मङ्गलायतन विश्वविद्यालय
॥ विश्वं ज्ञाने प्रतिष्ठितम् ॥

MANGALAYATAN
U N I V E R S I T Y

Learn Today to Lead Tomorrow

Business Statistics

MGO-1102

Edited By

Dr.Anurag Shakya

DIRECTORATE OF DISTANCE AND ONLINE EDUCATION

MANGALAYATAN
UNIVERSITY

CONTENTS

1. ROLE OF STATISTICS AND MEASURES OF CENTRAL TENDENCY

STRUCTURE

1.1. INTRODUCTION

NOTES

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 INFERENTIAL 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.

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 $\bar{x}$.

NOTES

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

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

i.e., $\bar{\mathbf{x}} = \frac{\Sigma \mathbf{x}}{\mathbf{n}}$

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

(*b*) **For a frequency distribution,**

$$\text{A.M.} = \frac{f_1x_1 + f_2x_2 + ... + f_nx_n}{f_1 + f_2 + ... + f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i} = \frac{\Sigma fx}{\Sigma f} = \frac{\Sigma fx}{\text{N}},$$

i.e., $\bar{\mathbf{x}} = \frac{\Sigma \mathbf{fx}}{\mathbf{N}}$

where f_i is the frequency of x_i ($1 \le i \le n$). For simplicity, Σf, *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 (Σfx) 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.

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

$= \mathbf{8.875}$ **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 $x_1, x_2, ..., x_9$.

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

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

Let x_{10} be the 10th item.

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

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

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

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

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

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 $= \dfrac{\text{Sum of marks obtained by 20 students}}{20}$

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

(*b*) New marks are:

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.

$\therefore$ New mean marks

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

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

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		Σfx = 1425

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

NOTES

1.8. STEP DEVIATION METHOD

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 $x_1, x_2, ..., x_n$, then the values of deviations are $d_1 = x_1 - A, d_2 = x_2 - A, ..., d_n = x_n - A$. 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

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

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

$$\therefore \quad \bar{x} = \frac{1}{N}\Sigma f_i x_i = \frac{1}{N}\Sigma f_i (A + u_i h) = \frac{1}{N}\Sigma f_i A + \frac{1}{N}\Sigma f_i u_i h$$

$$= A \cdot \frac{\Sigma f_i}{N} + \frac{1}{N}(\Sigma f_i u_i)h = A + \frac{\Sigma f_i u_i}{N} h \qquad (\because \Sigma f_i = N)$$

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

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

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

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:

$$\bar{\mathbf{x}} = \mathbf{A} + \frac{\Sigma \mathbf{d}}{\mathbf{n}} \qquad \textbf{(For Individual Series)}$$

$$\bar{\mathbf{x}} = \mathbf{A} + \frac{\Sigma \mathbf{fd}}{\mathbf{N}} \qquad \textbf{(For Frequency Distribution)}$$

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

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 'Σd' 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 :

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

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

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*	

NOTES

Solution. **Calculation of A.M.**

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				$\Sigma fu = 339$

Now $\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h = -5 + \left(\frac{339}{365}\right)10 = -5 + 9.2877 = 4.2877$°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 $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, ..., y_{n_2}$ be the items in the two groups respectively.

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

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

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

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

Now $$\bar{x} = \frac{\text{sum of items in both groups}}{n_1 + n_2}$$

$$= \frac{x_1 + x_2 + ... + x_{n_1} + y_1 + y_2 + ... + y_{n_2}}{n_1 + n_2} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

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

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.*

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

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

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

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

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.**

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

where $w_1, w_2, ..., w_n$ are the weights of the values $x_1, x_2, ..., x_n$ 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 $= \frac{\Sigma w x_1}{\Sigma w} = \frac{633}{10} = 63.3$

Weighted A.M. of B $= \frac{\Sigma w x_2}{\Sigma w} = \frac{624}{10} = 62.4$

Weighted A.M. of C $= \frac{\Sigma w x_3}{\Sigma w} = \frac{648}{10} = 64.8$

$\therefore$ 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

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

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

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 $(1 \le i \le n)$.

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.

NOTES

EXERCISE 1.1

1. Find the A.M. of the series 4, 6, 8, 10, 12.
2. 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.
3. Find the A.M. for the following frequency distribution:

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

4. 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

5. 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

6. 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

7. From the following information, find out:
   (*i*) Which of the factor pays larger amount as daily wages.
   (*ii*) 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

8. 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.
9. 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.
10. 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 nth root of the product of all the n values of the variable.

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

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

where x_1, x_2, ..., x_n are the values of the variable, under consideration. From the definition of G.M., we see that it involves the nth 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

$$\textbf{G.M.} = (\mathbf{x_1 x_2 \ldots\ldots x_n})^{1/n}$$

$$= \text{Antilog } [\log (x_1 x_2 \ldots\ldots x_n)^{1/n}]$$

$$= \text{Antilog}\left[\frac{1}{n}\log (x_1 x_2 \ldots\ldots x_n)\right]$$

$$= \text{Antilog}\left[\frac{1}{n}(\log x_1 + \log x_2 + \ldots\ldots + \log x_n)\right]$$

$$\therefore \quad \textbf{G.M.} = \textbf{Antilog}\left(\frac{\Sigma \log \mathbf{x}}{\mathbf{n}}\right)$$

For a frequency distribution,

$$\textbf{G.M.} = (\mathbf{x_1}^{f_1} \, \mathbf{x_2}^{f_2} \ldots\ldots \mathbf{x_n}^{f_n})^{1/N}$$

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

Proceeding on the same lines, we get

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

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 (Σ 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.*

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

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

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

Solution. **Calculation of G.M.**

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

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

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

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

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

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

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

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

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

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

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.

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

$$= \text{Antilog}\left(\frac{200 \log 700 + 300 \log 1000}{200 + 300}\right)$$

$$= \text{Antilog}\left(\frac{200\,(2.8451) + 300\,(3.0000)}{500}\right) = \text{Antilog}\left(\frac{569.0200 + 900}{500}\right)$$

$$= \text{Antilog}\,(2.9380) = \textbf{Rs. 867.}$$

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 nth period, then

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

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

$$\therefore \quad V_{10} = V_0\,(1 + r)^{10}$$

or $$200 = 100(1 + r)^{10} \quad \text{or} \quad (1 + r)^{10} = 2$$

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

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

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

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

$\therefore$ Average percentage rate of interest = 0.074 × 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.*

NOTES

Solution.

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
			$\Sigma \log x = 9.4067$

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

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

$\therefore$ 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 $w_1, w_2, \ldots\ldots, w_n$ are the weights of the values $x_1, x_2, \ldots\ldots, x_n$ 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

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

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

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

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

Now $$\text{Correct } \Sigma \log x = 16.1880 - \log 14 + \log 21$$

$$= 16.1880 - 1.1461 + 1.3222 = 16.3641.$$

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

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.

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

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

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

(*b*) **For a frequency distribution,**

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

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

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 (Σ(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*

Solution. **Calculation of H.M.**

NOTES

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

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

1.17. H.M. OF COMBINED GROUP

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

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

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

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

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

Now $$H = \frac{\text{no. of items in both groups}}{\text{sum of reciprocals of all the items in both groups}}$$

$$= \frac{n_1 + n_2}{\sum \frac{1}{x} + \sum \frac{1}{y}} \qquad \therefore \quad \mathbf{H} = \frac{\mathbf{n_1 + n_2}}{\dfrac{\mathbf{n_1}}{\mathbf{H_1}} + \dfrac{\mathbf{n_2}}{\mathbf{H_2}}}.$$

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.

$$\therefore \quad H = \frac{n_1 + n_2}{\dfrac{n_1}{H_1} + \dfrac{n_2}{H_2}} = \frac{10 + 12}{\dfrac{10}{29} + \dfrac{12}{35}}$$

$$= \frac{22}{0.3448 + 0.3429} = \frac{22}{0.6877} = \mathbf{31.9907.}$$

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 $w_1, w_2, ..., w_n$ are the weights of the values $x_1, x_2, ..., x_n$ 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.*

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
	$\sum w = 6$	$\sum \left(\frac{w}{x}\right) = 0.7116$

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

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,

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

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

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

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

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 friction, 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

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

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

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

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.

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

$\therefore$ Median = size of 5.5th item

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

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

The marks in Accountancy arranged in ascending order are:

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

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

Median = size of 5.5th item

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

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

$\therefore$ 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.

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

NOTES

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

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.

$$\therefore \quad \text{Median} = L + \left(\frac{\frac{N}{2} - c}{f}\right)h = 35 + \left(\frac{400 - 220}{180}\right)5$$

$$= 35 + 5 = \textbf{40 years.}$$

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.

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

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

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

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

or $$4273.86 = (398 - a)\ 25$$

or $$398 - a = 170.9544 \quad \text{or} \quad a = 227.0456 = 227$$

Also $$525 + a + b = 1000$$

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

∴ The missing frequencies are **227** and **228.**

NOTES

Merits of Median

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:

x	5	10	15	20	25
f	2	4	6	8	10

3. Find the median for the following frequency distribution:

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

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

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

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

Class Interval	10—20	10—30	10—40	10—50
Frequency	4	6	56	97
Class Interval	10—60	10—70	10—80	10—90
Frequency	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 densily. 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

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.

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

where L = lower limit of modal class

Δ_1 = difference of frequencies of modal class and pre-modal class,

Δ_2 = difference of frequencies of modal class and post-modal class

h = width of the modal class.

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

Mode = 3 Median – 2 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 Median – 2 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 median – 2 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.

NOTES

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*

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

Grouping Table

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

Analysis Table

Column	*30*	*15*	*20*	*10*	*25*
I	1				
II		1	1		
III		1		1	
IV	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:

mode = 3 median – 2 A.M.

Calculation of $\overline{X}$ 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				$\Sigma fu = 44$

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

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

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

$\therefore$ 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	15				
5—10	9		13	19		
10—15	4	6			15	
15—20	2		12			16
20—25	**10**	**18**		**20**		
25—30	**8**		**15**		**25**	
30—35	7	12				**20**
35—40	5		6	13		
40—45	1	4			9	
45—50	3					

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

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

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 median – 2(15.6)

$\Rightarrow$ 3 median = 16 + 31.2 = 47.2

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

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

Solution. The following are the relations between mathematical averages:

(I) A.M. ≥ G.M. ≥ H.M.

In particular, if all the items are identical, then

A.M. = G.M. = 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	

Calculation of Median

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

∴ Median = size of 17.5th item

∴ Median class is 10—15.

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

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

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

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.

NOTES

EXERCISE 1.5

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:

x	4	5	6	7	8	9	10	11	12	13
f	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:

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

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

Marks	*No. of candidates*	*Marks*	*No. of candidates*
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:

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

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

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

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.**

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

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

NOTES

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

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

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

- 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.**

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

  where $w_1, w_2, ..., w_n$ are the weights of the values $x_1, x_2, ..., x_n$ 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**

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

- 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.**

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

  where $w_1, w_2, ..., w_n$ are the weights of the values $x_1, x_2, ..., x_n$ 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

  Mode = 3 Median − 2 A.M.

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

1.26. REVIEW EXERCISES

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

NOTES

2. MEASURES OF DISPERSION

STRUCTURE

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.

$\therefore$ **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.

Example 2.1. *Find the range of the following distribution:*

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

NOTES

Solution. Here L = 28, S = 18

∴ Range = L – S = 28 – 18 = **10 years.**

It may be noted that S ≠ 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

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

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

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

∴ Coeff. of Range (A) > Coeff. of Range (B)

∴ 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.

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.	

