cdot \frac{e^{-4}(4)^0}{0!} = 100(0.0183) = 1.83 \approx 2$

$$f(1) = \frac{4}{1} f(0) = 4(1.83) = 7.32 \approx 7$$

$$f(2) = \frac{4}{2} f(1) = 2(7.32) = 14.64 \approx 15$$

$$f(3) = \frac{4}{3} f(2) = \frac{4(14.64)}{3} = 19.52 \approx 20$$

$$f(4) = \frac{4}{4} f(3) = \frac{4(19.52)}{4} = 19.52 \approx 20$$

$$f(5) = \frac{4}{5} f(4) = \frac{4(19.52)}{5} = 15.62 \approx 16$$

**NOTES**

## NOTES

$$f(6) = \frac{4}{6} f(5) = \frac{4(15.62)}{6} = 10.41 \approx 10$$

$$f(7) = \frac{4}{7} f(6) = \frac{4(10.41)}{7} = 5.95 \approx 6$$

$$f(8) = \frac{4}{8} f(7) = \frac{4(5.95)}{8} = 2.97 \approx 3$$

$$f(9) = \frac{4}{9} f(8) = \frac{4(2.97)}{9} = 1.32 \approx 1$$

$$f(10) = \frac{4}{10} f(9) = \frac{4(1.32)}{10} = 0.53 \approx 1$$

## EXERCISE 10.6

1. A typist commits the following number of mistakes per page in typing 100 pages. Fit a Poisson distribution and calculate theoretical frequencies:

Mistakes per page	0	1	2	3	4	5
Frequency	42	33	14	6	4	1

You are given,  $e^{-1} = 0.3679$ .

2. Below are given the number of vacancies of judges occurring in a High Court over a period of 96 years:

Fit a Poisson distribution to represent the frequencies of vacancies per year and find the expected frequencies:

No. of vacancies per year	0	1	2	3
No. of years	59	27	9	1

3. In 1,000 sets of trials for an event of small frequencies  $f_i$ , of the number of  $x_i$  successes are:

$x$	0	1	2	3	4	5	6	7
$f$	305	365	210	80	28	9	2	1

Fit a Poisson distribution to the above data and calculate the theoretical frequencies.

4. 5,000 television sets are inspected as they come off the production line and the number of defects per set is recorded below.

No. of defects	0	1	2	3	4
No. of sets	3680	720	520	70	10

Estimate the average number of defects per set and the expected frequencies of 0, 1, 2, 3 and 4 defects, assuming Poisson distribution.

## Answers

1. 37, 37, 18, 6, 2, 0                      2. 58, 29, 7, 1  
 3. 301, 361, 217, 87, 26, 6, 1, 0.  
 4. Average no. of defect per set = 0.402 ; 3351, 1341, 268, 36, 4.

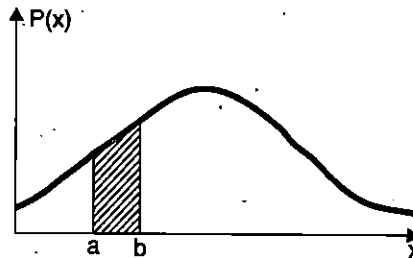


## 10.27. INTRODUCTION

In Binomial and Poisson distributions, we considered the probabilities of discrete random variables. Now we shall consider random variables which may take non-countably infinitely many possible values. Such a random variable is called a **continuous random variable**. The random variables corresponding to the measurement of height, weight, etc. are continuous. We have already discussed probability distributions of discrete random variables. We shall also be considering the probability function of a very important continuous random variable, namely *normal variable*.

## 10.28. PROBABILITY FUNCTION OF CONTINUOUS RANDOM VARIABLE

In discrete probability distributions, the probability is defined for each and every value of the variable and the sum of all these probabilities is one. On the other hand, continuous random variables are defined over intervals of real numbers which contains non-countably infinitely many numbers. Let  $x$  be a continuous random variable. The probability of  $x$  to take any particular value is generally zero. For example, if an individual is selected at random from a large group of males, then the probability that his weight ( $x$ ) is exactly 56 kg (i.e., 56.000 ..... kg) would be zero. On the other hand, the probability that weight ( $x$ ) lying between 55.600 ..... kg and 56.200 ..... kg need not be zero. Thus, we cannot define a probability function for a continuous random variable as we did in the case of a discrete random variable. In case of a continuous random variable ( $x$ ), the probability of  $x$  taking any particular value is generally zero and practically does not make any sense whereas the probability of  $x$  taking values between any two different values is meaningful.



For a continuous random variable,  $x$ , a function  $P(x)$  is called a **probability function** if:

(i)  $P(x) \geq 0$  and

(ii)  $\int_{-\infty}^{\infty} P(x) dx = 1$ .

If  $P(x)$  is a *probability function* of  $x$ , then we define:

$$P(a < x < b) = \int_a^b P(x) dx.$$

Thus, if  $P(x)$  is a *probability function* of  $x$ , then:

(i)  $P(x)$  is non-negative

(ii) area bounded by the curve and  $x$ -axis is equal to one



(iii) area bounded by the curve,  $x$ -axis and ordinates  $x = a$ ,  $x = b$  gives the measure of the probability that  $x$  lies between  $a$  and  $b$ .

**Remark.** Since the probability of  $x$  taking any particular value is generally zero, we have

$$P(a < x < b) = P(a \leq x < b) = P(a < x \leq b) = P(a \leq x \leq b).$$

## NOTES

### 10.29. NORMAL DISTRIBUTION

The normal distribution is a particular type of continuous probability distribution. This was discovered by **De Moivre (1667—1754)** in the year 1733. The normal distribution is obtained as a limiting case of a binomial distribution when  $n$ , the number of trials is indefinitely large and neither  $p$  nor  $q$  is very small.

### 10.30. DEFINITION

A continuous random variable  $x$  is said to have a **normal distribution (N.D.)** if its probability function is given by

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

where  $\mu$  and  $\sigma$  are the mean and standard deviation of the distribution respectively.

**Remark.** If  $x$  is a normal variable with mean  $\mu$  and variance  $\sigma^2$ , then we write symbolically as  $x \sim N(\mu, \sigma^2)$ .

### 10.31. STANDARD NORMAL DISTRIBUTION

Let  $x$  be a normal variable with probability function:

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

where  $\mu$  and  $\sigma$  are the mean and standard deviation of the distribution respectively.

We define  $z = \frac{x - \mu}{\sigma}$ .

It can be proved mathematically that  $z$  is also a normal variable with mean zero and variance one. A normal variable with mean zero and variance one is called a **standard normal variable (S.N.V.)**.

In terms of  $z$ , the probability function of  $x$  reduces to

$$P(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty.$$

**Remark.** Thus, if  $x \sim N(\mu, \sigma^2)$  and  $z = \text{S.N.V. of } x = \frac{x - \mu}{\sigma}$ , then  $z \sim N(0, 1)$ .

## 10.32. AREA UNDER NORMAL CURVE

Let  $x$  be a normal variable with mean  $\mu$  and standard deviation  $\sigma$ . Let  $z = \frac{x - \mu}{\sigma}$  be the corresponding S.N.V. We know that mean and standard deviation of the variable  $z$  are 0 and 1 respectively. Therefore, the curves of standard normal variables corresponding of normal variables are identical.

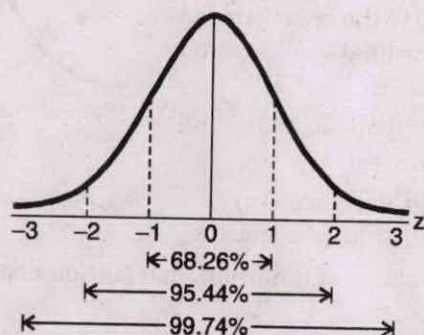
This is why the curve of a standard normal variable is called the **standard normal curve**. For the standard normal curve, we have:

(i) area between  $z = -1$  and  $z = 1$  is 68.26% of total area, which is one.

$$\therefore P(-1 \leq z \leq 1) = 0.6826.$$

(ii) area between  $z = -2$  and  $z = 2$  is 95.44% of total area.

$$\therefore P(-2 \leq z \leq 2) = 0.9544.$$



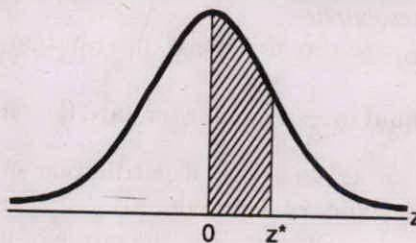
(iii) area between  $z = -3$  and  $z = 3$  is 99.74% of total area.

$$\therefore P(-3 \leq z \leq 3) = 0.9974.$$

The area bounded by the curve of a S.N.V.  $z$ ,  $z$ -axis and the ordinates at  $z = 0$  and any positive value of  $z$  is provided by a standard table. This knowledge of measure of area in case of S.N.V. is used to find the area bounded by the corresponding normal variable  $x$ ,  $x$ -axis and any two ordinates. This in turn would help us to find the probability of normal variable  $x$  lying between any two real numbers.

## 10.33. TABLE OF AREA UNDER STANDARD NORMAL CURVE

The table titled *area under standard normal curve* is given at the end. Let  $z^*$  be any arbitrary but fixed value of the variable  $z$ . The first column of the table provides for  $z$  values with one decimal digit and the second column gives areas bounded between  $z$ -curve and ordinates  $z = 0$  and  $z = z^*$ , which is equal to  $P(0 \leq z \leq z^*)$ .



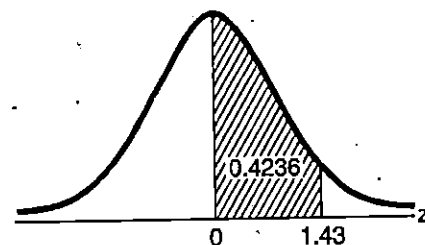
## NOTES

For example, from the table  $P(0 \leq z \leq 1.4) = 0.4192$ . The first row of the table provides for the second decimal digit of  $z^*$ . For example,  $P(0 \leq z \leq 1.43) = 0.4236$ .

## NOTES

**Remark.** Sometimes, the table for the probabilities  $P(-\infty < z \leq z^*)$  is given in the examination hall. In such a case, the students should find the value of  $P(0 \leq z \leq z^*)$  by using the following formula:

$$P(0 \leq z \leq z^*) = P(-\infty < z \leq z^*) - 0.5.$$



## 10.34. PROPERTIES OF NORMAL DISTRIBUTION

(i) The area bounded by the curve and the  $x$ -axis is equal to one.

(ii)  $P(a < x < b)$ , i.e. the probability that  $x$  lies between  $a$  and  $b$  is equal to the area bounded by the curve,  $x$ -axis and ordinates  $x = a$  and  $x = b$ .

(iii) The curve is bell-shaped and symmetrical about the line  $x = \mu$ .

(iv) The mean ( $\mu$ ) and variance ( $\sigma^2$ ) of a normal distribution are called its *parameters*.

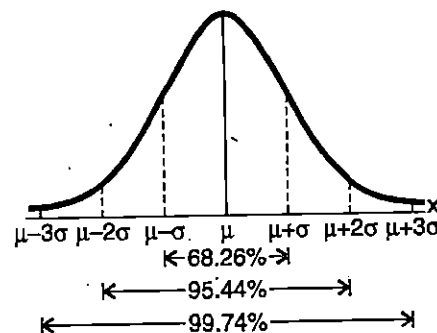
(v) The location and shape of a normal distribution depends upon the values of its parameters.

(vi) If the mean and S.D. of a normal distribution are  $\mu$  and  $\sigma$  respectively, then:

$$\begin{aligned} (a) P(\mu - \sigma \leq x \leq \mu + \sigma) &= P\left(\frac{\mu - \sigma - \mu}{\sigma} \leq \frac{x - \mu}{\sigma} \leq \frac{\mu + \sigma - \mu}{\sigma}\right) \\ &= P(-1 \leq z \leq 1) = 0.6826 \end{aligned}$$

$$\begin{aligned} (b) P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) &= P\left(\frac{\mu - 2\sigma - \mu}{\sigma} \leq \frac{x - \mu}{\sigma} \leq \frac{\mu + 2\sigma - \mu}{\sigma}\right) \\ &= P(-2 \leq z \leq 2) = 0.9544 \end{aligned}$$

$$\begin{aligned} (c) P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) &= P\left(\frac{\mu - 3\sigma - \mu}{\sigma} \leq \frac{x - \mu}{\sigma} \leq \frac{\mu + 3\sigma - \mu}{\sigma}\right) \\ &= P(-3 \leq z \leq 3) = 0.9974. \end{aligned}$$



(vii) The mean, median and mode of a normal distribution coincide with each other.

(viii) For a normal distribution,  $\gamma_1 = 0$  and  $\gamma_2 = 0$ . Therefore, normal distribution is *mesokurtic*.

(ix) In a normal distribution, mean deviation about mean is approximately equal to  $\frac{4}{5}$  time its standard deviation.

(x) In a normal distribution, quartile deviation is approximately equal to  $\frac{2}{3}$  times its standard deviation.

(xi) The tails of the curve of a normal distribution extend indefinitely on both sides of  $x = \mu$  and never touches the  $x$ -axis.

(xii) In a normal distribution,  $Q_1$  and  $Q_3$  are equidistant from the median.



**Example 10.21.** The mean and standard deviation of a normal variable  $x$  are 50 and 4 respectively. Find the values of the corresponding standard normal variable, when  $x$  is equal to 42, 84, 85, 32 and 40.

**Solution.** We have  $\mu = 50, \sigma = 4$ .

Let  $z$  be the standard normal variable corresponding to  $x$ .

$$\therefore z = \frac{x - \mu}{\sigma} = \frac{x - 50}{4}$$

$$\therefore \text{When } x = 42, \quad z = \frac{42 - 50}{4} = -2$$

$$\text{When } x = 84, \quad z = \frac{84 - 50}{4} = 8.5$$

$$\text{When } x = 85, \quad z = \frac{85 - 50}{4} = 8.75$$

$$\text{When } x = 32, \quad z = \frac{32 - 50}{4} = -4.5$$

$$\text{When } x = 40, \quad z = \frac{40 - 50}{4} = -2.5.$$

**Example 10.22.** Find the area under the standard normal curve which lies:

(i) to the right of  $z = 2.70$

(ii) to the left of  $z = 1.73$

(iii) to the right of  $z = -0.66$

(iv) to the left of  $z = -1.88$

(v) between  $z = 1.25$  and  $z = 1.67$

(vi) between  $z = -1.85$  and  $z = -0.90$

(vii) between  $z = -1.45$  and  $z = 1.45$

(viii) between  $z = -0.9$  and  $z = 1.58$ .

**Solution.** The variable  $z$  is a standard normal variable (S.N.V.).

$\therefore$  Total area under the curve of  $z$  and

$z$ -axis is one.