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

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

On comparing the values of coeff. of range for these series, one is likely to conclude that these 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 : $\frac{n+1}{4} = \frac{7+1}{4} = 2$ $\therefore$ Q_1 = size of 2nd item = ₹ 5

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

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

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

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

(i) the coefficient of range

(ii) interquartile range, and

(iii) 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*	

NOTES

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

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

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	

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

$\therefore$ Q_1 class is 14.5—19.5

$$\therefore \quad Q_1 = L + \left(\frac{N/4 - c}{f}\right)h = 14.5 + \left(\frac{6.25-4}{8}\right)5$$

$$= 14.5 + 1.4063 = 15.9063 \text{ marks}$$

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

$\therefore$ Q_3 class is 24.5—29.5

$$\therefore \quad Q_3 = L + \left(\frac{3\,(N/4) - c}{f}\right)h = 24.5 + \left(\frac{18.75-17}{4}\right)5$$

$$= 24.5 + 2.1875 = 26.6875 \text{ marks}$$

$\therefore$ Interquartile range

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

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

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

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

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

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

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

$$\therefore \quad P_{90} = L + \left(\frac{90\,(N/100) - c}{f}\right)h$$

$$= 29.5 + \left(\frac{22.5 - 21}{2}\right)5 = 29.5 + 3.75 = 33.25 \text{ marks}$$

NOTES

$\therefore$ 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

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

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

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

For a **frequency distribution**,

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

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

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.

NOTES

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

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

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

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

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

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

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

x	f	fx	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
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	$\Sigma fx = 273$			$\Sigma f\lvert x-\bar{x}\rvert = 74.4$

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

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

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)*	$\lvert x - med.\rvert$	$f\lvert x - med.\rvert$
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					$\Sigma f\lvert x - med.\rvert = 478$

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

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

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

NOTES

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	–7	7
2	6 $(\Sigma x)_B$	–5	5
3	10 = 20	–1	1
4	12	1	1
5	18 $(\Sigma x)_A$	7	7
6	19 = 49	8	8
$n = 6$			Σ| x – median | = 29

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

Direct Method

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

Short-cut Method

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

$$= \frac{49 - 20 - (3 - 3).11}{6} = \frac{29}{6} = \mathbf{4.8333.}$$

NOTES

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

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	$(\Sigma f)_B = 33$	– 2500	– 25000	$(\Sigma fx)_B = -30500$
– 2000 to – 1000	8		– 1500	– 12000	
– 1000 to 0	6		– 500	– 3000	
0—1000	4		500	2000	
1000—2000	5		1500	7500	
2000—3000	10	$(\Sigma f)_A = 65$	2500	25000	$(\Sigma fx)_A = 252500$
3000—4000	30		3500	105000	
4000—5000	15		4500	67500	
5000—6000	10		5500	55000	
	N = 98			$\Sigma fx = 222000$	

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

Now
$$\text{M.D.}(\bar{x}) = \frac{(\Sigma fx)_A - (\Sigma fx)_B - [(\Sigma f)_A - (\Sigma f)_B]\,\bar{x}}{N}$$

$$= \frac{252500 - (-30500) - (65 - 33)\,2265.3061}{98}$$

$$= \frac{210510.21}{98} = ₹\ 2148.0633$$

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

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.

NOTES

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.

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 'σ'.

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

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

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

For a **frequency distribution**,

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

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

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:

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

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

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

This measure is denoted as C.V.

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

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.

NOTES

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* $= (S.D.)^2$.

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	Σfx = 2000			$\Sigma f(x - \bar{x})^2$ = 19900

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{2800}{100} = 28.$$

NOTES

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

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

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.

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

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

$$\therefore \quad a + b = 11 \qquad \text{...(1)}$$

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

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

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

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

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

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

$$\Rightarrow \quad a^2 + b^2 + 41 = (21.2) \times 5 = 106 \quad \Rightarrow \quad a^2 + b^2 = 65 \qquad \text{...(2)}$$

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 $x_1, x_2, \ldots\ldots, x_n$, we have

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

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

NOTES

For a frequency distribution, this formula takes the form

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

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

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:

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

$$\text{S.D.} = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \qquad \textbf{(Frequency Distribution)}$$

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*

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

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: $\sqrt{\frac{\Sigma u^2}{n} - \left(\frac{\Sigma u}{n}\right)^2} \times h$.

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 Σfd and Σfd^2. Find S.D. by using the formula:*

$$\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

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 Σfu and Σfu². 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			$\Sigma fu = 33$	$\Sigma fu^2 = 111$

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

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

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

Calculation of C.V. for Batsman B

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		$\Sigma fu = 8$	$\Sigma fu^2 = 56$

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

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

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

$\therefore$ $\quad$ 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*

NOTES

Solution. **Calculation of C.V. for group A**

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				$\Sigma fu = -6$	$\Sigma fu^2 = 246$

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

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

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

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

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				$\Sigma fu = -25$	$\Sigma fu^2 = 209$

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

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

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

$$\bar{x}\ (A) = 58 \qquad \sigma\ (A) = 15$$
$$\bar{x}\ (B) = 48 \qquad \sigma\ (B) = 12$$
$$\bar{x}\ (C) = 12 \qquad \sigma\ (C) = 2$$

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

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

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

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.

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.

NOTES

EXERCISE 2.4

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.

NOTES

Answers

1. $\bar{x}$ = ₹ 40.52, S.D. = ₹ 17.41, Coeff. of S.D. = 0.4296
2. ₹ 14.5079
3. C.V. for A = 67.0738%, C.V. for B = 69.5120% } 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.

2.14. SUMMARY

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

  ∴ **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 'σ'.
- 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

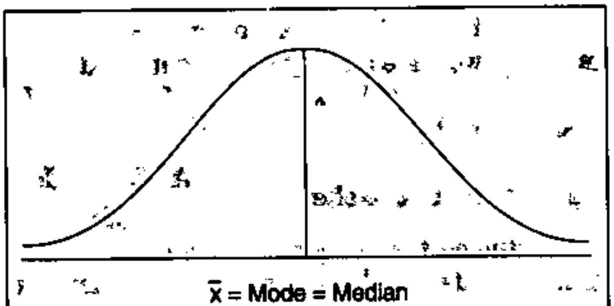
We have already seen that a single statistical measure is not capable of telling everything about a statistical distribution. A single measure cannot explore all the characteristics of a distribution. As we have already seen that an average of a distribution gives us an idea about the concentration of 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

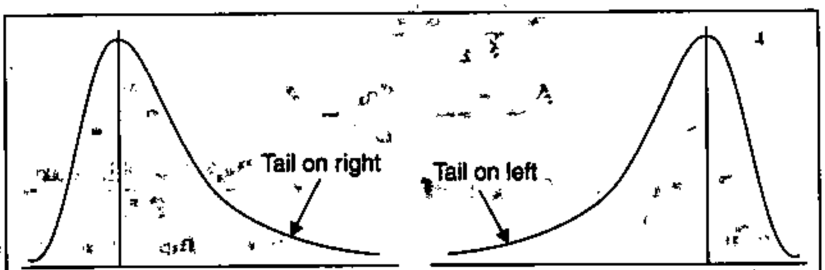
In symmetrical distribution, the values of mean, mode and median, would coincide. If the curve of the distribution is not symmetrical, it may admit of tail on either side of the distribution. Such a distribution lack in symmetry. **Skewness** is the word used for lack of symmetry. A distribution which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

NOTES

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



Symmetrical distribution



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

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

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.

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

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

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

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

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 $\bar{x}$, σ and mode are given, then find SK_P by using the formula:*

$$SK_P = \frac{\bar{x} - mode}{\sigma}.$$

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

$$SK_P = \frac{3(\bar{x} - median)}{\sigma}.$$

Rule III. *If the values of $\bar{x}$, σ 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 $SK_P = 0.32$, S.D. = 6.5, $\bar{x} = 29.6$

Now $$SK_P = \frac{\bar{x} - \text{Mode}}{\text{S.D.}}$$

$$\therefore\quad 0.32 = \frac{29.6 - \text{Mode}}{6.5}$$

$$\Rightarrow\quad 29.6 - \text{Mode} = 0.32 \times 6.5 = 2.08$$

$$\Rightarrow\quad \text{Mode} = 29.6 - 2.08 = \mathbf{27.52}.$$

NOTES

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

$$\text{coeff. of skewness} = -0.4, \text{ A.M.} = 45, \text{ median} = 48.$$

Now, $$\text{coeff. of skewness} = \frac{3(\bar{x} - \text{Median})}{\text{S.D.}}$$

$$\Rightarrow \quad -\frac{4}{10} = \frac{3(45-48)}{\text{S.D.}} = \frac{-9}{\text{S.D.}} \Rightarrow 4\text{ S.D.} = 90$$

$$\Rightarrow \quad \text{S.D.} = \frac{90}{4} = \mathbf{22.5}.$$

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.

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

$$= \frac{3(\bar{x} - \text{Median})}{\text{S.D.}} = \frac{3(15-15)}{5} = \frac{0}{5} = \mathbf{0}$$

$$\text{C.V.} = \frac{\text{S.D.}}{\bar{x}}(100) = \frac{5}{15} \times 100 = \mathbf{33.33\%}.$$

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.

∴ 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.

∴ After settlement, the wages of workers have increased. This is a gain to the workers.

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) $\bar{x} = \frac{\Sigma x}{n}$ ∴ $\Sigma x = n.\ \bar{x}$

∴ 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) C.V. before settlement $= \frac{120}{455} \times 100 = 26.374\%$

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

1. Find the coeff. of variation of a frequency distribution with the help of following information:

   A.M. = 50 Mode = 56

   Karl Pearson's coeff. of skewness = – 0.4.

2. 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	

NOTES

3. For the following data, calculate the coefficient of skewness based on mean, median and S.D.

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

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

$$\textbf{Bowley's coefficient of skewness} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}.$$

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 medium, Q_1 and Q_3 are given, then find SK_B by using the formula:*

$$SK_B = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$$

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 = size of 64th item

$\therefore$ Q_1 class is 200—400

$$Q_1 = L + \left(\frac{N/4 - c}{f}\right)h = 200 + \left(\frac{64 - 25}{40}\right) 200 = 200 + 195 = 395.$$

Q_3 : $3\left(\frac{N}{4}\right) = 3\left(\frac{256}{4}\right) = 192$

Q_3 = size of 192th item

$\therefore$ Q_3 class is 600—800.

NOTES

$$\therefore \quad Q_3 = L + \left(\frac{3(N/4) - c}{f}\right)h = 600 + \left(\frac{192 - 145}{75}\right) 200$$

$$= 600 + 125.33 = 725.33.$$

Median : $\frac{N}{2} = \frac{256}{2} = 128$

$\therefore$ Median = size of 128th item

$\therefore$ Median class is 400—600.

$$\therefore \quad \text{Median} = L + \left(\frac{N/2 - c}{f}\right)h = 400 + \left(\frac{128 - 65}{80}\right) 200$$

$$= 400 + 157.5 = 557.5.$$

$\therefore$ Bowley's coefficient of skewness

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

$$= \frac{725.33 + 395 - 2(557.5)}{725.33 - 395} = \frac{5.33}{330.33} = \mathbf{0.016.}$$

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).

$$\therefore \quad Q_1 = L + \left(\frac{N/4 - c}{f}\right) h = 10.5 + \left(\frac{30.75 - 7}{68}\right) 5 = 10.5 + 1.746 = 12.246.$$

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

$$\therefore \quad Q_3 = L + \left(\frac{3(N/4) - c}{f}\right)h = 15.5 + \left(\frac{92.25 - 75}{30}\right)5$$

$$= 15.5 + 2.875 = 18.375.$$

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)

$$\therefore \quad \text{Median} = L + \left(\frac{N/2 - c}{f}\right)h = 10.5 + \left(\frac{61.5 - 7}{68}\right)5 = 10.5 + 4.007 = 14.507.$$

Now, Bowley's coefficient of skewness

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

$$= \frac{18.375 + 12.246 - 2(14.507)}{18.375 - 12.246} = \frac{1.607}{6.129} = \mathbf{0.262}.$$

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

NOTES

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

$$\textbf{Kelly's coefficient of skewness} = \frac{P_{90} + P_{10} - 2\text{ Median}}{P_{90} - P_{10}}.$$

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* $\dfrac{D_9 + D_1 - 2\text{ Median}}{D_9 - D_1}$.

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

$$= \frac{P_{90} + P_{10} - 2\text{ Median}}{P_{90} - P_{10}} = \frac{22 + 5 - 2(12)}{22 - 5} = \frac{3}{17} = \mathbf{0.1765}.$$

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

P_{10} : $10\left(\frac{N}{100}\right) = 10\left(\frac{40}{100}\right) = 4$

$\therefore$ P_{10} = size of 4th item

$\therefore$ P_{10} class is 20—25

$$\therefore \quad P_{10} = L + \left(\frac{10\,(N/100) - c}{f}\right) h$$

$$= 20 + \left(\frac{4-0}{12}\right) 5 = 20 + 1.67 = ₹\ 21.67.$$

P_{90} : $90\left(\frac{N}{100}\right) = 90\left(\frac{40}{100}\right) = 36$

$\therefore$ P_{90} = size of 36th item

$\therefore$ P_{90} class is 35—40.

$$\therefore \quad P_{90} = L + \left(\frac{90\,(N/100) - c}{f}\right) h = 35 + \left(\frac{36-33}{4}\right) 5$$

$$= 35 + 3.75 = ₹\ 38.75.$$

Median: $\frac{N}{2} = \frac{40}{2} = 20$

$\therefore$ Median = size of 20th item

$\therefore$ Median class is 25—30

$$\therefore \quad \text{Median} = L + \left(\frac{N/2 - c}{f}\right) h = 25 + \left(\frac{20-12}{16}\right) 5$$

$$= 25 + 2.5 = ₹\ 27.50.$$

Now, Kelly's coefficient of skewness

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

$$= \frac{38.75 + 21.67 - 2\,(27.50)}{38.75 - 21.67} = \frac{5.42}{17.08} = \mathbf{0.3173.}$$

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

NOTES

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

Moment coefficient of skewness $= \dfrac{\mu_3}{\sqrt{\mu_2^3}}$.

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 $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$

$\therefore$ 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 μ_3. The moment coefficient of skewness is also denoted by γ_1.

The moment coefficient of skewness is generally denoted by 'SK_M'.

WORKING RULES FOR SOLVING PROBLEMS

Rule I. *If the values of μ_2 and μ_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 μ_1', μ_2' and μ_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 $SK_M = \dfrac{\mu_3}{\sqrt{\mu_2^3}}$.

Rule III. *If moments are not given, then first find μ_2 and μ_3 by using the given data and then use the formula:* $SK_M = \dfrac{\mu_3}{\sqrt{\mu_2^3}}$.

Rule IV. $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$ *and* $\gamma_1 = \dfrac{\mu_3}{\sqrt{\mu_2^3}}$.

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 $\mu_3 = -31$.

$\therefore$ 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:*

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^2	fu^3
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				Σfu = –20	Σfu^2 = 220	Σfu^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

$$\mu_2 = \mu_2' - \mu_1'^2 = 55 - (-1)^2 = \mathbf{54}$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -62.5 - 3(55)(-1) + 2(-1)^3$$

$$= -62.5 + 165 - 2 = \mathbf{100.5.}$$

∴ Moment coefficient of skewness

$$= \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{100.5}{\sqrt{(54)^3}} = \frac{100.5}{54 \times 7.35} = \mathbf{0.253.}$$

EXERCISE 3.4

1. 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.
2. 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.
3. 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

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

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

- 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

$$\textbf{Bowley's coefficient of skewness} = \frac{\mathbf{Q_3 + Q_1 - 2\ Median}}{\mathbf{Q_3 - Q_1}}$$

- 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

$$\textbf{Kelly's coefficient of skewness} = \frac{\mathbf{P_{90} + P_{10} - 2\ Median}}{\mathbf{P_{90} - P_{10}}}$$

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

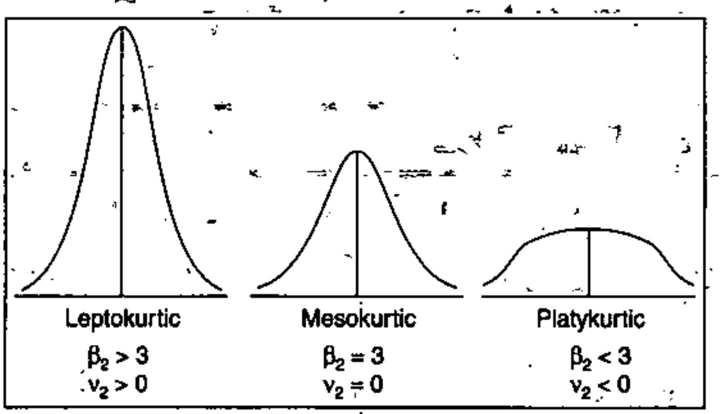
STRUCTURE

4.1. INTRODUCTION

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 or 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.

NOTES

4.3. MEASURE OF KURTOSIS

The measure of kurtosis is denoted by β_2 and is defined as

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

where μ_2 and μ_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 'ν_2', which is defined as $\nu_2 = \beta_2 - 3$.

$\therefore \quad \nu_2 > 0 \Rightarrow \beta_2 - 3 \geq 0 \Rightarrow \beta_2 > 3 \Rightarrow$ the distribution is Leptokurtic.

Similarly, if $\nu_2 = 0$, then $\beta_2 = 3$

$\therefore$ The distribution is Mesokurtic.

$\nu_2 < 0 \Rightarrow \beta_2 < 3 \Rightarrow$ the distribution is Platykurtic.

WORKING RULES FOR SOLVING PROBLEMS

Rule I. *If the values of μ_2 and μ_4 are given, then find β_2 by using the formula:*

$$\beta_2 = \frac{\mu_4}{\mu_2^2}.$$

Rule II. *If raw moments μ_1', μ_2', μ_3' and μ_4' are given, then calculate:*

$\mu_2 = \mu_2' - \mu_1'^2$ *and* $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$

Now, find $\beta_2 = \frac{\mu_4}{\mu_2^2}$.

Rule III. *If moments are not given, then first find μ_2 and μ_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 $\mu_4 = 35000$.

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 $\mu_4' = 108$.

$\therefore \quad \mu_2 = \mu_2' - (\mu_1')^2 = 17 - (-1.5)^2 = 14.75$

NOTES

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$
$$= 108 - 4(-1.5)(-30) + 6(17)(-1.5)^2 - 3(-1.5)^4 = 142.3125.$$

Now, $$\beta_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{142.3125}{(14.75)^2} = \frac{142.3125}{217.5625} = 0.654 < 3.$$

∴ 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*	*fu²*	*fu³*	*fu⁴*
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			Σfu = – 30	Σfu^2 = 224	Σfu^3 = – 6	Σfu^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 μ_2, μ_3, μ_4

$$\mu_2 = \mu_2' - \mu_1'^2 = 298.667 - (-4)^2 = 282.667$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -80 - 3(298.667)(-4) + 2(-4)^3 = 3376.004$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$
$$= 276266.667 - 4(-80)(-4) + 6(298.667)(-4)^2 - 3(-4)^4 = 302890.7.$$

Skewness

Moment coefficient of skewness,

$$\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{3376.004}{\sqrt{(282.667)^3}} = \frac{3376.004}{282.667\sqrt{282.667}} = \mathbf{0.71}.$$

∴ 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 β_2, the measure of kurtosis, we will have to find the values of μ_2 and μ_4.

Class	*Frequency* f	*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				Σfu = – 32	Σfu^2 = 104	Σfu^3 = – 146	Σfu^4 = 596

Now,

$$\mu_1' = \left(\frac{\Sigma fu}{N}\right)h = \left(\frac{-32}{35}\right)7 = -6.4$$

$$\mu_2' = \left(\frac{\Sigma fu^2}{N}\right)h^2 = \left(\frac{104}{35}\right)(7)^2 = 145.6$$

$$\mu_3' = \left(\frac{\Sigma fu^3}{N}\right)h^3 = \left(\frac{-146}{35}\right)(7)^3 = -1430.8$$

$$\mu_4' = \left(\frac{\Sigma fu^4}{N}\right)h^4 = \left(\frac{596}{35}\right)(7)^4 = 40885.6$$

∴

$$\mu_2 = \mu_2' - (\mu_1')^2 = 145.6 - (-6.4)^2 = 104.64$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2\mu_2' - 3(\mu_1')^4$$

$$= 40885.6 - 4(-6.4)(-1430.8) + 6(-6.4)^2(145.6) - 3(-6.4)^4$$

$$= 40885.6 - 36628.48 + 35782.656 - 5033.1648 = 35006.612.$$

$$\therefore \quad \beta_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{35006.612}{(104.64)^2} = 3.1971 > 3.$$

$\therefore$ 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 $\quad \bar{x} = 10$, variance $= 16$, $\gamma_1 = 1$, $\beta_2 = 4$.

We know $\quad \mu_1 = 0$ (always), $\mu_2 =$ variance $= 16$

$$\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^{\,3}}} \Rightarrow 1 = \frac{\mu_3}{\sqrt{(16)^3}} \Rightarrow \mu_3 = 16 \times 4 = \mathbf{64}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} \Rightarrow 4 = \frac{\mu_4}{(16)^2} \Rightarrow \mu_4 = 4 \times 256 = \mathbf{1024}.$$

Moments about origin

$$\gamma_1 = \bar{x} = \mathbf{10}$$

$$\gamma_2 = \mu_2 + \bar{x}^2 = 16 + (10)^2 = \mathbf{116}$$

$$\gamma_3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3 = 64 + 3(16)(10) + (10)^3 = \mathbf{1544}$$

$$\gamma_4 = \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4$$

$$= 1024 + 4(64)(10) + 6(16)(10)^2 + (10)^4 = \mathbf{23184}.$$

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 β_2 and is defined as

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}}$$

where μ_2 and μ_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 'ν_2', which is defined as $\nu_2 = \beta_2 - 3$.

4.5. REVIEW EXERCISES

1. Explain the term 'kurtosis'.
2. How does kurtosis differ from skewness?
3. Explain the method of studying kurtosis.
4. What are Skewness and Kurtosis? Give formula for measuring them.
5. Define 'Leptokurtic' distribution.
6. 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 μ_2 and μ_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 μ_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 β_2 and discuss the kurtosis.
12. Compute the value of β_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 β_1 and β_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

We know that a **time series** is a collection of values of a variable taken at different time periods. If $y_1, y_2, \ldots\ldots, y_n$ be the values of a variable y taken at time periods $t_1, t_2, \ldots\ldots t_r$, then we write this time series as $\{(t_i, y_i); i = 1, 2, \ldots\ldots, n\}$. 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

NOTES

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.

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 accustumed 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

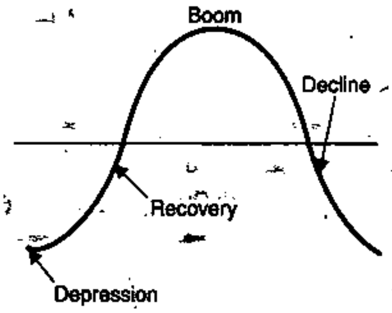
The **cyclical variations** in a time series counts for the swings of graph of time series about its trend line (curve). Cyclical variations are seldom periodic and they may or may not follow same pattern after equal interval of time.

NOTES

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.

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 components T is in abosolute value where as 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

NOTES

This is a graphic method. Let $\{(t_i, y_i): i = 1, 2, ..., n\}$ 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), ..., (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.

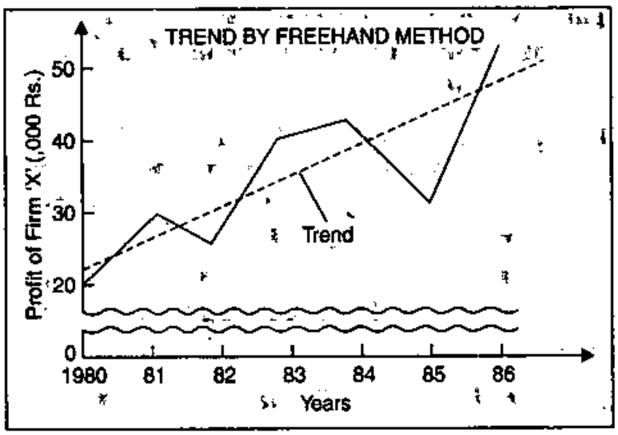
(*iii*) The area above the trend is equal to area below it.