(i) Area to the right of  $z = 2.7$

$$= 0.5 - \text{area between } z = 0$$

and

$$z = 2.7$$

$$= 0.5 - 0.4965$$

(Using area Table)

$$= 0.0035.$$

(ii) Area to the left of  $z = 1.73$

$$= 0.5 + \text{area between } z = 0$$

and

$$z = 1.73$$

$$= 0.5 + 0.4582 = 0.9582.$$

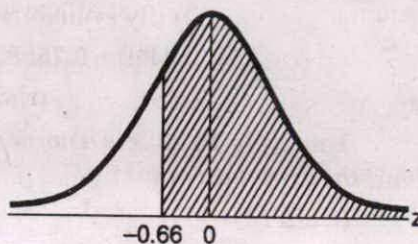
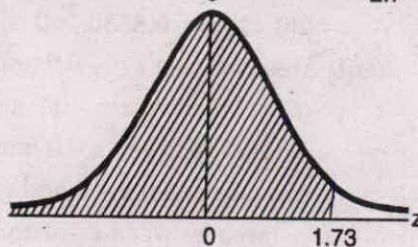
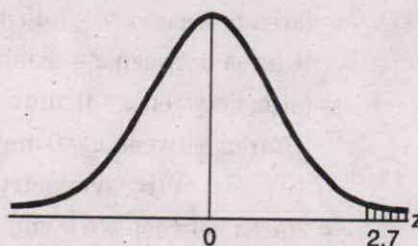
(iii) Area to the right of  $z = -0.66$

$$= (\text{area from } z = -0.66 \text{ to } z = 0) + 0.5$$

$$= (\text{area from } z = 0 \text{ to } z = 0.66) + 0.5$$

(By symmetry of curve)

$$= 0.2454 + 0.5 = 0.7454. \quad (\text{Using Table})$$

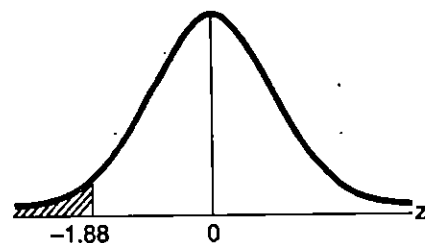


## NOTES

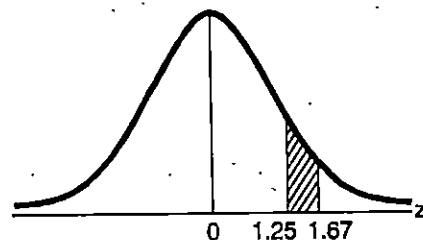
## NOTES

(iv) Area to the left of  $z = -1.88$   
 $= 0.5 - \text{area between } z = -1.88$

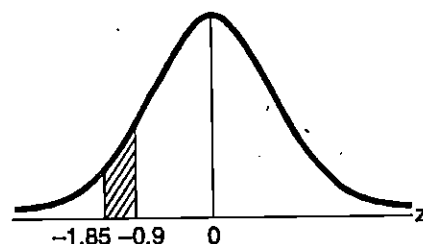
and  $z = 0$   
 $= 0.5 - \text{area between } z = 0$   
 and  $z = 1.88$  (By symmetry of curve)  
 $= 0.5 - 0.4699 = 0.0301$ .  
 (Using Table)



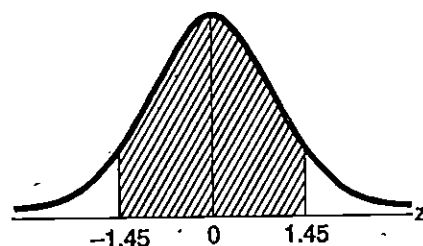
(v) Area between  $z = 1.25$  and  $z = 1.67$   
 $= (\text{area between } z = 0 \text{ and } z = 1.67)$   
 $- (\text{area between } z = 0 \text{ and } z = 1.25)$   
 $= 0.4525 - 0.3944 = 0.0581$ .  
 (Using Table)



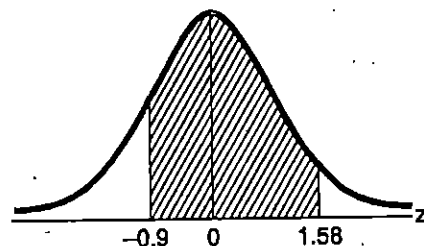
(vi) Area between  $z = -1.85$  and  $z = -0.90$   
 $= (\text{area between } z = -1.85 \text{ and } z = 0)$   
 $- (\text{area between } z = -0.90 \text{ and } z = 0)$   
 $= (\text{area between } z = 0 \text{ and } z = 1.85)$   
 $- (\text{area between } z = 0 \text{ and } z = 0.90)$   
 (By symmetry of curve)  
 $= 0.4678 - 0.3159 = 0.1519$ .  
 (Using Table)



(vii) Area between  $z = -1.45$  and  $z = 1.45$   
 $= (\text{area between } z = -1.45 \text{ and } z = 0)$   
 $+ (\text{area between } z = 0 \text{ and } z = 1.45)$   
 $= (\text{area between } z = 0 \text{ and } z = 1.45)$   
 $+ (\text{area between } z = 0 \text{ and } z = 1.45)$   
 (By symmetry of curve)  
 $= 2(\text{area between } z = 0 \text{ and } z = 1.45)$   
 $= 2(0.4265) = 0.8530$ . (Using Table)



(viii) Area between  $z = -0.9$  and  $z = 1.58$   
 $= (\text{area between } z = -0.9 \text{ and } z = 0)$   
 $+ (\text{area between } z = 0 \text{ and } z = 1.58)$   
 $= (\text{area between } z = 0 \text{ and } z = 0.9)$   
 $+ (\text{area between } z = 0 \text{ and } z = 1.58)$   
 (By symmetry of curve)  
 $= 0.3159 + 0.4429 = 0.7588$ .  
 (Using Table)



**Example 10.23.**  $x$  is a normal variable with mean 25 and standard deviation 5. Find the probability that :

(i)  $x \leq 10$   
 (iii)  $|x - 30| \geq 10$ .

(ii)  $15 \leq x \leq 30$



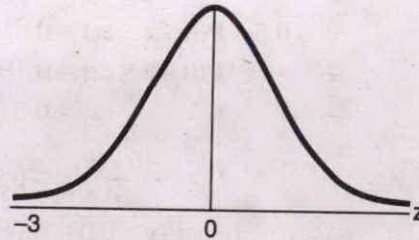
**Solution.** We have mean = 25, S.D. = 5.

Let  $z$  be the S.N.V. corresponding to  $x$ .

$$\therefore z = \frac{x - \mu}{\sigma} = \frac{x - 25}{5}$$

(i) When  $x = 10$ ,  $z = \frac{10 - 25}{5} = -3$

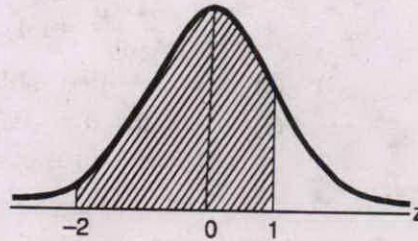
$$\begin{aligned} \therefore \text{Required probability} &= P(x \leq 10) \\ &= P(z \leq -3) = P(z \geq 3) \\ &= 0.5 - P(0 \leq z \leq 3) \\ &= 0.5 - 0.4987 \\ &= \mathbf{0.0013.} \end{aligned}$$



(ii) When  $x = 15$ ,  $z = \frac{15 - 25}{5} = -2$

When  $x = 30$ ,  $z = \frac{30 - 25}{5} = 1$

$$\begin{aligned} \therefore \text{Required probability} &= P(15 \leq x \leq 30) \\ &= P(-2 \leq z \leq 1) \\ &= P(-2 \leq z \leq 0) + P(0 \leq z \leq 1) \\ &= P(0 \leq z \leq 2) + P(0 \leq z \leq 1) \\ &= 0.4772 + 0.3413 \\ &= \mathbf{0.8185.} \end{aligned}$$



(iii) Required probability =  $P(|x - 30| \geq 10)$

$$= 1 - P(|x - 30| < 10)$$

$$= 1 - P(30 - 10 < x < 30 + 10)$$

$$= 1 - P(20 < x < 40)$$

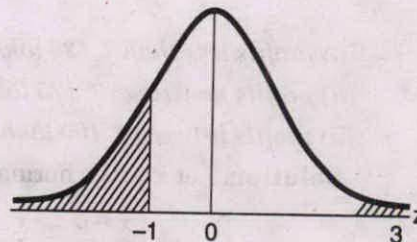
$$= 1 - P\left(\frac{20 - 25}{5} < \frac{x - 25}{5} < \frac{40 - 25}{5}\right)$$

$$= 1 - P(-1 < z < 3)$$

$$= 1 - \{P(-1 \leq z \leq 0) + P(0 \leq z \leq 3)\}$$

$$= 1 - \{P(0 \leq z \leq 1) + P(0 \leq z \leq 3)\}$$

$$= 1 - \{0.3413 + 0.4987\} = \mathbf{0.16.}$$



$$[\because P(z = -1) = P(z = 3) = 0]$$

**Example 10.24.** In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution.

**Solution.** Let  $x$  be the normal variable and  $z$  be its S.N.V. Let  $\mu$  and  $\sigma$  be the mean and standard deviation of  $x$ .

By the given conditions,

$$P(x < 45) = \frac{31}{100} = 0.31 \quad \text{and} \quad P(x > 64) = \frac{8}{100} = 0.08.$$

Since  $P(x < 45) < 0.5$ ,  $\therefore x = 45$  lies on the left of  $x = \mu$  and so the value of the corresponding  $z$ -variable is -ve.

## NOTES

## NOTES

When  $x = 45$ , let  $z = \frac{45 - \mu}{\sigma} = -z_1$  ( $z_1 > 0$ )

When  $x = 64$ , let  $z = \frac{64 - \mu}{\sigma} = z_2$  ( $z_2 > 0$ )

$$\therefore P(x < 45) = 0.31$$

$$\Rightarrow P(z < -z_1) = 0.31$$

$$\Rightarrow P(z > z_1) = 0.31$$

$$\Rightarrow 0.5 - P(0 \leq z \leq z_1) = 0.31$$

$$\Rightarrow P(0 \leq z \leq z_1) = 0.19$$

$$\Rightarrow z_1 = 0.5 \text{ (From area table)}$$

$$\therefore \frac{45 - \mu}{\sigma} = -0.5 \quad \text{or} \quad 45 - \mu = -0.5 \sigma \quad \dots(1)$$

Also  $P(x > 64) = 0.08$

$$\Rightarrow P(z > z_2) = 0.08 \quad \Rightarrow 0.5 - P(0 \leq z \leq z_2) = 0.08$$

$$\Rightarrow P(0 \leq z \leq z_2) = 0.42 \quad \Rightarrow z_2 = 1.4 \text{ (From area table)}$$

$$\therefore \frac{64 - \mu}{\sigma} = 1.4 \quad \text{or} \quad 64 - \mu = 1.4 \sigma \quad \dots(2)$$

$$(1) - (2) \Rightarrow -19 = -1.9\sigma \quad \Rightarrow \sigma = 10$$

$$\therefore (1) \Rightarrow 45 - \mu = -0.5(10) = -5$$

$$\Rightarrow \mu = 45 + 5 = 50$$

$$\therefore \mu = 50 \text{ and } \sigma = 10.$$

**Example 10.25.** The profits of 400 companies are normally distributed with mean ₹ 150 lakhs and standard deviation ₹ 20 lakhs. Estimate the number of companies with:

(i) profits less than ₹ 128 lakhs

(ii) profits more than ₹ 175 lakhs

(iii) profits between ₹ 100 lakhs and ₹ 138 lakhs.

**Solution.** Let  $x$  be the normal variable 'profit'. Let  $z$  be the corresponding S.N.V.

$$\therefore z = \frac{x - \mu}{\sigma} = \frac{x - 150}{20}$$

$$(i) P(\text{profit less than ₹ 128 lakhs}) = P(x < 128)$$

$$= P(x - 150 < 128 - 150)$$

$$= P\left(\frac{x - 150}{20} < \frac{-22}{20}\right)$$

$$= P(z < -1.1) = P(z > 1.1)$$

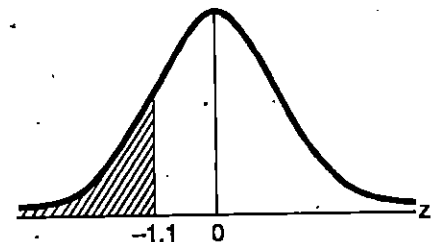
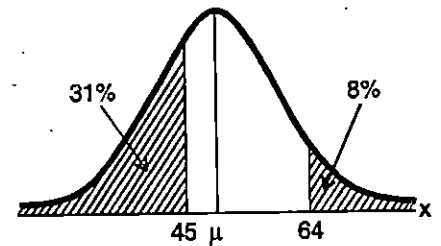
(By symmetry)

$$= 0.5 - P(0 < z \leq 1.1)$$

$$= 0.5 - 0.3643 = 0.1357.$$

No. of companies with profit less than ₹ 128 lakhs

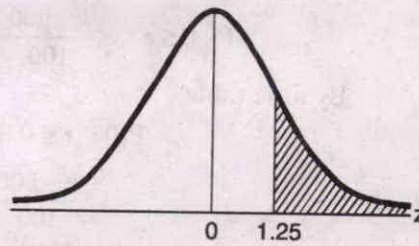
$$= 400 P(x < 128) = 400 \times 0.1357 = 54.$$





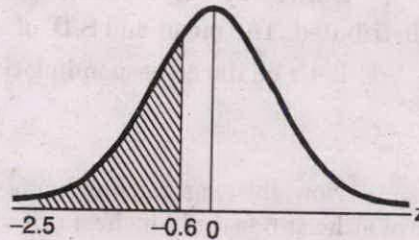
## NOTES

$$\begin{aligned}
 \text{(ii) } P(\text{profit more than ₹ 175 lakhs}) &= P(x > 175) = P(x - 150 > 175 - 150) \\
 &= P\left(\frac{x - 150}{20} > \frac{25}{20}\right) = P(z > 1.25) \\
 &= 0.5 - P(0 < z \leq 1.25) \\
 &= 0.5 - 0.3944 = 0.1056.
 \end{aligned}$$



$$\begin{aligned}
 \therefore \text{No. of companies with profit more than ₹ 175 lakhs} &= 400 P(x > 175) \\
 &= 400 \times 0.1056 = 42.
 \end{aligned}$$

$$\begin{aligned}
 \text{(iii) } P(\text{profit between ₹ 100 lakhs and ₹ 138 lakhs}) &= P(100 < x < 138) \\
 &= P(100 - 150 < x - 150 < 138 - 150) \\
 &= P\left(\frac{-50}{20} < \frac{x - 150}{20} < \frac{-12}{20}\right) \\
 &= P(-2.5 < z < -0.6) \\
 &= P(0.6 < z < 2.5) \quad (\text{By symmetry}) \\
 &= P(0 < z < 2.5) - P(0 < z < 0.6) \\
 &= 0.4938 - 0.2257 = 0.2681.
 \end{aligned}$$



$$\begin{aligned}
 \therefore \text{No. of companies with profits between ₹ 100 lakhs and ₹ 138 lakhs} &= 400 P(100 < x < 138) \\
 &= 400 \times 0.2681 = 107.
 \end{aligned}$$

**Example 10.26.** The marks obtained by students in a degree examination are normally distributed. The mean marks and S.D. of the distribution are 500 and 100 respectively. If 674 appeared in the examination and out of these, 550 are to be declared passed, what should be the minimum pass marks?

**Solution.** Let  $x$  denote the normal variable marks. Let  $z$  be the S.N.V. of  $x$ .

$$z = \frac{x - 500}{100}.$$

Let minimum pass marks be  $k$ .

$$\therefore 674 P(x \geq k) = 550$$

$$\Rightarrow P(x \geq k) = \frac{550}{674} = 0.816$$

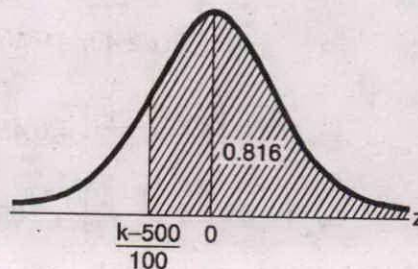
$$\text{or } P\left(\frac{x - 500}{100} \geq \frac{k - 500}{100}\right) = 0.816$$

$$\text{or } P\left(z \geq \frac{k - 500}{100}\right) = 0.816$$

$$\text{or } P\left(\frac{k - 500}{100} \leq z\right) = 0.816$$

$$\text{or } P\left(\frac{k - 500}{100} \leq z \leq 0\right) + 0.5 = 0.816$$

$$\text{or } P\left(\frac{k - 500}{100} \leq z \leq 0\right) = 0.316$$



or

$$P\left(0 \leq z \leq -\frac{k-500}{100}\right) = 0.316$$

(By symmetry of normal curve)

By area table,

$$P(0 \leq z \leq 0.9) = 0.316 \text{ (Approx.)}$$

$$\therefore -\frac{k-500}{100} = 0.9$$

$$\text{i.e., } 500 - k = 90 \quad \text{or } k = 500 - 90 = 410.$$

$\therefore$  Minimum pass marks = 410.

**Example 10.27.** Assuming the mean height of soldiers to be 68.22 inches with a variance of 10.8 (inches)<sup>2</sup>, find how many soldiers in a regiment of 10,000 would you expect to be over 6 feet tall?

**Solution.** Let  $x$  denote the variable height. We assume that  $x$  is normally distributed. The mean and S.D. of  $x$  are 68.22 and  $\sqrt{10.8} = 3.2863$ .