(*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.

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.

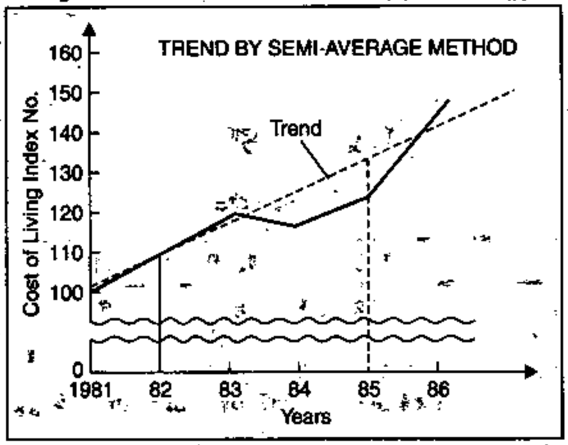
Example 5.2. *Fit a straight line trend to the following data by using semi-average method :*

Year	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*
Cost of Living Index No.	*100*	*110*	*120*	*118*	*130*	*159*

NOTES

Solution. **Trend Line by Semi-Average Method**

Year	Cost of Living Index		Year	Cost of Living Index	
1981	100		1984	118	
1982	110	$\frac{330}{3} = 100$	1985	130	$\frac{407}{3} = 135.67$
1983	120		1986	159	



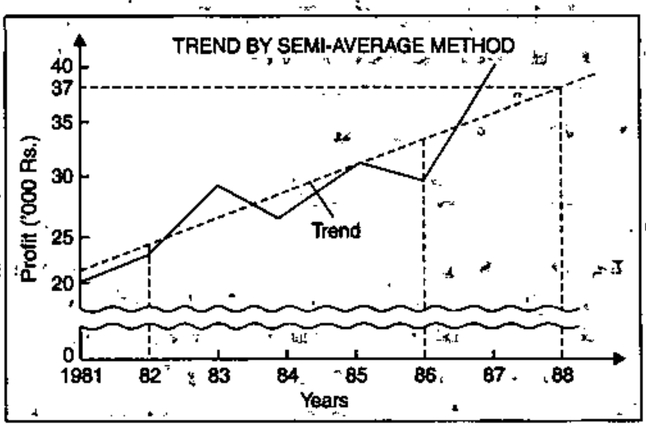
Example 5.3. *Fit a straight line trend by using the following data:*

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	$\frac{69}{3} = 23$	1986	29	$\frac{99}{3} = 33$
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 $\{(t_i, y_i): i = 1, 2, ..., n\}$ be the given time series. Here $y_1, y_2, ..., y_n$ are the values of the variable (y) corresponding to time periods $t_1, t_2, ..., t_n$ respectively.

We define **moving totals of order m** as $y_1 + y_2 + ... + y_m$, $y_2 + y_3 + ... + y_{m+1}$, $y_3 + y_4 + ... + y_{m+2}$, ...

The **moving averages of order m** are defined as

$$\frac{y_1 + y_2 + ... + y_m}{m}, \quad \frac{y_2 + y_3 + ... + y_{m+1}}{m}, \quad \frac{y_3 + y_4 + ... + y_{m+2}}{m}, ...$$

These moving averages will be called **m yearly moving averages** if the values, $y_1, y_2, ... y_n$ 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.

NOTES

Let the order of moving averages be 3. The moving averages will be

$$\frac{y_1+y_2+y_3}{3},\ \frac{y_2+y_3+y_4}{3},\ \frac{y_3+y_4+y_5}{3},\ldots,\frac{y_{n-2}+y_{n-1}+y_n}{3}.$$

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

$$\frac{y_1+y_3+y_3+y_4+y_5}{5},\ \frac{y_2+\ldots\ldots+y_6}{5},\ldots\ldots,\frac{y_{n-4}+\ldots+y_n}{5}.$$

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:

$$\frac{y_1+y_2+y_3+y_4}{4},\ \frac{y_2+y_3+y_4+y_5}{4},\ldots\ldots,\frac{y_{n-3}+y_{n-2}+y_{n-1}+y_n}{4}.$$

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:

$$\{(t_i, y_i)\}: i = 1, 2, \ldots\ldots, n\}.$$

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.

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.t.*	*5 Yearly m.a.*	*7 Yearly m.t.*	*7 Yearly m.a.*	*9 Yearly m.t.*	*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.

NOTES

Solution. **Trend by Moving Average Method**

Year	*Output*	*4 yearly moving total*	*2 yearly moving total of column 3*	*4 yearly centered moving average*
1967	72		—	—
1968	68		—	—
		264		
1969	64		524	**65.5**
		260		
1970	60		524	**65.5**
		264		
1971	68		536	**67**
		272		
1972	72		560	**70**
		288		
1973	72		580	**72.5**
		292		
1974	76		580	**72.5**
		288		
1975	72		—	—
1976	68		—	—

Merits of Moving Average Method

1. This method is rigidily 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:

Year	Production (Lakh tonnes)	Year	Production (Lakh tonnes)
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:

Year	No. of failures	Year	No. of failures
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:

Year	Tonnage of cargo cleared	Year	Tonnage of cargo cleared
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):

Year	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962
Bank Clearance (in crores of rupees)	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:

Year	1980	1981	1982	1983	1984	1985	1986	1987
Sale (in lakhs of rupees)	20	22	25	24	26	30	35	40

NOTES

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		

Answers

1. 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
2. 21.857, 20, 17.571, 15.429, 12.429, 11.571, 11.571, 11.143, 10.143,9
3. 1256.75, 1278.875, 1321.25, 1368.875, 1429.25, 1495.875, 1537.375, 1563.625
4. 68.6, 76.8, 82, 84.2, 86.8, 93.8, 94.4, 91.8
5. 23.5, 25.25, 27.5; 30.75
6. 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 mathemetical method. Let $\{(t_i, y_i): i = 1, 2, \ldots\ldots, n\}$ 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 ($= y_1, y_2, \ldots\ldots, y_n$) 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 $(t_1, t_2, \ldots\ldots, t_n)$ 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.

NOTES

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:*

$$\Sigma y = an + b\Sigma x \quad \text{...(1)}$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{...(2)}$$

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 = *year* – *1976*	x^2	xy	$y_e = a + bx$
1	1975	800	–1	1	–800	873.572 + 20.357 (–1) = **853.215**
2	1976	900	0	0	0	873.572 + 20.357(0) = **873.572**
3	1977	920	1	1	920	873.572 + 20.357(1) = **893.929**
4	1978	930	2	4	1860	873.572 + 20.357(2) = **914.286**
5	1979	940	3	9	2820	873.572 + 20.357(3) = **934.643**
6	1980	980	4	16	3920	873.572 + 20.357(4) = **955.000**
$n = 7$	1981	930	5	25	4650	873.572 + 20.357(5) = **975.357**
Total		6400	14	56	13370	

Let the equation of the trend line by $y_e = a + bx$.

The normal equations are:

$$\Sigma y = an + b\Sigma x \quad \text{...(1)}$$

and $$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{....(2)}$$

(1) $\Rightarrow$ $6400 = 7a + 14b$ ...(3)

(2) $\Rightarrow$ $13370 = 14a + 56b$ ...(4)

(3) × 2 $\Rightarrow$ $12800 = 14a + 28b$ ...(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.

NOTES

For 1983, $x = 1983 - 1976 = 7$.

$\therefore \quad y_e\,(1983) = 873.572 + (20.357)7 =$ ₹ **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

$$\Sigma y = an + b.0 \quad \text{and} \quad \Sigma xy = a.0 + b\Sigma x^2.$$

$$\therefore \qquad \mathbf{a} = \frac{\Sigma \mathbf{y}}{\mathbf{n}} \quad \text{and} \quad \mathbf{b} = \frac{\Sigma \mathbf{xy}}{\Sigma \mathbf{x}^2}.$$

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 = 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:

$$\Sigma y = an + b\Sigma x \qquad \text{...(1)}$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2. \qquad \text{...(2)}$$

(1) $\Rightarrow \quad 630 = 7a + b.0 \quad \Rightarrow \quad a = 90$

(2) $\Rightarrow \quad 56 = a.0 + 28b \quad \Rightarrow \quad b = 2$

∴ 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 represent 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 = 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:

$$\Sigma y = an + b\Sigma x \quad \text{...(1)}$$

and $$\Sigma xy = a\Sigma x + b\Sigma x^2. \quad \text{...(2)}$$

(1) ⇒ $$84 = 6a + b(0) \Rightarrow a = \frac{84}{6} = 14$$

(2) ⇒ $$54 = a(0) + b(17.5) \Rightarrow b = \frac{54}{17.5} = 3.0857.$$

∴ 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 + (3.085)(2.5) = **21.7142.**

NOTES

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

Fit a straight line by the least squares method and calculate the trend values.

Solution. We define x = year – 1980 and y = production.

Trend Line by Least Squares Method

S. No.	Year	y	x = 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:

$$\Sigma y = an + b\Sigma x \quad ...(1)$$

and
$$\Sigma xy = a\Sigma x + b\Sigma x^2. \quad ...(2)$$

(1) $\Rightarrow$ $623 = 7a + b$ ...(3)

(2) $\Rightarrow$ $159 = a + 51b$ ...(4)

(4) × 7 $\Rightarrow$ $1113 = 7a + 357b$ ...(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.**

NOTES

EXERCISE 5.4

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. Your 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 anlysis.

   The following are the figures of production (in thousand tonnes) of a sugur 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.

NOTES

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

1. $y_e = 5.12 + 3.49\,x$ where origin = 1981, x unit = 1 year
2. $y_e = 13 + 0.75\,x$, with origin = 1975 and x unit = 1 year. Trend values are 10.75, 11.5, 12.25, 13, 13.75, 14.5, 15.25
3. Trend values (in lakh units) : 6.2857, 9.3714, 12.4571, 15.5428, 18.6285, 21.7142
4. $y_e = 135 + 3.1\,x$, where origin = 1987 and x unit = 1 year; 153.6
5. $y_e = 26 + x$, where origin = 1973 and x unit = 1 year
6. ₹ 95428.40
7. $y_e = 87.8 + 2.1\,x$, where origin = 1983 and x unit = 1 year
8. $y_e = 555.625 + 0.4167\,x$, where origin = 1977.5 and x unit = 1 year. y_e (1973) = 553.7498
9. 20.071, 21.714, 23.357, 25, 26.643, 28.286, 29.929 ; estimated value for 1982 = 34.858
10. $y_e = 62.2 + 1.27\,x$ 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 then the straight line trends.

Let $\{(t_i, y_i): i = 1, 2, \ldots\ldots, n\}$ be the given time series. Let x denote the deviations of the time periods $(t_1, t_2 \ldots\ldots, t_n)$ 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*:

$$\Sigma y = an + b\Sigma x + c\Sigma x^2 \quad \ldots(1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \ldots(2)$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4. \quad \ldots(3)$$

Here also, it is very important to mention the origin and the x unit with the trend curve equation.

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 $t_1, t_2, \ldots\ldots t_n$ 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

$$\Sigma y = an + c\Sigma x^2 \qquad \ldots(1')$$

$$\Sigma xy = b\Sigma x^2 \qquad \ldots(2')$$

$$\Sigma x^2 y = a\Sigma x^2 + c\Sigma x^4. \qquad \ldots(3')$$

(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:

$$\Sigma y = an + b\Sigma x + c\Sigma x^2 \qquad \ldots(1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \qquad \ldots(2)$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4. \qquad \ldots(3)$$

or

$$72.7 = 5a + b.0 + 1000c$$

$$184 = a.0 + 1000b + c.0$$

$$14700 = 1000a + b.0 + 340000c$$

or

$$72.7 = 5a + 1000c \qquad \ldots(4)$$

$$184 = 1000b \qquad \ldots(5)$$

$$14700 = 1000a + 340000c \qquad \ldots(6)$$

NOTES

(5) $\Rightarrow$ $b = 184/1000 = 0.184$

(4) × 200 $\Rightarrow$ $14540 = 1000a + 200000c$ ...(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

$\mathbf{y_e = 14.3114 + 0.184x + 0.001143x^2}$, **with origin = 1941 and x unit = 1 year.**

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%.**

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		

(*i*) Find the equation of parabolic trend.

(*ii*) Find the trend values for the years 1990—2002.

(*iii*) Plot the original data and trend values on a graph paper.

(*iv*) Estimate the production figure for the years 2003 and 2004.

Answers

1. $y_e = 30 - 12x + 2x^2$, where origin = 1983 and x unit = 1 year.
2. $y_e = 10.4286 + 3.3x + 0.7857x^2$, 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 $\{(t_i, y_i): i = 1, 2, ..., n\}$ be the given time series. Let x denote the deviations of the time periods $\{t_1, t_2, ..., t_n\}$ 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

$$y_e = ab^x \quad ...(1)$$

where a, b are constants.

$$(1) \Rightarrow \quad \log y_e = \log a + x \log b. \quad ...(2)$$

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*:

$$\Sigma \log y = (\log a)n + (\log b)\Sigma x \quad ...(3)$$

$$\Sigma x \log y = (\log a)\,\Sigma x + (\log b)\,\Sigma x^2. \quad ...(4)$$

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
$$\Sigma x \log y = (\log a).0 + (\log b)\,\Sigma x^2.$$

$$\therefore \quad \log a = \frac{\Sigma \log y}{x} \quad \text{and} \quad \log b = \frac{\Sigma x \log y}{\Sigma x^2}.$$

$$\therefore \quad \mathbf{a = AL\left(\frac{\Sigma \log y}{n}\right) \quad and \quad b = AL\left(\frac{\Sigma x \log y}{\Sigma x^2}\right)}.$$

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.

NOTES

Exponential Trend by Least Squares Method

S. No.	Census year	Population (in crores)	log y	$x = \frac{year - 1941}{10}$	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

Let the equation of the exponential trend be $y = ab^x$.

$\therefore \quad \log y = \log a + x \log b \quad$ ...(1)

The normal equations are:

$\Sigma \log y = (\log a)n + (\log b)\,\Sigma x \quad$ ..(2)

and $\quad \Sigma x \log y = (\log a)\,\Sigma x + (\log b)\,\Sigma x^2. \quad$ ...(3)

(2) $\Rightarrow$ $10.6850 = 7 \log a + (\log b).0 \quad \Rightarrow \quad \log a = \frac{10.6850}{7} = 1.5264$

(3) $\Rightarrow$ $1.6178 = (\log a).0 + (\log b).28 \quad \Rightarrow \quad \log b = \frac{1.6178}{28} = 0.0578$

$\therefore$ (1) $\Rightarrow$ $\log y_e = 1.5264 + 0.0578x \quad$ ...(4)

Also $\quad \log a = 1.5264 \Rightarrow a = \text{AL } 1.5264 = 33.60$

and $\quad \log b = 0.0578 \Rightarrow b = \text{AL } 0.0578 = 1.142$

$\therefore \quad y_e = ab^x \Rightarrow \mathbf{y_e = 33.6 \times (1.142)x}$**, where** $\mathbf{x = \frac{year - 1941}{10}}$.

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, $x = \frac{1991 - 1941}{10} = 5.$

$\therefore$ $y_e(1991) = 33.6 \times (1.142)^5 = 33.6 \times (1.142)^3 \times (1.142)^2$

$= 50.0423 \times (1.142)^2 = \mathbf{65.2634}$ **crores.**

For 2001, $x = \frac{2001 - 1941}{10} = 6.$

$\therefore$ $y_e(2001) = 33.6 \times (1.142)^6$

$= 33.6 \times (1.142)^5 \times 1.142 = 65.2634 \times 1.142$

$= \mathbf{74.5408}$ **crores.**

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 $y = AB^u$ 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 tones)*
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

1. $y = 13.79\ (2.977)^x$, where x = year – 2000
2. $y = 133\ (1.43)^x$, where x = year – 1998
3. $y = 3.07\ (1.06)^u$, 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 $y_1, y_2, ..., y_n$ be the values of a variable y taken at time periods $t_1, t_2, ... t_n$, then we write this time series as $\{(t_i, y_i); i = 1, 2, ..., n\}$.

- 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.

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

NOTES

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.

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.

6.4. TYPES OF INDEX NUMBERS

NOTES

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 is 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) Dorbish 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

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

where 0 and 1 suffixes stand for base period and current period respectively.

P_{01} = price index number for the current period

Σp_1 = sum of prices of commodities per unit in the current period

Σp_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:

NOTES

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 $$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{151}{161} \times 100 = \mathbf{93.79.}$$

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 $\Sigma p_1 = 7 + 4 + 50 = 61$

$$\therefore \quad P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{61}{54} \times 100 = \mathbf{112.96.}$$

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 $\Sigma p_1 = 700 + 4 + 50 = 754$

$$\therefore \quad P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{754}{648} \times 120 = \mathbf{116.36.}$$

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.

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,

$$\textbf{Price Relative (P)} = \frac{p_1}{p_0} \times 100.$$

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,

$$P_{01} = \frac{\sum\left(\frac{p_1}{p_0} \times 100\right)}{n} \qquad \text{(if A.M. is used)}$$

i.e.,

$$\mathbf{P_{01} = \frac{\Sigma P}{n}}$$

where P_{01} is the required price index number,

$\frac{p_1}{p_0} \times 100$ = Price relative = P.

n = no. of commodities under consideration.

In averaging price relatives, geometric mean is also used. In this case, the formula is

$$\mathbf{P_{01} = Antilog\left(\frac{\Sigma \log P}{n}\right)}$$

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
			$\Sigma P = 745$

NOTES

$$\therefore \qquad P_{01} = \frac{\Sigma P}{n} = \frac{745}{6} = 124.17.$$

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
				ΣP = 361.11

$$\therefore \qquad P_{01} = \frac{\Sigma P}{n} = \frac{361.11}{3} = 120.37.$$

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
				ΣP = 361.11

$$\therefore \qquad P_{01} = \frac{\Sigma P}{n} = \frac{361.11}{3} = 120.37.$$

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**

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

$$= 2 \text{ kg per rupee}$$

∴ Price of wheat per quintal

$$= \frac{100}{2} = ₹\ 50$$

Similarly, we shall express the prices of other commodities per quintal.

NOTES

Commodity	Unit	Ist year		2nd year		3rd year		Average price
		p_1	P	p_1	P	p_1	P	p_0
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 by Simple Aggregative Method

Index no. for 1st year $= \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{400}{497.77} \times 100 = \mathbf{83.36}$

Index no. for 2nd year $= \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{437.5}{497.77} \times 100 = \mathbf{87.89}$

Index no. for 3rd year $= \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{633.33}{497.77} \times 100 = \mathbf{127.23}$

Index numbers by Simple A.M. of Price Relatives Method

Index no. for 1st year $= \frac{\Sigma P}{n} = \frac{230.88}{3} = \mathbf{76.96}$

Index no. for 2nd year $= \frac{\Sigma P}{n} = \frac{267.78}{3} = \mathbf{89.26}$

Index no. for 3rd year $= \frac{\Sigma P}{n} = \frac{372.63}{3} = \mathbf{124.21.}$

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

NOTES

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

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

where '0' and '1' suffixes stand for base period and current period respectively.

$\Sigma p_1 q_0$ = sum of products of prices of the commodities in the current period with their corresponding quantities used in the base period.

$\Sigma p_0 q_0$ = 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

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

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

$$P_{01} = \frac{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right)}{2} \times 100$$

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.

We have $$P_{01} = \frac{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right)}{2} \times 100 = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100}{2}$$

$$= \frac{\text{Laspeyre's index no + Paasche's index no.}}{2}$$

$\therefore$ 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.

If P_{01} is the required index number for the current period, then

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

where symbols p_0, q_0, p_1, q_1 have their usual meaning.

We have $$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100\right)\left(\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100\right)}$$

$$= \sqrt{\begin{pmatrix}\text{Laspeyre's} \\ \text{Index no.}\end{pmatrix}\begin{pmatrix}\text{Paasche's} \\ \text{Index no.}\end{pmatrix}}$$

∴ 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

$$P_{01} = \frac{\Sigma p_1(q_0 + q_1)}{\Sigma p_0(q_0 + q_1)} \times 100$$

where p_0, q_0, p_1, q_1 have their usual meaning.

We can also write this index numbers as

$$P_{01} = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

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

$$P_{01} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

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.**

NOTES

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

$$P_{01} = \frac{\Sigma WP}{\Sigma W}, \text{ where } P = \frac{p_1}{p_0} \times 100.$$

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

$$P_{01} = \text{Antilog}\left(\frac{\Sigma W \log P}{\Sigma W}\right).$$

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_0q_0	p_1q_1	p_0q_1	p_1q_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

Laspeyre's price index number $= \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100 = \frac{200}{160} \times 100 = \mathbf{125.}$

Paasche's price index number $= \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100 = \frac{130}{103} \times 100 = \mathbf{126.21}$

NOTES

Bowley's price index number

$$= \frac{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right)}{2} \times 100 = \frac{\left(\frac{200}{160} + \frac{130}{103}\right)}{2} \times 100 = \mathbf{125.607.}$$

Fisher's price index number

$$= \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{200}{160} \times \frac{130}{103}} \times 100 = \mathbf{125.605.}$$

Marshall Edgeworth's price index number

$$= \frac{\Sigma p_1 (q_0 + q_1)}{\Sigma p_0 (q_0 + q_1)} \times 100 = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

$$= \frac{200 + 130}{160 + 103} \times 100 = \mathbf{125.47.}$$

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_0q_1	p_1q_0
	p_0	p_0q_0	q_0	p_1	p_1q_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

$$= \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{102}{122} \times \frac{112}{134}} \times 100 = \mathbf{83.59.}$$

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.

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

Kelly's price index number $= \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100 = \frac{31030}{20965} \times 100 = \mathbf{148.}$

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.

$$= \frac{\Sigma WP}{\Sigma W} = \frac{815}{7} = \mathbf{116.43.}$$

Price index no. by weighted G.M.

$$= AL\left(\frac{\Sigma W \log P}{\Sigma W}\right) = AL\left(\frac{14.4314}{7}\right)$$

$$= AL\ (2.0616) = \mathbf{115.3.}$$

NOTES

Example 6.7. *Prepare Index Number from the following information for the year 1980 taking the prices of 1975 as base:*

	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

$$\therefore \text{ Price index no. by weighted A.M.} = \frac{\Sigma WP}{\Sigma W} = \frac{1400}{10} = \mathbf{140.}$$

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

NOTES

3. Use the data given below and calculate Fisher's ideal price index number for the year 1993 with 1990 as base:

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

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	2002	(2001)
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.**

NOTES

Mathematically,

$$\text{Link Relative (L.R.)} = \frac{\text{Price in current period}}{\text{Price in preceding period}} \times 100$$

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 withdrawned 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	$\frac{400}{3} = \mathbf{133.33}$	$\frac{575}{3} = \mathbf{191.67}$	$\frac{625}{3} = \mathbf{208.33}$	$\frac{925}{3} = \mathbf{308.33}$

∴ F.B.I. for years 1987, 1988, 1989, 1990 with base 1986 are **133.33, 191.67, 208.33, 308.33** respectively.