Let  $z$  be the corresponding S.N.V.

$$\therefore z = \frac{x - 68.22}{3.2863}$$

Now, the expected number of soldiers, who are at least 6 feet (72 inches) tall

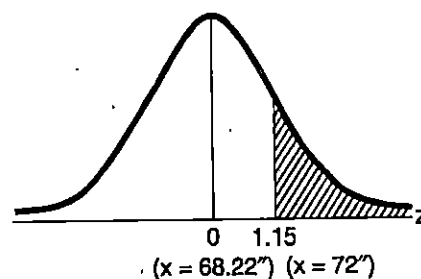
$$= 10,000 P(x > 72)$$

$$= 10,000 P\left(\frac{x - 68.22}{3.2863} > \frac{72 - 68.22}{3.2863}\right)$$

$$= 10,000 P(z > 1.15)$$

$$= 10,000 (0.5 - P(0 \leq z \leq 1.15))$$

$$= 10,000 (0.5 - 0.3749) = 1251.$$



**Example 10.28.** In a certain examination, the percentages of passes and distinctions were 45 and 9 respectively. Estimate the average marks obtained by the candidates, the minimum pass and distinction marks being 40 and 75. (Assume the distribution of marks to be normal).

**Solution.** Let  $x$  denote the normal variable marks. Let  $\mu$  and  $\sigma$  be the mean and standard deviation of  $x$  respectively. Let  $z$  be the S.N.V. of  $x$ .

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

Now, percentage of passes = 45%

$$\therefore 100 P(x \geq 40) = 45$$

$$\Rightarrow P(x \geq 40) = 0.45$$

$$\Rightarrow P\left(\frac{x - \mu}{\sigma} \geq \frac{40 - \mu}{\sigma}\right) = 0.45$$

$$\Rightarrow P\left(z \geq \frac{40 - \mu}{\sigma}\right) = 0.45$$

$$\Rightarrow 0.5 - P\left(0 \leq z \leq \frac{40 - \mu}{\sigma}\right) = 0.45$$

$$\Rightarrow P\left(0 \leq z \leq \frac{40 - \mu}{\sigma}\right) = 0.05.$$

Also from the table,

$$P(0 \leq z \leq 0.12) \doteq 0.05$$



$$\therefore \frac{40 - \mu}{\sigma} = 0.12 \quad \dots(1)$$

Also, percentage of distinctions = 9%

$$\therefore 100 P(x \geq 75) = 9 \Rightarrow P(x \geq 75) = 0.09$$

$$\Rightarrow P\left(\frac{x - \mu}{\sigma} \geq \frac{75 - \mu}{\sigma}\right) = 0.09 \Rightarrow P\left(z \geq \frac{75 - \mu}{\sigma}\right) = 0.09$$

$$\Rightarrow 0.5 - P\left(0 \leq z \leq \frac{75 - \mu}{\sigma}\right) = 0.09 \Rightarrow P\left(0 \leq z \leq \frac{75 - \mu}{\sigma}\right) = 0.41$$

Also, from the table,

$$P(0 \leq z \leq 1.34) \doteq 0.41$$

$$\therefore \frac{75 - \mu}{\sigma} = 1.34 \quad \dots(2)$$

Dividing (1) by (2), we get

$$\frac{40 - \mu}{75 - \mu} = \frac{0.12}{1.34} \Rightarrow \mu = 36.5576 \doteq 37$$

$\therefore$  Average marks = 37.

### EXERCISE 10.7

- The mean and standard deviation of a normal variable are 35 and 5 respectively. Find the values of the corresponding S.N.V., when  $x = 10, 15, 22, 34, 35, 55$ .
- If  $z$  is a standard normal variable, then find the following probabilities:
  - $P(1 \leq z \leq 2)$
  - $P(z \geq 3)$ .
- $x$  is a normal variable with mean 50 and standard deviation 8. Find the probabilities:
  - $x \leq 60$
  - $10 \leq x \leq 40$
  - $x > 60$ .
- A normal curve has  $\bar{x} = 40$  and  $\sigma = 15$ . Find the area between  $x_1 = 25$  and  $x_2 = 60$ .
- The income of a group of 5000 persons was found to be normally distributed with mean ₹ 700 and standard deviation ₹ 50. Find the expected number of persons getting (i) less than Rs. 680 (ii) more than ₹ 750 and (iii) between ₹ 680 and ₹ 750.
- The marks obtained by a large group of students in a final examination in statistics have mean 68 and standard deviation 9. If these marks are normally distributed, what percentage of students can you expect to have secured marks between 60 and 65, both inclusive?
- In a sample of 120 workers in a factory, the mean and standard deviation of wages were ₹ 11.35 and ₹ 3.03 respectively. Find the percentage of workers getting wages between ₹ 9 and ₹ 17 in the whole factory, assuming that the wages to be normally distributed.
- 5000 candidates appeared in a certain examination paper carrying a maximum of 100 marks. It was found that the marks were normally distributed with mean 39.5 and standard deviation 12.5. Determine the approximate number of candidates who secured a first class for which a minimum of 60 is necessary.
- In a large group of persons, it is found that 5% are under 60 inches and 40% are between 60 and 65 inches in height. Assuming the distribution to be normal, find the mean and standard deviation of the height.

### NOTES



## Answers

1. -5, -4, -2.6, -0.2, 0, 4  
 2. (i) 0.1359 (ii) 0.0013  
 3. (i) 0.8944 (ii) 0.1056 (iii) 0.1056  
 4. 0.7495 5. (i) 1723 (ii) 793 (iii) 2484  
 6. 18.4% 7. 100 P(9 < x < 17) = 75.09% 8. 252  
 9. 65.41, 3.28.

## NOTES

### 10.35. FITTING OF A NORMAL DISTRIBUTION

Let  $x$  be a normal variable. Let the probability density function  $P(x)$  of  $x$  be given by

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

where  $\mu$  and  $\sigma$  are the mean and standard deviation of the distribution respectively. For calculating the expected normal frequencies, corresponding to an observed frequency distribution, the following steps are taken:

- (i) The values of mean ( $\mu$ ) and S.D. ( $\sigma$ ) are calculated by usual methods.
- (ii) The values of the standard normal variable  $z = \frac{x-\mu}{\sigma}$  are calculated corresponding to each lower limit of classes in the given distribution.
- (iii) Corresponding to these values of  $z$ , the areas under the normal curve to the left of these ordinates are calculated. This is done by using the table given at the end.
- (iv) The areas for the successive classes are obtained by subtracting the corresponding areas calculated in step (iii).
- (v) The areas for the successive classes are multiplied by  $N$  to get the required expected frequencies.

**Example 10.29.** Fit a normal curve to the following data:

Class	0—5	5—10	10—15	15—20	20—25	25—30	30—35	35—40	40—45
No. of items	20	24	32	28	20	16	34	10	16

**Solution.**

**Calculation of Mean and S.D.**

Class	$f$	$x$	$d = x - A$ $A = 22.5$	$u = d/h$ $h = 5$	$fu$	$fu^2$
0—5	20	2.5	-20	-4	-80	320
5—10	24	7.5	-15	-3	-72	216
10—15	32	12.5	-10	-2	-64	128
15—20	28	17.5	-5	-1	-28	28
20—25	20	22.5	0	0	0	0
25—30	16	27.5	5	1	16	16
30—35	34	32.5	10	2	68	136
35—40	10	37.5	15	3	30	90
40—45	16	42.5	20	4	64	250
Total	200				-66	1190

$$\text{Mean } (\mu) = A + \left( \frac{\Sigma fu}{N} \right) h = 22.5 + \left( \frac{-66}{200} \right) 5 = 20.85$$

$$\text{S.D. } (\sigma) = \sqrt{\frac{\Sigma fu^2}{N} - \left( \frac{\Sigma fu}{N} \right)^2} \times h = \sqrt{\frac{1190}{200} - \left( \frac{-66}{200} \right)^2} \times 5 = 12.084.$$

## NOTES

## Calculation of Expected Frequencies

Class	Lower class limit $x$	$z = \frac{x - \mu}{\sigma} = \frac{x - 20.85}{12.084}$	Area under normal curve to the left of ordinate at $z$	Area corresponding to class	Expected frequency $= N \times \text{Area}$
0—5	0	-1.72	0.0427	0.0524	$10.48 \div 10$
5—10	5	-1.31	0.0951	0.0890	$17.80 \div 18$
10—15	10	-0.90	0.1841	0.1315	$26.30 \div 26$
15—20	15	-0.48	0.3156	0.1565	$31.30 \div 31$
20—25	20	-0.07	0.4721	0.1610	$32.20 \div 32$
25—30	25	0.34	0.6331	0.1433	$28.66 \div 29$
30—35	30	0.76	0.7764	0.1026	$20.52 \div 21$
35—40	35	1.17	0.8790	0.0639	$12.78 \div 13$
40—45	40	1.58	0.9429	0.0338	$6.76 \div 7$
45—50	45	1.99	0.9767	—	—

## EXERCISE 10.8

1. Fit a normal curve to the following data:

Variable	60—62	63—65	66—68	69—71	72—74
Frequency	5	18	42	27	8

2. Fit a normal curve to the following data:

Class	60—65	65—70	70—75	75—80
Frequency	3	21	150	335
	80—85	85—90	90—95	95—100
	326	135	26	4

## Answers

1. 4, 21, 39, 28, 8      2. 3, 31, 148, 322, 319, 144, 30, 3.

## 10.36. SUMMARY

- The binomial distribution is a particular type of probability distribution. This was discovered by **James Bernoulli (1654—1705)** in the year 1700. This

## NOTES

distribution mainly deals with attributes. An attribute is either present or absent with respect to elements of a population.

- A random variable which counts the number of successes in a random experiment with trials satisfying above four conditions is called a **Binomial variable**.
- The shape of the binomial distribution depends upon the probability of success ( $p$ ) and the number of trials in the experiment. If  $p = q = \frac{1}{2}$ , then the distribution will be symmetrical for every value of  $n$ . If  $p \neq q$ , then the distribution would be asymmetrical i.e., skewed. The magnitude of skewness varies as the difference between  $p$  and  $q$ .
- As number of trials ( $n$ ) in the binomial distribution increases, the number of successes also increases. If neither  $p$  nor  $q$  is very small, then as  $n$  approaches infinity, the skewness in the distribution disappears and it becomes continuous. We shall see that such a continuous, bell shaped distribution is called a *normal distribution*.
- The Poisson distribution is also a discrete probability distribution. This was discovered by French mathematician **Simon Denis Poisson** (1781 – 1840) in the year 1837. This distribution deals with the evaluation of probabilities of rare events such as “no. of car accidents on road”, “no. of earthquakes in a year”, “no. of misprints in a book”, etc.
- The Poisson distribution is derived as a limiting case of binomial distribution.
- A random variable which counts the number of successes in a random experiment with trials satisfying above conditions is called a **Poisson variable**.
- The normal distribution is a particular type of continuous probability distribution. This was discovered by **De Moivre** (1667—1754) in the year 1733. The normal distribution is obtained as a limiting case of a binomial distribution when  $n$ , the number of trials is indefinitely large and neither  $p$  nor  $q$  is very small.

---

### 10.37. REVIEW EXERCISES

---

1. What are the conditions under which Binomial probability model is appropriate
2. Explain the utility of Poisson distribution in practical life.
3. What is a normal probability distribution? What are the salient features of a normal curve?
4. Explain the distinctive features of Binomial and Poisson distributions.
5. What is binomial distribution? Under what conditions will it tend to a normal distribution?
6. What is Poisson distribution? Point out its role.
7. Explain the characteristics of Poisson distribution.
8. Explain the properties of a Binomial distribution. What is its relationship with Poisson distribution?
9. Write short note on Normal distribution.
10. Explain the properties of Normal distribution.
11. How does a normal distribution differ from a binomial distribution? What are the important properties of a normal distribution?
12. Discuss the conditions for the Binomial distribution. What are its important properties?
13. Define binomial distribution and explain its important features.
14. What is meant by theoretical frequency distribution? Discuss the salient features of the Binomial and Normal distributions.
15. Differentiate between Normal and Binomial distributions.
16. What is meant by Theoretical Frequency Distribution? Discuss the main features of Binomial, Poisson and Normal distributions.



# 11. ESTIMATION THEORY AND HYPOTHESIS TESTING

Estimation Theory and  
Hypothesis Testing

NOTES

## STRUCTURE

- 11.1. Introduction
- 11.2. Null Hypothesis and Alternative Hypothesis
- 11.3. Level of Significance and Confidence Limits
- 11.4. Type I Error and Type II Error
- 11.5. Power of the Test

### I. Test of Significance for Small Samples

- 11.6. Student's  $t$ -Test
- 11.7. Assumptions for Student's  $t$ -Test
- 11.8. Degree of Freedom
- 11.9. Test for Single Mean
- 11.10.  $t$ -test for Difference of Means
- 11.11. Paired  $t$ -test for Difference of Means
- 11.12. F-Test
- 11.13. Properties of F-Distribution
- 11.14. Procedure to F-Test
- 11.15. Critical Values of F-Distribution

### II. Test of Significance for Large Samples

- 11.16. Test of significance for Proportion
- 11.17. Test of Significance for Single Mean
- 11.18. Test of Significance for Difference of Means
- 11.19. Chi-square Test
- 11.20. Chi-square Test to Test the Goodness of Fit
- 11.21. Chi-square Test to Test the Independence of Attributes
- 11.22. Conditions for  $\chi^2$  Test
- 11.23. Uses of  $\chi^2$  Test
- 11.24. Summary
- 11.25. Review Exercises

## 11.1. INTRODUCTION

To describe a set of data or observations, we use statistics such as mean and standard deviation. These statistics are estimated from samples. Sample is nothing but a small section selected from the population and the process of drawing or selecting a sample from the population is called 'sampling'. It is essential that a sample must be a random

## NOTES

selection so that each member of the population has the equal chance of being selection in the sample. A statistical population consists of observations of some characteristic of interest associated with the individuals concerned and not the individual items or persons themselves.

A statistical measure based only on all the units selected in a sample is called 'statistic', e.g., sample mean, sample standard deviation, proportion of defectives, etc. whereas a statistical measure based on all the units in the population is called 'parameter'. The terms like mean, median, mode, standard deviation are called parameters when they describe the characteristics of the population and are called statistic when they describe the characteristics of the sample.

A very important aspect of the sampling theory is the study of the tests of significance which enables us to decide on the basis of the sample results whether to accept or reject the hypothesis. A test of significance can be used to compare the characteristics of two samples of the same type. Some of the well known tests of significance for small samples are *t*-test and *F*-test.

## 11.2. NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

A statistical hypothesis is a statement about a population parameter. There are two types of statistical hypothesis, null hypothesis and alternative hypothesis.

The hypothesis formulated for the sake of rejecting it under the assumption that it is true, is called the null hypothesis and is denoted by  $H_0$ . Null hypothesis asserts that there is no significant difference between the sample statistic and the population parameter and whatever difference is observed that is merely due to fluctuations in sampling from the same population.

Rejecting null hypothesis implies that it is rejected in favour of some other hypothesis which is accepted. A hypothesis which is accepted when  $H_0$  is rejected is called the alternative hypothesis and is denoted by  $H_1$ . What we intend to conclude is stated in the alternative hypothesis.

## 11.3. LEVEL OF SIGNIFICANCE AND CONFIDENCE LIMITS

The probability level below which we reject the hypothesis is known as the 'level of significance'. The region in which a sample value falling is rejected, is known as the 'critical region' or the 'rejection region'. We, generally, take two critical regions which cover 5% and 1% areas of the normal curve.

Depending on the nature of the problem, we use a single-tail test or double-tail test to estimate the significance of a result. In a single-tail test, only the area on the right of an ordinate is taken into consideration whereas in a double-tail test, the areas of both the tails of the curve representing the sampling distribution are taken into consideration.

For example, a test for testing the mean of a population

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis  $H_1 : \mu > \mu_0$  (right tailed) or  $H_1 : \mu < \mu_0$  (left tailed) is a single tailed test. In the right tailed test ( $H_1 : \mu > \mu_0$ ), the critical region lies entirely



in the right tail of the sampling distribution; while for the left tail test ( $H_1: \mu < \mu_0$ ), the critical region is entirely in the left tail of the sampling distribution.

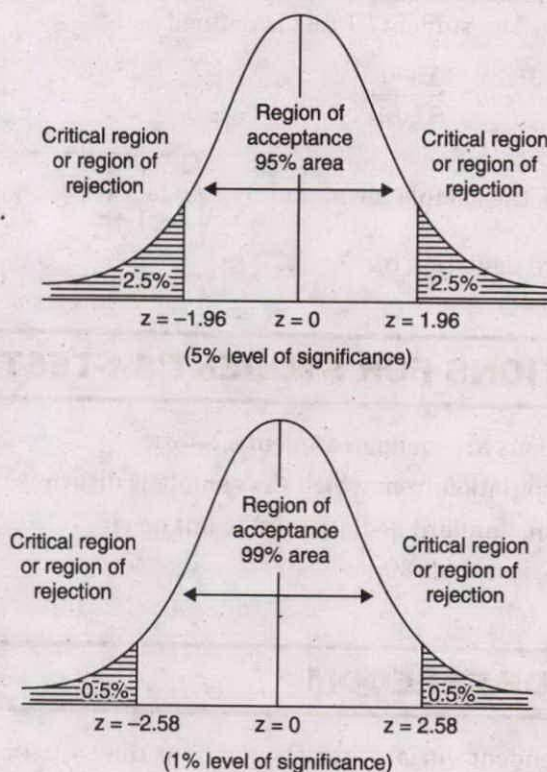
A test of statistical hypothesis where the alternative hypothesis is two tailed such as:

$H_0: \mu = \mu_0$  against the alternative hypothesis

$H_1: \mu \neq \mu_0$  ( $\mu > \mu_0$  and  $\mu < \mu_0$ ) is known as two tailed test and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of the test statistic.

The value of  $z$  corresponding to 5% level of significance is  $\pm 1.96$  and corresponding to 1% level of significance value of  $z$  is  $\pm 2.58$ . The set of  $z$ -scores outside the range  $\pm 1.96$  and  $\pm 2.58$  constitute the critical region of the hypothesis (or the region of rejection) at 5% and 1% level of significance respectively.

The following figure showing region of acceptance and rejection for 5% and 1% level of significance.



## 11.4. TYPE I ERROR AND TYPE II ERROR

The error of rejecting  $H_0$  when  $H_0$  is true is called the type I error and the error of accepting  $H_0$  when  $H_0$  is false ( $H_1$  is true) is called the type II error. The probability of type I error is denoted by  $\alpha$  and the probability of type II error is denoted by  $\beta$ .

$P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) = \alpha$

$P(\text{accepting } H_0 \text{ when } H_1 \text{ is true}) = \beta$

## NOTES

## NOTES

**11.5. POWER OF THE TEST**

A good test should accept the null hypothesis when it is true and reject the null hypothesis when it is false.  $1 - \beta$  (i.e., 1-probability of type II error) measures how well the test is working and is called the power of the test.

Power of the test =  $1 - \beta$ .

**I. TEST OF SIGNIFICANCE FOR SMALL SAMPLES****11.6. STUDENT'S t-TEST**

Let  $x_1, x_2, \dots, x_n$  be a random sample of size  $n$  ( $n < 30$ ) from a normal population with mean  $\mu$  and variance  $\sigma^2$ . The student's  $t$ -test is defined as

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}},$$

where  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , is the sample mean and  $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$  is an unbiased estimate of the standard deviation  $\sigma$ .

**11.7. ASSUMPTIONS FOR STUDENT'S t-TEST**

The following assumptions are made in student's  $t$ -test:

- (i) The parent population from which the sample is drawn is normal.
- (ii) The population standard deviation ( $\sigma$ ) is unknown.
- (iii) Sample size is less than 30.

**11.8. DEGREE OF FREEDOM**

The number of independent variates which make up the statistic is known as the degree of freedom (d.f.) and is denoted by  $\nu$  (the letter 'Nu' of the Greek alphabet).

In general the degree of freedom is defined as

d.f. = number of frequencies - number of independent constraints on them.

**11.9. TEST FOR SINGLE MEAN**

Suppose we want to test

- (i) If a random sample  $x_i$  ( $i = 1, 2, \dots, n$ ) of size  $n$  has been drawn from a normal population with a specified mean say  $\mu$  or

(ii) If the sample mean differs significantly from the hypothetical value  $\mu$  of the population mean.

Under null hypothesis  $H_0$ :

(i) The sample mean has been drawn from the population with mean  $\mu$  or

(ii) There is no significant difference between the sample mean  $\bar{x}$  and the population mean  $\mu$ , the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}},$$

$$\text{where } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

follows Student's  $t$ -distribution with  $(n - 1)$  degrees of freedom.

We now compare the calculated value of  $t$  with the tabulated value at certain level of significance. If calculated  $|t| >$  tabulated  $t$ ,  $H_0$  is rejected and if calculated  $|t| <$  tabulated  $t$ ,  $H_0$  may be accepted.

**Note.** We know, the sample variance,

$$s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$\therefore ns^2 = (n - 1) S^2$$

$$\text{or } \frac{S^2}{n} = \frac{s^2}{n-1} \Rightarrow \frac{S}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$$

Hence, the test statistic becomes

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}.$$

**Example 11.1.** The mean weekly sales of soap bars in departmental stores was 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful?

**Solution.** Here,  $n = 22$ ,  $\bar{x} = 153.7$ ,  $s = 17.2$

Null hypothesis  $H_0 : \mu = 146.3$ , i.e., the advertising campaign is not successful.

Alternative hypothesis  $H_1 : \mu > 146.3$  (Right tail)

Under  $H_0$ , the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \text{ with } (22 - 1) = 21 \text{ d.f.}$$

$$t = \frac{153.7 - 146.3}{17.2/\sqrt{22-1}} = \frac{7.4 \times \sqrt{21}}{17.2} = 9.$$

Since calculated value of  $t = 9$  is greater than the tabulated value of  $t = 1.72$  for 21 d.f. at 5% level of significance. It is highly significant. So  $H_0$  is rejected, i.e., the advertising campaign was successful in promoting sales.

**Example 11.2.** Ten individuals are chosen at random from a normal population and the heights are found to be in inches 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71. Test if the sample belongs to the population whose mean height is 66 inches. (Given  $t_{0.05} = 2.26$  for 9 d.f.)

## NOTES

## NOTES

## Solution.

$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
63	-4.8	23.04
63	-4.8	23.04
66	-1.8	3.24
67	-0.8	0.64
68	0.2	0.04
69	1.2	1.44
70	2.2	4.84
70	2.2	4.84
71	3.2	10.24
71	3.2	10.24
$\Sigma x_i = 678$		$\Sigma(x_i - \bar{x})^2 = 81.6$

Here,  $n = 10$

$$\bar{x} = \text{sample mean} = \frac{\Sigma x_i}{n} = \frac{678}{10} = 67.8 \text{ inches}$$

$$S = \sqrt{\frac{1}{n-1} \Sigma(x_i - \bar{x})^2} = \sqrt{\frac{1}{9} \times 81.6}$$

$$= \sqrt{9.0667} = 3.011$$

Null hypothesis  $H_0: \mu = 66$ , i.e., population mean is 66 inches  
Under  $H_0$ , the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{67.8 - 66}{3.011/\sqrt{10}} = \frac{1.8 \times \sqrt{10}}{3.011}$$

$$= \frac{5.692}{3.011} = 1.8904$$

$$\text{degree of freedom} = n - 1 = 10 - 1 = 9$$

$$t_{0.05} = 2.26 \text{ for } 9 \text{ d.f.}$$

As the calculated value of  $|t|$  is less than  $t_{0.05}$ , the difference between  $\bar{x}$  and  $\mu$  may be due to fluctuations of random sampling.  $H_0$  may be accepted. In other words, the data does not provide any significant evidence against the hypothesis that the population mean is 66 inches.

**Example 11.3.** A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Show that the assumption of a mean of 43.5 inches for the population is not reasonable. (Given  $t_{0.05} = 2.13$ ,  $t_{0.01} = 2.95$  for 15 degrees of freedom)

**Solution.** Here,  $\bar{x} = 41.5$  inches,  $n = 16$ ,  $\Sigma(x_i - \bar{x})^2 = 135$  sq. inches

$$S = \sqrt{\frac{1}{n-1} \Sigma(x_i - \bar{x})^2} = \sqrt{\frac{1}{15} \times 135} = \sqrt{9} = 3$$

Null hypothesis  $H_0: \mu = 43.5$  inches, i.e., the data are consistent with an assumption that the mean height in population is 43.5 inches.

Alternative hypothesis  $H_1: \mu \neq 43.5$  inches

Under  $H_0$ , the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$|t| = \frac{|41.5 - 43.5|}{3/\sqrt{16}} = \frac{2 \times 4}{3} = 2.667$$

$$\text{degrees of freedom} = n - 1 = 16 - 1 = 15$$



We are given  $t_{0.05} = 2.13$  and  $t_{0.01} = 2.95$  for 15 degrees of freedom.

Since calculated  $|t|$  is greater than  $t_{0.05} = 2.13$ , null hypothesis  $H_0$  is rejected at 5% level of significance and we conclude that the assumption of mean 43.5 inches for the population is not reasonable.

**Remark.** Since calculated  $|t|$  is less than  $t_{0.01} = 2.95$ , null hypothesis  $H_0$  may be accepted at 1% level of significance.

## 11.10. t-TEST FOR DIFFERENCE OF MEANS

Given two independent random samples  $x_i$  ( $i = 1, 2, \dots, n_1$ ) and  $y_j$  ( $j = 1, 2, \dots, n_2$ ) of sizes  $n_1$  and  $n_2$  with means  $\bar{x}$  and  $\bar{y}$  and standard deviations  $S_1$  and  $S_2$  from normal populations with the same variance, we have to test the hypothesis that the population means are same. In other words, since a normal distribution is completely specified by its mean and variance, we have to test the hypothesis that the two independent samples come from the same normal population.

The statistic is given by

$$t = \frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i; \bar{y} = \frac{1}{n_2} \sum_{j=1}^{n_2} y_j$$

and

$$S^2 = \frac{1}{(n_1 + n_2 - 2)} [(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2]$$

or

$$S^2 = \frac{1}{n_1 + n_2 - 2} \left[ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right]$$

follows Student's  $t$ -distribution with  $(n_1 + n_2 - 2)$  degrees of freedom.

If the calculated value of  $|t|$  be  $>$  tabulated  $t$ , the difference between the sample means is said to be significant at certain level of significance; otherwise the data are said to be consistent with the hypothesis.

## 11.11. PAIRED t-TEST FOR DIFFERENCE OF MEANS

If the size of the two samples is the same, say equal to  $n$ , and the data are paired,  $(x_i, y_i)$ , ( $i = 1, 2, \dots, n$ ) corresponds to the same  $i$ th sample unit. The problem is to test if the sample means differ significantly or not.

Here, we consider the increments,  $d_i = x_i - y_i$ , ( $i = 1, 2, \dots, n$ ).

Under the null hypothesis  $H_0$  that increments are due to fluctuations of sampling, the statistic

$$t = \frac{\bar{d}}{S/\sqrt{n}}$$

where

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$$

follows Student's  $t$ -distribution with  $(n - 1)$  degrees of freedom. If  $\sum d_i$  is negative, we may consider  $|\bar{d}|$ . This test is generally one tailed test. Therefore, the alternative hypothesis is  $H_1: \mu_1 > \mu_2$  or  $H_1: \mu_1 < \mu_2$ .



**Example 11.4.** The following data related to the heights (in cms) of two different varieties of wheat plants.

NOTES

Variety 1	63	65	68	69	71	72				
Variety 2	61	62	65	66	69	69	70	71	72	73

Test the null hypothesis that the mean heights of plants of both varieties are the same.

**Solution.** Given  $n_1 = 6$ ,  $n_2 = 10$

Null hypothesis  $H_0: \mu_1 = \mu_2$

Alternative hypothesis  $H_1: \mu_1 > \mu_2$  (right tail)

Under  $H_0$  the test statistic is given by

$$t = \frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Variety 1

$x$	$x - \bar{x} = x - 68$	$(x - \bar{x})^2$
63	-5	25
65	-3	9
68	0	0
69	1	1
71	3	9
72	4	16
$\Sigma x = 408$		$\Sigma(x - \bar{x})^2 = 60$

Variety 2

$y$	$y - \bar{y} = y - 67$	$(y - \bar{y})^2$
61	-6	36
62	-5	25
65	-2	4
65	-2	4
66	-1	1
66	-1	1
70	3	9
70	3	9
72	5	25
73	6	36
$\Sigma y = 670$		$\Sigma(y - \bar{y})^2 = 150$

$$\bar{x} = \frac{1}{n_1} \Sigma x_i = \frac{408}{6} = 68 \quad \bar{y} = \frac{1}{n_2} \Sigma y_i = \frac{670}{10} = 67$$

$$S^2 = \frac{1}{n_1 + n_2 - 2} [\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2]$$

$$= \frac{1}{6 + 10 - 2} [60 + 150] = \frac{210}{14} = 15 \Rightarrow S = 3.873$$

$$\therefore t = \frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{68 - 67}{3.873 \sqrt{\frac{1}{6} + \frac{1}{10}}} = \frac{1}{3.873 \times 0.5164} = 0.499$$

Tabulated  $t_{0.05}$  for 14 degrees of freedom for single tail-test is 1.76.

Since calculated value of  $t$  is less than 1.76, it is not at all significant at 5% level of significance. Hence,  $H_0$  may be accepted and we conclude that the height of the plants are not different at 5% level of significance.

**Example 11.5.** The mean values of birth weight with standard deviations and sample sizes are given below by socio-economic status. Is the mean difference in birth weight significant between socio-economic group?

## NOTES

	High socio-economic group	Low socio-economic group
Sample size	$n_1 = 15$	$n_2 = 10$
Birth weight (kg)	$\bar{x} = 2.91$	$\bar{y} = 2.26$
Standard deviation	$S_1 = 0.27$	$S_2 = 0.22$

**Solution.** Given  $n_1 = 15$ ,  $n_2 = 10$ ,  $\bar{x} = 2.91$ ,  $\bar{y} = 2.26$

$$S_1 = 0.27 \text{ and } S_2 = 0.22$$

Null hypothesis  $H_0 : \mu_1 = \mu_2$

Alternative hypothesis  $H_1 : \mu_1 > \mu_2$  (right tail), i.e. high socio-economic group is superior to low socio-economic group.

Under  $H_0$  the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\begin{aligned} S^2 &= \frac{1}{n_1 + n_2 - 2} [(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2] \\ &= \frac{1}{15 + 10 - 2} [(15 - 1) \times (0.27)^2 + (10 - 1) \times (0.22)^2] \\ &= \frac{1.0206 + 0.4356}{23} = \frac{1.4562}{23} = 0.063 \end{aligned}$$

$\Rightarrow$

$$S = 0.25$$

$\therefore$

$$t = \frac{2.91 - 2.26}{0.25 \sqrt{\frac{1}{15} + \frac{1}{10}}} = \frac{0.65 \times \sqrt{150}}{0.25 \times \sqrt{25}} = \frac{0.65 \times 2.45}{0.25} = 6.37$$

Tabulated value of  $t$  for 23 degrees of freedom at 5% level of sign. Finance for right tailed test is 1.71. Since calculated  $t$  is much greater than tabulated  $t$ , it is highly significance and  $H_0$  is rejected and conclude that mean of high group is greater than low group.

**Example 11.6.** Memory capacity of 8 students was tested before and after training. State at 5% level of significance whether the training was effective from the following scores:

Student	1	2	3	4	5	6	7	8	Total
Before	49	53	51	52	47	50	52	53	407
After	52	55	52	53	50	54	54	53	423

Use paired  $t$ -test for your answer.

**Solution.** Let  $x$  denotes the scores before training and  $y$  denotes the scores after training.

Null hypothesis  $H_0 : \mu_1 = \mu_2$ , i.e. there is no significant difference in the scores before and after the training. In other words, the given increments are just by chance (fluctuations of sampling).

Alternative hypothesis  $H_1 : \mu_1 < \mu_2$  (to conclude that training has been effected)  
(One tail)

## NOTES

Student	Score before training (x)	Score after training (y)	$d = x - y$	$d^2$
1	49	52	-3	9
2	53	55	-2	4
3	51	52	-1	1
4	52	53	-1	1
5	47	50	-3	9
6	50	54	-4	16
7	52	54	-2	4
8	53	53	0	0
			$\Sigma d = -16$	$\Sigma d^2 = 44$

Under  $H_0$  the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}}$$

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i = \frac{-16}{8} = -2$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 = \frac{1}{n-1} [\Sigma d_i^2 - n(\bar{d})^2]$$

$$= \frac{1}{7} [44 - 8 \times (-2)^2] = \frac{44 - 32}{7} = \frac{12}{7} = 1.714$$

$\Rightarrow$

$$S = 1.31$$

$$\therefore |t| = \frac{|d|}{S/\sqrt{n}} = \frac{|-2|}{1.31/\sqrt{8}} = \frac{2 \times 2.83}{1.31} = 4.32$$

Tabulated  $t_{0.05}$  for  $(8 - 1) = 7$  degrees of freedom for one tail test is 1.90.