NOTES

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**

∴ 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	$\frac{10}{8} \times 100 = 125$	$\frac{16}{10} \times 100 = 180$	$\frac{20}{16} \times 100 = 125$	$\frac{24}{20} \times 100 = 120$
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
IInd	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

1. 100, 133.33, 183.70, 229.62, 275.54
2. 100, 133.33, 183.7, 229.63, 275.56
3. 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

$$Q_{01} = \frac{\Sigma q_1}{\Sigma q_0} \times 100.$$

2. Simple Average of Quantity Relative Method

$$Q_{01} = \frac{\Sigma Q}{n} \qquad \text{(Using A.M.)}$$

$$= \text{Antilog}\left(\frac{\Sigma \log Q}{n}\right) \qquad \text{(Using G.M.)}$$

where $\quad Q = \text{quantity relative} = \frac{q_1}{q_0} \times 100.$

3. Laspeyre's Method

$$Q_{01} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100.$$

4. Paasche's Method

$$Q_{01} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100.$$

5. Dorbish and Bowley's Method

$$Q_{01} = \frac{\left(\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} + \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}\right)}{2} \times 100.$$

6. Fisher's Ideal Method

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100.$$

7. Marshall Edgeworth's Method

$$Q_{01} = \frac{\Sigma q_1 (p_0 + p_1)}{\Sigma q_0 (p_0 + p_1)} \times 100.$$

8. Kelly's Method

$$Q_{01} = \frac{\Sigma q_1 p}{\Sigma q_0 p} \times 100.$$

9. Weighted Average of Quantity Relative Method

$$Q_{01} = \frac{\Sigma WQ}{\Sigma W} \qquad \text{(Using A.M.)}$$

$$= \text{Antilog}\left(\frac{\Sigma W \log Q}{\Sigma W}\right) \qquad \text{(Using G.M.)}$$

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.

NOTES

In all the above formulae, suffixes '0' and '1' stand for base period and current period respectively and

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_0p_0	q_1p_1	q_0p_1	q_1p_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

$$\text{Fisher's quantity index number} = \sqrt{\frac{\Sigma q_1p_0}{\Sigma q_0p_0} \times \frac{\Sigma q_1p_1}{\Sigma q_0p_1}} \times 100$$

$$= \sqrt{\frac{4880}{2820} \times \frac{6440}{3700}} \times 100 = \mathbf{173.55}.$$

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_0p_0	q_0	p_1	Value q_1p_1	q_1	q_1p_0	q_0p_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

$$= \frac{\Sigma q_1}{\Sigma q_0} \times 100 = \frac{37}{35} \times 100 = \mathbf{105.71}$$

(*ii*) Laspeyre's quantity index no.

$$= \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100 = \frac{450}{426} \times 100 = \mathbf{105.63}$$

(*iii*) Paasche's quantity index no.

$$= \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100 = \frac{558}{528} \times 100 = \mathbf{105.68}$$

(*iv*) Dorbish and Bowley's quantity index no.

$$= \frac{\left(\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} + \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}\right)}{2} \times 100 = \frac{\left(\frac{450}{426} + \frac{558}{528}\right)}{2} \times 100 = \mathbf{105.66}$$

(*v*) Fisher's quantity index no.

$$= \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100 = \sqrt{\frac{450}{426} \times \frac{558}{528}} \times 100 = \mathbf{105.66}$$

(*vi*) Marshall Edgeworth's quantity index no.

$$= \frac{\Sigma q_1(p_0 + p_1)}{\Sigma q_0(p_0 + p_1)} \times 100 = \frac{\Sigma q_1 p_0 + \Sigma q_1 p_1}{\Sigma q_0 p_0 + \Sigma q_0 p_1} \times 100 = \frac{450 + 558}{426 + 528} \times 100 = \mathbf{105.66.}$$

III. VALUE INDEX NUMBERS

NOTES

6.18. SIMPLE AGGREGATIVE METHOD

The simple aggregative method of computing value index number (V_{01}) is given by

$$V_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times 100$$

where $\Sigma p_1 q_1$ = sum of values of items in the current period

$\Sigma p_0 q_0$ = 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_0q_0	p_1q_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

Value index number $= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times 100 = \frac{380}{290} \times 100 = \mathbf{131.03}.$

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

NOTES

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.

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 $I_1, I_2, \ldots\ldots, I_n$ 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

$$I = \frac{I_1 + I_2 + \ldots + I_n}{n} \quad i.e., \quad I = \frac{\Sigma I}{n}.$$

If $W_1, W_2, \ldots\ldots, W_n$ be the weights of index numbers $I_1, I_2, \ldots\ldots, I_n$ respectively, then

$$I = \frac{W_1I_1 + W_2I_2 + \ldots + W_nI_n}{W_1 + W_2 + \ldots + W_n} \quad \text{or} \quad I = \frac{\Sigma WI}{\Sigma W}.$$

If G.M. is to be used for finding index number of combined group, then

$$I = AL\left(\frac{W_1 \log I_1 + W_2 \log I_2 + \ldots + W_n \log I_n}{W_1 + W_2 + \ldots + W_n}\right) \text{ or } I = AL\left(\frac{\Sigma W \log I}{SW}\right).$$

NOTES

Example 6.13. *Construct the index number of business activity in India for the following data:*

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

Index No. of combined group $= \frac{\Sigma WI}{\Sigma W} = \frac{21520}{100} = \mathbf{215.2}$.

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	$259 + a + b = 350$		$36400 + 150a + 100b$

Now $\quad 259 + a + b = 350$

$\therefore \quad a + b = 350 - 259 = 91$

$\therefore \quad b = 91 - a.$

NOTES

Now, cost of living index = $\frac{\Sigma WI}{\Sigma W}$

$$\therefore \quad 136 = \frac{36400 + 150a + 100b}{350}$$

$$\therefore \quad 47600 = 36400 + 150a + 100(91 - a)$$

$$\therefore \quad 11200 = 150a + 9100 - 100a$$

$$\therefore \quad 2100 = 50a$$

$$\therefore \quad a = 42$$

$$\therefore \quad b = 91 - a = 91 - 42 = 49.$$

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.

We have $$P_{01} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \quad \text{and} \quad P_{10} = \sqrt{\frac{\Sigma p_0q_1}{\Sigma p_1q_1} \times \frac{\Sigma p_0q_0}{\Sigma p_1q_0}}$$

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.

Now $$P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times \sqrt{\frac{\Sigma p_0q_1}{\Sigma p_1q_1} \times \frac{\Sigma p_0q_0}{\Sigma p_1q_0}}$$

$$= \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times \frac{\Sigma p_0q_1}{\Sigma p_1q_1} \times \frac{\Sigma p_0q_0}{\Sigma p_1q_0}} = \sqrt{1} = 1.$$

$\therefore \quad P_{01} \times P_{10} = 1.$

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.*

NOTES

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*

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	$\frac{14.2}{16.1} \times 100 = 80.20$	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
$n = 6$				$\Sigma \log P = 11.8993$

$$\therefore \text{ Price index no. for 1996} = \text{AL}\left(\frac{\Sigma \log P}{n}\right) = \text{AL}\left(\frac{11.8993}{6}\right) = \text{AL } 1.9832 = \mathbf{96.20.}$$

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
$n = 6$				$\Sigma \log P_1 = 12.1011$

$$\therefore \text{ Price index no. for 1990} = \text{AL}\left(\frac{\Sigma \log P}{n}\right) = \text{AL}\left(\frac{12.1011}{6}\right) = \text{AL } 2.0169 = 104.$$

$\therefore$ Product of index numbers = 96.20 × 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.

NOTES

$\therefore$ The index numbers are consistent.

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

$$P_{01} \times Q_{01} = V_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

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

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \quad \text{and} \quad Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

Now

$$P_{01} \times Q_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

$$= \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

$$= \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}}$$

$$= \sqrt{\frac{\Sigma p_1 q_1 \times \Sigma p_1 q_1}{\Sigma p_0 q_0 \times \Sigma p_0 q_0}} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

= Value index number.

$\therefore$ Fisher's method satisfies this test.

6.24. CIRCULAR TEST (C.T.)

An index number method is said to satisfy the **circular test** if $I_{01}, I_{12}, I_{23}, \ldots\ldots, I_{n-1n}$ and I_{n0} are the index numbers for the periods $t_1, t_2, t_3, \ldots\ldots, t_n, t_0$ corresponding to base periods $t_0, t_1, t_2, \ldots\ldots, t_{n-1}, t_n$ respectively, then

$$I_{01} \times I_{12} \times I_{23} \times \ldots\ldots \times I_{n-1n} \times I_{n0} = 1.$$

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.

NOTES

If $n = 2$, we have

$$I_{01} \times I_{12} \times I_{20} = 1 \quad \text{or} \quad I_{01} \times I_{12} = I_{02} \qquad (\because \; I_{02} \times I_{20} = 1)$$

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 $$P_{01} = \frac{\Sigma p_1}{\Sigma p_0}, \; P_{12} = \frac{\Sigma p_2}{\Sigma p_1}, \; P_{20} = \frac{\Sigma p_0}{\Sigma p_2}$$

$$\therefore \quad P_{01} \times P_{12} \times P_{20} = \frac{\Sigma p_1}{\Sigma p_0} \times \frac{\Sigma p_2}{\Sigma p_1} \times \frac{\Sigma p_0}{\Sigma p_2} = 1$$

∴ 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

$$= P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{188}{84} \times \frac{141}{67}} \times 100 = \mathbf{217.03}$$

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)

$$= \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} = \sqrt{\frac{67}{84} \times \frac{141}{188}} = 0.7734 \qquad \text{(Not as \%)}$$

V_{01} = Value index number of 1982 with base 1975 (= 1)

$$= \frac{\Sigma V_1}{\Sigma V_0} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} = \frac{141}{84} = 1.6786 \qquad \text{(Not as \%)}$$

Now, $P_{01} \times Q_{01} = 2.1703 \times 0.7734 = 1.6785 = V_{01}$ (nearly)

∴ F.R.T. is verified.

NOTES

EXERCISE 6.6

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 is 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.

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.

$$\textbf{Consumer Price Index No.} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

where '0' and '1' suffixes stand for base period and current period respectively.

$\Sigma p_1 q_0$ = sum of the products of the prices of commodities in the current period with their corresponding quantities used in the base period.

$\Sigma p_0 q_0$ = 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	$2\frac{1}{2}$ *Qtl.* × *12*	*12*	*25*
Pulses	*3 kg* × *12*	*0.4*	*0.6*
Oil	*2 kg* × *12*	*1.5*	*2.2*
Clothing	*6 mt.* × *12*	*0.75*	*1*
Housing		*20 P.M.*	*30 P.M.*
Miscellaneous		*10 P.M.*	*15 P.M.*

NOTES

Solution. **Calculation of Cost of Living Index Number**

Item	q_1	p_0	p_1	p_1q_1	p_0q_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

$$\text{Cost of living index no.} = \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100 = \frac{1436.4}{824.4} \times 100 = \mathbf{174.24.}$$

6.31. FAMILY BUDGET METHOD

In this method, the expenditure on different commodities in the base period, are used as weights.

$$\textbf{Consumer Price Index No.} = \frac{\Sigma PW}{\Sigma W}$$

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.

$W = p_0q_0.$

$$\text{We have C.P.I.} = \frac{\Sigma PW}{\Sigma W} = \frac{\Sigma\left(\frac{p_1}{p_0} \times 100\right)p_0q_0}{\Sigma p_0q_0} = \frac{\Sigma(p_1 \times 100)q_0}{\Sigma p_0q_0} = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100.$$

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*

NOTES

Solution. Let 'x' be the retail price index for rent, fuel and light group:

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	$13770 + 18x$

$$\text{Cost of living index for 1938} = \frac{\Sigma IW}{\Sigma W}$$

$$\therefore \quad 168.12 = \frac{13770 + 18x}{100}$$

$$16812 = 13770 + 18x$$

$$\therefore \quad x = \frac{16812 - 13770}{18} = \mathbf{169.}$$

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

$$\text{Cost of living index for 1989} = \frac{\Sigma IW}{\Sigma W} = \frac{35320}{100} = \mathbf{353.20}$$

$$\text{Cost of living index for 1990} = \frac{\Sigma IW}{\Sigma W} = \frac{35158}{100} = \mathbf{351.58.}$$

The worker should not be given any extra allowance, because the cost of living index has not increased in 1990.

NOTES

EXERCISE 6.7

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

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

NOTES

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

7. MEASURES OF CORRELATION

STRUCTURE

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.

NOTES

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.

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

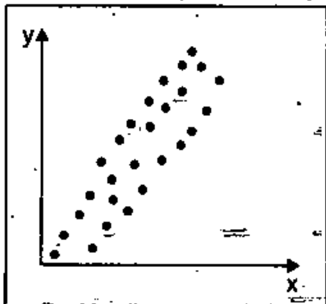
Here, a decrease in the value of x is accompanied by an increase in the value of y, except for $x = 105$. The correlation between x and y is negative.