Since calculated value of  $t$  is greater than the tabulated  $t$ ,  $H_0$  is rejected at 5% level of significance. Hence, we conclude that the scores differ significantly before and after the training, i.e. training was effected.

## EXERCISE 11.1

- A brand of matches is sold in boxes on which it is claimed that the average contents are 40 matches. A check on a pack of 5 boxes gives the following results:  
41, 39, 37, 40, 38  
(i) Test the manufacturer's claim keeping the interests of both the manufacturer and the customer in mind.  
(ii) As a customer test the manufacturer's claim.
- A sample of size 10 drawn from a normal population has a mean 31 and a variance 2.25. Is it reasonable to assume that the mean of the population is 30? (Use 1% level of significance).
- A random sample of size 10 from a normal population with mean  $\mu$  gives a sample mean of 40 and sample standard deviation of 6. Test the hypothesis that  $\mu = 44$  against  $\mu \neq 44$  at 5% level of significance.

## NOTES

4. A new drug manufacturer wants to market a new drug only if he could be quite sure that the mean temperature of a healthy person taking the drug could not rise above 98.6°F otherwise he will withhold the drug. The drug is administered to a random sample of 17 healthy persons. The mean temperature was found to be 98.4°F with a standard deviation of 0.6°F. Assuming that the distribution of the temperature is normal and  $\alpha = 0.01$ , what should the manufacturer do?
5. The marks of students in two groups were obtained as

I	18	20	36	50	49	36	34	49	41
II	29	28	26	35	30	44	46		

Test whether the groups were identical.  
(Given  $t_{0.05} = 2.14$  for 14 degrees of freedom)

6. Two different types of drugs A and B were tried on certain patients for increasing weight. 5 persons were given drug A and 7 persons were given drug B. The increase in weight in pounds is given below:

Drug A	8	12	13	9	3		
Drug B	10	8	12	15	6	8	11

Do the two drugs differ significantly with regard to their effect in increasing weight.  
(Given  $t_{0.05} = 2.23$  for 10 degrees of freedom)

7. The mean life of a sample of 10 electric light bulbs was found to be 1456 hours with standard deviation of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with standard deviation of 398 hours. Is there a significant difference between the means of the two batches?  
(Given  $t_{0.05} = 2.06$  for 25 degrees of freedom)
8. To verify whether a course in Statistics improved performance, a similar test was given to 12 participants both before and after the course. The original marks recorded in alphabetical order of the participants were 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. After the course, the marks were in the same order 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 78. Was the course useful?  
(Given  $t_{0.05} = 2.201$  for 11 degrees of freedom)
9. A certain medicine given to each of the 9 patients resulted in the following increase of blood pressure. Can it be concluded that the medicine will in general be accompanied by an increase in blood pressure.  
7, 3, -1, 4, -3, 5, 6, -4, -1  
(Given  $t_{0.05} = 2.306$  for 8 degrees of freedom)

### Answers

- |  |   |
|--|---|
| 1. (i) Accept manufacturer's claim         | (ii) manufacturer's claim is justified. |
| 2. Yes                                     | 3. Accept null hypothesis               |
| 4. The manufacturer should market the drug | 5. Two groups are identical             |
| 6. No                                      | 7. No                                   |
| 8. Yes                                     | 9. No                                   |

## 11.12. F-TEST

This test uses the variance ratio to test the significance of difference between two sampled variances. F-test which is based on F-distribution is called so in honour of a great statistician Prof. R.A. Fisher.

Let  $x_1, x_2, \dots, x_{n_1}$  and  $y_1, y_2, \dots, y_{n_2}$  be the values of two independent random samples drawn from the same normal population with variance  $\sigma^2$ . Then, we define variance ratio  $F$  as follows:

$$F = \frac{S_1^2}{S_2^2}; S_1 > S_2,$$

where

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2$$

and  $\bar{x}, \bar{y}$  are the sample means.

The distribution of variance ratio  $F$  with  $v_1$  and  $v_2$  degrees of freedom is given by

$$y = \frac{y_0 F^{\left(\frac{v_1-2}{2}\right)}}{\left(1 + \frac{v_1}{v_2} F\right)^{\left(\frac{v_1+v_2}{2}\right)}},$$

where  $y_0$  is so chosen that the total area under the curve is unity.

The parameters  $v_1$  and  $v_2$  represent degrees of freedom. For samples of sizes  $n_1$  and  $n_2$ , we have

$$v_1 = n_1 - 1 \quad \text{and} \quad v_2 = n_2 - 1.$$

### 11.13. PROPERTIES OF F-DISTRIBUTION

(i) The value of  $F$  cannot be negative as both terms of  $F$ -ratio are the squared values.

(ii) The range of the values of  $F$  is from 0 to  $\infty$ .

(iii) The  $F$ -distribution is independent of the population variance  $\sigma^2$  and depends on  $v_1$  and  $v_2$  only.

The  $F$ -distribution for various degrees of freedom  $v_1$  and  $v_2$  is given in the following table:

Table: Values of  $F$  for 5% and 1% level, where  $v_1$  is the number of degree of freedom for greater estimate of variance and  $v_2$  for the smaller estimate of variance.

### 11.14. PROCEDURE TO F-TEST

(i) Set up the null hypothesis  $H_0 = \sigma_1^2 = \sigma_2^2 = \sigma^2$ , i.e. the independent estimates of the common population variance do not differ significantly.

(ii) Find the degrees of freedom  $v_1$  and  $v_2$  given by  $v_1 = n_1 - 1$  and  $v_2 = n_2 - 1$  respectively.

(iii) Calculate the variances of two samples and then calculate  $F$ .

(iv) From  $F$ -distribution table note the value of  $F$  for  $v_1, v_2$  degrees of freedom at the desired level of significance.

(v) Compare the calculated value of  $F$  with tabulated value of  $F$  at the desired level of significance. If the calculated value of  $F$  is less than the tabulated value, then the difference is not significant and we may conclude that the same could have come from two populations with the same variance i.e., accept  $H_0$ , otherwise reject  $H_0$ .

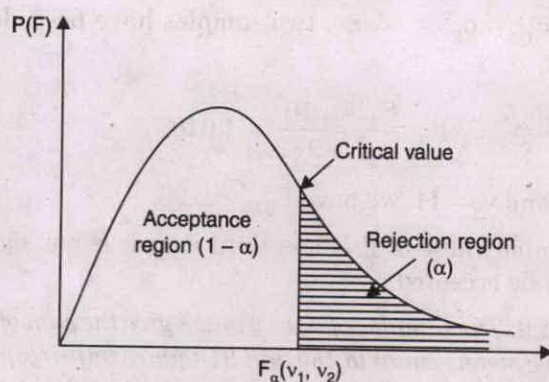


## 11.15. CRITICAL VALUES OF F-DISTRIBUTION

### NOTES

The available F-table give the critical values of F for the right-tailed test, i.e. the critical region is determined by the right-tail areas. Thus, the significance value  $F_{\alpha}(v_1, v_2)$  at level of significance and  $(v_1, v_2)$  degrees of freedom is determined by

$$P[F > F_{\alpha}(v_1, v_2)] = \alpha, \text{ as shown below:}$$



**Example 11.7.** In one sample of size 8 the sum of the squares of deviations of the sample values from the sample mean is 84.4 and in the other sample of size 10 it is 102.6. Test whether this difference is significance at 5% level. Given that for  $v_1 = 7$  and  $v_2 = 9$ ;  $F_{0.05} = 3.29$ .

**Solution.** Here,  $n_1 = 8$ ,  $n_2 = 10$

and

$$\Sigma(x - \bar{x})^2 = 84.4, \Sigma(y - \bar{y})^2 = 102.6$$

$$S_1^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{1}{7} \times 84.4 = 12.057$$

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{1}{9} \times 102.6 = 11.4$$

Under  $H_0$ :  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , i.e. the estimates of  $\sigma^2$  given by the samples are homogeneous,

$$F = \frac{S_1^2}{S_2^2} = \frac{12.057}{11.4} = 1.057$$

For  $v_1 = 7$  and  $v_2 = 9$ , we have  $F_{0.05} = 3.29$ . Since calculated value of F is less than  $F_{0.05}$ ,  $H_0$  may be accepted at 5% level of significance.

**Example 11.8.** Two random samples gave the following information:

Sample	Size	Sample mean	Sum of squares of deviations from the mean
1	10	15	90
2	12	14	108

Test whether the samples have been drawn from the same normal population. Given that for  $v_1 = 9$  and  $v_2 = 11$ ;  $F_{0.05} = 2.90$  (approx.).

**Solution.** Here,  $n_1 = 10$ ,  $n_2 = 12$ ,  $\bar{x} = 15$ ,  $\bar{y} = 14$

$$\Sigma(x - \bar{x})^2 = 90; \Sigma(y - \bar{y})^2 = 108$$

$$S_1^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{1}{9} \times 90 = 10$$

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{1}{11} \times 108 = 9.82$$

Under  $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ , i.e. two samples have been drawn from the same normal population.

$$F = \frac{S_1^2}{S_2^2} = \frac{10}{9.82} = 1.018$$

For  $v_1 = 9$  and  $v_2 = 11$ , we have  $F_{0.05} = 2.90$ .

Since calculated value of  $F$  is less than  $F_{0.05}$  it is not significant. Hence, null hypothesis  $H_0$  may be accepted.

**Example 11.9.** The samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 and 91 square units respectively. Test whether the samples have been drawn from the same normal population. Given that for  $v_1 = 8$  and  $v_2 = 7$ ;  $F_{0.05} = 3.73$ .

**Solution.** Here,  $n_1 = 9$ ,  $n_2 = 8$ ,  $\Sigma(x - \bar{x})^2 = 160$ ,  $\Sigma(y - \bar{y})^2 = 91$

$$S_1^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{1}{8} \times 160 = 20$$

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{1}{7} \times 91 = 13$$

Under  $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ , i.e. two samples have been drawn from the same normal population.

$$F = \frac{S_1^2}{S_2^2} = \frac{20}{13} = 1.54 \text{ (approx.)}$$

For  $v_1 = 8$  and  $v_2 = 7$ , we have  $F_{0.05} = 3.73$

Since calculated value of  $F$  is less than  $F_{0.05}$  it is not significant. Hence,  $H_0$  may be accepted.

**Example 11.10.** Two samples are drawn from two normal populations. From the following data test whether the two samples have the same variances at 5% level of significance.

Sample I	60	65	71	74	76	82	85	87		
Sample II	61	66	67	85	78	88	86	85	63	91

**Solution.** Here,  $n_1 = 8$ ,  $n_2 = 10$

Under  $H_0 : S_1^2 = S_2^2$ , i.e. two samples have the same variance.

$$H_1 : S_1^2 \neq S_2^2$$



Sample-I

Sample-II

## NOTES

$x$	$x - \bar{x}$	$(x - \bar{x})^2$	$y$	$y - \bar{y}$	$(y - \bar{y})^2$
60	60-75 = -15	225	61	61-77 = -16	256
65	65-75 = -10	100	66	66-77 = -11	121
71	71-75 = -4	16	67	67-77 = -10	100
74	74-75 = -1	1	85	85-77 = 8	64
76	76-75 = 1	1	78	78-77 = 1	1
82	82-75 = 7	49	88	88-77 = 11	121
85	85-75 = 10	100	86	86-77 = 9	81
87	87-75 = 12	144	85	85-77 = 8	64
			63	63-77 = -14	196
			91	91-77 = 14	196
$\Sigma x = 600$		$\Sigma(x - \bar{x})^2 = 636$	$\Sigma y = 770$		$\Sigma(y - \bar{y})^2 = 1200$

$$\bar{x} = \frac{\Sigma x}{n_1} = \frac{600}{8} = 75 \quad \bar{y} = \frac{\Sigma y}{n_2} = \frac{770}{10} = 77$$

$$\text{Variance of sample-I} = S_1^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{636}{8 - 1} = 90.857$$

$$\text{Variance of sample-II} = S_2^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{1200}{10 - 1} = 133.33$$

$$F = \frac{S_2^2}{S_1^2} = \frac{133.33}{90.857} = 1.467$$

For  $v_1 = 7$  and  $v_2 = 9$ , we have  $F_{0.05} = 3.29$ .

Since calculated value of  $F$  is less than  $F_{0.05}$ ,  $H_0$  may be accepted, i.e. the samples I and II have the same variance.

## EXERCISE 11.2

1. In a sample of 8 observations, the sum of squared deviations of items from the mean was 94.5. In another sample of 10 observations, the value was found to be 101.7. Test whether the difference is significant at 5% level.
2. The following are the values in thousands of an inch obtained by two engineers in 10 successive measurements with the same micrometer. Is one engineer significantly more consistent than the other?

Engineer A	503	505	497	505	495	502	499	493	510	501
Engineer B	502	497	492	498	499	495	497	496	498	

3. The nicotine content (in milligrams) of two samples of tobacco were found to be as follows:

Sample A	24	27	26	21	25	
Sample B	27	30	28	31	22	36

Can it be said that the two samples come from the same normal population?

## NOTES

4. The daily wages in ₹ of skilled workers in two cities are as follows:

City	Size of sample of workers	S.D. of wages in the sample
A	16	25
B	13	32

Test at 5% level of significance the equality of variances of the wage distribution in the two cities.

5. The time taken by workers in performing a job by methods I and II is given below:

Method I	20	16	26	27	23	22	—
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution from population from which these samples are drawn do not differ significantly?

6. Two random samples drawn from two normal populations are given below:

Sample I	63	65	68	69	71	72	—	—	—	—
Sample II	63	62	65	66	69	69	70	71	72	73

Test whether the two populations have the same variance at 5% level of significance.

## Answers

- |             |                    |         |
|-------------|--------------------|---------|
| 1. No       | 2. Not significant | 3. yes  |
| 4. Accepted | 5. Not significant | 6. Yes. |

## II. TEST OF SIGNIFICANCE FOR LARGE SAMPLES

For practical purposes a sample is taken as a large sample if  $n > 30$ . Under large sample test there are some important tests to test the significance. These tests are as follows:

- Test of significance for proportion
  - Single proportion
  - Difference of proportions
- Test of significance for single mean.
- Test of significance for differences of
  - Means
  - Standard deviations.

### 11.16. TEST OF SIGNIFICANCE FOR PROPORTION

(i) **Single proportion:** This test is used to test the significant difference between proportion of the sample and the population.

Let  $X$  be the number of successes in  $n$  independent trials with constant probability  $P$  of success for each trial.

We have  $E(X) = nP$  and  $V(X) = nPQ$ , where  $Q = 1 - P =$  probability of failure

Now, 
$$p = \frac{X}{n} \quad (p = \text{observed proportion of success})$$



## NOTES

Now, 
$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{nP}{n} = P$$

$$V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{nPQ}{n^2} = \frac{PQ}{n}$$

$$S.E. (p) = \sqrt{\frac{PQ}{n}}$$

$$Z = \frac{p - E(p)}{S.E. (p)} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0, 1)$$

where  $E \rightarrow$  expected value,  $V \rightarrow$  Variance and  $S.E. \rightarrow$  Standard error

$Z$  is called a test statistic which is used to test the significant difference of the sample and population proportion.

**Note 1.** The probable limits for the observed proportion of success are  $E(p) \pm Z_{\alpha} \sqrt{V(p)}$

i.e.,  $P \pm Z_{\alpha} \sqrt{\frac{PQ}{n}}$ , where  $Z_{\alpha}$  is the significant value at the level of significance  $\alpha$ .