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.

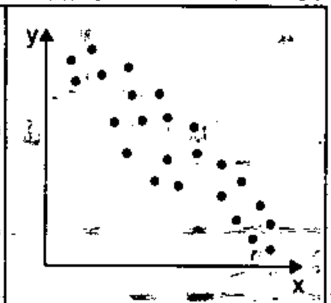
7.5. LINEAR AND NON-LINEAR CORRELATION

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.

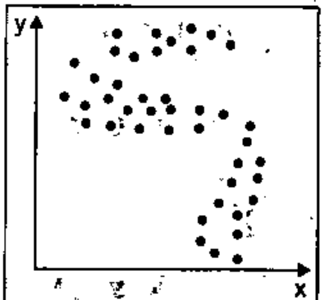
The extent of linear correlation is found by using Karl Pearson's method, Spearman's rank correlation method and concurrent deviation method.



Positive linear correlation



Negative linear correlation



Non-linear correlation

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 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 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,
$$\mathbf{r} = \frac{\Sigma(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{y}})}{\sqrt{\Sigma(\mathbf{x} - \bar{\mathbf{x}})^2} \sqrt{\Sigma(\mathbf{y} - \bar{\mathbf{y}})^2}}$$

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 $-1 \le r \le 1$.

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
$0.75 \le r < 1$	High degree positive correlation
$0.50 \le r < 0.75$	Moderate degree positive correlation
$0 < r < 0.50$	Low degree positive correlation
$r = 0$	No correlation
$-0.50 < r < 0$	Low degree negative correlation
$-0.75 \le r \le -0.50$	Moderate degree negative correlation
$-1 < r \le -0.75$	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 ρ. The letter ρ is the Greek letter 'rho'.

Remark 3. The square of Karl Pearson's coefficient of correlation is called the **coefficient of determination.**

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.

NOTES

Remark 4. $$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} \text{ implies}$$

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}}$$

$$\therefore \quad \mathbf{r} = \frac{\boldsymbol{\Sigma(x-\bar{x})(y-\bar{y})}}{\boldsymbol{n\sigma_x \sigma_y}}$$

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,

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} = \frac{24}{\sqrt{36}\sqrt{44}} = \frac{24}{39.7995} = \mathbf{+\ 0.603.}$$

$\therefore$ 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 $Y = y - \bar{y}$.

Also $$r(x, y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} \quad \therefore \quad r(x, y) = \frac{\Sigma XY}{\sqrt{\Sigma X^2}\sqrt{\Sigma Y^2}} \quad \ldots(1)$$

NOTES

Calculation of r(x, y)

S. No.	X	Y	XY	X^2	Y^2
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
$n = 7$	$\Sigma X = 0$	$\Sigma Y = 4$	$\Sigma XY = 22$	$\Sigma X^2 = 28$	$\Sigma Y^2 = 28$

$$\therefore \quad \text{(1) implies } r(x, y) = \frac{22}{\sqrt{28}\sqrt{28}} = \frac{22}{28} = \mathbf{0.7857.}$$

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 $n = 10, \Sigma xy = 120, \Sigma x^2 = 90, \sigma_Y = 8.$

Also $x = X - \overline{X}$ and $y = Y - \overline{Y}$.

$$\therefore \quad \Sigma(X - \overline{X})(Y - \overline{Y}) = \Sigma xy = 120, \; \Sigma(X - \overline{X})^2 = \Sigma x^2 = 90$$

$$\sigma_Y = 8 \text{ implies } \sqrt{\frac{\Sigma(Y - \overline{Y})^2}{n}} = 8 \text{ or } \Sigma(Y - \overline{Y})^2 = (8)^2 \times 10 = 640.$$

$$\therefore \quad r(X, Y) = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma(X - \overline{X})^2}\sqrt{\Sigma(Y - \overline{Y})^2}}$$

$$= \frac{120}{\sqrt{90} \times \sqrt{640}} = \frac{120}{3\sqrt{10} \times 8\sqrt{10}}$$

$$= \frac{120}{240} = \frac{1}{2} = \mathbf{0.5.}$$

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 $\overline{x}$ and $\overline{y}$ in the data. Suppose for some data, we get $\overline{x} = 27.374$ and $\overline{y} = 14.873$, 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 $\overline{x}$ and $\overline{y}$ 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:

We have $$r = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2}\sqrt{\Sigma(y_i - \overline{y})^2}}.$$

NOTES

Now, $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})$

$$= \Sigma x_i y_i - (\Sigma x_i)\,\bar{y} - \bar{x}\,(\Sigma y_i) + n\bar{x}\,\bar{y}$$

$$= \Sigma x_i y_i - \Sigma x_i \left(\frac{\Sigma y_i}{n}\right) - \left(\frac{\Sigma x_i}{n}\right) \Sigma y_i + n\left(\frac{\Sigma x_i}{n}\right)\left(\frac{\Sigma y_i}{n}\right)$$

$$= \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n}.$$

Also $\Sigma(x_i - \bar{x})^2 = \Sigma(x_i^2 + \bar{x}^2 - 2x_i\bar{x}) = \Sigma x_i^2 + n\bar{x}^2 - 2(\Sigma x_i)\,\bar{x}$.

$$= \Sigma x_i^2 + n\left(\frac{\Sigma x_i}{n}\right)^2 - 2(\Sigma x_i)\left(\frac{\Sigma x_i}{n}\right)$$

$$= \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = \frac{n\Sigma x_i^2 - (\Sigma x_i)^2}{n}.$$

Similarly, $\Sigma(y_i - \bar{y})^2 = \dfrac{n\Sigma y_i^2 - (\Sigma y_i)^2}{n}$.

$$\therefore \quad r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}} \text{ implies}$$

$$r = \frac{\dfrac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n}}{\sqrt{\dfrac{n\Sigma x_i^2 - (\Sigma x_i)^2}{n}}\sqrt{\dfrac{n\Sigma y_i^2 - (\Sigma y_i)^2}{n}}}$$

$$\therefore \quad \mathbf{r} = \frac{\mathbf{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}}{\sqrt{\mathbf{n\Sigma x_i^2 - (\Sigma x_i)^2}}\sqrt{\mathbf{n\Sigma y_i^2 - (\Sigma y_i)^2}}}.$$

For simplicity, we write

$$\mathbf{r} = \frac{\mathbf{n\Sigma xy - (\Sigma x)(\Sigma y)}}{\sqrt{\mathbf{n\Sigma x^2 - (\Sigma x)^2}}\sqrt{\mathbf{n\Sigma y^2 - (\Sigma y)^2}}}.$$

Example 7.4. *Find the coefficient of correlation for the following data:*

$$n = 10,\ \Sigma x = 50,\ \Sigma y = -30,\ \Sigma x^2 = 290,\ \Sigma y^2 = 300,\ \Sigma xy = -115.$$

Solution. $r = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

$$= \frac{10(-115) - (50)(-30)}{\sqrt{10(290) - (50)^2}\sqrt{10(300) - (-30)^2}}$$

$$= \frac{350}{\sqrt{400}\sqrt{2100}} = \frac{35}{\sqrt{8400}} = \text{AL}\left[\log\left(\frac{350}{\sqrt{8400}}\right)\right]$$

$$= \text{AL}\left[\log 35 - \frac{1}{2}\log 8400\right] = \text{AL}\left[1.5441 - \frac{1}{2}(3.9243)\right]$$

$$= \text{AL}\,(-0.4181) = \text{AL}\,(\bar{1}.5819) = \mathbf{0.3819}.$$

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*

NOTES

Solution.

Calculation of 'r'

S. No.	x	y	xy	x^2	y^2
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
$n = 5$	$\Sigma x = 39$	$\Sigma y = 27$	$\Sigma xy = 250$	$\Sigma x^2 = 337$	$\Sigma y^2 = 209$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{5(250) - (39)(27)}{\sqrt{5(337) - (39)^2}\sqrt{5(209) - (27)^2}}$$

$$= \frac{197}{\sqrt{164}\sqrt{316}} = \frac{197}{227.6488} = \mathbf{0.8654.}$$

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^2	y^2
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
$n = 5$	$\Sigma x = 39$	$\Sigma y = 27$	$\Sigma xy = 250$	$\Sigma x^2 = 337$	$\Sigma y^2 = 209$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{5(250) - (39)(27)}{\sqrt{5(337) - (39)^2}\sqrt{5(209) - (27)^2}}$$

$$= \frac{197}{\sqrt{164}\sqrt{316}} = \text{AL}\left[\log \frac{197}{\sqrt{164}\sqrt{316}}\right]$$

$$= \text{AL}\left[\log 197 - \frac{1}{2}(\log 164 + \log 316)\right]$$

$$= \text{AL}\left[2.2945 - \frac{1}{2}(2.2148 + 2.4997)\right]$$

$$= \text{AL}\,(2.2945 - 2.3573) = \text{AL}\,(-0.0628)$$

$$= \text{AL}\,(-1 + 1 - 0.0628) = \text{AL}\,(\bar{1}.9372) = \mathbf{0.8654.}$$

Example 7.7. *Calculate coefficient of correlation between Density of population and Death rate for the following data :*

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

$$\text{Density of population}^* = \frac{\text{Population}}{\text{Area}} \quad \text{and} \quad \text{Death rate}^* = \frac{\text{No. of deaths}}{\text{Population}} \times 100.$$

$\therefore$ For region A, $x = \dfrac{40000}{200} = 200,\ y = \dfrac{480}{40000} \times 100 = 1.2.$

For region B, $x = \dfrac{75000}{150} = 500,\ y = \dfrac{1200}{75000} \times 100 = 1.6.$

For region C, $x = \dfrac{72000}{120} = 600,\ y = \dfrac{1080}{72000} \times 100 = 1.5.$

For region D, $x = \dfrac{20000}{80} = 250,\ y = \dfrac{270}{20000} \times 100 = 1.35.$

Correlation between x and y

S.No.	x_i	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
$n = 4$	$\Sigma x = 1550$	$\Sigma y = 5.65$	$\Sigma xy = 2277.5$	$\Sigma x^2 = 712500$	$\Sigma y^2 = 8.0725$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{4(2277.5) - (1550)(5.65)}{\sqrt{4(712500) - (1550)^2}\ \sqrt{4(8.0725) - (5.65)^2}}$$

$$= \frac{352.5}{\sqrt{447500}\sqrt{0.3675}} = \frac{352.5}{405.532} = \mathbf{0.8692.}$$

Example 7.8. *In two sets of variables of X and Y with 50 observations of each, the following data were observed:*

$$\overline{X} = 10,\ S.D.\ of\ X = 3,\ \overline{Y} = 6,\ S.D.\ of\ Y = 2,\ r_{XY} = +0.3.$$

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 $n = 50, \overline{X} = 10, \sigma_X = 3, \overline{Y} = 6, \sigma_Y = 2, r_{XY} = 0.3.$

$$\overline{X} = \frac{\Sigma X}{n} \Rightarrow 10 = \frac{\Sigma X}{50} \Rightarrow \Sigma X = 500$$

$$\overline{Y} = \frac{\Sigma Y}{n} \Rightarrow 6 = \frac{\Sigma Y}{50} \Rightarrow \Sigma Y = 300$$

$$\sigma_X = 3 \Rightarrow \sqrt{\frac{\Sigma X^2}{n} - \overline{X}^2} = 3 \Rightarrow \frac{\Sigma X^2}{50} - (10)^2 = 9$$

$$\Rightarrow \quad \Sigma X^2 = 109 \times 50 = 5450$$

$$\sigma_Y = 2 \Rightarrow \sqrt{\frac{\Sigma Y^2}{n} - \overline{Y}^2} = 2 \Rightarrow \frac{\Sigma Y^2}{50} - (6)^2 = 4$$

$$\Rightarrow \quad \Sigma Y^2 = 40 \times 50 = 2000.$$

Also
$$r_{XY} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$\therefore \quad 0.3 = \frac{50\Sigma XY - (500)(300)}{\sqrt{50 \times 5450 - (500)^2}\sqrt{50 \times 2000 - (300)^2}}$$

$$\therefore \quad = \frac{50\,\Sigma XY - 150000}{150 \times 100}$$

$$\Rightarrow \quad 0.3 \times 15000 = 50\,\Sigma XY - 150000.$$

$$\Rightarrow \quad 50\,\Sigma XY = 4500 + 150000 \Rightarrow \Sigma XY = 3090.$$

After dropping the incorrect pair (X = 10, Y = 6), we have 49 pairs of values. Now we find correct values of ΣX, ΣY, ΣX^2, ΣY^2 and ΣXY.