2. If  $P$  is not known then the probable limits for the proportion in the population are

$$p \pm Z_{\alpha} \sqrt{\frac{pq}{n}}$$

3. If  $\alpha$  is not given, then we can use  $3\sigma$  limits. Hence, probable limits for the observed proportion of success are  $P \pm 3\sqrt{\frac{PQ}{n}}$  and probable limits for the proportion in the population are

$$p \pm 3\sqrt{\frac{pq}{n}}$$

4. A set of four selected values is commonly used for  $\alpha$ . Each  $\alpha$  and corresponding  $Z_{\alpha}$  and  $Z_{\alpha/2}$  values are given in the following table:

For two-tailed test		For one-tailed test	
$\alpha$	$Z_{\alpha/2}$	$\alpha$	$Z_{\alpha}$
0.20	1.282	0.10	1.282
0.10	1.645	0.05	1.645
0.05	1.960	0.025	1.960
0.01	2.576	0.01	2.326

**(ii) Difference of Proportions:** This test is used to test the difference between the sample proportions.

Let two samples  $X_1$  and  $X_2$  of sizes  $n_1$  and  $n_2$  respectively taken from two different populations, then  $p_1 = \frac{X_1}{n_1}$  and  $p_2 = \frac{X_2}{n_2}$ .

To test the significance of the difference between the sample proportions  $p_1$  and  $p_2$  we set the null hypothesis  $H_0$ , that there is no significant difference between the two sample proportion.



Under the null hypothesis  $H_0$ , the test statistic is

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \text{ and } Q = 1 - P$$

## NOTES

If sample proportions are not given, we set the null hypothesis

$$H_0 : p_1 = p_2$$

under  $H_0$  the test statistic is

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}, \text{ where } Q_1 = 1 - P_1 \text{ and } Q_2 = 1 - P_2.$$

**Example 9.11.** A coin is tossed 324 times and the head turned up 175 times. Test the hypothesis that the coin is unbiased.

**Solution.** Null hypothesis  $H_0$  : the coin is unbiased i.e.,

$$P = \frac{1}{2}$$

Here,  $n = 324$ ,  $X = \text{Number of heads} = 175$

$$P = \text{prob. of getting a head in a toss} = \frac{1}{2}$$

$$Q = 1 - P = 1 - \frac{1}{2} = \frac{1}{2}$$

$$\begin{aligned} \therefore Z &= \frac{X - E(X)}{\text{SE of } X} = \frac{X - nP}{\sqrt{nPQ}} = \frac{175 - 324 \times \frac{1}{2}}{\sqrt{324 \times \frac{1}{2} \times \frac{1}{2}}} \\ &= \frac{13}{9} = 1.44 < 1.96 \end{aligned}$$

Since  $|Z| < 1.96$ , null hypothesis is accepted at 5% level of significance. Hence the coin is unbiased.

**Example 9.12.** A die is thrown 1000 times and a throw of 5 or 6 was obtained 420 times. On the assumption of random throwing do the data indicate an unbiased die?

**Solution.** Null hypothesis  $H_0$  : the die is unbiased

Under  $H_0$ ,  $P = \text{probability of getting 5 or 6}$

$$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

$$Q = 1 - P = 1 - \frac{1}{3} = \frac{2}{3}$$

Here,  $n = 1000$ ,  $X = \text{Number of success} = 420$

$$Z = \frac{X - nP}{\sqrt{nPQ}} = \frac{420 - 1000 \times \frac{1}{3}}{\sqrt{1000 \times \frac{1}{3} \times \frac{2}{3}}} = \frac{420 - 333.33}{\sqrt{222.222}} = \frac{86.67}{14.91} = 5.813$$

Since  $|Z| = 5.813 > 3$  (Maximum value of  $Z$ ),  $H_0$  is rejected i.e., the die is biased.

**Example 11.13.** 500 apples are taken at random from a large basket and 65 are found to be bad. Find the S.E. of the proportion of bad ones in a sample of this size and assign limits within which the percentage of bad apples most probably lies.

**Solution.** Here,  $n = 500$ ,  $X =$  number of bad apples in the sample  $= 65$

$$p = \text{proportion of bad apples in the sample} = \frac{65}{500} = 0.13 \text{ and}$$

$$q = 1 - p = 1 - 0.13 = 0.87$$

$\therefore$  The proportion of bad apples  $P$  in the population is not known.

$\therefore$  We can take  $P = p = 0.13$ ,  $Q = q = 0.87$  and  $N = n = 500$

$$\text{S.E. of proportion} = \sqrt{\frac{PQ}{N}} = \sqrt{\frac{0.13 \times 0.87}{500}} = 0.015$$

Limits for proportions of bad apples in the population is

$$P \pm 3\sqrt{\frac{PQ}{N}} = 0.13 \pm 3\sqrt{\frac{0.13 \times 0.87}{500}} = 0.13 \pm 0.045 = 0.175 \text{ and } 0.085$$

$$= 17.5\% \text{ and } 8.5\%.$$

**Example 11.14.** Before an increase in excise duty on tea, 400 people out of a sample of 500 persons were found to be tea drinkers. After an increase in the excise duty, 400 persons were known to be tea drinkers in a sample of 600 people. Do you think that there has been a significant decrease in the consumption of tea after the increase in the excise duty?

**Solution.** Here  $n_1 = 500$ ,  $n_2 = 600$   
 $X_1 = 400$ ,  $X_2 = 400$

$$p_1 = \text{proportion of drinkers in first sample} = \frac{400}{500} = \frac{4}{5} = 0.8$$

$$p_2 = \text{proportion of drinkers in second sample} = \frac{400}{600} = \frac{2}{3} = 0.67$$

Since proportion  $P$  of the population is not given, it can be estimated by using

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{400 + 400}{500 + 600} = \frac{800}{1100} = \frac{8}{11}$$

and  $Q = 1 - P = 1 - \frac{8}{11} = \frac{3}{11}$

Null hypothesis  $H_0 : P_1 = P_2$  (there is no significant difference in the consumption of tea before and after increase of excise duty)

Alternative hypothesis  $H_1 : P_1 > P_2$  (right tailed test), under  $H_0$  the test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.8 - 0.67}{\sqrt{\frac{8}{11} \times \frac{3}{11} \left(\frac{1}{500} + \frac{1}{600}\right)}} = \frac{0.13}{0.027} = 4.815$$

Since  $|Z| = 4.815 > 1.645$  also  $|Z| = 4.815 > 2.33$  at both the significant values of  $Z$  at 5% and 1% level of significant respectively,  $H_0$  is rejected i.e., there is a significant decrease in the consumption of tea due to increase in excise duty.

## NOTES

**Example 11.15.** 500 articles from a factory are examined and found to be 2% defective. 800 similar articles from a second factory are found to have only 1.5% defectives. Can it reasonably be concluded that the products of the first factory are inferior to those of second?

**NOTES**

**Solution.** Here,  $n_1 = 500$ ,

$$p_1 = \text{proportion of defectives from first factory} = \frac{2}{100} = 0.02$$

$$n_2 = 800,$$

$$p_2 = \text{proportion of defectives from second factory} = \frac{1.5}{100} = 0.015$$

Since proportion  $P$  of the population is not given it can be estimated by using

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{10 + 12}{500 + 800} = \frac{22}{1300} = 0.017$$

and

$$Q = 1 - P = 1 - 0.017 = 0.983$$

Null hypothesis  $H_0 : P_1 = P_2$  (there is no significant difference between the products of first and second factory)

Alternative hypothesis  $H_1 : P_1 \neq P_2$  (two tailed test)

Under  $H_0$  the test statistic

$$\begin{aligned} Z &= \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.02 - 0.015}{\sqrt{0.017 \times 0.983\left(\frac{1}{500} + \frac{1}{800}\right)}} \\ &= \frac{0.005}{0.00737} = 0.678 \end{aligned}$$

Since  $|Z| = 0.678 < 1.96$ , null hypothesis is accepted at 5% level of significance. Hence there is no significant difference between the products of first and second factory i.e., the products of the first factory are not inferior to those of second.

**Example 11.16.** In two large populations there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1400 and 1000 respectively from the two populations.

**Solution.** Here,  $n_1 = 1400$ ,  $n_2 = 1000$

$$P_1 = \text{proportion of fair haired in the first population} = \frac{30}{100} = 0.3$$

$$P_2 = \text{proportion of fair haired in the second population} = \frac{25}{100} = 0.25$$

$$Q_1 = 1 - P_1 = 1 - 0.3 = 0.7, Q_2 = 1 - P_2 = 1 - 0.25 = 0.75$$

Null hypothesis  $H_0 : p_1 = p_2$  (Sample proportions are equal) i.e., the difference in population proportions is likely to be hidden in sampling.

Alternative hypothesis  $H_1 : p_1 \neq p_2$  (two tailed test)

Under  $H_0$  the test statistic is

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} = \frac{0.30 - 0.25}{\sqrt{\frac{0.3 \times 0.7}{1400} + \frac{0.25 \times 0.75}{1000}}} = \frac{0.05}{0.01837} = 2.72$$

Since  $|Z| = 2.72 > 1.96$ , null hypothesis is rejected at 5% level of significance. Hence at 5% level of significance these samples will exhibit the difference in the population proportions.

### EXERCISE 11.3

### NOTES

1. A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is unbiased.
2. In a hospital 525 female and 475 male babies were born in a month. Do these figures confirm the hypothesis that females and males are born in equal number?
3. A die is thrown 10000 times and a throw of 3 or 4 was obtained 4200 times. On the assumption of random throwing do the data indicate an unbiased die?
4. Given that on the average 4% of insured men of age 65 die within a year and that 60 of a particular group of 1000 such men (age 65) died within a year. Can this group be regarded as a representative sample?
5. 325 men out of 600 men chosen from a big city were found to be smokers. Does this information support the conclusion that the majority of men in the city are smokers?
6. A random sample of 400 apples is taken from a large basket and 40 are found to be bad. Estimate the proportion of bad apples in the basket and assign limits within which the percentage most probably lies.
7. A manufacturer claimed that at least 95% of the equipments which he supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipments revealed that 18 were faulty. Test the manufacturer's claim at a level of significance (i) 5% (ii) 1%.
8. 1000 articles from a factory are examined and found to be 2.5% defective. 1500 similar articles from a second factory are found to have only 2% defectives. Can it reasonably be concluded that the products of the first factory are inferior to those of second?
9. A manufacturing firm claims that its brand A product outsells its brand B product by 8%. If it is found that 42 out of a sample of 200 persons prefer brand A and 18 out of another sample of 100 persons prefer brand B. Test whether the 8% difference is valid claim.
10. In a survey on a particular matter in a college, 850 males and 560 females voted. 500 males and 320 females voted yes. Does this indicate a significant difference of opinion between male and female on this matter at 1% level of significance?
11. Two samples of sizes 1200 and 900 respectively drawn from two large populations. In the two large populations there are 30% and 25% respectively of fair haired people. Test whether these two samples will reveal the difference in the population proportions.
12. Before an increase in excise duty on tea 800 persons out of a sample of 1000 persons were found to be tea drinkers. After an increase in excise duty 800 people were tea drinkers in a sample of 1200 people. Test whether there is a significant decrease in the consumption of tea after the increase in excise duty.

### Answers

1.  $H_0$  is accepted at 5% level of significance.
2. Yes,  $H_0$  is accepted at 5% level of significance.
3.  $H_0$  is rejected.
4.  $H_0$  is rejected.
5.  $H_0$  is rejected at 5% level of significance.
6. 8.5 : 11.5
7. Using left tailed test,  $H_0$  is rejected at both 5% and 1% level of significance.
8. No,  $H_0$  is accepted.
9.  $H_0$  is accepted.
10.  $H_0$  is accepted.
11.  $H_0$  is rejected at 5% level of significance.
12.  $H_0$  is rejected.

## 11.17. TEST OF SIGNIFICANCE FOR SINGLE MEAN

This test is used to test the significant difference between sample mean and population mean.

Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a normal population with mean  $\mu$  and variance  $\sigma^2$ .

The standard error (S.E.) of mean of a random sample of size  $n$  from a population is given by

## NOTES

S.E. ( $\bar{x}$ ) =  $\frac{\sigma}{\sqrt{n}}$ , where  $\sigma$  is the standard deviation of the population.

We set the null hypothesis  $H_0$  that the sample has been drawn from a large population with mean  $\mu$  and variance  $\sigma^2$  i.e., there is no significant difference between the sample mean ( $\bar{x}$ ) and population mean ( $\mu$ ).

Under the null hypothesis  $H_0$  the test statistic is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

If standard deviation of the population ( $\sigma$ ) is not known, we use the test statistic given as

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \text{ where } s \text{ is the standard deviation of the sample.}$$

**Note.** The limits of the population mean  $\mu$  are given by  $\bar{x} \pm Z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$  i.e.,

$$\bar{x} - Z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

These limits are called the confidence limits for  $\mu$ .

**Example 11.17.** A normal population has a mean of 6.8 and standard deviation of 1.5. A sample of 400 members gave a mean of 6.75. Is the difference significant?

**Solution.** Here,  $\mu = 6.8$ ,  $\bar{x} = 6.75$ ,  $\sigma = 1.5$ ;  $n = 400$

Null hypothesis  $H_0$ :  $\bar{x} = \mu$  (there is no significant difference between  $\bar{x}$  and  $\mu$ )

Alternative hypothesis  $H_1$ : there is a significant difference between  $\bar{x}$  and  $\mu$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{6.75 - 6.8}{1.5/\sqrt{400}} = -\frac{0.05}{0.075} = -0.67$$

Since  $|Z| = 0.67 < 1.96$   $H_0$  is accepted at 5% level of significance. Hence there is no significant difference between  $\bar{x}$  and  $\mu$ .

**Example 11.18.** A random sample of 400 members has a mean 99. Can it be reasonably regarded as a sample from a large population of mean 100 and standard deviation 8 at 5% level of significance?

**Solution.** Here,  $\mu = 100$ ,  $\bar{x} = 99$ ,  $\sigma = 8$ ,  $n = 400$

Null hypothesis  $H_0$ : the sample is drawn from a large population with mean 100 and standard deviation 8.

Alternative hypothesis  $H_1$ :  $\mu \neq 100$  (two tailed test)

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{99 - 100}{8/\sqrt{400}} = -\frac{1}{0.4} = -2.5$$

Since  $|Z| = 2.5 > 1.96$ ,  $H_0$  is rejected at 5% level of significance. Hence there is a significant difference between  $\bar{x}$  and  $\mu$  i.e., it can not be regarded as a sample from a large population.



## NOTES

**Example 11.19.** The management of a company claims that the average weekly income of their employees is ₹ 900. The trade union disputes this claim stressing that it is rather less. An independent sample of 150 randomly selected employees estimated the average to be ₹ 856 with standard deviation of ₹ 354. Would you accept the view of the management?

**Solution.** Here,  $\mu = 900$ ,  $\bar{x} = 854$ ,  $s = 354$ ,  $n = 150$

Null hypothesis  $H_0$ : there is no significant difference between  $\bar{x}$  and  $\mu$  i.e., the view of management is correct.

Alternative hypothesis  $H_1$ :  $\mu \neq 900$  (two-tailed test)

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{854 - 900}{354/\sqrt{150}} = -\frac{46}{28.904} = -1.59$$

Since  $|Z| = 1.59 < 1.96$ ,  $H_0$  is accepted at 5% level of significance. Hence the view of management is correct.

**Example 11.20.** In a population with a standard deviation of 14.8, what sample size is needed to estimate the mean of population within  $\pm 1.2$  with 95% confidence?

**Solution.** Here,  $\bar{x} - \mu = \pm 1.2$ ,  $\sigma = 14.8$ ,  $Z = 1.96$

We know that  $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Using this, we have

$$1.96 = \frac{\pm 1.2}{14.8/\sqrt{n}} = \frac{\pm 1.2\sqrt{n}}{14.8}$$

On squaring both the sides we have

$$(1.96)^2 = \left(\frac{\pm 1.2}{14.8}\right)^2 \times n \quad \text{or} \quad n = \left(\frac{1.96 \times 14.8}{\pm 1.2}\right)^2 = 584.35 \approx 584.$$

**Example 11.21.** A random sample of 900 measurements from a large population gave a mean value of 64. If this sample has been drawn from a normal population with standard deviation of 20, find the 95% and 99% confidence limits for the mean in the population.

**Solution.** Here,  $n = 900$ ,  $\bar{x} = 64$ ,  $\sigma = 20$

At 95% confidence  $Z = 1.96$

At 99% confidence  $Z = 2.58$

The confidence limits for the population mean  $\mu$  is given by

$$\bar{x} \pm Z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

The confidence limits for 95% confidence are

$$64 \pm 1.96 \times \frac{20}{\sqrt{900}} = 64 \pm 1.307 = 62.693 \text{ and } 65.307$$

The confidence limits for 99% confidence are

$$64 \pm 2.58 \times \frac{20}{\sqrt{900}} = 64 \pm 1.72 = 62.28 \text{ and } 65.72.$$

## EXERCISE 11.4

## NOTES

1. A random sample of 900 members has a mean 3.4 cms. Can it be reasonably regarded as a sample from a large population of mean 3.2 cms and standard deviation 2.3 cms?
2. A random sample of 400 male students is found to have a mean height of 160 cms. Can it be reasonably regarded as a sample from a large population with mean height 162.5 cms and standard deviation 4.5 cms?
3. A random sample of 200 measurements from a large population gave a mean value of 50 and a standard deviation of 9. Determine 95% confidence interval for the mean of population.
4. A random sample of 400 measurements from a large population gave a mean value of 82 and a standard deviation of 18. Determine 95% confidence interval for the mean of population.
5. A company manufacturing electric bulbs claims that the average life of its bulbs is 1600 hours. The average life and standard deviation of random sample of 100 such bulbs were 1570 hours and 120 hours respectively. Should we accept the claim of the company?
6. An insurance agent has claimed that the average age of policy holders who insure through him is less than the average for all agents which is 30.5 years. A random sample of 100 policy holders who had insured through him reveal that the mean and standard deviation are 28.8 years and 6.35 years respectively. Test his claim at 5% level of significance.
7. The guaranteed average life of a certain type of bulbs is 1000 hours with a standard deviation of 125 hours. It is decided to sample the output so as to ensure that 90% of the bulbs do not fall short of the guaranteed average by more than 2.5%. What must be the minimum size of the sample?

## Answers

- |  |                            |
|--|----------------------------|
| 1. Yes, $H_0$ is accepted.                   | 2. Yes, $H_0$ is accepted. |
| 3. 48.8 and 51.2                             | 4. 80.24 and 83.76         |
| 5. No, rejected at 5% level of significance. | 6. Claim is valid.         |
| 7. $n = 4$                                   |                            |

## 11.18. TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS

(i) This test is used to test the significant difference between the means of two large samples.

Let  $\bar{x}_1$  be the mean of a sample of size  $n_1$  from a population with mean  $\mu_1$  and variance  $\sigma_1^2$  and let  $\bar{x}_2$  be the mean of an independent sample of size  $n_2$  from another population with mean  $\mu_2$  and variance  $\sigma_2^2$ .

We set the null hypothesis  $H_0$  that there is no significant difference between the sample means i.e.,  $\mu_1 = \mu_2$ .

Under the null hypothesis  $H_0$  the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$



If the samples are drawn from the same population with common standard deviation ( $\sigma$ ), then under the null hypothesis the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (\because \sigma_1 = \sigma_2 = \sigma)$$

## NOTES

**Note. 1.** If  $\sigma_1 \neq \sigma_2$  and  $\sigma_1$  and  $\sigma_2$  are not known, the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

**2.** If common standard deviation ( $\sigma$ ) is not known and  $\sigma_1 = \sigma_2$  then  $\sigma$  can be obtained by using

$$\sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}}$$

The test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

**(ii) Standard Deviations.** This test is used to test the significant difference between the standard deviations of two populations.

Let two independent random sample of sizes  $n_1$  and  $n_2$  having standard deviations  $s_1$  and  $s_2$  be drawn from the two normal population with standard deviation  $\sigma_1$  and  $\sigma_2$  respectively.

We set the null hypothesis  $H_0$  that the sample standard deviations do not differ significantly i.e.,  $\sigma_1 = \sigma_2$ .

Under the null hypothesis  $H_0$  the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$$

If  $\sigma_1$  and  $\sigma_2$  are unknown then the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}}$$

**Example 11.22.** Examine whether there is any significant difference between the two samples for the following data:

Sample	Size	Mean
1	50	140
2	60	150

Standard deviation of the population = 10.

**Solution.** Here,  $n_1 = 50$ ,  $n_2 = 60$ ,  $\bar{x}_1 = 140$ ,  $\bar{x}_2 = 150$ ,  $\sigma = 10$

Null hypothesis  $H_0 : \mu_1 = \mu_2$  i.e., samples are drawn from the same normal population.

Alternative hypothesis  $H_1 : \mu_1 \neq \mu_2$ Under  $H_0$  the test statistics is

## NOTES

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{140 - 150}{10 \sqrt{\frac{1}{50} + \frac{1}{60}}} = -\frac{10}{1.915} = -5.22$$

Since  $|Z| = 5.22 > 3$ ,  $H_0$  is rejected. Hence the samples are not drawn from the same normal population.

**Example 23.** Intelligence tests on two groups of boys and girls gave the following results.

	Mean	S.D.	Size
Girls	70	10	70
Boys	75	11	100

Examine if the difference between mean scores is significant.

**Solution.** Here,  $n_1 = 70$ ,  $n_2 = 100$ ,  $\bar{x}_1 = 70$ ,  $\bar{x}_2 = 75$ ,  $s_1 = 10$ ,  $s_2 = 11$

Null hypothesis  $H_0$  : There is no significant difference between mean scores i.e.,  $\bar{x}_1 = \bar{x}_2$ .

Alternative hypothesis  $H_1 : \bar{x}_1 \neq \bar{x}_2$  (two-tailed test)

Under  $H_0$  the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{70 - 75}{\sqrt{\frac{10^2}{70} + \frac{11^2}{100}}} = -\frac{5}{2.639} = -1.895$$

Since  $|Z| = 1.895 < 1.96$ ,  $H_0$  is accepted at 5% level of significance. Hence there is no significant difference between mean scores.

**Example 11.24.** The means of two large samples of 1000 and 2000 members are 168.75 cms and 170 cms respectively. Can the samples be regarded as drawn from the same population of standard deviation 6.25 cms?

**Solution.** Here,  $n_1 = 1000$ ,  $n_2 = 2000$ ,  $\bar{x}_1 = 168.75$ ,  $\bar{x}_2 = 170$ ,  $\sigma = 6.25$

Null hypothesis  $H_0 : \mu_1 = \mu_2$  i.e., samples are drawn from the same population.

Alternative hypothesis  $H_1 : \mu_1 \neq \mu_2$  (two-tailed test)

Under  $H_0$  the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{168.75 - 170}{6.25 \sqrt{\frac{1}{1000} + \frac{1}{2000}}} = -\frac{1.25}{0.242} = -5.165$$

Since  $|Z| = 5.165 > 1.96$ ,  $H_0$  is rejected at 5% level of significance. Hence the samples are not drawn from the same population.

**Example 11.25.** Two random samples of sizes 1000 and 2000 farms gave an average yield of 2000 kg and 2050 kg respectively. The variance of wheat farms in the country may be taken as 10 kg. Examine whether the two samples differ significantly in yield.



**Solution.** Here,  $n_1 = 1000$ ,  $n_2 = 2000$ ,  $\bar{x}_1 = 2000$ ,  $\bar{x}_2 = 2050$ ,  $\sigma^2 = 100$  i.e.,  $\sigma = 10$   
Null hypothesis  $H_0 : \mu_1 = \mu_2$  i.e., samples are drawn from the same population.  
Alternative hypothesis  $H_1 : \mu_1 \neq \mu_2$  (two tailed test)  
Under  $H_0$  the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{2000 - 2050}{10 \sqrt{\frac{1}{1000} + \frac{1}{2000}}} = -\frac{50}{0.387} = -129.20$$

Since  $|Z| = 129.20 > 3$  (maximum value of  $Z$ ), highly significant,  $H_0$  is rejected.  
Hence the samples are not drawn from the same normal population.

**Example 11.26.** Random samples drawn from two large cities gave the following information relating to the heights of adult males:

	Mean height (in inches)	Standard deviation	No. in samples
City 1	67.42	2.58	1000
City 2	67.25	2.50	1200

Test the significance of difference in standard deviations of the samples at 5% level of significance.

**Solution.** Here,  $n_1 = 1000$ ,  $n_2 = 1200$ ,  $\bar{x}_1 = 67.42$ ,  $\bar{x}_2 = 67.25$ ,  $s_1 = 2.58$ ,  $s_2 = 2.50$ ,  $\sigma$  is not known.

Null hypothesis  $H_0 : \sigma_1 = \sigma_2$  i.e., the sample standard deviations do not differ significantly.

Alternative hypothesis  $H_1 : \sigma_1 \neq \sigma_2$  (two-tailed test)

Under  $H_0$  the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{2.58 - 2.50}{\sqrt{\frac{(2.58)^2}{2000} + \frac{(2.50)^2}{2400}}} = \frac{0.08}{0.077} = 1.039$$

Since  $|Z| = 1.039 < 1.96$ ,  $H_0$  is accepted. Hence sample standard deviations do not differ significantly.

**Example 11.27.** In a survey of incomes of two classes of workers of two random samples gave the following data:

	Size of sample	Mean annual income in ₹	Standard deviation in ₹
Sample 1	100	582	24
Sample 2	100	546	28

Examine whether the difference between

- Mean and
- The standard deviations significant.

## NOTES



**Solution.** Here,  $n_1 = 100$ ,  $n_2 = 100$ ,  $\bar{x}_1 = 582$ ,  $\bar{x}_2 = 546$ ,  $s_1 = 24$ ,  $s_2 = 28$

(i) Null hypothesis  $H_0 : \mu_1 = \mu_2$  i.e., sample means do not differ significantly.

Alternative hypothesis  $H_1 : \mu_1 \neq \mu_2$  (two tailed test)

Under  $H_0$  the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{582 - 546}{\sqrt{\frac{(24)^2}{100} + \frac{(28)^2}{100}}} = \frac{36}{3.6878} = 9.762$$

Since  $|Z| = 9.762 > 1.96$ , highly significant,  $H_0$  is rejected at 5% level of significance. Hence sample means differ significantly.

(ii) Null hypothesis  $H_0 : \sigma_1 = \sigma_2$  i.e., sample standard deviations do not differ significantly.

Alternative hypothesis  $H_1 : \sigma_1 \neq \sigma_2$  (two-tailed test)

Under  $H_0$  the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{24 - 28}{\sqrt{\frac{(24)^2}{200} + \frac{(28)^2}{200}}} = \frac{-4}{2.6077} = -1.53$$

Since  $|Z| = 1.53 < 1.96$ ,  $H_0$  is accepted at 5% level of significance. Hence sample standard deviations do not differ significantly.

### EXERCISE 11.5

- The number of accidents per day were studied for 144 days in city A and for 100 days in city B. The mean numbers of accidents and standard deviations were respectively 4.5 and 1.2 for city A and 5.4 and 1.5 for city B. Is city A more prone to accidents than city B.
- The mean yields of a crop from two places in a district were 210 kgs and 220 kgs per acre from 100 acres and 150 acres respectively. Can it be regarded that the sample were drawn from the same district which has the standard deviation of 11 kgs per acre?
- Given the following data:

	No. of cases	Mean wages in ₹	Standard deviation of wages in ₹
Sample 1	400	47.4	3.1
Sample 2	900	50.3	3.3

Examine whether the two mean wages differ significantly.

- A sample of heights of 6400 soldiers has a mean of 67.85 inches and a standard deviation of 2.56 inches. While another sample of heights of 1600 sailors has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the sailors are on the average taller than soldiers?
- Intelligence tests on two groups of boys and girls gave the following results:

	Mean	S.D	Size
Girls	75	8	60
Boys	73	10	100

Examine if the difference between mean scores is significant.

### NOTES

## NOTES

6. The yield of a crop in a random sample of 1000 farms in a certain area has a standard deviation of 192 kgs. Another random sample of 1000 farms gives a standard deviation of 224 kgs. Are the standard deviations significantly different?
7. The standard deviation of a random sample of 900 members is 4.6 and that of another random sample of 1600 is 4.8. Examine if the standard deviations are significantly different.
8. The mean yield of two sets of plots and their variability are as follow:

	Set of 40 plots	Set of 60 plots
Mean yield per plot	1258 kgs	1243 kgs
S.D. per plot	34	28

Examine whether

- (i) the difference in the variability in yields is significant,
- (ii) the difference in the mean yields is significant.

### Answers

1. No
2. No
3. Yes, highly significant
4. Highly significant
5. Not significant at 5%
6. Yes
7. Not significant
8. (i) Not significant (ii) significant.

## 11.19. CHI-SQUARE TEST

In test of hypothesis of parameters, it is usually assumed that the random variable follows a particular distribution. To confirm whether our assumption is right, Chi-square test is used which measures the discrepancy between the observed (actual) frequencies and theoretical (expected) frequencies, on the basis of outcomes of a trial or observational data. Chi-square is a letter of the Greek alphabet and is denoted by  $\chi^2$ . It is a continuous distribution which assumes only positive values.

## 11.20. CHI-SQUARE TEST TO TEST THE GOODNESS OF FIT

The value of  $\chi^2$  is used to test whether the deviations of the observed (actual) frequencies from the theoretical (expected) frequencies are significant or not. Chi-square test is also used to test whether a set of observations fit a given distribution or not. Therefore, chi-square provides a test of goodness of fit.

If  $O_1, O_2, \dots, O_n$  is a set of observed (actual) frequencies and  $E_1, E_2, \dots, E_n$  is the corresponding set of theoretical (expected) frequencies, then the statistic  $\chi^2$  is given by

$$\chi^2 = \sum_{i=1}^n \left\{ \frac{(O_i - E_i)^2}{E_i} \right\}$$

is distributed with  $(n - 1)$  degrees of freedom.

Here, we test the null hypothesis.

$H_0$ : There is no significant difference between the observed (actual) values and the corresponding expected (theoretical) values.

v.s.,  $H_1 : H_0$  is not true.

If  $\chi^2_{\text{cal}} \geq \chi^2_{\text{tab}}$  (or  $\chi^2_{\alpha, n-1}$ ) then  $H_0$  is rejected otherwise  $H_0$  is accepted.

**Note.** If the null hypothesis  $H_0$  is true, the test statistic  $\chi^2$  follow chi-square distribution with  $(n-1)$  degrees of freedom, where

$$\sum_{i=1}^n O_i = \sum_{i=1}^n E_i ; \text{ i.e. } \sum_{i=1}^n (O_i - E_i) = 0.$$

## NOTES

### 11.21. CHI-SQUARE TEST TO TEST THE INDEPENDENCE OF ATTRIBUTES

The value of  $\chi^2$  is used to test whether two attributes are associated or not, i.e. independence of attributes. To test the independence of attributes contingency table is used.

A contingency table is a two-way table in which rows are classified according to one attribute or criterion and columns are classified according to the other attribute or criterion. Each cell contains that number of items  $O_{ij}$  possessing the qualities of the  $i$ th row and  $j$ th column, where  $i = 1, 2, \dots, r$  and  $j = 1, 2, \dots, s$ . In such a case contingency table is said to be of order  $(r \times s)$ . Each row or column total is known as

marginal total. Also we have the sum of row totals  $\sum_{i=1}^r R_i$  is equal to the sum of column

totals  $\sum_{j=1}^s C_j$ , i.e.

$$\sum_i R_i = \sum_j C_j = N, \text{ where } N \text{ is the total frequency.}$$

Let us consider the two attributes A and B, where A divided into  $r$  classes  $A_1, A_2, \dots, A_r$  and B divided into  $s$  classes  $B_1, B_2, \dots, B_s$ . If  $R_i$  represents the number of persons possessing the attributes  $A_i$ ;  $C_j$  represents the number of persons possessing the attributes  $B_j$  and  $O_{ij}$  represent the number of persons possessing attributes  $A_i$  and  $B_j$  respectively. The contingency table of order  $(r \times s)$  is shown in the following table:

Columns Rows					Total
	$B_1$	$B_2$	.....	$B_s$	
$A_1$	$O_{11}$	$O_{12}$	.....	$O_{1s}$	$R_1$
$A_2$	$O_{21}$	$O_{22}$	.....	$O_{2s}$	$R_2$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$A_r$	$O_{r1}$	$O_{r2}$	.....	$O_{rs}$	$R_r$
Total	$C_1$	$C_2$	.....	$C_s$	N



Corresponding to each  $O_{ij}$  the expected frequency  $E_{ij}$  in a contingency table is calculated by

$$E_{ij} = \frac{R_i \times C_j}{N} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Here, we test the null hypothesis.