Corrected sums

$\Sigma X = 500 - 10 = 490,$ $\quad\Sigma Y = 300 - 6 = 294,$

$\Sigma X^2 = 5450 - (10)^2 = 5350,$ $\quad\Sigma Y^2 = 2000 - (6)^2 = 1964,$

$\Sigma XY = 3090 - (10 \times 6) = 3030.$

$$\therefore \quad \text{Correct } r_{XY} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{49(3030) - (490)(294)}{\sqrt{49(5350) - (490)^2}\sqrt{49(1964) - (294)^2}}$$

$$= \frac{4410}{\sqrt{22050}\sqrt{9800}} = \frac{4410}{14700} = \mathbf{0.3}.$$

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.

Let $$u = \frac{x - A}{h} \quad \text{and} \quad v = \frac{y - B}{k}.$$

∴ 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

$$r(x, y) = r(u, v).$$

$$\therefore \quad r(x, y) = \frac{\Sigma(u - \bar{u})(v - \bar{v})}{\sqrt{\Sigma(u - \bar{u})^2}\sqrt{\Sigma(v - \bar{v})^2}}$$

On simplification, we get

$$\mathbf{r(x, y)} = \frac{\mathbf{n\Sigma uv - (\Sigma u)(\Sigma v)}}{\sqrt{\mathbf{n\Sigma u^2 - (\Sigma u)^2}}\sqrt{\mathbf{n\Sigma v^2 - (\Sigma v)^2}}}.$$

The values of u and v are called the **step deviations** of the values of x and y respectively. In the above form:

Σu is the sum of step deviations of the items of x-series.

Σv is the sum of step deviations of the items of y-series.

Σuv 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.

Σu^2 is the sum of the squares of the step deviations of items of x-series.

Σv^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,

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

$$\mathbf{r(x, y)} = \frac{\mathbf{n\Sigma uv - (\Sigma u)(\Sigma v)}}{\sqrt{\mathbf{n\Sigma u^2 - (\Sigma u)^2}}\sqrt{\mathbf{n\Sigma v^2 - (\Sigma v)^2}}}.$$

Example 7.9. *Find the correlation coefficient between 'height of father' and 'height of son', for the following data:*

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

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
$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{8(24) - 0 \times 0}{\sqrt{8(36) - 0^2}\ \sqrt{8(44) - 0^2}} = \frac{8(24)}{\sqrt{8}\ \sqrt{36}\sqrt{8}\sqrt{44}}$$

$$= \frac{24}{6 \times \sqrt{44}} = \frac{4}{\sqrt{4} \times \sqrt{11}} = \frac{2}{\sqrt{11}}$$

$$= \text{AL}\left\{\log 2 - \frac{1}{2}\log 11\right\} = \text{AL}\left\{0.3010 - \frac{1}{2}(1.0414)\right\}$$

$$= \text{AL}\ \{0.3010 - 0.5207\} = \text{AL}\ \{-0.2197\}$$

$$= \text{AL}\ \{-1 + 1 - 0.2197\} = \text{AL}\ \{\bar{1}.7803\} = 0.6030.$$

$\therefore$ **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

NOTES

Solution. Let x and y denote the variables I.R. and A.R. respectively.

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 $$r = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{10(72) - (-10)(20)}{\sqrt{10(180) - (-10)^2}\sqrt{10(180) - (20)^2}}$$

$$= \frac{720 + 200}{\sqrt{1800 - 100}\sqrt{1800 - 400}} = \frac{920}{\sqrt{1700}\sqrt{1400}}$$

$$= \text{AL}\left\{\log 920 - \frac{1}{2}(\log 1700 + \log 1400)\right\}$$

$$= \text{AL}\left\{2.9638 - \frac{1}{2}(3.2304 + \log 3.1461)\right\}$$

$$= \text{AL}\{-0.2244\} = \text{AL}\{\bar{1}.7756\} = 0.5965.$$

$\therefore$ **r = 0.5965.**

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

$\Sigma u = -170$, $\Sigma v = -20$, $\Sigma u^2 = 8288$,

$\Sigma v^2 = 2264$, $\Sigma uv = 3044$.

Let 'r' be the coefficient of correlation between the variables x and y.

$$\therefore \quad r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{10(3044) - (-170)(-20)}{\sqrt{10(8288) - (-170)^2}\sqrt{10(2264) - (-20)^2}}$$

$$= \frac{30440 - 3400}{\sqrt{82880 - 28900}\sqrt{22640 - 400}} = \frac{27040}{\sqrt{53980}\sqrt{22240}}$$

$$= \text{AL}\left\{\log 27040 - \frac{1}{2}(\log 53980 + \log 22240)\right\}$$

$$= \text{AL}\left\{4.4320 - \frac{1}{2}(4.7322 + 4.3472)\right\} = \text{AL}\left\{4.4320 - \frac{1}{2}(9.0794)\right\}$$

$$= \text{AL}\,\{4.4320 - 4.5397\} = \text{AL}\,\{-0.1077\} = \text{AL}\,\{\bar{1}.8923\} = 0.7803.$$

$$\therefore \quad \mathbf{r = 0.7803.}$$

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^2	v^2
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
$n = 6$					$\Sigma u = 3$	$\Sigma v = 20$	$\Sigma uv = -100$	$\Sigma u^2 = 19$	$\Sigma v^2 = 850$

NOTES

Now $$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{6(-100) - (3)(20)}{\sqrt{6(19) - (3)^2}\sqrt{6(850) - (20)^2}}$$

$$= \frac{-660}{\sqrt{105}\sqrt{4700}} = \frac{-660}{702.4956} = -0.9395.$$

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

NOTES

5. Calculate the coefficient of correlation and its probable error between the heights of fathers and sons for the following data:

Height of Father (in inches)	65	66	67	68	69	70	71
Height of Son (in inches)	67	68	66	69	72	72	69

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 $u = x - 10$, $v = y - 15$.

$\therefore \quad \Sigma u^2 = 180, \Sigma v^2 = 215, \Sigma uv = 60.$

Now $\quad \Sigma u = \Sigma(x - 10) = \Sigma x - n(10) = 140 - 10 \times 10 = 40$ 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:

$$r_k = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)},$$

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*	*D²*
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
$n = 12$				$\Sigma D^2 = 416$

Coefficient of rank correlation,

$$r_k = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)} = 1 - \frac{6(416)}{12(12^2 - 1)} = 1 - 1.4545 = -\mathbf{0.4545}.$$

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
$n = 10$							$\Sigma D_{12}^2 = 74$	$\Sigma D_{23}^2 = 44$	$\Sigma D_{13}^2 = 156$

We have $r_{12} = 1 - \dfrac{6\Sigma D_{12}^2}{n(n^2-1)} = 1 - \dfrac{6(74)}{10(10^2-1)} = \mathbf{0.5515.}$

$$r_{23} = 1 - \frac{6\Sigma D_{23}^2}{n(n^2-1)} = 1 - \frac{6(44)}{10(10^2-1)} = \mathbf{0.7333.}$$

$$r_{13} = 1 - \frac{6\Sigma D_{13}^2}{n(n^2-1)} = 1 - \frac{6(156)}{10(10^2-1)} = \mathbf{0.0545.}$$

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

Calculation of 'r_k'

S. No.	R_1	R_2	$D = R_1 - R_2$	D^2
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
$n = 16$				$\Sigma D^2 = 136$

Coefficient of rank correlation,

$$r_k = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)} = 1 - \frac{6(136)}{16((16)^2 - 1)} = 1 - 0.2 = \mathbf{0.8}.$$

(*ii*) The value of $r_k = 0.8$ 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
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
$n = 16$	$\Sigma x = 136$	$\Sigma y = 136$	$\Sigma xy = 1428$	$\Sigma x^2 = 1496$	$\Sigma y^2 = 1496$

NOTES

∴ Karl Pearson's coefficient of correlation,

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{16(1428) - (136)(136)}{\sqrt{16(1496) - (136)^2}\sqrt{16(1496) - (136)^2}}$$

$$= \frac{4352}{\sqrt{5440}\sqrt{5440}} = \frac{4352}{5440} = 0.8.$$

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.

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,

$$r_k = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m^3 - m) + ...\right\}}{n(n^2 - 1)}$$

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.

NOTES

Calculation of 'r_k'

Roll No.	*Marks in Hindi*	*Marks in English*	R_1	R_2	$D = R_1 - R_2$	D^2
1	45	40	5.5	4	1.5	2.25
2	56	36	2.5	6	– 3.5	12.25
3	39	30	8	9	– 1	1
4	54	44	4	2	2	4
5	45	36	5.5	6	– 0.5	0.25
6	40	32	7	8	– 1	1
7	56	45	2.5	1	1.5	2.25
8	60	42	1	3	– 2	4
9	30	20	10	10	0	0
10	36	36	9	6	3	9
$n = 10$						$\Sigma D^2 = 36$

Now $$r_k = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m^3 - m) + ...\right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6\left\{36 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right\}}{10(10^2 - 1)}$$

$$= 1 - \frac{6\left\{36 + \frac{1}{2} + \frac{1}{2} + 2\right\}}{990} = 1 - \frac{39}{165} = \mathbf{0.7636.}$$

It shows that there is a high degree positive linear correlation between the variables.

Example 7.17. *Find the coefficient of correlation between x and y by method of rank differences.*

x	*48*	*33*	*40*	*9*	*16*	*16*	*65*	*24*	*16*	*57*
y	*13*	*13*	*24*	*6*	*15*	*4*	*20*	*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	13	3	5.5	– 2.5	6.25
2	33	13	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
$n = 10$						$\Sigma D^2 = 41$

Now, the coefficient of rank correlation is

$$r_k = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m^3 - m) + ...\right\}}{n(n^2 - 1)}$$

NOTES

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 ΣD^2, with the values of m as 3, 2, 2.

$$\therefore \quad r_k = 1 - \frac{6\left\{41 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right\}}{10(10^2 - 1)}$$

$$= 1 - \frac{6\left\{41 + 2 + \frac{1}{2} + \frac{1}{2}\right\}}{990} = 1 - \frac{44}{165} = \mathbf{0.7333.}$$

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

Incorrect $r_k = 0.5$

$n = 10$

Incorrect difference of ranks (D) = 3

Correct difference of rank (D) = 7

We know that $r_k = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$

$$\therefore \quad \text{Incorrect } r_k = 1 - \frac{6(\text{incorrect } \Sigma D^2)}{n(n^2 - 1)}$$

$$\therefore \quad 0.5 = 1 - \frac{6(\text{incorrect } \Sigma D^2)}{10(10^2 - 1)}$$

$\therefore$ Incorrect $\Sigma D^2 = 82.5$.

Now Correct ΣD^2 = incorrect ΣD^2 – (incorrect D^2) + (correct D^2)

$$= 82.5 - (3)^2 + (7)^2 = 82.5 - 9 + 49 = 122.5.$$

$$\text{Correct } r_k = 1 - \frac{6(\text{correct } \Sigma D^2)}{n(n^2 - 1)} = 1 - \frac{6(122.5)}{10(10^2 - 1)}$$

$$= 1 - 0.7424 = \mathbf{0.2575.}$$

Merits

1. This method is applicable to both qualitative and quantitative variables.
2. Only this method in applicable when ranks are given.
3. This method involves less calculation work as compared to Karl Pearson's method.

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

NOTES

7. Find the coefficient of correlation between x and y by the method of rank differences:

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. $r_k = -0.0556$.

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

1. Explain the meaning of the term 'Correlation'. Does it always signify cause and effect relationship?
2. What is correlation? Distinguish between positive and negative correlation.
3. 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 interference, 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.