$H_0$  : There is no association between the attributes under study, i.e. attributes A and B are independent.

v.s.,  $H_1$  : attributes are associated, i.e., attributes A and B are not independent.

$H_0$  can be tested by the statistic

$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$  is distributed with  $(r - 1)(s - 1)$  degrees of freedom.

If  $\chi_{\text{cal}}^2 \geq \chi_{\text{tab}}^2$  (or  $\chi_{\alpha, (r-1)(s-1)}^2$ ), then  $H_0$  is rejected otherwise  $H_0$  is accepted.

**Note 1.** For a contingency table with  $r$  rows and  $s$  columns, the degrees of freedom =  $(r - 1)(s - 1)$ .

**2.** For a  $2 \times 2$  contingency table  $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$  we use the following formula to calculate the value of statistic  $\chi^2$  as

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(b + d)(a + c)(c + d)},$$

where  $N = a + b + c + d$

$\chi^2$  has  $(2 - 1)(2 - 1) = 1$  degree of freedom.

**3. Yate's correction.** In a  $2 \times 2$  contingency table, if any of cell frequency is less than 5, we make a correction to make  $\chi^2$  continuous. Decrease by  $\frac{1}{2}$  those cell frequencies which are greater than expected frequencies and increase by  $\frac{1}{2}$  those cell frequencies which are less than expected frequencies. This will affect the marginal totals. This correction is known as Yate's correction.

After applying the Yate's correction, the corrected value of  $\chi^2$  is given by

$$\chi^2 = \frac{N \left( \left| ad - bc \right| - \frac{N}{2} \right)^2}{(a + b)(b + d)(a + c)(c + d)}.$$

## 11.22. CONDITIONS FOR $\chi^2$ TEST

1. The number of observations collected must be large, i.e.  $n \geq 30$ .
2. No theoretical frequency should be very small.
3. The sample observations should be independent.
4.  $N$ , the total of frequencies should be reasonably large, say, greater than 50.

## NOTES

## 11.23. USES OF $\chi^2$ TEST

### NOTES

1. To test the goodness of fit.
2. To test the discrepancies between observed and expected frequencies.
3. To determine the association between attributes.

**Example 11.28.** The following table gives the number of accidents that took place in an industry during various days of the week. Test whether the accidents are uniformly distributed over the week.

Days	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
No. of accidents	16	20	14	13	17	16

**Solution.** Here,  $n = 6$ , total number of accidents = 96

Null hypothesis  $H_0$ : the accidents are uniformly distributed over the week.

Under  $H_0$ , the expected number of accidents of each of these days

$$= \frac{\text{Total no. of accidents}}{\text{No. of days}} = \frac{96}{6} = 16$$

The observed and expected number of accidents are given below:

$O_i$	16	20	14	13	17	16
$E_i$	16	16	16	16	16	16
$(O_i - E_i)^2$	0	16	4	9	1	0

$$\chi^2 = \sum_{i=1}^6 \frac{(O_i - E_i)^2}{E_i} = \frac{0 + 16 + 4 + 9 + 1 + 0}{16} = \frac{30}{16} = 1.875.$$

Tabulated value of  $\chi^2$  for 5 ( $6 - 1 = 5$ ) degrees of freedom at 5% level of significance is 11.07.

Since calculated value of  $\chi^2$  is less than tabulated value of  $\chi^2$ , so  $H_0$  is accepted, i.e., the accidents are uniformly distributed over the week.

**Example 11.29.** A die is thrown 120 times and the result of these throws are given as:

No. appeared on the die	1	2	3	4	5	6
Frequency	16	30	22	18	14	20

Test whether the die is biased or not.

**Solution.** Here,  $n = 6$ , total frequency = 120

Null hypothesis  $H_0$ : die is unbiased

Under  $H_0$ , the expected frequencies for each digit =  $\frac{120}{6} = 20$



## NOTES

The observed and expected frequencies are given below :

$O_i$	16	30	22	18	14	20
$E_i$	20	20	20	20	20	20
$(O_i - E_i)^2$	16	100	4	4	36	0

$$\chi^2 = \sum_{i=1}^6 \frac{(O_i - E_i)^2}{E_i} = \frac{16 + 100 + 4 + 4 + 36 + 0}{20} = \frac{160}{20} = 8$$

Tabulated value of  $\chi^2$  for 5 ( $6 - 1 = 5$ ) degrees of freedom at 5% level of significance is 11.07. Since calculated value of  $\chi^2$  is less than tabulated value of  $\chi^2$ , so  $H_0$  is accepted, i.e. the die is unbiased.

**Example 11.30.** The following table shows the distribution of digits in numbers chosen at random from a telephone directory:

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	1026	1107	997	966	1075	933	1107	972	964	853

Test at 5% level whether the digits may be taken to occur equally frequently in the directory.

**Solution.** Here,  $n = 10$ , total frequency = 10,000

Null hypothesis  $H_0$  : all the digits occur equally frequently in the directory

Under  $H_0$ , the expected frequency of each of the digits =  $\frac{10,000}{10} = 1000$

The observed and expected frequencies are given below:

$O_i$	1026	1107	997	966	1075	933	1107	972	964	853
$E_i$	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
$(O_i - E_i)^2$	676	11449	9	1156	5625	4489	11449	784	1296	21609

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} = \frac{676 + 11449 + \dots + 21609}{1000} \\ &= \frac{58542}{1000} = 58.542\end{aligned}$$

Tabulated value of  $\chi^2$  for 9 ( $10 - 1 = 9$ ) degrees of freedom at 5% level of significance is 16.92.

Since calculated value of  $\chi^2$  is greater than tabulated value of  $\chi^2$ , so  $H_0$  is rejected, i.e., all the digits in the numbers in the telephone directory do not occur equally frequently.

**Example 11.31.** Fit a Poisson distribution for the following data and test the goodness of fit.

No. of defects ( $x$ )	0	1	2	3	4	5
Frequency	6	13	13	8	4	3

## NOTES

**Solution.** Null hypothesis  $H_0$ : Poisson distribution is a good fit to the data.  
We first find the Poisson distribution for the above data.

$$\text{Mean of given distribution} = \frac{\sum f_i x_i}{\sum f_i} = \frac{94}{47} = 2$$

Here,  $\lambda = 2$  (For a Poisson distribution mean  $= \lambda$ )

$$N = \sum f_i = 47$$

The expected frequencies of the Poisson distribution are given by

$$E(r) = N \times e^{-\lambda} \frac{\lambda^r}{r!} = 47 \times e^{-2} \frac{2^r}{r!}; r = 0, 1, 2, 3, 4, 5$$

The expected frequencies are as:

$$E(0) = 47 \times e^{-2} \cdot \frac{2^0}{0!} = 6.36 \approx 6 \quad (e^{-2} = 0.1353)$$

$$E(1) = 47 \times e^{-2} \cdot \frac{2^1}{1!} = 12.72 \approx 13$$

$$E(2) = 47 \times e^{-2} \cdot \frac{2^2}{2!} = 12.72 \approx 13$$

$$E(3) = 47 \times e^{-2} \cdot \frac{2^3}{3!} = 8.48 \approx 9$$

$$E(4) = 47 \times e^{-2} \cdot \frac{2^4}{4!} = 4.24 \approx 4$$

$$E(5) = 47 \times e^{-2} \cdot \frac{2^5}{5!} = 1.696 \approx 2$$

$x$	0	1	2	3	4	5
$O_i$	6	13	13	8	4	3
$E_i$	6.36	12.72	12.72	8.48	4.24	1.696
$(O_i - E_i)^2$	0.1296	0.0784	0.0784	0.2304	0.0576	1.7004

$$\begin{aligned} \chi^2 &= \sum_{i=0}^5 \frac{(O_i - E_i)^2}{E_i} = \frac{0.1296}{6.36} + \frac{0.0784}{12.72} + \frac{0.0784}{12.72} + \frac{0.2304}{8.48} + \frac{0.0576}{4.24} + \frac{1.7004}{1.696} \\ &= 0.02038 + 0.00616 + 0.00616 + 0.02717 + 0.01358 + 1.0026 \\ &= 1.07605 \end{aligned}$$

Tabulated value of  $\chi^2$  for 4 ( $6 - 2 = 4$ ) degrees of freedom at 5% level of significance is 9.488.

Since calculated value of  $\chi^2$  is less than tabulated value of  $\chi^2$ , so  $H_0$  is accepted, i.e., Poisson distribution is a good fit to the data.

**Example 11.32.** Find the expected frequencies of  $2 \times 2$  contingency table  $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$ .

**Solution.**

Attributes	$B_1$	$B_2$	Total
$A_1$	$a$	$b$	$a + b$
$A_2$	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$N = a + b + c + d$

The expected frequencies are

$$E(a) = E(A_1, B_1) = \frac{(a+b)(a+c)}{a+b+c+d}$$

$$E(b) = E(A_1, B_2) = \frac{(a+b)(b+d)}{a+b+c+d}$$

$$E(c) = E(A_2, B_1) = \frac{(c+d)(a+c)}{a+b+c+d}$$

$$E(d) = E(A_2, B_2) = \frac{(c+d)(b+d)}{a+b+c+d}$$

## NOTES

**Example 11.33.** In a locality 100 persons were randomly selected and asked about their educational achievements. The results are given below:

Sex	Education		
	Middle	High school	College
Male	10	15	25
Female	25	10	15

Based on this information can you say the education depends on sex.

**Solution.** Null hypothesis  $H_0$  : Education is independent of sex.

Under the null hypothesis expected frequencies can be calculated by using

$$E_{ij} = \frac{R_i \times C_j}{N}$$

( $i = 1, 2; j = 1, 2, 3$ )

Sex	Education			Total
	Middle	High school	College	
Male	10	15	25	50 ( $R_1$ )
Female	25	10	15	50 ( $R_2$ )
Total	35 ( $C_1$ )	25 ( $C_2$ )	40 ( $C_3$ )	$N = 100$

Expected frequencies are:

Sex	Education			Total
	Middle	High school	College	
Male	$\frac{50 \times 35}{100} = 17.5$	$\frac{50 \times 25}{100} = 12.5$	$\frac{50 \times 40}{100} = 20$	50
Female	$\frac{50 \times 35}{100} = 17.5$	$\frac{50 \times 25}{100} = 12.5$	$\frac{50 \times 40}{100} = 20$	50
Total	35	25	40	100

## NOTES

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(10 - 17.5)^2}{17.5} + \frac{(15 - 12.5)^2}{12.5} + \frac{(25 - 20)^2}{20} + \frac{(25 - 17.5)^2}{17.5} + \frac{(10 - 12.5)^2}{12.5} + \frac{(15 - 20)^2}{20}$$

$$= 3.214 + 0.5 + 1.25 + 3.214 + 0.5 + 1.25 = 9.928$$

Tabulated value of  $\chi^2$  for 2 [(2 - 1) (3 - 1) = 2] degrees of freedom at 5% level of significance is 5.991. Since calculated value of  $\chi^2$  is greater than tabulated value of  $\chi^2$ , so  $H_0$  is rejected, i.e., education is not independent of sex or there is a relation between education and sex.

**Example 11.34.** The following table gives the number of good and bad parts produced by each of the three shifts in a factory.

	Good parts	Bad parts	Total
Day shift	960	40	1000
Evening shift	940	50	990
Night shift	950	45	995
Total	2850	135	2985

Test whether the production of bad parts is independent of the shifts on which they were produced.

**Solution.** Null hypothesis  $H_0$  : The production of bad parts is independent of the shift on which they were produced, i.e. production and shifts are independent.

Under the null hypothesis expected frequencies can be calculated by using

$$E_{ij} = \frac{R_i \times C_j}{N} \quad (i = 1, 2, 3; j = 1, 2)$$

Expected frequencies are:

	Good parts	Bad parts	Total
Day shift	$\frac{1000 \times 2850}{2985} = 954.774$	$\frac{1000 \times 135}{2985} = 45.226$	1000
Evening shift	$\frac{990 \times 2850}{2985} = 945.226$	$\frac{990 \times 135}{2985} = 44.774$	990
Night shift	$\frac{995 \times 2850}{2985} = 950.000$	$\frac{995 \times 135}{2985} = 45.000$	995
	2850	135	2985

$$\chi^2 = \sum_{i=1}^3 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(960 - 954.774)^2}{954.774} + \frac{(40 - 45.226)^2}{45.226} + \frac{(940 - 945.226)^2}{945.226}$$

$$+ \frac{(50 - 44.774)^2}{44.774} + \frac{(950 - 950)^2}{950} + \frac{(45 - 45)^2}{45}$$

$$= 0.0286 + 0.6039 + 0.0289 + 0.6099 + 0 + 0 = 1.2713$$

Tabulated value of  $\chi^2$  for 2 [(3 - 1) (2 - 1) = 2] degrees of freedom at 5% level of significance is 5.991

Since calculated value of  $\chi^2$  is less than tabulated value of  $\chi^2$ , so  $H_0$  is accepted, i.e., the production of bad parts is independent of the shift on which they were produced.

## NOTES

### 11.24. SUMMARY

- A statistical measure based only on all the units selected in a sample is called 'statistic', e.g., sample mean, sample standard deviation, proportion of defectives, etc. whereas a statistical measure based on all the units in the population is called 'parameter'. The terms like mean, median, mode, standard deviation are called parameters when they describe the characteristics of the population and are called statistic when they describe the characteristics of the sample.
- A statistical hypothesis is a statement about a population parameter. There are two types of statistical hypothesis, null hypothesis and alternative hypothesis.
- The hypothesis formulated for the sake of rejecting it under the assumption that it is true, is called the null hypothesis and is denoted by  $H_0$ . Null hypothesis asserts that there is no significant difference between the sample statistic and the population parameter and whatever difference is observed that is merely due to fluctuations in sampling from the same population.
- The number of independent variates which make up the statistic is known as the degree of freedom (d.f.) and is denoted by  $\nu$  (the letter 'Nu' of the Greek alphabet).
- In test of hypothesis of parameters, it is usually assumed that the random variable follows a particular distribution. To confirm whether our assumption is right, Chi-square test is used which measures the discrepancy between the observed (actual) frequencies and theoretical (expected) frequencies, on the basis of outcomes of a trial or observational data. Chi-square is a letter of the Greek alphabet and is denoted by  $\chi^2$ . It is a continuous distribution which assumes only positive values.

### 11.25. REVIEW EXERCISES

1. The frequency distribution of the digits on a set of random numbers was observed to be:

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Test the hypothesis that the digits are uniformly distributed.

2. The following table gives the number of accidents that took place in an industry during various days of the week:

Days	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
No. of accidents	14	18	12	11	15	14

Test if accidents are uniformly distributed over the week.



## NOTES

3. A die is thrown 276 times and the results of these throws are given below:

No. appeared on the die	1	2	3	4	5	6
Frequency	40	32	29	59	57	59

Test whether the die is biased or not.

4. A sample analysis of examination results of 500 students was made. It was found that 220 had failed; 170 had secured a third class; 90 were placed in second class; 20 got first class. Are these results commensurable with the general examination result which is in the ratio of 4 : 3 : 2 : 1 for the above said categories respectively.
5. Four dice were thrown 112 times and the number of times 1, 3 or 5 was thrown were as under:

No. of dice throwing 1, 3 or 5	0	1	2	3	4
Frequency	10	25	40	30	7

Test the hypothesis that all dice were fair.

6. Fit a Poisson distribution for the following data and test the goodness of fit.

No. of defects ( $x$ )	0	1	2	3	4
Frequency	109	65	22	3	1

7. For the data given in the following table use  $\chi^2$ -test to test the effectiveness of inoculation in preventing the attack of smallpox.

	Attacked	Not attacked
Inoculated	25	220
Not inoculated	90	160

8. Two investigators draw samples from the same town in order to estimate the number of persons falling in the income groups 'poor', middle class' and 'well to do'. Their results are as follows:

Investigator	Income groups		
	Poor	Middle class	Well to do
A	140	100	15
B	140	50	20

Test whether the sampling techniques of the two investigators are significantly dependent of the income groups of people.

## Answers

- Yes
- Biased
- Yes
- Poisson distribution is a good fit to the data.
- Inoculation against smallpox is a preventive measure.
- Sampling techniques are dependent of the income groups.
- Yes
- No



**SWAMI VIVEKANAND**  
**SUBHARTI**  
**UNIVERSITY**  
UGC Approved Meerut  
*Where Education is a Passion ...*

